Investigating Prediction Modelling of Academic Performance for Students in Rural Schools in Kenya

MVURYA MGALA

MSc. ISE (Sunderland), PGDE (Egerton), BSc. (KU)

Thesis

Submitted in Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science, Faculty of Science
UNIVERSITY OF CAPE TOWN

Supervised by: A. Prof. Hussein Suleman & A.Prof A. Mbogho

August 2016

Declaration of Authorship

I, Mvurya Mgala , declare that this thesis titled, 'Investigating Prediction Modelling of Academic Performance for Students in Rural Schools in Kenya' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: M.M.

Date: 01-09-2016

Publications Undertaken

Some content used in this desertation, in the form of ideas, figures, and tables have appeared in the following four publications.

- Mgala, M., and Mbogho, A. (2014). Selecting relevant features for classifier optimization. In Advanced Machine Learning Technologies and Applications (pp. 211-222). Springer International Publishing.
- Mgala, M., and Mbogho, A. (2015, May). Data-driven intervention-level prediction modeling for academic performance. In Proceedings of the Seventh International Conference on Information and Communication Technologies and Development (p. 2). ACM.
- Mgala, M., Mbogho, A., Mwatelah, Z., and Suleman, H. (2015). Investigating mobile academic performance prediction system's predictive ability. CAPA Scientifc Journal 3 (2015), 15-25.
- Mgala, M., Suleman, H., and Mbogho, A. (2016, November). Undereducation, Motivating Intervention in Rural Schools with MAPPS. In Proceedings of the First African Conference on Human Computer Interaction (pp. 203-207). ACM.

"Setting a goal is not the main thing. It is deciding how you will go about achieving it and staying with that plan."

Tom Landry

Abstract

Academic performance prediction modelling provides an opportunity for learners' probable outcomes to be known early, before they sit for final examinations. This would be particularly useful for education stakeholders to initiate intervention measures to help students who require high intervention to pass final examinations. However, limitations of infrastructure in rural areas of developing countries, such as lack of or unstable electricity and Internet, impede the use of PCs. This study proposed that an academic performance prediction model could include a mobile phone interface specifically designed based on users' needs. The proposed mobile academic performance prediction system (MAPPS) could tackle the problem of underperformance and spur development in the rural areas.

A six-step Cross-Industry Standard Process for Data Mining (CRISP-DM) theoretical framework was used to support the design of MAPPS. Experiments were conducted using two datasets collected in Kenya. One dataset had 2426 records of student data having 22 features, collected from 54 rural primary schools. The second dataset had 1105 student records with 19 features, collected from 11 peri-urban primary schools. Evaluation was conducted to investigate: (i) which is the best classifier model among the six common classifiers selected for the type of data used in this study; (ii) what is the optimal subset of features from the total number of features for both rural and peri-urban datasets; and (iii) what is the predictive performance of the Mobile Academic Performance Prediction System in classifying the high intervention class. It was found that the system achieved an F-Measure rate of nearly 80% in determining the students who need high intervention two years before the final examination. It was also found that the system was useful and usable in rural environments; the accuracy of prediction was good enough to motivate stakeholders to initiate strategic intervention measures.

This study provides experimental evidence that Educational Data Mining (EDM) techniques can be used in the developing world by exploiting the ubiquitous mobile technology for student academic performance prediction.

"No one who achieves success does so without the acknowledging the help of others. The wise and the confident acknowledge this help with gratitude."

-Alfred North Whitehead

Acknowledgements

Many people have walked with me in this adventure. They have made the journey known to be difficult, manageable, known to be unbearable, bearable, known to be a steep hill a gentle incline. These people have held me when I slipped, guided me when I missed the point, encouraged me in those moments I would have called quites. They surely deserve my greatest gratitude.

First, I would like to thank the almighty God for being my present help in times of need and a strong tower to turn to, He is the Rock of my salvation.

I thank Hasso Plattner Institute (HPI) for the generous scholarship in the first two years of my journey. These were the foundation years of my journey and without the financial support, it would have been impossible to even get started. Many thanks to the late director of HPI professor Gary Marsden who brought me on board and was my core supervisor for one year.

My sincere thanks to my Supervisors: A. professor Hussein Suleman of University of Cape Town, and A. Professor Audrey Mbogho of Pwani university. Thank you for your excellent guidance, patience, motivation, encouragement and for providing an excellent atmosphere for doing this study.

Beside my supervisors, I thank the rest of the staff members of the department of computer science at the university of Cape Town for their insightful comments and questions that awakened my brain to think. Thanks to the system administrators, Sammy and Craig.

My appreciation goes to Martha my wife for being so understanding and an encouragement in this journey. Thanks to my children, Deborah, Isaac, Winnie and Ivan for their understanding and honest prayers.

My sincere thanks go to the Kenyan National Commission for Science Technology and Innovation (NACOSTI) for clearing me to carry out the data collection. Thanks to the County Director of Education Mr Juma Mwatenga for the authority to conduct the research in Kwale County. Thanks to the DEOs and DQASOs who were very

cooperative. Thanks to the head teachers and teachers who worked with me very closely from data collection to testing the tool.

I thank my fellow colleagues in the ICT4D lab, university of Cape Town who made this journey possible in one way or another, Chao, Ntwa, Haji, Ronke, Nini, Thomas, Pierre, Tsabi, Fiona, Yamiko, Phiri, James, George, Sarah.

Lastly, may the good God reward all those persons whose names I have not mentioned here for the support during this journey. You played a significant role and I am grateful to you all. "Thank you"...

Contents

List of Tables

List of Figures

Abbreviations

AEO	**A**rea **E**ducation **O**fficer
CDE	**C**ounty **D**irector **E**ducation
DEO	**D**istrict **E**ducation **O**fficer
DQASO	**D**istrict **Q**uality **A**ssurance **S**tandards **O**fficer
FPE	**F**ree **P**rimary **E**ducation
FS	**F**eature **S**election
ICT	**I**nformation and **C**ommunication **T**echnology
ICT4D	**I**nformation and **C**ommunication **T**echnology **F**or **D**evelopment
KCPE	**K**enya **C**ertificate **P**rimary **E**ducation
MDG	**M**illenium **D**evelopment **G**oals
ML	**M**achine **L**earning
MAPPS	**M**obile **A**ccademic **P**erformance **P**rediction **S**ystem
UPE	**U**niversal **P**rimary **E**ducation

Dedicated to my late father Mr Mabengo, my mother Ruth Mkambe, my dear wife Martha and my children, Deborah, Isaac, Winnie and Ivan...

Chapter 1

Introduction

1.1 Background

The high failure rate of primary school students in rural areas of developing nations is a big challenge. Although the challenge is global, it has been experienced more in resource constrained areas of developing countries (Van de Grift and Houtveen, 2006). High failure rate, or underperformance, is when students score marks that are below a defined threshold in standardised exit examinations at the end of a learning cycle (Suryadarma et al., 2006). This definition of underperformance is adopted in this study. The problem has been associated with undereducation; where a primary certificate is the highest level of education. Prediction is adopted in this study, which is a positivist theory referring to the ability to control and predict (Braa and Vidgen, 1999). For example features that are most predictive of a target could be identified so that a model is built that will be used to determine the target for new records. Efforts to tackle the problem in developed countries led to the development of academic performance prediction models (Romero et al., 2013), where students' final marks are estimated early enough for appropriate intervention measures to be put in place to assist weak students (Tamhane et al., 2014). The intervention measures could be remedial classes offered in school or extra tuition organised by the parents. Additionally, students determined as potential failures could be taken through counseling sessions or be given meals if determined to be needy. Clearly, what is needed is adequate motivation on the education stakeholders to initiate the strategic intervention.

The academic performance prediction models have been built using educational data. Such data contains useful information that could be extracted in order to give insight to

the causes affecting students' academic performance. This knowledge is what could be used to initiate intervention for the students who need it. The use of educational data to discover useful knowledge about students is known as Educational Data Mining (EDM). Its focus is to develop, research, and apply data mining techniques to discover knowledge that is otherwise difficult to extract because of the large amounts of data involved (Romero et al., 2010, Scheuer and McLaren, 2012). In the developed countries, EDM has been used to identify students who could be assisted early enough to avoid dropping out of school (Fall and Roberts, 2012). Primary school dropouts are undereducated because they lack professional qualifications, they can only undertake unskilled jobs (Von Wachter et al., 2011).

As a contribution towards tackling the problem of underperforming in academic work, students in this study have been classified into two groups: high intervention and low intervention. Those that need high intervention are the students that if nothing is done to assist them early enough are likely to score below a determined threshold for passing. The process of classifying the students into such classes is known as binary classification. It is an EDM technique in which a student record is categorised to be in either of the two classes using a defined criteria during the model training process (Romero and Ventura, 2010). The high and low intervention levels are more meaningful than predicting the actual marks that a student will score in the final examination, the objective is to motivate strategic intervention (Vandamme et al., 2007). The high intervention level only suggests that a student requires intervention, it does not predetermine whether or not the student will pass. Studies on building prediction models, especially in the developed world, have used PCs. Such studies are: predicting whether a student will complete their university education or will drop out (Dekker et al., 2009); predicting the success or failure of a university student (Guruler et al., 2010); and predicting high or low performance of a university student (Luo et al., 2015).

The difference between the previous studies and our study is the recognition that PCs cannot be used in some rural areas of developing nations. There is poor infrastructure and scarcity of resources to buy and maintain PCs. Electricity is still either unavailable or unstable, especially in Kenya, one of the countries in Sub-Saharan Africa. PCs are still seen as a technology for the developed world that may not be appropriate in the rural areas of developing countries (Brewer et al., 2005). The challenges of rural areas of the developing world are the main reason for a large fraction of the human race to miss out on technologies such as Educational Data Mining.

The mobile phone has proved to be the only technology that has spread everywhere, in both developed and developing countries, even in the rural areas (Kumar et al., 2015).

We therefore, proposed to incorporate mobile phones in the academic performance prediction system. However, the mobile phone that would be readily available and affordable to everyone has unique challenges such as, a small screen size, small memory, and low processing power, this makes model training with big datasets impossible (Matyila et al., 2013).

The combination of challenges faced by the rural population together with the challenges of the type of mobile phone that could be affordable in rural areas of developing nations determined the design process adopted in this study. One key design process that was useful for designing the system in a small mobile screen was the determination of an optimal feature subset. This feature subset has similar predictive power for the target class as the complete dataset (Yu and Liu, 2003).

The objective of this research was to contribute towards tackling the problem of under-education in resource constrained environments of developing nations. Our contribution was the design and development of the academic performance prediction system that incorporated a mobile phone interface. With this system, it was possible to predict the students that would need high intervention early enough to motivate strategic intervention. A User-Centred Design (UCD) approach was adopted in the interface design process to improve the chances for system acceptance (Marsden et al., 2008).

The education system followed in Kenya as the country where the study was conducted is discussed next.

1.2 System of Education in Kenya

Kenya is one of the countries in Sub-Saharan Africa. It was chosen as the research area because it is the home country of the researcher. Currently, the country follows a system where a student spends eight years in primary school, four years in secondary school and a minimum of four years in university (8-4-4 system). The system was introduced during the reign of the second president in 1985 (Kimosop et al., 2015, Sifuna, 1992, Wycliffe et al., 2013). The first eight years begin in Class 1 after a child has spent 3 years of early childhood education. The student learns through for eight years in primary school. At the end of the eighth year, students sit for a standardised exit national examination, known as Kenya Certificate of Primary Education (KCPE) (Lucas and Mbiti, 2012).

Class 8 candidates sit for the following papers (Kimosop et al., 2015):

- English grammar, reading, comprehension and composition writing

- Swahili grammar, reading, comprehension and composition writing
- Mathematics
- Science
- Social Studies which combines History, Civics, Geography, Religion

The students sit for a total of five papers, each paper has equal weight of 100 marks. A candidate who scores 100% in each paper will get a total of 500 marks. The marks a student scores at this level are used to determine whether or not they will proceed to secondary education (Lucas and Mbiti, 2014). Those that score below a determined threshold become dropouts. This threshold is usually 250 marks. Those who score below this mark are considered dropouts of the school system. A small number may get admission in youth polytechnics; but most of them end up doing unskilled jobs in urban centers.

The total score also determines the type of secondary school a student will attend. Those who score above 400 marks are likely to be admitted in prestigious schools known as national schools. These schools receive more financial assistance from the government, they have better facilities and adequate staff. The country has a total of 105 national schools, at least 2 in each of the 47 counties. In each county, about 430 students are admitted in these national schools. The students who score below 400 marks are admitted in Intra-County, County, and Sub-County schools. The students who score higher marks are admitted to the Intra-county schools while those who score about 250 marks are admitted in Sub-County schools. The Kenyan system of education is such that only bright students are allowed to progress with education, the weak students drop out (Wycliffe et al., 2013).On average, 25% of the students who complete Class Eight do not get admission in secondary schools. These students are mostly from rural primary schools. This study focuses on such students.

1.3 Study Motivation

Education is a basic human right, according to UNESCO it is important to all human race. UNICEF was formed with the objective of supporting Education for All (EFA) (Brundrett, 2011). When the United Nations (UN) came up with the millennium development goals, they declared attainment of primary education the second most important goal (Bruns and Rakotomalala, 2003). Although much effort was put to achieve the goal, it was seen as impossible to achieve it in Sub-Saharan Africa (Easterly, 2009), the latest

statistics stand at over 60% (Poverty, 2015). The initiative to offer Free Primary Education (FPE) was the main cause for the higher enrollment rate (İşcan et al., 2015). However, the higher enrollment also brought about the challenge of ensuring quality education. The problem of poor academic performance could therefore be associated with the large enrollment due to FPE. Educationists have realised there is need to balance the emphasis on having every child attend school with acquiring quality education (Poverty, 2015). This is because, poor academic performance causes students to drop out of school (Rogers, 2014). Therefore, UNESCO has challenged all governments to shift their emphasis to improving the quality of education regardless of their status, whether rich or poor, male or female, or whether they live in urban areas or rural areas (Brundrett, 2014).

Further, governments in developing countries need to change their emphasis on financial support from secondary and tertiary education to tackling the problem of poor academic performance in primary schools. Primary school education forms the foundation. A weak foundation has been identified as the main reason why there is a shortage of skilled human resource in developing countries (Masino and Niño-Zarazúa, 2015).

An early survey conducted at the beginning of this study provided substantial motivation for it to be carried out. The focus in the survey was to establish the views of education stakeholders in Kwale County. The survey results established the fact that the problem has existed for many years. These findings motivated the design and development of the Mobile Academic Performance Prediction System (MAPPS) that could classify students into the high and low intervention levels. The survey results are presented next.

1.3.1 Survey Results

This survey was conducted with officers in the education office, head teachers and teachers of primary schools in Kwale County. The goal was to get the participants' perspective of the high failure rate in the County.

1.3.1.1 Survey Results with Education Officers and Head Teachers

The officers who participated include the County Director of Education (CDE), District Education Officers (DEOs), District Quality Assurance and Standards Officers (DQASOs), Area Education Officers (AEOs). Head Teachers were from 14 primary schools. A semi-structured interview was conducted with each one of them. The researcher visited them in their offices after making a phone call appointment.

Distribution of Participants

Table 1.1 shows the distribution of the participants in the survey.

Participant	**Total Number**
County Director of Education	1
District Education Officer	2
District Quality Assurance and Standards Officer	1
Area Education Officer	3
Head Teachers	14
TOTAL	21

TABLE 1.1: Participants of the semi-structured survey

A total of 21 participants took part in the survey, 14 head teachers and 7 education officers. The head teachers were more available and had a lot more to say about the problem since they are in touch with what happens in the schools. The CDE is charged with the responsibility of overseeing education activities in the whole County, such as all matters pertaining to education quality. Some of the tasks they undertake include administration, inspection, and supervision in the schools. They delegate some of the tasks to DEOs who work with AEOs to enforce quality education in the whole County (Sisungo, 2012). Those charged with the responsibility of interacting with teachers to enforce best teaching practices and syllabus coverage are the District Quality Assurance and Standards Officers' (DQASOs). They organise capacity building seminars and workshops that equip teachers with skills to make them better teachers (King'oina, 2011).

Although head teachers are more informed about the problem of academic performance, the education officers also have responsibilities that force them to be informed of what happens in the schools.

Participants' Working Experience

The working period of participants in the County give insight about how much they known the problem. The results of the experience survey are presented in Table 1.2.

Experience	**Participants**
Up to one year	2
Up to 3 years	5
Up to 5 years	5
Over 5 years	9

TABLE 1.2: Participants' working experience in the County

The results show that 9 participants had worked for over 5 years in the County. The rest had worked for the following number of years: up to 5 years, and up to 3 years, five participants each; up to 1 year, two participants. These results show that 14 participants

had an experience of working for more than 5 years in the County. These were capable of giving a true picture of the problem.

Participants' Report on How Long the Problem has Existed

The participants reported on what they knew about how long the problem of high failure rate had existed in the County. Table 1.3 shows the survey results.

Period	**Participants**
lasted for 2 years	1
lasted for 5 years	1
lasted for over 5 years	19

TABLE 1.3: Period the problem of high failure rate had existed

The results show that, out of the 21 participants, 19 suggested the problem had been in existence for over 5 years.

One head teacher had this to say, *".....the problem of high failure rate in this County has lasted for as long as 15 years since I started teaching.....".*

One participant suggested the problem had lasted for 5 years. Another participant thought it had lasted for only two years. The participants did not want to project what they knew, they had worked in the County for the number of years they reported the problem had existed. The survey shows that most of the participants suggest the problem has been in existence for many years in the County.

Proposed Solutions by Participants

The participants were asked to suggest solutions to the problem. A sample of the suggested solutions are presented:

The CDE had this to say *"....all the education stakeholders have to work together to reduce the high rate of failure among our students, when everyone does their responsibility well, we shall overcome.....".*

The DQASO said *".....the problem is that the teachers have a negative attitude towards the students' ability in the County. Head teachers are also failing in their administrative responsibility. All education stakeholders have to participate in looking for solutions to this problem.....".*

An AEO said *".....the education office in the County has to put more emphasis in inspecting the schools. Where there is a scarcity of teachers, more must be employed. Parents as key stakeholders should be encouraged to participate more through sensitisation.....".*

The head teachers collectively said *".....there is need for the school environments to be improved; teachers in the County need to be exposed to what other teachers in passing Counties do; teachers should be given some financial intensives; there is need to feed students in schools; and parents and communities should be educated on the importance of education....."*.

The wide range of suggestions on the possible solutions show that the problem of high failure rate exists in the schools in Kwale County. The proposed solutions do not necessarily suggest the type of solution proposed in this study; however, it is clear the participants are searching for a solution. The CDE suggested the need for education stakeholders to come together and work towards solving the problem. The proposed prediction system in this study could motivate the stakeholders to come up with suitable intervention for the problem. The DQASO's suggestions that teachers need to change their negative attitude, and that head teachers should become better managers, may only be realised if they find good reasons for the change. It is proposed in this study that a change of attitude by teachers and improved management by head teachers could be motivated if large numbers of students in their schools are predicted as requiring high intervention early enough. Similarly, the AEO suggested the need for education officers to increase the rate of inspection in schools, the system designed in this study could facilitate the education officers during their inspection visits. Lastly, the head teachers proposed a list of things they feel could go along way to improve the situation, these suggestions could be put in place if education stakeholders are motivated enough. The proposal to determine the week students early enough in this study could refocus the energies of the education stakeholders to act on some of the intervention measures suggested by the head teachers.

The education officers suggested a long list of causes of poor academic performance in the County. This list was verified by the teachers' survey discussed next.

1.3.1.2 Teachers' Survey

Teachers are key stakeholders in education, therefore, this survey sought to discover important insights from the teachers about the problem of high failure rate among students in Kwale County. The objective of the survey was to establish from the teachers' perspective whether the problem exists or not. A total of 124 teachers from 13 primary schools in Kwale County participated in the survey using questionnaires (Appendix G). The questionnaires were filled and all were returned because they were filled in the presence of the researcher while he visited the schools. The survey results are presented

in the following order: first, the teachers' characteristics are presented; secondly, the perspective of teachers on the possible causes of the problem are given.

Teachers' Work Experience

The teachers work experience is an important characteristic in relation to their knowledge of the problem being studied. The survey sought to know the rating of teachers in terms of the number of years they had worked. Results are presented in Table 1.4.

Experience	Number of teachers	Percentage
Over five years	76	61.29%
Between 2 and 5 years	28	22.58%
Less than 2 years	20	16.13%

TABLE 1.4: Teaching experience of the teachers

The results show that 61.29% of the teachers had a teaching experience of over five years, 22.58% had and experience of up to five years and 16.13% had an experience of less than two years. Experience in teaching is associated with teacher quality (Harris and Sass, 2011). Therefore, it made sense to rely on the teachers' experience in the process of data collection and user requirement gathering.

Teachers' Professional Qualification

In addition to teaching experience, teachers need to have professional qualifications. A survey was conducted to gather information about the teachers' qualifications as shown in Table 1.5.

Type of training	Number of teachers	Percentage
Bachelor of Education	4	3.23%
Diploma in Education	28	22.58%
P1	73	58.87%
untrained	19	15.32%

TABLE 1.5: Teacher professional qualifications

As shown, 58.87% had P1 professional qualification. P1 is the least primary school teacher qualification required. 22.58% had a diploma in education. This is a professional teaching qualification obtained from university, it is higher than the P1 qualification. Only 3.23% of the teachers had a degree level professional qualification. The remaining 15.32% had no professional qualification. They were mainly school leavers engaged by the schools to relieve the professional teachers from the heavy work load. A total of 84.68% of the teachers had professional qualification, meaning, their suggestions concerning the possible causes of the high failure rate of student were plausible. These results show that a large percentage of the teachers had some professional training and therefore

capable of making a reasonable judgment about the causes of academic failure among the students.

Rate of Teacher Transfer

The survey sought to understand whether or not teachers are rapidly transfered from one school to another. This was measured in terms of the number of years a teacher had stayed in the same school. The results are presented in Table 1.6.

Number of years	Number of teachers	Percentage
Over five years	49	39.52%
Between 3 and 5 years	21	16.94%
Between 1 and 3 years	30	24.19%
Less than 1 years	22	17.74%
Other	2	1.61%

TABLE 1.6: The rate of teacher transfer

The results show that 56.46% of the teachers had stayed in one school for over 3 years. This shows that at least they were acquainted with their environment and the academic performance situation in their schools.

Class Size

The number of students a teacher teaches in a class is known to directly impact on the teachers' performance. This survey sought to know the class sizes the participation teachers handled. Table 1.7 presents the results.

Number of students	Number of teachers	Percentage
Above 60	114	91.94%
Between 41 and 60	9	7.26%
Below 40	0	0.00%
No answer	1	0.81%

TABLE 1.7: Number of students in a Class

Results show that 91.94% of the teachers had large classes of over 60 students. A crowded class is defined as one having over 40 students (Duflo et al., 2015). Among the participating teachers, only 7.26% taught classes with less than 40 students. The number of overcrowded classes show that there is a shortage of teachers.

Syllabus Completing Rate

The commitment of a teacher could be associated with syllabus coverage. This survey sought to find out the participants' rate of their syllabus coverage. Table 1.8 presents the findings.

Among the participants, 58.06% reported that they managed to complete their syllabi, 29.84% said they were not able to complete the syllabi. The remaining, 12.10% , were

Completed Syllabus	Number of teachers	Percentage
Yes	72	58.06%
No	37	29.84%
Uncertain	15	12.10%

TABLE 1.8: Syllabus completion rate

uncertain. This finding implies that only slightly over half of the teachers completed the syllabus. This suggest that the level of commitment among the teachers is wanting. A previous study established a correlation between syllabus coverage and academic performance (Nakhanu, 2012).

Opinion of Teachers on the Possible Causes of the Problem

The participating teachers were presented with a set of opinion questions to find out their opinion about the possible causes of the problem of high rate of failure in Kwale County (Appendix G). The opinion questions were extracted from a list of causes obtained from literature (Etsey, 2005, Orodho et al., 2014) and from an interview conducted earlier with education officers and head teachers. A Likert scale with the following options was used: strongly agree; agree; neutral; disagree; strongly disagree. The findings are presented in Table 1.9.

The percentage of teachers who responded to each item suggests the possible causes of the high rate of failure in Kwale County. The items with the highest ratings were: inadequate teacher salaries; unfavourable working conditions; poor classroom environment; inadequate text books; shortage of teachers; and parents not following up on students' progress. Inadequate teacher salaries had a rating of 84%, meaning, they disagreed with the statement that their salaries were adequate. Teachers need to be motivated with higher salaries, they help in motivating the teachers to be regular in school and improve academic performance (Duflo et al., 2012). This high rate of dissatisfaction suggests that teachers may be contributing to the high rate of failure among students in Kwale County.

Parents lack of interest in the students' academic progress is the second factor. Children's good academic performance has been associated with the help parents give them at home (Karbach et al., 2013). The survey shows that 73.39% of the teachers disagree that parents follow up their children's progress. Some teachers commented that most parents are not educated, this could explain why they cannot do the follow up.

The rate for teacher shortage was 71%, teachers agreed that there is a shortage. Incidentally, teachers shortage is a common problem in rural areas, a study conducted in Australia revealed the same. The study results indicated that teacher shortage is an indicator of high failure rate (Sullivan et al., 2013). Therefore, our survey results suggest

Opinion Statement	**Strongly agree**	**Agree**	**Neutral**	**Disagree**	**Strongly disagree**
1. Students do not work hard	25.00%	48.39%	16.13%	8.87%	1.61%
2. Our salaries are adequate	3.23%	5.65%	6.45%	40.32%	44.35%
3.Work conditions are favourable	1.61%	20.16%	19.35%	43.55%	15.33%
4. Students have low abilities	4.03%	20.97%	18.55%	45.97%	10.48%
5.Students have the responsibility to understand the lesson	4.84%	14.52%	8.06%	45.16%	27.42%
6. We have an appropriate classroom environment	4.84%	25.81%	11.29%	41.13%	16.94%
7.We get enough text and exercise books	5.65%	28.23%	13.71%	43.55%	8.87%
8. We have a good working relationship with our Head teacher	31.45%	49.19%	10.48%	6.45%	2.42%
9. There is close supervision by the school administration	23.39%	61.29%	9.68%	4.03%	1.61%
10. There is close supervision by education officers	16.94%	42.74%	12.90%	21.77%	5.65%
11. There is a conducive teaching/learning environment	25.81%	54.84%	13.71%	3.23%	2.42%
12. There is a shortage of teachers	29.03%	41.94%	8.87%	17.74%	2.42%
13. Parents do follow up on students' progress	3.23%	12.10%	11.29%	29.84%	43.55%
14. Students are undisciplined	6.45%	23.39%	22.58%	41.13%	6.54%

TABLE 1.9: Teachers' opinion on the causes of Poor Academic Performance in Kwale County

that the teacher shortage reported could be one of the possible causes of the high rate of failure in the County.

Lastly, inadequate resources also turned to be one of the possible indicators of high rate of failure. Both classroom resources and teaching and learning resources were found to be inadequate. Scarcity of resources such as classrooms, teaching and learning materials have been associated with achieving poorly in academic work (Sullivan et al., 2013). The survey in this study recorded a rating of 58% of the teachers saying that their classrooms lacked suitable learning environment. Similarly, 52% responded that the actual teaching and learning materials were not sufficient. These ratings show that over half of the participating teachers do not have adequate resources in their schools. Therefore, lack of resources could also be an indicator of the high rate of failure in Kwale County.

The two survey findings, the first, by education officers and head teachers, and the second, by 124 teachers in 13 primary schools have indicated that there is a problem of high rate of failure among primary school students in Kwale County.

1.4 Problem Statement

The aim of the study was to identify the best classifier model for the type of data used, select the optimal feature subset from the total number of features, and design a system that integrates the classifier model with a mobile phone interface to make the prediction of students requiring high and low intervention possible in rural areas of developing countries, and to carry out an evaluation of the system.

The scope of this study was on the academic performance of primary school students in rural areas of developing countries. The study was conducted in Kenya. Data for the experiments was collected in Kwale County, Kenya. This County was chosen because it is one of the Counties that has many rural schools which suffer from high failure rate. It is also the County where the researcher was born and schooled to primary level. Kwale County has for many years been ranked among the bottom five Counties out of all the 47 Counties in the KCPE examinations.

An initial survey with a an education officer confirmed the problem, the officer said the following:

".....in this Sub-County about 30% of the candidates who sit for KCPE score above 250 marks and get admission in secondary schools, the rest drop out and end up doing unskilled labour. If no intervention is initiated to reduce the failure rate, this region will not develop because its citizens are undereducated .....".

The study developed a system that could be used to reduce the number of students who score below 250 marks in the Class Eight final examinations. Students were classified into two categories: those who require high intervention and those who require low intervention. The EDM classification process was carried out in a remote server using a logistic regression classifier. A mobile phone acted as an interface to the system. The reviewed literature suggests that this approach has not been applied in Sub-Saharan countries of developing nations.

The data used in the experiments was manually collected because of lack of digital data. The data collection process took up to three months to gather 3000 student records from 54 primary schools.

1.5 General Research Questions

How can a prediction system be designed so that it motivates strategic intervention in primary schools in rural areas of developing countries?

1.5.1 Research Questions

The following are the specific research questions addressed in this study:

1. Which is the best classifier model among the six common classifiers selected for the type of data used in this study?
   (a) To compare the prediction performance of the six selected classifier models on the rural dataset in terms of numbers of correctly classified and incorrectly classified students.
   (b) To compare the prediction performance of the six classifier models using the six selected metrics.
   (c) To determine the best classifier model according to the classifier performance results obtained using peri-urban data.
2. What is the optimal subset of features from the total number of features for both rural and peri-urban datasets?
   (a) To determine the most predictive features from the three lists that have been ranked using ranking algorithms.
   (b) To determine the optimal number of features that achieve the highest predictive performance of the selected classifier models in the two datasets.
3. What is the predictive performance of the Mobile Academic Performance Prediction System in classifying the high intervention class?
   (a) To compare the MAPPS classification performance between the rural schools dataset and the peri-urban dataset.
   (b) To compare the prediction performance of MAPPS in the two student datasets; one for students two years and the other for students one year before they sit for the final examination.

1.6 Research Design

The Cross-Industry Standard Process for Data Mining (CRISP-DM) was used together with User-Centered Design (UCD) to accomplish the system design. Figure 1.1 illustrates the complete design process.

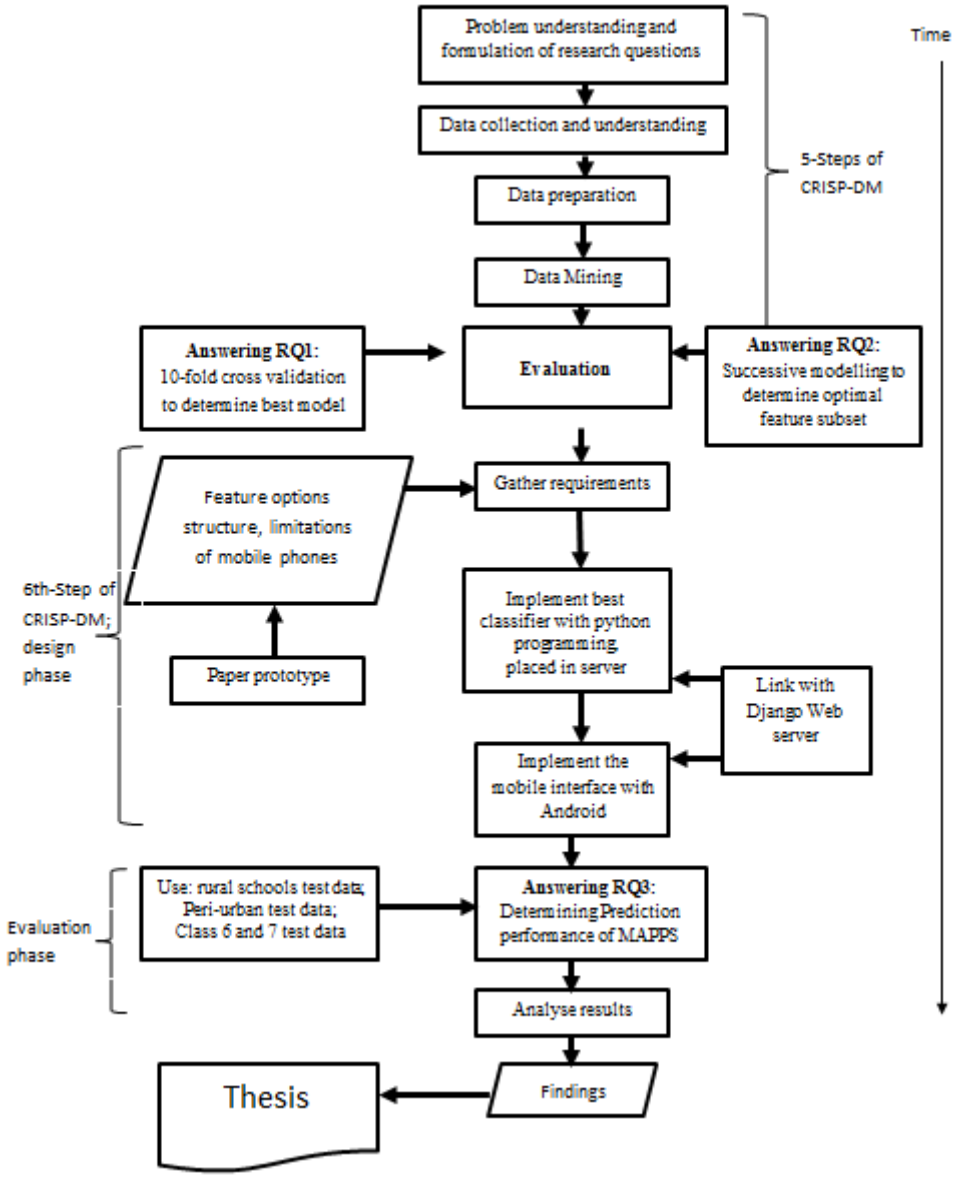


FIGURE 1.1: The combined CRISP-DM and UCD processes followed in this study

The CRISP-DM process steps followed in this study are: problem understanding; data understanding; data preparation, data mining; evaluation; and using the discovered knowledge (Kurgan and Musilek, 2006).

The study aimed to contribute in tackling the high rate of failure in rural primary schools of a developing country. Step one was to understand the problem and identify the factors or features that indicate failure in academic performance. The features were gathered through surveys and from literature. The features compiled were then used in the data collection step. The data which was manually collected was entered into excel worksheets to create meaningful datasets. In the preprocessing step, data was transformed into the format that would allow it to be acceptable classifier modelling. Six common classifiers algorithms were selected including logistic regression, MLP, SMO, Naïve Bayes, J48, and Random Forest.

The first and second research questions in this study were answered in the first 5 phases (see figure 1.1) of the research design. Answering the first question entailed building six classifier models and determining the best model through evaluation using the 10-fold cross validation criteria. Six metrics were adopted in validating the results: ROC area, F-Measure, Cohen Kappa, RMSE, sensitivity and specificity. The second question was answered in parts, firstly, features were ranked using three filter algorithms (Bolón-Canedo et al., 2013), next, the ranked feature lists were used to successively build models in order to determine the optimal subset of features (Hassanien et al., 2014, Ramaswami and Bhaskaran, 2009).

The third question was addressed by first implementing the best classifier model using Python, it was then placed in remote server in the Computer science department of the University of Cape Town. The mobile phone interface was designed and implemented with Android, it was then linked to the classifier model via Django Web server Figure 1.1. Android was preferred because it is open source and is rapidly getting popularity as an operating system for mobile phones (Erturk, 2012). The resulting system was given the name Mobile Academic Performance Prediction System (MAPPS). The evaluation process for MAPPS was conducted in three phases: using 30% of the rural schools' dataset; using 40% of peri-urban schools' dataset; and using data collected directly from Class 6 and 7 students in 15 primary schools as teachers used the system for three weeks. A number of metrics including sensitivity, specificity, F-Measure, and accuracy, were generated and used from a confusion matrix (Hempel et al., 2012). The evaluation process conducted led to the analysis of research findings being reported in Chapter 7.

1.7 Thesis Structure

The remaining thesis is structured as follows:

Chapter 2: Educational Data Mining
The chapter discusses the theoretical perspective of Educational Data Mining to establish the theoretical basis adopted in this study.

Chapter 3: Literature Review
Literature related to this study is reviewed. The chapter has been structured into the following sections: causes of school dropout; causes of student underperforming in developing countries; features selection; academic performance prediction modelling; adoption of mobile phones in ICT4D; and mobile phone use in education. The review gives insight into the existing gaps and opportunities that this study proposes to fill.

Chapter 4: Methodology
The chapter presents the design of the academic performance prediction model. The Cross-Industry Standard Process for Data Mining is used to guide the design process. This framework steps which were used in this Chapter include data collection and preprocessed, determination of the optimal feature subset, and determination of the best classifier model. Finally, it is proposed that the last step of CRISP-DM, using the discovered knowledge, be the mobile interface design, implementation and finally evaluation of MAPPS.

Chapter 5: Classifier Model and Optimal Feature Subset
The chapter presents a discussion of the EDM process followed to address the first two research questions. This discussion is split into how experiments were conducted to determine the best classifier model, and to determine the optimal feature subset. Lastly, the chapter presents a summary of the experimental results and a discussion on how these results address the first two research questions.

Chapter 6: Design and Implementation of MAPPS
The chapter discusses the design process for the mobile phone interface of MAPPS. A discussion on the findings from a contextual inquiry are presented. These findings help in understanding the problem from the users perspective. Next, the user-centered design approach that was followed to design the mobile phone interface is discussed. The chapter ended with a summary of the design process followed.

Chapter 7: Results and Discussion of MAPPS
The chapter presents a discussion of the evaluation criteria and the subsequent evaluation of MAPPS carried out to address the third research question. The experiments, how they were conducted using MAPPS and the results are discussed in details. The chapter concludes by presenting a summary of the user evaluation of MAPPS usability and usefulness.

Chapter 8: Conclusions

The chapter begins by restating the three research questions and presents an elaborate discussion on how the study has answered each one them. Next, a summary of the conclusions is presented, followed by a discussion of the limitations of the research. Finally, future work is discussed.

Chapter 2

Educational Data Mining

2.1 Introduction

The first chapter presented the introduction and motivated the study. It also briefly mentioned that EDM process supports building of academic performance prediction models. This chapter describes the theoretical perspective of EDM and its formal application in building the Mobile Academic Performance Prediction System. Firstly, a justification discussion for the use of EDM as opposed to the Learning Analytics is presented. Thereafter, EDM is discussed as a way of aligning the theory to this research.

2.2 Choice of Educational Data Mining

The need for a theoretical underpinning of this research was influenced by two factors: to select a theoretical concept that supports the building of prediction models; and a theoretical concept that underlies the goal of motivating the initiation of strategic intervention for the students so that they score above average marks in national standardised examinations. There are two fields that have been identified as focused on analysing educational data with a view to understand learners and their learning environment. These fields are: Educational data mining, and Learning analytics. Although the two fields have similarities from their definitions, a number of points justify the use of EDM in this study. The two fields have been defined as cited in (Siemens and Baker, 2012, p.1):

"EDM is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in."

"LA is the measurement, collection, analysis and reporting of data about learners and their contexts for the purpose of understanding and optimising learning and the environments in which it occurs."

Although the two fields have a similar focus, to understand the students and the environments in which they learn, this study has chosen EDM for the same reasons as those given by Siemens and Baker (2012):

First, EDM focuses more on automated discoveries. From this point of view, EDM is preferred since the knowledge discovery is purely data driven after the humans have contributed the initial set of features (Mgala and Mbogho, 2015).

Secondly, EDM follows a preferred framework where features are studied together but also in smaller groups called feature subsets. This way, it is possible to also determine the features that have the greatest impact on the target class (Bratu et al., 2008). This could also be beneficial to education stakeholders who would like to come up with strategic measures; focusing on a subset of features can be more meaningful.

Further, EDM is preferred because of its popularity with the community that have conducted student academic performance prediction (Bhardwaj and Pal, 2012, Golding and Donaldson, 2006, Kotsiantis et al., 2002).

Lastly, EDM is more focused on automation that could empower more educational stakeholders; it has a bias towards automated adaptation. This is in line with the goal of this study, where a system that is built will be used by the education stakeholders to predict the future performance of new students. That way, the automation will allow for the technology to be available and usable by a wider population of the education stakeholders. Additionally, EDM uses approaches and methods relevant to this study, such as binary classification techniques that groups students into the desired categories (Bhardwaj and Pal, 2012). In this study, the students were grouped into those that require high intervention and those that require low intervention.

Educational Data Mining has been looked at as a process of applying Data Mining techniques to data originating from the education sector with a view to resolving educational issues (Romero and Ventura, 2010). EDM uses Data mining as a means to detect useful and meaningful patterns from data (Romero and Ventura, 2010). The aspect of being an emerging field comes about because the DM techniques are lately being used in the area of education.

2.3 Theoretical Perspective of EDM

The goal of EDM is to understand learners and provide information about the learning process (Romero et al., 2010). The research in learning known as human learning that has existed for over a century is considered the origin of EDM (Shanks and John, 1994). The difference between EDM and human learning is that EDM uses accumulated data from students over a period of the learning process, whereas human learning utilises experiments where students are exposed to a designed experimental setup in a laboratory (Romero et al., 2010).

In this study, prediction was used in line with EDM; data mining (DM) techniques were used on the data obtained from the rural schools. These data were analysed for the purpose of solving the problem of under-performance in academic achievement among primary school students. DM itself is the process of analysising data in order to extract useful patterns (Fayyad et al., 1996). Lately, EDM, which is the application of DM methods that are specifically used with educational data has been used to achieve an understanding of the students and their learning environment (Baker et al., 2010). The main catalysts to the development of EDM is the increase in data generated from student learning processes that is stored in state databases (Koedinger et al., 2008). Additionally, the widespread use of e-learning and web-based education, especially in developed countries, has created large amounts of student data (Castro et al., 2007).

EDM in this research aims to utilise the data from an educational environment, specifically, rural areas of developing countries. The objective is not only to better understand the students, but also come up with an approaches that combine data and theory to motivate initiation of strategic intervention that could benefit the students.

2.3.1 Data Mining in Education

EDM systems follow a process that is similar to those followed in other application areas where DM is used, such as in business, medicine, genetics and others (Romero et al., 2004). DM has been extensively applied to e-commerce systems with the objective of increasing sales (Raghavan, 2005). In education, the pace of utilising DM has been much slower, although the situation is improving (Romero et al., 2010).

Data Mining is one of the most common applications of Machine Learning (Kotsiantis et al., 2007). In education, Data mining may seem very similar to its application in other domains; however, their are some three key areas that make the difference (Romero and Ventura, 2007):

Objective- The objective of applying DM in education is different from that of its application in other domain areas. For example, in business, the objective is to increase profits. This is a measurable quantity that is determined by the increase in the amount of money made. EDM on the contrary has applied objectives, such as, improving the student learning. Additionally, EDM has pure research objectives, such as searching for a deeper understanding of an issue in education. Such objectives are not quantifiable and may require unique measurement approaches.

Data- The various types of data in education are unique to that area of education; they have formats and relationships that are unique. For example, data drawn from Intelligent Tutoring Systems have a specific structure and relationships different from data from other educational systems such as e-learning or face-to-face education. This makes the application of DM in education a special case that cannot be generalised with other domains. It is an area that requires its own approaches.

Techniques- The special characteristics of educational data call for different data mining approaches. Some DM techniques may be applied directly, while, others have to be adapted to the unique problem that is presented.

2.3.2 Areas of EDM Research

EDM research has been proposed to exist in three main areas (Romero et al., 2010):

One of these areas is developing tools and techniques and determining the ones most suitable for a given educational dataset, including the determination of best practices for evaluation metrics and model fitting. Currently majority of EDM research is in this area leaving the other two areas unexploited.

The second area entails finding out suitable questions whose answers could be extracted from the data. Some of these questions have been asked by teachers for many years without answers. However, the questions lately have been answered more accurately using EDM. Examples of such questions are: do any students require extra tuition classes to score better grades? Or, which students will need counseling to avoid dropping out of school? Seemingly EDM has proven itself capable of utilising data in answering such questions. There is however, room to ask the data many more interesting questions that would contribute in growing the EDM field. Our research falls in this category where we ask the data, which students require high intervention in order for them to score better marks in national standardised examinations.

The third area of possible EDM research is finding out the beneficiaries. The obvious beneficiaries have been the teachers and the students; however, research could be conducted to find out how best parents could also benefit directly or indirectly. Similarly, other beneficiaries could be school principals and education officers. This area too has room for further research, especially in the area of expanding the list of beneficiaries of EDM.

2.3.2.1 Educational Systems in EDM

Educational systems are categorised according to the type of data source generating the data to carry out EDM. Researchers use many data sources that are mainly grouped into two types of educational systems: traditional classroom and distance education (Romero and Ventura, 2007). The traditional classroom has many variations including primary education, and higher education. Our study focuses on primary education. Compared to higher education, less research has been conducted with primary school datasets (see reviewed literature in Chapter 3). The sources of data in the traditional classroom is traditional datasets that contain: student information; educator information; and class information (Ma et al., 2000) . Distance education on the other hand combines all the systems where the student does not have face- to-face interaction with the teacher, such as: e-learning, or web-education. Web-education is the most commonly used type of distance education where education is delivered over the Internet (Johnson et al., 2000). There are three types of web-based education: particular web-based courses; learning content management systems such as Moodle; and adaptive intelligent web-based educational systems (Romero and Ventura, 2007). Our study falls under the traditional classroom system. The area where the study was undertaken lacks electricity and Internet connectivity that are necessary for distance education.

2.3.2.2 Beneficiaries of EDM systems

EDM has various stakeholders - different groups of users that have different mission, vision and objectives of using data mining (Hanna, 2004). These users include students, teachers and education officers (Romero and Ventura, 2010). The three selected users and their possible objectives are discussed next.

Students- the EDM system recommends activities and facilitates allocation of resources that could improve their academic performance (Romero and Ventura, 2010).

Educators/Teachers- to help teachers manage their students; to find out which student requires intervention; and understand the student learning process, and evaluate their teaching methods (Merceron and Yacef, 2005).

Academics responsible/Education officers- to facilitate them in allocating the school resources such as human and teaching materials; to facilitate their decision in utilising the available resources; and to evaluate teachers and the curriculum (Romero and Ventura, 2010).

This study identified these three stakeholders as the main beneficiaries of the EDM system that was built, but parents and guardians together with school administrators are also indirect beneficiaries.

2.3.2.3 EDM Cycle

When data mining is applied to educational systems, the process needs to be iterative, where a hypothesis is formed, tested, and refined (Romero and Ventura, 2007) as illustrated in Figure 2.1.

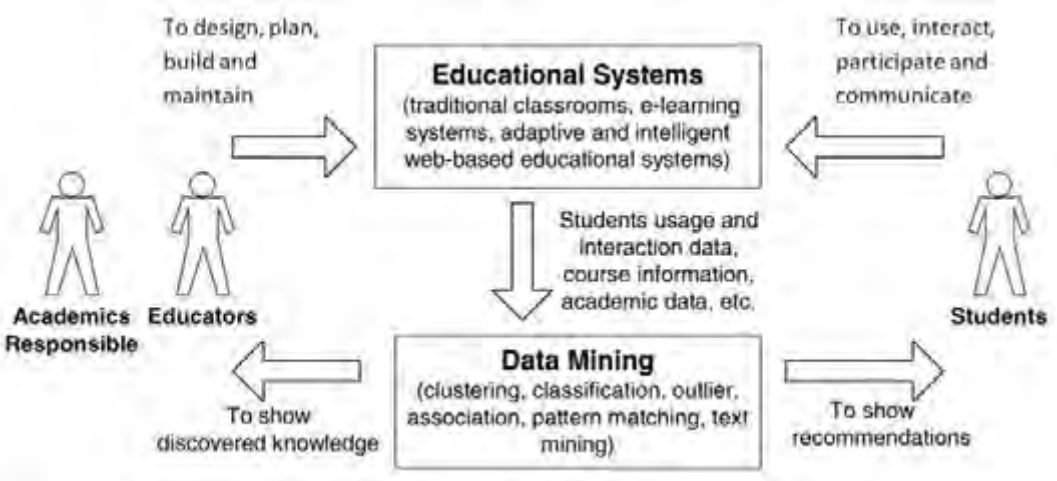


FIGURE 2.1: The iterations of applying data mining in education (Romero and Ventura, 2007)

Figure 2.1 shows that educators and academics are responsible for designing, planning, building and maintaining the educational systems (Romero and Ventura, 2007). Examples of these systems include: traditional classrooms, and e-learning. Students use, interact and participate with these systems. As they interact with the systems, data about their usage, interaction, and test marks are gathered. Different data mining techniques, such as clustering and classification, are then applied on the data in order to extract useful patterns that could be used to help the students achieve better

performance. The discovered knowledge is beneficial to all the education stakeholders, especially the teachers and the students.

2.3.3 Applications of EDM

Literature has suggested a number of applications of educational data mining. These applications are: those that focus on student and domain model improvement; those that determine pedagogical support while using educational software, and the general goal of understanding learners and the learning process (Baker et al., 2010, Baker and Yacef, 2009). Six applications were also proposed in another study, which include assessing learning progress, providing adaptation and recommendations of courses with reference to the student's capability, evaluating the learning curriculum, obtaining feedback required by teachers and students in web-based education, and finding out the distinctive qualities of a student's learning behaviour (Castro et al., 2007). The applications suggested by these authors have a greater emphasis on distant education, such as, e-learning. Another broad set of applications has been suggested which has eleven application areas: analysis and visualisation; providing feedback; recommendation; predicting performance; student modelling; detecting behaviour; grouping students; social network analysis; developing concept maps; planning and scheduling; and constructing course ware (Romero and Ventura, 2010).

The application area of this study falls under the category of predicting student performance. Prediction's objective is to find an unknown value about a student that points to performance, knowledge or score (Romero and Ventura, 2010). A value that is numerical is determined using a regression technique, while one that is categorical is determined using classification. Regression analysis aims at determining a function that defines a relationship between one or more features and a target (Draper and Smith, 2014). Classification on the other hand categorises records into groups depending on some identified characteristics that are discovered from the training set during model building (Espejo et al., 2010). Our study uses the classification method where students are grouped into two groups: those that require high intervention; and those that require low intervention to pass the standardised national primary school exit examination.

2.3.3.1 Classification for Prediction

As mentioned earlier, prediction is one of the EDM application that can be achieved using the classification technique. Classification has been defined as a process of allocating items into target classes depending on some determined characteristics of each item predetermined using a training set that has complete records (Espejo et al., 2010).

Classification has been defined as the task of placing a new record (test data) in their correct target classes based on some training data that are complete with the target classes (Tang et al., 2014). Therefore, to classify items in data mining is to predict that the given items belong to a given category of items (Krishnaiah et al., 2014).

Classification as a data mining task assigns records in a dataset to target classes with a goal of accurately predicting the target class for each record in the dataset (Archana and Elangovan, 2014). The first phase in the classification process is to classify data with known class targets. For example, in the case of this study, a model that predicts high intervention students is developed by using students' complete records - with the final examination results. The model is then used to predict the students that have not yet sat for the final examination.

The training process or the use of the training dataset to build classifier models finds a function between the independent variables (the predictors) and the dependent variable (the target) (Krishnaiah et al., 2014). It is this function which is the model, and is applied to a new dataset with unknown target classes. Testing the accuracy of a built model is done by comparing the output of the model on the new data with known target values for those records. Originally, the data for a classification task is divided into two datasets as suggested: the train set; and the test set. These datasets are used in each of the two phases of the classification process as described next.

Model building

Classifier model building is divided into two types: supervised - where records have known labels, the correct targets; and unsupervised - where records are unlabeled, called clustering (Kotsiantis et al., 2007). In clustering, the goal is to find unknown and useful groupings of the records (Jain et al., 1999). The present study concentrates on the supervised classification technique where a portion of students' records are utilised in the model building. The accuracy of the models built are tested as discussed next.

Model Testing

Once the model has been built, the next phase is for it to be tested with records in a dataset that have not been seen by the classifier, where the known target of every test record is compared with the model's predicted result (Krishnaiah et al., 2014). The classifier performance is determined by the percentage of correctly classified records in the test data. The test data has to conform with the train data in terms of the type and number of attributes. Both the train data and the test data must originate from the same database. A common technique used in determining classifier performance is cross-validation (Kohavi et al., 1995). In this technique, the data is divided into subsets of equal size. All but one of these subsets are used for training, and the remaining one is used for testing. This is repeated multiple times with a different subset being used

for testing each time. The mean error rater for all the data subsets becomes the error of the model (Krishnaiah et al., 2014).

2.3.3.2 Classification Techniques

The common techniques that have been used in classifying student academic performance are: Neural Networks, Bayesian Networks, Rule Based systems, and Regression and correlation Analysis (Romero and Ventura, 2010). Most of these have been used together in order to compare and determine the best technique for a given dataset. Some examples of studies that used classification include comparison of classification techniques to predict "pass" or "fail" in an Intelligent Tutoring System (ITS) (Hämäläinen and Vinni, 2006); prediction of students' final marks with Moodle Usage data (Romero et al., 2008); prediction of final grades on logged data (Minaei-Bidgoli et al., 2003); and comparison of artificial neural network, decision tree and linear regression to predict university students' academic performance (Ibrahim and Rusli, 2007). The present study selected six classification algorithms that fall in the categories of rule based, regression, neural networks, and Bayesian networks. These were chosen because they are the most commonly used algorithms in many studies, including those cited here. The classification performance of these classifiers were compared to determine which one among them is the most suitable for the data set used in this study.

2.4 Summary

This chapter has described the theoretical perspective of educational data mining, its origin and application in the educational field. The choice of EDM was justified by comparing it with Learning Analytics. Justification for using EDM was further established in four ways: (i) the aim of the present study was to classify students into two categories - high intervention and low intervention. Classification is a data mining technique that has been used with educational data; (ii) the objectives of applying DM techniques in education is different from applying it in the other domains. In EDM the objectives are not directly quantifiable - such as to improve student learning - hence the need for a unique treatment of the field; (iii) educational data is also unique depending on the environment or area. For example, primary school student data is different from university student data. Similarly, traditional education data is different from web-based education data. This further emphasises the need for EDM to be seen as a unique field of study; (iv) although the same techniques of DM may be used in EDM, some of the educational data with special characteristics will need different DM approaches. The

existence of different areas of research as mentioned by Romero and others in the handbook of EDM and the eleven application areas Romero and Ventura mentioned further confirms that EDM is an established area of research. Further, Romero, Espejo, and Tang and other researchers demonstrated that classification is a technique that can be applied with educational data. In these previous studies, students were grouped into different binary groups such as, pass and fail. In our study, early determination of the students that require high intervention will motivate initiation of strategic intervention by the educational stakeholders. This could reduce the number of students who fail and drop out of the school system.

Chapter 3

Literature Review

3.1 Introduction

Education Data Mining techniques have been used to build academic performance prediction models to predict student's future performance in academic institutions. The aim of this work is to provide a means to find out early enough those students that need help so that strategic intervention can be initiated. The research hopes to contribute towards solving the problem of poor academic performance. This chapter focuses on how previous studies have tackled the problem.

The chapter begins by reviewing the causes of school dropout and, relates school dropout to poor academic performance. This is followed by a review on how previous studies reduced the many causes of school dropout or poor academic performance to an optimal subset. Thereafter, the chapter reviews studies related to academic performance prediction modelling. Significant work has been done on academic performance modelling with different target classes, in different levels of education, and using different sources of data. The review aims is to see how these various models supported the prediction of student academic performance.

Thereafter, mobile phone use in education is reviewed, especially in the developing world. Their limitations to be used as hand-held computers is discussed. The chapter concludes by presenting a summary of the gaps this research can fill and looking at any opportunities to be exploited.

3.2 School Dropout and Poor Academic Performance

Students do not drop out of school for no reason. A study conducted to understand this complex phenomenon concluded that there are several causes that could possibly contribute. These causes, however, first contribute to the student's poor academic performance which in turn provokes the student to drop out (Fall and Roberts, 2012). Therefore, there is need to to understand which factors cause students to under-perform in their academic work and end up dropping out.

Several factors have been identified and presented in a conceptual model as illustrated in Figure 3.1

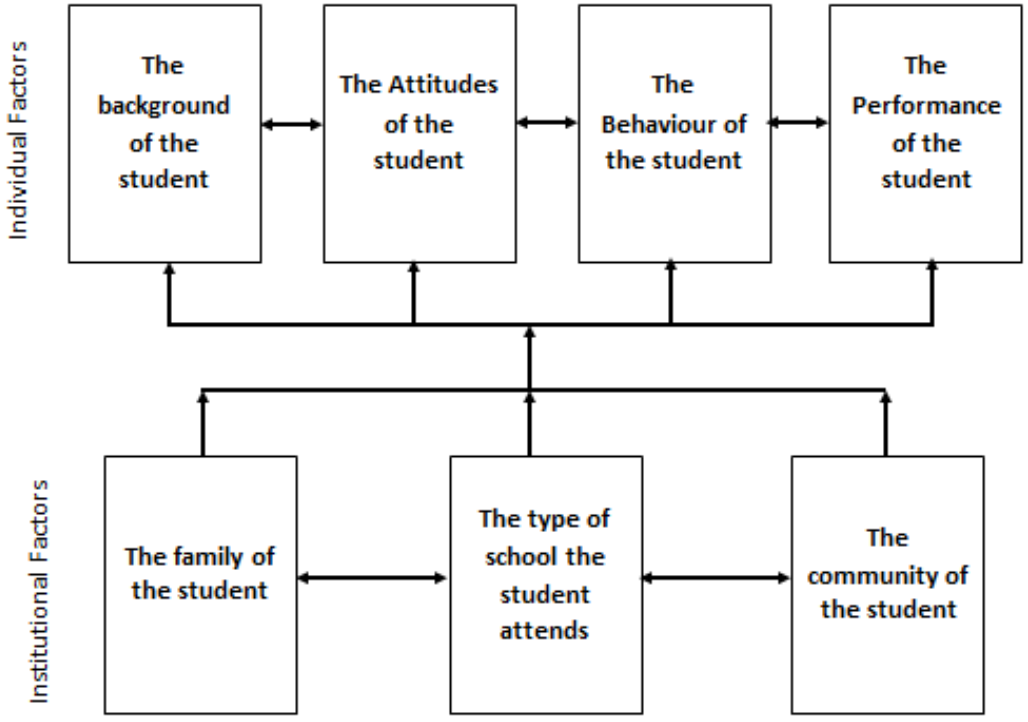


FIGURE 3.1: Conceptual model of student performance (Rumberger and Lim, 2008)

As illustrated in Figure 3.1, the factors that cause students to drop out have been categorised as: individual students' characteristics; and institutional characteristics. Findings suggest that: no factor in itself causes students to drop out; the students' academic performance together with their behavior directly affects their decision to dropout; the process of dropping out of school is pointed to by an accumulation of poor academic performance together with bad behaviour (Rumberger and Lim, 2008). These findings suggest that students drop out of school because they have performed poorly in their academic work. Selected factors in each category are discussed next.

3.2.1 Individual Student Factors

Students' individual characteristics include performance, behaviours, attributes, and background. As seen in the figure, these factors receive input from the institutional factors

Performance

Performance is linked to achievement, school transfers, and attainment. Academic achievement was investigated in an analytical study with Grade Six student data among urban poor schools and seen to be a cause for student dropout (Balfanz et al., 2007). The study determined absenteeism, getting poor marks in Maths and English, and poor rating on behaviour as the factors that make students seek to stop their education career prematurely. School transfer by students has also been identified as a cause for poor performance (Mehana and Reynolds, 2004). Similarly, an investigation on the effect of school transfers found that those students who changed schools, especially in the early grades, had a greater tendency of performing poorly in their academic work and were more likely to repeat classes (Turner and Thompson, 2015). The investigation also correlated school transfers to such factors as low family income, living with one parent, and parents with low academic attainment.

A student who progresses from one level of study to the next level is considered to make an academic attainment (Rumberger and Lim, 2008). A study found that many students who were retained in the same grade and hence become over-age were more likely to drop out and more unlikely to graduate (Silver et al., 2008) .

Behaviours

Student engagement, is one form of behaviour that has also been identified to have a relationship with academic performance (Finn and Zimmer, 2012). It refers to the students being committed to their school work, including attending classes and completing their assignments, and participating in core curriculum school activities. The study suggests that student engagement activities could be incorporated in the school policy as intervention measures for students who are likely to fail or drop out. Student behaviour is termed as deviant when they indulge in drug abuse, engage in romantic actions, and start families (Battin-Pearson et al., 2000). Such behaviour has been seen to increase the possibility of students dropping out of school. When bad behaviour is tolerated in school, studies have shown it to increase student dropout especially among elementary school children (Ou et al., 2007). Likewise, outside school bad habits have also been associated with dropout (Sweeten, 2006). Abuse of drugs such as cannabis (Verweij et al., 2013) and becoming parents while at school (Basch, 2011) has also been seen to contribute to the problem of poor academic performance and dropping out of school.

Further, peer pressure has been associated with bad behaviour; when the friends engage in bad behavior, it is a strong influence on the students (Chattopadhay, 2014). Lastly, when students are employed as teenagers, their time to be engaged in school work and homework is adversely affected, resulting in poor academic performance and school dropout (Monahan et al., 2011).

Attitudes Characteristics

Attitudes include goals, values and self-perception. Thoughts or feelings have been known to affect student value for school which determines the level of students' academic performance (Conley, 2012). Similarly, educational expectation has been seen as the most common indicator that represent an educational goal; students who aim for higher levels of education are less likely to drop out of school (Rumberger and Lim, 2008).

Background Characteristics

The final category within the individual factors is background: demographics, health and past experiences. Gender as one of the demographics has been investigated to determine its contribution to student dropout. Findings suggest that male students have a higher rate of dropping out than females students (Laird et al., 2007). However, gender has been associated with a number of other factors such as: family, academic background, attitudes, and behaviours (Rumberger and Lim, 2008). In the United States it was found that among the White population, female students had lower rates of dropout compared to the Black population (Crowder and South, 2003). A study on immigration status found that students who migrated to the United States had a higher dropout rate (Laird et al., 2007). This high rate has been explained by the low social capital in the families, the schools they attend and the communities they live in (Perreira et al., 2006). There was no direct correlation between ethnicity and race and student dropping out, though family background or educational performance was seen to explain some effects. Related was an investigation on the effect of proficiency in the English language. The study shows that students who had a good command of English also had lower chances of dropping out (Perreira et al., 2006). Lastly, in this category, it was found that forms of disability that impact on learning were associated with higher dropout rates (Reschly and Christenson, 2006).

Health is a key factor for determining whether or not students will drop out of school. One study found that students with perfect all round health were less likely to dropout of school (Roebuck et al., 2004). Another study found that students with mental symptoms indicated higher chances of dropping out of school (Daniel et al., 2006). Further, past experience, especially preschool education was associated with future students' academic performance and school dropout. The study found that preschool reduced the

tendency for student dropout and also enhanced academic output in higher levels of education (Barnett and Belfield, 2006). Preschool was also found to improve the chances of graduating by 22 percent (Gorey, 2001).

3.2.2 Institutional Factors

Institutional factors include families, schools, and communities. As mentioned earlier, they contribute to the impact that individual factors have on students' academic achievement or dropping out of school.

Families

Family background has been identified as an important determining factor in academic achievement. Findings show that when parents get involved with their children's schooling, the children improve in school achievement (Pomerantz et al., 2007). Family background can be split in to three phases: "family structure", "family resources", and "family practices". Family structure refers to the composition of the family in terms of the number of persons and whether there are both parents or a single parent. In cases where there is only one parent, the student is likely to miss out on financial support together with the close monitoring and supervision that both parents could offer (Martin, 2012). Such families that have only one parent increased the chances of dropping out of school (Perreira et al., 2006). Similarly, separated parents increase the chances of a student dropping out (Pong and Ju, 2000). Likewise, any form of stress on the students as a result of a parent's death or sickness has also been identified to increase the probability of dropping out (Alexander et al., 2001). It has also been established that any form of change in family structure could be a cause of changing the place of residence and school - denying the student valuable social relations and hence reducing the chances of academic achievement (Ream, 2005). Finally, family size was also seen to have a negative contribution to academic success - a large number of children in the family may mean limited resources distributed to each child (Rumberger and Lim, 2008).

Family resources is a key attribute in the families category. Resources have been categorised into financial, human, and social resources (Bornstein and Bradley, 2014). Socioeconomic status has been identified as a key indicator of school drop out (Bornstein and Bradley, 2014). Further, Parents' education level and hence the level of assistance has been identified as an indicator of academic achievement (Fall and Roberts, 2012). School dropout has been seen to be lower when parents have higher levels of education. Finally, family income was also identified as an important indicator, findings showing that students from high income parents are less likely to dropout of school and had better chances graduating (Dahl and Lochner, 2012).

Family practices are the actions that parents should engage in to improve chances of graduation for their children, parental expectation is the most outstanding parental practice indicator. Findings have shown that parental expectations have a reasonable positive correlation with student scores, and that higher parental expectations reflected lower dropout rates (Bowen et al., 2012). Parent ambitions for their children, parent availability when called upon by the school, parent interaction level with the school, and parents' network with others are key indicators. It has been established that the chances of dropout are significantly reduced when parents are close to their children (Perreira et al., 2006). Additionally, it was found that students with siblings who had dropped out of school were also more likely to drop out (Jacob, 2001).

Schools

School characteristics is the second category of institutional factors, which is classified into student characteristics, school structure, school resources, school practices and policy (Rumberger and Lim, 2008). The student characteristic indicators that have been identified as influential to academic achievement include the average socio-economic status (Bornstein and Bradley, 2014), the number of students who are at-risk of failing, the population of marginalised students, the proportion of transfer cases, and the proportion of students with problematic families. School structure factors have been identified as: school location - whether urban, rural, or suburban; School size - in terms of the number of students; and type of school - whether public or private (Rumberger and Lim, 2008). Similarly, the availability or non availability of free lunch program; the number of marginalised students; and the ratio of teachers to students have been associated with school structure factors (Dong et al., 2015). Incidentally, students from rural schools were found to have higher dropout rates than those in urban school (Heck and Mahoe, 2006). Further, school size mattered - the larger the school, the higher the chances of students dropping out (Rumberger and Palardy, 2005). The study also found out that public schools had a higher dropout rate than privately managed schools.

School resources have been identified to have the following indicators: the amount of money the government spends per child, the average salaries paid to teachers, the ratio of teachers to students, and the quality of teachers (Rumberger and Lim, 2008). Higher graduation rate has already been correlated with more money spend per child in the rural schools (Roscigno and Crowle, 2001). Teachers salaries have also been associated with student dropout; the higher the average teachers' salary, the lower the rate of student dropout (Rumberger and Palardy, 2005). Additionally, it was found that when the class is oversize, the chances of students dropping out increases (Rumberger and Thomas, 2000). The probability of students dropping out was particularly high in schools with oversize classes if the schools were located in rural areas (Finn et al., 2005).

School practices are an important category of school characteristics. They include such practices as: the teaching methods used, and the climate created for effective student engagement to promote learning (Hoy et al., 2006). Investigation into the correlation between positive school climate and student dropout found there is a reduction in dropout when there is a higher positive climate (Worrell and Hale, 2001). Likewise, reduced student absenteeism was seen to reduce the number of students who dropped out of school (Rumberger and Thomas, 2000). Further, it was reported that a well monitored learning environment that ensured students spend more hours in homework also reduced dropouts(Rumberger and Palardy, 2005). Parallel results from the same study found a relationship between poor discipline and increased dropouts. Additionally, it was found that students who did not have a good relationship with their teachers were more likely to drop out of school (Stearns et al., 2007), and that schools where teachers were involved in disciplining students and in curriculum issues had fewer cases of dropout compared to schools where teachers were not involved (Rumberger and Palardy, 2005).

Communities

Communities is the third category of institutional factors; studies have considered community composition and community resources (Rumberger and Lim, 2008). The influence of communities was categorised as: the ability for the community to avail services and opportunities to the children and youth; availability of influential family friends in the neighbourhood; and mutual relationships that can enhance positive input to the youth through supervising and monitoring their behaviours (Leventhal and Brooks-Gunn, 2000). Living in an affluent neighbourhood was seen to contributed positively to the academic success of the children (Chung et al., 2011). On the contrary, students who get exposed to violence have a higher rate of dropping out of school (Patton et al., 2012). It was also found that when there are employment opportunities that attract the youth, the rate of students dropping out increases (Warren et al., 2006).

3.2.3 Dropping out and Academic Performance

Academic performance and dropping out are seen as related outputs of a combination of factors as illustrated in the self-system model of motivational development in Figure 3.2

As seen in Figure 3.2, the model links institutional factors including family support, and teacher support, to individual factors such as: perceived identification with school; and perceived control. Further, these factors are seen to influence the engagement behaviour factors that in turn directly contribute to the two educational outcomes: academic achievement and dropping out. The two educational outcomes, the model suggests, are

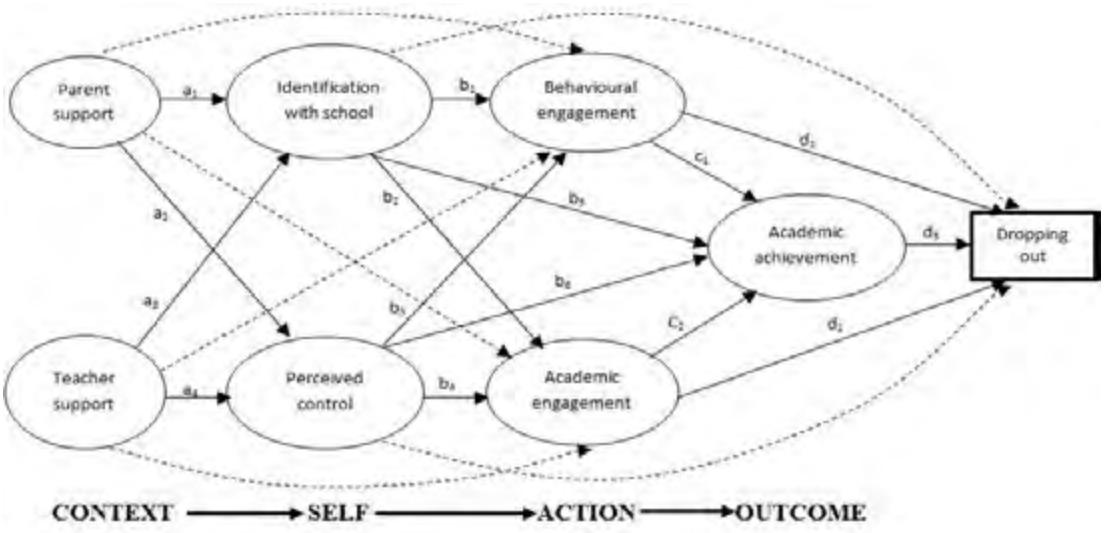


FIGURE 3.2: Self-System model of motivational development linking academic performance to dropping out, the solid lines show the direct contribution while the curved dotted lines show indirect contribution (Fall and Roberts, 2012)

equally affected by the engagement factors as indicated by the solid lines. Additionally, academic performance directly contributes to dropping out (Fall and Roberts, 2012).

Based on the model discussed, it can be concluded that the factors that cause dropping out also cause poor academic achievement to a large extent. This study adopted the view that poor academic performance is a cause of students dropping out of school. Specifically, in the developing world context, the study focuses on students in primary schools who drop out of school because they did not pass the standardised national exit examination. In this context, the students do not willingly dropout but are thrown out of the school system.

3.2.4 Causes of Poor Academic Performance in Developing Countries

A focus on Free Primary Education

In many Sub-Saharan countries, the problem of poor academic performance came about with the introduction of Free Primary Education (FPE). FPE has been defined as a policy of eliminating or abolishing school fees to a section or the whole of primary school cycle (Abuya et al., 2015). It is an initiative that was motivated by the inclusion of Universal Primary Education (UPE) in the Millennium Development Goals (MDG) (Poverty, 2015). Many sub-Saharan African countries adopted the policy, some examples are: Malawi(1994), Uganda(1997), Tanzania and Lesotho in 2000, Burundi, Rwanda, Ghana, Cameroon and Kenya in 2003 (Unesco, 2001).

Kenya introduced FPE in 2003, the third initiative after two previous attempts: the first initiative in 1974 by the first president had a massive impact of 150% increase in Class One intake and the second with more than 60% increase. This third initiative resulted in an increase in Class One enrollment of 35% from 0.969 million in 2002 to 1.312 million in 2003. However, in all the three cases, the impressive initial gains were quickly eroded, one to two years later, the enrollment gradually reducing and the dropout rate increasing, largely because of the increased poor academic performance (Somerset, 2010).

Therefore, although FPE had the positive effect of increasing enrollment which is a step towards attaining UPE, an important MDG, it brought about a number of negative effects (Somerset, 2010). Further, findings by the same study suggest that the most outstanding negative effect is the lowering of the quality of education. This is a phenomenon that was noticed across a number of nations that adopted the FPE policy, such as Kenya, Lesotho, Malawi, and Uganda. The literature attributes the weakening quality to the large influx of students which cause "access shock". Classes become overcrowded and are forced to learn in double and triple shifts. Further, there is a severe shortage of teachers, and both teaching and learning resources. Additionally, there were many overage students who were only fit to join adult classes (Abuya et al., 2015). These challenges turned what was meant to be a good and positive policy to become a paradox by compromising the quality of education in public schools (Orodho et al., 2014).

What causes poor academic performance?

Several reasons explain poor academic performance of students, as stated earlier. In Africa, some of these causes emanate from the introduction of FPE. These reasons are said to include school based factors, home based factors, Government policies and an overlap of the three (Orodho et al., 2014). Other findings suggest that the causes of poor academic performance in schools are directly related to: school factors (lack of teaching materials, textbooks and trained teachers); teacher factors (lateness, absenteeism, using local language, and poor syllabus coverage); pupil characteristics (truancy, lack of help on studies at home, low interest in lessons); and parent characteristics (failure to provide breakfast, text books, basic school needs, and low interaction with teachers) (Etsey, 2005).

Among the school factors, teacher effectiveness is the most important indicator in student achievement. Literature indicates that an effective teacher has qualities including knowledge and ability to organize the subject matter, skills of instruction, a positive attitude, clear communication with students, is respectful and fair, has concern for student learning, gives fair assignments and assessments and gives timely feedback. Additionally, the teachers' effectiveness is closely related to factors such as: adequate and relevant

teaching and learning materials, and a good environment within the school (Orodho et al., 2014).

Parents' socio-economic status has also been associated with poor academic achievement. For example, concepts are seen in Korea such as the "tiger mom" phenomenon or "academic zeal" of parents; that means parents have increased involvement in children's academic performance (Pears et al., 2008). The quality and extent of parental involvement and investment in children's academic performance could be limited by the socio-economic status of the parents. Parents have to pay for extra classes to tutor the children. Children with parents who do not have money to pay for extra classes are therefore at risk of remaining behind in the academic progress (Bae and Wickrama, 2014).

Therefore, parental support and performance expectations could be among the most important factors that influence student behaviour and academic performance (Chen and Gregory, 2009) . Findings of a study on the relationship between parent expectations and end-of-grade performance in Maths and reading supported the concept that, higher expectations predict higher performance (Bowen et al., 2012). Further, communication and collaboration between parents and teachers could improve parental expectations for their children and the expectations of the school settings (Wegmann and Bowen, 2010).

Negative stereotype contributes to poor achievement in academics, as it diverts useful cognitive resources that would be utilised in the learning process (Bowen et al., 2013). Stereotype, has been defined as a belief associated with negative meaning towards a community. Social psychology research has established the power of negative stereotype to hinder academic performance of students from stereotyped regions or groups (Bowen et al., 2013). Stereotype threat is defined as a fear of doing something that would possibly confirm what people say; it has the capacity of causing under-performing and also hinder the learning process (Mangels et al., 2012).

A focus on Kenya

Literature focused on the Kenyan case has revealed specific and related factors that affect academic performance in primary schools (Gakure et al., 2013). The causes have been identified as the following: inefficient leadership in the school administration; lack of educational facilities; failure to meet the educational needs of students; no preparation by teachers and no homework given to students; poor characteristics of teachers in terms of behavior and professionalism; negative social influence; large class sizes; having several streams per Class; small amount of time allocated to teaching and learning; lack of teacher commitment in class; lack of parental care and advice; lack of teacher supervision by head-teachers; negative attitude by teachers and other stakeholders; lack of teamwork amongst teachers; absenteeism and lack of commitment by students; insufficient learning

materials; lack of the a spirit of competition amongst students and schools; external and political influence of appointment and transfer of head-teachers; and over enrolled classes due to Free Primary Education (Gakure et al., 2013). These factors could effectively be grouped into five categories: school management and administration factors, teacher factors, student factors, parent factors, and environmental factors.

3.2.5 Summary of Causes of Dropout

So far in this section, we have looked at the many causes of school dropout or poor academic achievement in the literature. A large list of these causes have been extracted from the literature reviewed. These are categorised as individual causes and Institutional causes. In Kenya, the causes poor academic performance could effectively be grouped into five categories: school management and administration factors, teacher factors, student factors, parent factors, and environmental factors. Clearly, the list of the individual causes is large, and there is need to select only those causes that are most predictive of academic failure in a given environment. Initially, we conducted a survey in the study area to understand the causes of poor academic performance on the ground. However, the list was still large and could not be implemented on a small screen mobile phone. The next section therefore discusses feature selection, a preprocessing step that reduces attributes to an optimal subset.

3.3 Optimal Feature Subset Selection

3.3.1 Why Feature Selection?

This subsection presents literature on feature subset selection as a preprocessing step within the data preparation step of the CRISP-DM process. It is defined as a process of reducing the size of a dataset by eliminating some of the irrelevant features; it improves classifier performance and enhances clarity of results (Yu and Liu, 2003). Feature selection has also been defined as as a process of selecting the smallest size subset of features that achieved the best possible classifier performance (Kohavi and John, 1997). Therefore, the objective of feature selection is to determine a subset of features, also called attributes, or independent variables, that will enable a classifier to perform optimally (Jain and Zongker, 1997). Further, It has also been seen as the process of finding a subset that does not adversely reduce the performance of the classifier (Pudil et al., 1994). Additionally, more recently, it has been defined as the process of determining a smaller number of predictive features from all the features available (Liu and Yu, 2005).

Findings have shown that classifier performance could be increased by eliminating irrelevant or redundant features in the list (Collins et al., 2005). These early studies establish the fact that feature selection is an importance step in EDM.

Feature selection, therefore, determines the most predictive minimum number of attributes that will be used to predict the target attribute. The process improves the effectiveness of classifier training and enhances the prediction performance, in addition to eliminating difficulties of understanding the results (Tang et al., 2014). For example, an improved classifier accuracy and simplification of results was achieved when a small dataset was used after feature selection (Kumar and Kumar, 2011). Further, the study reported improved effectiveness in terms of computer resource saving. Other identified advantages of feature selection include reducing the classifier training time, enhancing the ability for the classifier to generalise, and to facilitate a greater comprehension of the area of study (Peteiro-Barral et al., 2013). Therefore, it is important to use only the features that are indicative of the target classes (Cawley and Talbot, 2010)

3.3.2 Feature Selection Techniques

Feature selection techniques aim to select an optimal subset of features; one where all the features contribute to predicting the target class (Huang, 2015). There is a possibility of having several selected subsets with different sizes. A general guideline is to select a subset with the least number of features.

Common techniques that have been used for feature subset selection include filters, and wrappers (Saeys et al., 2007). All the techniques follow similar steps to achieve the selection process. These steps are illustrated in Figure 3.3.

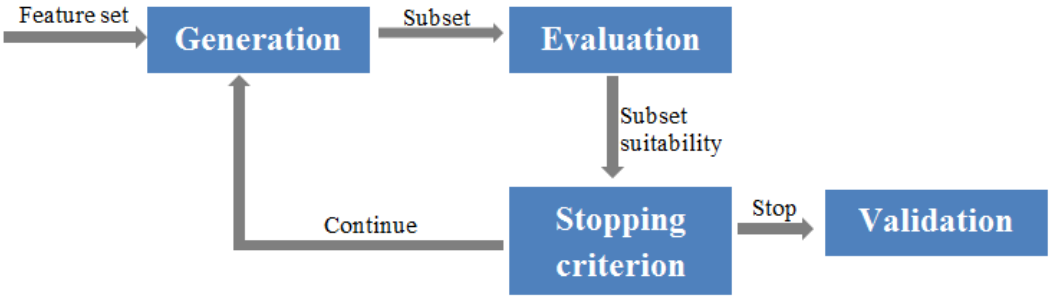


FIGURE 3.3: General feature selection process (Harb and Moustafa, 2012)

Figure 3.3 illustrates the proposed four feature subset selection steps: 1) subset generation - a sequence of steps that generates the subsets to be assessed according to a selected procedure; 2) evaluation - each subset is assessed and contrasted with the foregoing best

subset, the new subset is adopted if it is superior to the previous one; 3) stopping criteria - regulates the point at which the search process for a best feature subset may terminate; 4) results validation - achieved by observing the difference in classifier performance with the different feature subsets (Harb and Moustafa, 2012).

Filters

Filters are a technique that performs subset selection independent of the classifiers; they rate the predictive ability of the feature subset using the characteristics of the data (Bolón-Canedo et al., 2013). Filter algorithms rank the features in order of their predictive ability. Filters are preferred because they have the following advantages: they take a much shorter time to determine subsets as they utilise less computer resources; they can be used with increased volumes of data; and they are not attached to any classifier (Saeys et al., 2007). Their shortcoming is the assumption that features are separate entities and the fact that they operate independently of the classifiers. Filters have been identified as having an ability to operate using two approaches: features can be ranked using some predictive scale -"ranking method" or, they can be determined by the process of maximising a determined cost function - "space search method" (Lazar et al., 2012). Ranking is the more common approach in which, the optimal subset is achieved by picking the high ranked features and eliminating the ones that are low in the rank. The ranking approach is in four steps: 1) determining a scoring function for ranking the features from the most predictive to the least predictive; 2) approximate the "statistical significance" scores; 3) picking out the highly predictive features; and 4) validate the determined feature subset (Tang et al., 2014).

The steps are illustrated in the Figure 3.4

The space search approach is achieved in three steps: 1) determine the cost function that needs to be optimised; 2) determine the feature subset that optimises the cost function using an optimisation algorithm; and 3) validate the determined feature subset (Tang et al., 2014).

An example of a feature selection study in educational data used six filter techniques to determine an optimal subset (Ramaswami and Bhaskaran, 2009). The techniques used include Correlation Based (Lazar et al., 2012), Chi-Square (Jantawan and Tsai, 2014), Gain Ratio(Jantawan and Tsai, 2014), Information Gain (Jantawan and Tsai, 2014), ReliefF (Lazar et al., 2012), and Symmetric Uncertainty (Lazar et al., 2012). The dataset used consisted of higher secondary school students in India. The initial set had 33 features. The features were ranked using each of the six techniques, in order of their predictive ability, from the most predictive to the least predictive. Different techniques achieved different ranking. To determine the best subset, the features were systematically input into a Naïve Bayes classifier (Feng et al., 2013), beginning with two

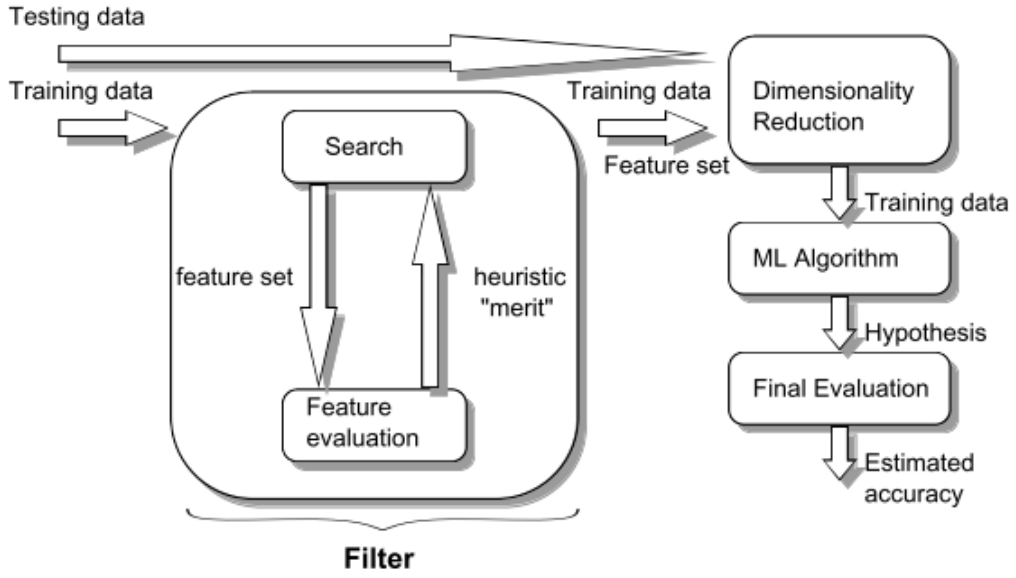


FIGURE 3.4: The filter feature selection approach (Hall, 1999)

of the top features in each of the six ranked sets. Iteratively, features were added one at a time. The classifier performance for the cumulative feature subsets were determined using Receiver Operating Characteristic (ROC) metric (Pampaka, 2011) and F1-Measure metric (Powers, 2011). Results indicated that Correlation Based and Information Gain techniques attained the highest ROC value. Information Gain, however, attained the highest value with only 7 features compared to 9 features of the Correlation Based technique. The results as indicated by F1-Measure achieved the highest classification value with three techniques: Information Gain, Symmetric Uncertainty, and Chi-Square. The highest value was attained with 12 features. Finally, each of the optimal subsets as determined by the accuracy of the Naïve Bayes classifier was used as input into four classifiers: Naïve Bayes, Voted Perceptron (Du and Swamy, 2014), OneR (Suganya and Sumathi, 2014), and PART (Oliver and Hand, 2014). The results were highest for all the classifiers when the 7 feature subset obtained using Information Gain technique was used. The results support what is known; that performance of classifiers is enhanced with an optimal number of features and that the reduction of the database size implies less computational resources in both training and prediction stages.

The effectiveness of filter techniques in reducing the dataset size was also demonstrated in a study where, using 10 filter algorithms, 77 features in an educational dataset of 670 records were reduced to only 15 (Marquez-Vera et al., 2010). The experiments were conducted in WEKA (Hall et al., 2009). The algorithms ranked the attributes and the top attributes were picked by their count of selection by each algorithm. The total 15

selected optimal features were determined by their predictive performance using several classifier models.

Wrappers

Wrappers have been described as a technique that uses the predictive accuracy of a selected classifier to assess the suitability of a determined feature subset (Tang et al., 2014). Wrappers process is in three steps: 1) searching for the feature subset from a collection of all the generated subsets ; 2) determine the accuracy of the chosen subset with the selected classifier; and 3) iterating the first and second steps until the anticipated level of performance is attained (Tang et al., 2014). The selection process in the classifier is hidden; that is the feature evaluation cross validation as seen in Figure 3.5. Step one produces the subsets whose performance is given by the classifier before being returned to search stage for the next iteration. The process is terminated when a feature set with the highest performance is determined.

The process is illustrated in the Figure 3.5

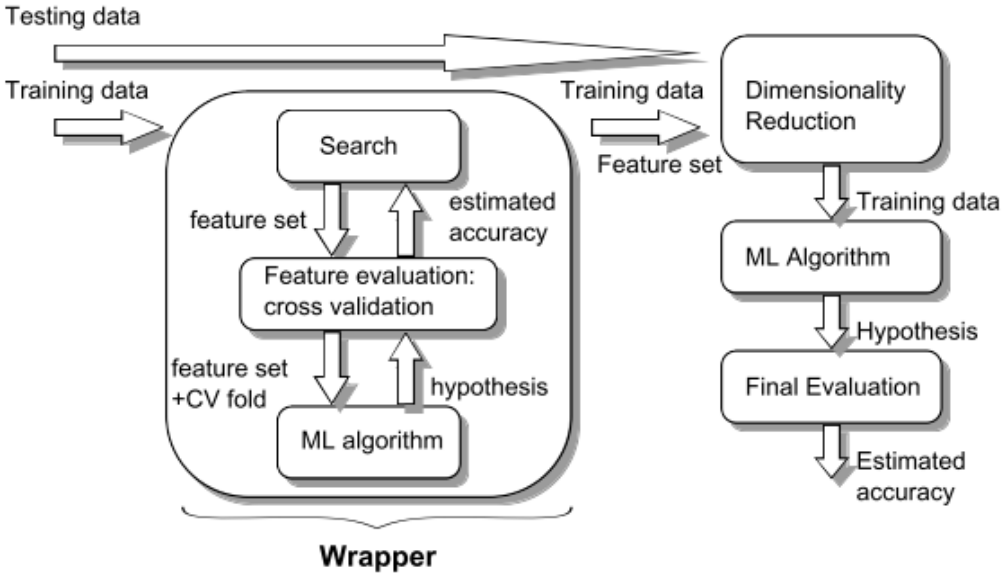


FIGURE 3.5: The wraper feature selection approach (Hall, 1999)

A few studies have used wrappers on education data. One study extracted the most predictive features from a set of 15 features obtained in the ASSISTments platform database (Harb and Moustafa, 2012). The dataset was obtained from grades 4 to 10 students' records in suburban schools in Massachusetts. Six classifiers were used: J48 (Bhargava et al., 2013), Instance-based K-nearest neighbour (Garcia et al., 2012), KMeans clustering (Celebi et al., 2013), Naïve Bayes updateable (Panda et al., 2010), OneR, and Voting Frequency Interval VFI (Malviya and Umrao, 2014). The three standard steps of subset selection were followed. First, the six classifiers were used to rank the 15

features, such that the top three features in each ranked set were taken as the initial subset. Second, the classifiers were used on successive features as they were increased from the three features to the last one. Next, the best prediction accuracy was noted and the prediction performance of the six classifiers compared. The best classification results were obtained using a 7 feature subset ranked using the KMeans clustering algorithm. The highest accuracy was attained by three classifiers: VFI (87.48%), Naïve Bayes updateable (87.41%), and ONER (86.94%). This study emphasised the benefits of feature selection in terms of achieving nearly 80 percent reduction in the data used for training while maintaining good accuracy.

The next section presents a review of the use of educational data to predict academic performance or probability of students at-risk of failing. Some of the studies employed feature selection techniques as part of the preprocessing step.

3.4 Academic Performance Prediction Modelling

This section presents a discussion of the studies related to academic performance prediction modelling of students. Significant work has been done in building prediction models on PCs. The section reviews how these models have supported the prediction of student performance. The discussion will focus on studies that aimed to classify students into binary target classes. Further categorisation will be on models that were built for different levels of students and using different datasets from different sources.

3.4.1 Binary Classifier Prediction Models

Binary classification separates items, like the student records in this study, into two categories: high intervention and low intervention. A significant number of studies have been conducted using binary classification on student datasets to achieve various objectives. To focus the discussion, this section presents the studies conducted using various types of datasets: (i) traditional classroom dataset; (ii) e-learning dataset; (iii) MOOCs datasets; and (iv) intelligent tutoring system datasets. Each of these categories may also have different levels of education, including primary schools, and higher education.

3.4.1.1 Traditional Classroom Dataset

A significant number of studies have used traditional classroom data to build academic performance models. This section discusses two categories of studies. The first category is studies that built academic performance prediction models using primary school data

or a combination of primary school and secondary school datasets. The second are studies that built models using higher educational data.

Primary school level

There is a limited number of studies that have focused on data mining models to predict primary school student performance. Many of these studies used PCs and were mostly conducted in developed countries. To focus the discussion, a number of studies are reviewed.

Predicting students into a binary class of, 'at risk' and 'no risk' of failing Grade 8 assessment in Maths and Science was conducted in the USA (Tamhane et al., 2014). Features were generated from: test scores from earlier grades, demographics, and behavioural factors. Being a longitudinal study, they were able to obtain a large dataset. The data was available in state databases. A number of models were built using SPSS (Xiao et al., 2015) and WEKA (Durrant et al., 2014). The accuracy of predicting at risk students was about 90% using the Receiver operation characteristics (ROC) metric. Logistic Regression was found to be the better classifier compared to Na´ive Bayes and Decision Tree. The model was found to be useful for deployment to predict students in the fail or pass categories as early as Grade 5. Similarly, a study predicted a binary class of 'success' or 'failure' using Grade Seven and Grade Eight data of 5000 records in secondary schools placement tests (Şen et al., 2012). The study used features that were identified as the most indicative of the target class, they included end of year exam scores for the two years in primary school, scholarship, size of family, scores in language exams, and mean grade points in Grade Seven and Eight. Four classifiers were built including Artificial Neural Network, Support Vector Machine, Decision Tree (C5), and Multinomial Logistic Regression (Neupane et al., 2015). The accuracy metric predicted the failing class up to 95% accurately using the decision tree (C5) algorithm.

In a related study, a combination of both elementary and secondary school student data was used with a target class of 'pass' or 'fail', to predict secondary school students likely to fail in their exams (Marquez-Vera et al., 2010). The study used a small dataset of 670 students with 77 attributes. Data sources included surveys conducted with the students for demographic information, the national evaluation centre to gather admission and socioeconomic data, and end of year examination scores. Feature Selection was conducted to reduce the features to15. Ten classifiers were built using algorithms found in WEKA (Han et al., 2011); five were rule based and five were tree algorithms. This dataset was, however, too small and imbalanced. To improve the imbalance in the classes, SMOKE (Chawla et al., 2002) was used. The improved results were obtained using JRip algorithm at a True negative rate of 93.3% and an F-Measure of 94.6%.

These studies indicate the possibility of building binary classifiers with different sizes of dataset, ranging from a large dataset obtained in a longitudinal study to a small dataset, where data was collected using questionnaires. Different techniques were applied, indicating, no single technique did well in all types of datasets. A weakness of the studies is that they did not use various metrics measures to validate their results, they simply selected the metric that attained the highest value. Additionally, they did not develop a tool that could be used in developing countries.

Higher education level

A significant number of studies have built binary classifier models for higher education academic performance prediction. This section discusses such works. The aim of the review is to see how binary classification was achieved and applied in different settings.

Investigating the prediction of 'passing' or 'failing' of university students was achieved by comparing the prediction performance of Classification and Regression (Strecht et al., 2015). The study used 5779 records with attributes such as: demographics (age, gender, marital status, nationality, displaced or not, special needs); enrollment data (admission type, student type, status of student, year of admission, course nature, dedication type); and financial matters (scholarship, debt situation). The classifiers and regression models selected include K-Nearest Neighours, Random Forest (Liaw and Wiener, 2002), AdaBoost (Rätsch et al., 2001), Classification and Regression Tree (Lewis, 2000), Support Vector Machine (Tong and Koller, 2002), and Naïve Bayes. Evaluation was done with the F-Measure metric on a 10-fold cross validation (Refaeilzadeh et al., 2009). Results indicate that the classification models achieved reasonable results. However, the regression models that aimed to predict a student's grade did not achieve good results. Similarly, students 'at risk' or 'not' of proceeding to the second year of their degree program were successfully predicted (Agnihotri and Ott, 2014). The study used 25 attributes that comprised of data from admission records, university entry exam data, survey data from the students, and financial data. The participating students numbered 1453, of which 983 proceeded to second year and 470 dropped out after the first year. The data was split into 70% training and 30% testing. Four classifiers were used: Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression (MacKenzie and Peng, 2014). Prediction performance was measures using recall and precision metrics. Results indicate that Logistic Regression attained the best Recall value of 62%. It was used to build an ensemble model, which attained a recall value of 74%; the model was capable of predicting 74% of the students who would not proceed to second year. The model was useful to the counseling department; it could identify students with various challenges as indicated by the attributes. Appropriate intervention measures could therefore be put in place.

An investigation to predict university students that would drop out or not, after semester one, was conducted before students enrolled in the university (Dekker et al., 2009). The goal of the study was to determine the students that are likely to drop out early enough to strategise on how to keep them in the university. Three datasets were used including pre-university, university grades, and a combined dataset. The total sample contained 1527 records. Six classifiers found in WEKA (Hall et al., 2009) were used, including decision trees, Bayesian Network, logistic regression, rule based learners, Random Forest, and OneR. OneR was selected as the baseline classifier (Jiang et al., 2005). Using 10-fold cross-validation, three pre-university features - science score, main score, and maths score - achieved the highest information gain value. An accuracy of 71% was achieved with the J48 classifier while an accuracy of 81% was achieved with the university grades dataset using the CART classifier. The combined dataset attained 80% accuracy. The study demonstrated the usefulness of pre-university and/or the early university data for predicting students who could drop out. In another study, the causes affecting passing and failing of first year college students were investigated (Gray et al., 2013). The study used features that have been known to affect academic performance including age, gender, previous exam results, and psychometric features relevant to performance. A total of 636 student records were used, 296 of which were passes and 340 were failures. A set of six classifiers were used to group the students into binary classes - poor achievers and strong achievers. Further, to overcome the distribution imbalance in the data, oversampling the minority class was used. Several experiments were conducted with the complete dataset and splits of the dataset. The highest result was attained with Support Vector Machine when the dataset was split according to students' age. When the complete dataset was used, a performance of 73.82% was attained.

An investigation of successful and unsuccessful university students was conducted with a focus to identify individual student factors that contribute most to the passing of university students (Guruler et al., 2010). A student knowledge discovery system was built as a workbench for the process. The study used the following attributes: information on registration, previous school information, university entry exam scores, family status, and family finances. The target attribute had two classes: successful or unsuccessful. A filter feature selection, correlation based method was used to determine the most predictive features. The ranked features below a threshold value of 0.01 were removed from the list. The highest prediction rate was achieved with the decision tree classifier, while the most useful attributes were identified as those generated from information on registration and family financial status.

In Australia, the possibility of predicting final exam performance in terms of high performance or low performance was investigated using 220 first year university student records in a computer programming unit (Luo et al., 2015). The study used attributes

such as: scores in assignments, degree type, local or a foreign; and some demographic attributes. Experiments were conducted with both standard decision tree (DT) and association rules(AR) classifiers and the discrimination aware (DA) versions of the two. Results indicate that the standard classifier attained the highest accuracy of 83.46% even though all the other three classifiers attained accuracies that were nearly similar to the best classifier.

In a different approach, study habits in programming were used as indicators to predict whether college students will pass or fail a mathematics course (Vihavainen et al., 2013). A total of 52 students participated. Snapshots of their habits while programming were collected using a plug-in. Thousands of snapshots, 48, 000, were collected and used to generate features. Modelling was carried out using Bayesian networks. Accuracy, recall, precision and F-Measure metrics were calculated. The study was able to record an accuracy of 84.6% likelihood of a student failing their mathematics course only after five weeks of programming lessons, giving time for possible intervention. In a another study university students who might fail and hence drop out were predicted (Bayer et al., 2012). The students' social behaviour data in addition to demographics, semester related data, and data from other studies was used. A list of 30 features were compiled. Initial classification results gave poor results, which prompted the feature selection process resulting in seven most relevant features. Eight feature selection algorithms were used, all based on the filter method. Classifier models were built in the WEKA environment: decision trees, lazy learners, rule learners, support vector machines, and Naïve Bayes. In an attempt to improve accuracy, ensemble techniques (Banfield et al., 2007), such as bagging; a technique used to improve the stability and accuracy of algorithms by minimizing variance and avoiding overfitting (Bühlmann, 2012) and voting; a strategy where classifier results of classifiers' decisions are put together and the class that appears more time is selected (Site and Mishra, 2013) were used. Results indicated the best performance was obtained using the rule learner. The combined dataset that included the social behaviour data, attained an accuracy of 93.67% and a True Positive Rate (TPR) of 92.30%. Experiments using the dataset without the social behaviour data attained an accuracy of 82.53% and a TPR of 78.50%.

The number of academic performance prediction models built with traditional education data for higher level education is much higher than that of primary schools. Probably most researchers find higher education data easily available than primary school data, given that most research work is conducted in universities. Clearly there is need for more data mining studies to be done with primary school data. This level is more critical for a student to drop out that at higher levels. A disadvantage that developing nations have is the lack of national databases where researchers can access student data. With such facilities missing doing EDM work is difficult, which explains the scarcity of literature on

EDM in developing countries. EDM work in rural schools is therefore still unexploited and this study aims to contribute towards filling that gap.

3.4.1.2 E-Learning Dataset

There has been a significant number of studies that fall in the category of prediction of student academic performance using binary classification techniques on e-learning and learning management systems data. However, these mainly fall in the higher education category. These studies have mostly been conducted in developed countries (Baker et al., 2015).

Higher education level

An investigation was conducted to detect at risk and no risk students in a university using learning management system activities (Romero and Ventura, 2010). The activities consisted of accessing relevant materials, doing assignments, and scores in assignments. A dataset of 4002 student records was used, having 24.7% at risk students. Classification task was conducted using algorithms including Logistic Regression, Step Regression (Jurečková and Picek, 2005), W-Kstar, W-J48 (Elmadani et al., 2015), and Naïve Bayes. Results were determined using the kappa measure (Greer and Mark, 2015), precision, recall, and accuracy. Logistic Regression attained a recall value of 59.5%, a precision of 56.8%, a kappa value of 34.4% and an accuracy of 66.2% in detecting the at risk students. The authors suggested the accuracy was sufficiently high for the purpose of initiating intervention for the at risk students. In a related study, an investigation to predict dropout or retained student target classes for American universities in an e-learning platform was conducted (Tan and Shao, 2015). The study used a large number of students' records (62,375 in total). Attributes were derived from personal characteristics and academic performance. Three classifiers were built, including Artificial Neural Network, Decision tree, and Bayesian network. Performance was measured on the test data (30% of the total). A confusion matrix generated the matrices that included accuracy, precision, recall, and F-Measure. Decision Tree attained the best results, with an accuracy of 94.63%, F-Measure of 71.91%, recall of 82.22%, and a precision of 63.89%. The authors concluded that the results indicate a reasonable ability in identifying the dropout class.

Further, an investigation to predict whether a student will pass or fail an important exam using online activities was conducted (Macfadyen and Dawson, 2010). A dataset of 118 undergraduate students pursuing an online Biology course was used. To determine the students likely to fail, fifteen attributes were identified and grouped, such as:

the number of discussion messages the students posted, the frequency of emails sent, and the overall number of assessments successfully completed. The classification process was carried out using logistic regression. Results indicate the fail and pass target class was identified with an average accuracy of 73.7%, while an accuracy of 80.9% was achieved for the fail category alone. The fail category was the focus of the study; this is the category of students that require intervention. Similarly, an investigation was conducted to predict students at risk of dropping out using data mining in Turkey on an online education program (Yukselturk and Ozekes, 2014). A total of 189 student records having 10 attributes was gathered for the purpose of the study. Data was collected with the help of questionnaires. The 10 attributes were: gender, age, level of education, earlier experience of online classes, type of employment, self-efficacy - a belief in ones' capacity to learn student, online learning readiness - the qualities that one has that are necessary for taking online classes , prior knowledge - the knowledge acquired by a student before the join a class, and locus of control - the quality of feeling in charge of situations. The target class was whether the student will drop out or not. Four classifiers were employed for the classification task: K-Nearest Neighbour, Decision Tree, Naïve Bayes, and Neural Network. Feature selection was conducted using Genetic Algorithms (Sivaraj and Ravichandran, 2011). Three features were identified as the most predictive: self-efficacy on doing courses online; students' online learning readiness, and prior experience of taking online courses. Results indicate K-Nearest Neighbour attained the highest sensitivity of 87% prediction of the target class. Finally, an early study predicted whether or not university students will drop out from a distance course (Kotsiantis et al., 2003). The goal was to reduce the dropout rate for university students taking a distance learning course. The study determined the best performing algorithm and a web-based system was implemented. Data was obtained from Hellenic Open University. The sample contained 354 records. The attributes were categorised as: background information (sex, age, marital status, e.t.c); academic performance (test scores, attendance of meetings with tutors). The target attribute was binary (dropout or not). The selected classifiers included decision tree (C4.5), Artificial Neural Network (Back Propagated), Naïve Bayes, Instance based (3-Nearest Neighbour), Logistic Regression, and Sequential Minimal Optimisation (SMO). The data was divided into five parts. Experiments were conducted with each part to determine the most influential group of factors. A new set of data gathered by the tutors was used as the test data. Results indicate the usefulness of the techniques in determining the students likely to drop out early enough with an accuracy of 83% attained by Naïve Bayes.

The use of e-learning data to develop academic performance prediction models has been successfully done for higher education. This is because e-learning is more applicable in higher education that primary school students who need close monitoring. Clearly,

e-learning has benefited students and professionals in developed countries because of availability of Internet connectivity (Baker et al., 2015). Unfortunately, most developing countries have missed out on such opportunities because of scarcity of resources. Internet that would be used for e-learning is completely not available in most rural areas of African countries.

3.4.1.3 MOOCs Dataset

Massive Open Online Courses (MOOCs) offer distance education through online courses that are free, unlimited and support interaction between learners and professors (Lewin, 2013). The data generated as students take courses has been used in mainly in higher education academic performance prediction modelling.

Higher education level

A recent study used natural language processing to detect students that will not successfully complete an enrolled course (Crossley et al., 2015). A dataset of 320 students with 132 unsuccessful and 188 successful students was used. Three NLP tools were used to analyse students' texts in a forum; the quality of words in the forum formed the attributes that were used to predict the binary target class. Results indicate an accuracy of 67.8%, indicative of students passing and hence completing the enrolled course. The study was motivated by the large numbers of students who enroll and do not successfully complete the courses in MOOCs. Such findings therefore could add to the existing methods of success prediction in MOOCs. The study proposed that a possible early warning could be to send regular emails to encourage and guide the students before they opt to drop out. In a related study, MOOC data was used to build models that predicted the final performance and successful completion of students (Jiang et al., 2014). The students who registered for a four week course were classified as successful or not in the first week. Those that were classified as unsuccessful were to be assisted. The study utilised data from student performance in the first weeks' assignments, interaction data within the MOOC, and external motivation. Two models were built: one that classified the students who attained a distinction and a normal pass; the second determined whether a student will attain the normal pass or not. The models used a Logistic Regression classifier. Results indicate a 92.6% accuracy of prediction for the first model, and 79.6% for the second model. The results confirm the predictors selected were strongly indicative of performance. External motivation was specifically responsible for successful completion of the course.

The studies have shown that MOOC data generated as students interact with the online courses can be used to build binary classifiers that are useful for initiating early intervention. However, most of the beneficiaries of MOOCs are from North America and Europe, while there are very few users from Asia and Africa (Chen, 2013). Further, poor infrastructure in Africa would be a major hindrance to their use. The short term nature of MOOC courses and the fact that they are distant education makes them unsuitable for tutoring.

3.4.1.4 Intelligent Tutoring System Dataset

Intelligent Tutoring Systems are computer systems that aim to facilitate learning in formal education by engaging the students in activities that help them to reason (Corbett et al., 1997). Lately, many studies have made use of Intelligent Tutoring Systems data to predict students' academic performance especially in the developed countries (Graesser et al., 2012). This category of datasets have been used to create models for both primary school level and higher education level.

Primary school level

Intelligent Tutoring Systems (ITS) were used to predict end of year standardised examination scores for grade seven students in the United States (Kelly et al., 2013). The aim was to ensure the at-risk students are helped early enough because of the high-stakes state standardised exams. A total of 129 students used ASSISTment in their maths lessons to do their classwork, homework and assessments. The generated attributes include total questions answered, ratio of questions correctly answered the first time, ratio of help used, and average trials for every question attempted. Both regression and classification techniques were used in the analysis. Regression was conducted using linear regression - it attained an accuracy of 75%. Classification with J48 decision tree achieved 68.4% accuracy using the whole year dataset. Further experiments improved the classification accuracy to 76% when a selected portion of the same dataset was used.

With the need to fill the gap of Science, Technology, Engineering, and Mathematics (STEM) personnel (Hill et al., 2010), some studies have been conducted to predict whether or not students will pursue a STEM career. One such study predicted whether a student will choose a career in STEM while still in middle school in the USA (Pedro et al., 2014). In the study, a total of 363 college students participated, they were selected from among those who used ASSISTments software in their mathematics classes . The study focused on identifying early enough, while in the middle school, those students who will not study STEM courses for early intervention. The purpose is to train enough manpower in the STEM related careers. The main predictor attribute

was action logs as students interacted with the system. The features generated from the action logs were reduced using the backward elimination technique; the feature search begins with the complete set and the irrelevant features are iteratively removed (Lee and Moore, 2014). Logistic Regression was used to build the models. Results obtained from a 6-fold cross-validation indicated an accuracy of 66.2% and a kappa value of 0.257. The experiments demonstrate a possibility of identifying a student who will pursue a STEM-related career in college while still in middle school. Further, attributes generated through students' interaction with the ASSISTment educational system while in middle school were used to predict those who will progress to college or not (Pedro et al., 2013). The study noted that factors related to family, finances, career ambitions, and ability have been known to correlate with academic performance among college students though not immediately actionable. They therefore proposed attributes generated through students' interaction with ASSISTment educational system. A total of 3,737 students participated in the study. Their interaction with the system generating 2,107,108 actions; answering questions and asking for help, from which features were generation. Two metrics were used to measure performance: ROC measure (A') (Hanley and McNeil, 1982), and Cohen's Kappa (Cohen et al., 1960). Logistic regression models were built using the features that were found to be statistically significant. These results indicated a prediction accuracy of 68.6%, using the ROC value. The Kappa value predicted those students that will register for college education with an accuracy of 23.9% higher than a normal guess. The insight obtained could help in initiating intervention for those who will not go to college.

Higher education level

A study involving university students investigated which ones may fail their final examination early enough so that intervention can be put in place to assist them (Koprinska et al., 2015). The study focused on predicting whether a student will pass or fail an end of semester course. Data was generated from three sources: an automatic device that marks programs and gives instant feedback - called PASTA - where data such as number of assignments done and passed or failed can be accessed; a collaborative platform - where student- to- student and student- to- lecturer interact through asking questions and getting feedback; and from assessment scores. A combination of correlation and wrapper feature selection methods were used to reduce the full set of attributes. A decision tree classifier was used in the classification task. Results indicate 87% accuracy of detecting whether a learner will pass or fail. The authors suggest the rules generated by the decision tree are simple enough and hence useful to provide warning signs to lecturers and students early enough during the term.

The studies that used Intelligent Tutoring System Data have a bias towards primary school. This is expected because ITS is an attempt to model intelligent computer based

Instruction (Corbett et al., 1997). Primary school students benefit more from such systems. However, the students in developing regions have not benefited from ITS since most of them do not have access to PCs and electricity.

3.4.2 Section Summary

The studies on academic performance prediction modelling have indicated a scarcity of research conducted with primary school data. Most of the work is done in developed countries where student data is available in state databases or where ITS have been used. This work contributes to the studies conducted using primary school data. Further, a review on studies conducted with the other dataset types reveal that most work was done using higher education data. Additionally, most data was obtained through the use of PCs and the Internet. These resources are scarce in developing countries. Mobile phones, on the other hand, are a category of computing technology that has penetrated everywhere (Traxler and Leach, 2006), hence the proposed Mobile Academic Performance Prediction System that utilises a mobile phone interface. The next section presents the opportunities offered by mobile technology, and also considers the limitations.

3.5 Technology for Developing Nations

With the rapid increase of mobile use in developing countries, it is not surprising to propose their integration in the design of academic performance prediction models. In the developing world, mobile technology has been accepted as easy to use. Further, the scarcity of infrastructure, missing or irregular power supply, lack of skills and resources to support a network have been a catalyst to the penetration of mobile phones (Traxler and Leach, 2006).

General Mobile Phone Use in Developing Countries

Literature has shown mobile phones have been used extensively in developing countries such as in India (Grönlund et al., 2008). Early findings claim that where the mobile phone has been embraced, goals of development such as enhanced trade and health provision have improved (Feldmann, 2003). The developing countries have been the main beneficiaries of the "leapfrogging" phenomenon (Donner, 2008). Mobile phones are more convenient - accessible, cheaper, and hence capable of closing the "digital divide" (Wade, 2002). Mobile phones have been rated as technologies capable of transforming lives (Donner, 2008). They have been reported to have penetrated even among the rural poor in the developing countries where the communities' social capital is reported to

have improved (Goodman, 2005). Other benefits of the mobile phone include less travel, running a business remotely, and maintaining connections with people (Samuel et al., 2005). In India, mobile phones have been reported to improve profits in the fishing industry (Jensen, 2007), and in farming to assist workers find market for their products (Islam and Grönlund, 2011). In Kenya, mobile phones have been used in a number of ways: to reduce travel costs by day-lobourers in search for temporary jobs (Chepken, 2012); to transfer money, what is known as mobile money or M-PESA by safaricom (Buku and Meredith, 2012); and M-finance, which is now an accepted term, referring to the use of mobile phone for money transfer, and "mobile banking" (Porteous, 2011). In Niger, a study of mobile use found it possible to reduce the number of trips one makes, saving time and hence lowering of travel expenses (Aker, 2008). In Tanzania a study on the use of mobile phones reported better services in social and production activities (Sife et al., 2010). In Uganda, mobile phones have been used to enhance the working of personal health record system; specifically to make the system usable among people with little education among the rural poor (Ssembatya, 2014).

Generally, the mobile technology has positively affected even the poor communities in rural areas (Duncombe, 2012). Mobile phones have supported entrepreneurs, small scale farmers, and business people in Africa (Donner, 2009). A good example is the advisory mobile system in Africa (Gakuru et al., 2009). In Kenya, mobile phones have helped improve many lives (Hughes and Lonie, 2007). These studies point to the fact that the miobile technology has helped to narrow the digital divide between the developed and developing countries. There is, therefore, need to investigate the possibility of explointing the technology even in the area of academic performance prediction.

Mobile Phone Use in Education

Developing countries have been seen to have the highest potential of utilising the mobile technology to facilitate teaching and learning (Kafyulilo, 2014). The mobile phone is a solution to the many challenges including limited electricity supply and scarce Internet connectivity (Traxler and Kukulska-Julme, 2005). Even before the mobile phone became widespread, an early study strongly suggested its advantages including simplicity and ease of use, being available and accessible, and offering flexible learning process (Collis and Moonen, 2001). The positive mobile phone characteristics for use in education are also echoed in another study as including enhanced portability, increased range of operations, the high penetration, and improved connectivity (Pachler et al., 2009). Students and teachers can carry a mobile phone in their pockets and handbags to be available whenever needed - they have therefore been used to overcome the shortage of technology as a teaching tool in education (Kafyulilo, 2014).

Mobile phones have enormous computing capabilities that can be used in learning. They also have the advantages of being able to operate in areas with an irregular supply of electricity or none, need less maintenance, are generally easy to use, and quite affordable and accessible (Masters, 2005). This makes them suitable for rural schools. Previous studies in Kenya (Traxler and Kukulska-Julme, 2005) and South Africa (Ford and Batchelor, 2007) suggested the mobile phone is as useful in education as the personal computer. A study conducted in Tanzania agreed with a previously stated advantage of the mobile phone - as an accessible technology that can be used in all levels of education in both rural and urban areas (Kafyulilo, 2014). In Ghana, mobile phones were utilised to assist school head teachers acquire leadership skills through Short Message Service (SMS) (Swaffield et al., 2013). The ubiquitous nature of the mobile technology was the key factor to the successful utilisation of the text messages.

Mobile learning was proposed because of the rapid increase in the processing ability of the mobile phone, its affordability and ubiquitous nature as the key justification for its use in the learning environment (Hashemi et al., 2011). The study specifically talked of the introduction of smart phones and increased wireless networks as important in learning institutions. Similarly, mobile learning has been seen as an extension of e-learning, where the mobile phone facilitates access learning materials from anywhere (Harichandan, 2009) .

The main catalysts as cited include the fact that mobile phones maintain power for a longer time compared to laptops and notebooks, and having touch screen interfaces that help users achieve higher levels of interaction (Goundar, 2011). Current statistics on global mobile phone sales indicate that 90 percent are Smart phones (Evjemo et al., 2014). Smart phones are becoming as ubiquitous as the normal phones (Philip and Garcia, 2015). Smart phones have in recent times become widespread and a necessity for everyone especially in education (Page, 2014). Smart phones have become preferred to desktop computers because they can be carried anywhere (So et al., 2009). They have features of a normal mobile phone in addition to: a touchscreen, capability of WiFi, and numerous functions depending of the user's application installed, similar to personal computers (Evjemo et al., 2014).

In developing nations, attempts have been made to use the mobile technology in education. A number of these attempts have been made that used smart phones in mobile learning (Nsofor et al., 2015). Their use was motivated by the limited access to computers in developing countries (Thao and Nam, 2014). Smart phones have a great potential to be used in education to facilitate student learning (Muhammad et al., 2015). Additionally, the increase in the number of applications being developed for smart phones

has been noted to facilitate the global trend of carrying out computing tasks anywhere and anytime including in developing rural regions (Srithar and Selvaraj, 2015).

Despite the extensive use of mobile phones, their integration in academic performance prediction models is missing in the literature so far reviewed. The studies that have attempted to use mobile phones in education are in supporting learning (Mbogo, 2015). As learning devices, especially in higher education, mobile phones have found acceptance (Kafyulilo, 2014, Raisamo, 2014). However, this is not so with their use in academic performance prediction models. This shows there is a gap in providing prediction modelling that integrates mobile phones to be usable in rural regions of developing countries where PCs are too expensive to be used. However, the design of systems that incorporate mobile phones has a number of challenges as presented next.

3.5.1 Limitations of Mobile Phones

Despite their many advantages, mobile phones and smart phones have their limitations. Some of these are: small screen; input limitations; limitation of access to the Internet; and small memory (Shudong and Higgins, 2005). The most prominent limitation is the small screen size (Churchill and Hedberg, 2008). This limitation could easily be overcome by mobile manufacturers making bigger devices; however, then the devices will loose their advantage of being portable.

One way in which this study overcame the screen size limitation was to select and use only the most predictive features in building the prediction model. The idea was to achieve an optimal subset that could have nearly equal capacity of predicting the target as the complete dataset. This optimal number of features identified was convenient for use on the mobile interface small screen. This study adopted the feature selection process in designing the Mobile Academic Performance Prediction System.

The second limitation is the input limitation, even on smart phones, touch input can be tedious especially in small screens (Dotenco et al., 2014). To overcome this limitation, we designed a mobile interface that used forms; the forms increased the speed of student record entry, and the accuracy of data.

As for the small memory, we adopted the client server model. The model allocates tasks to between the mobile interface and the server. The server contained the classifier model, this is also where the training of the model took place. The client side was the mobile interface that allowed a student's record to be entered and sent to the server for prediction.

The limitation of Internet to be used on the mobile phones is gradually being overcome by the mobile service providers such as, Safaricom, Orange and Airtel who have spread their network to the rural areas.

Lastly, the reasons for choosing mobile phones in this study can be summarised as:

- Mobile phones have penetrated the developing world more than any other information and communication technology.
- Mobile phones do not require expensive infrastructure.
- Mobile phones are capable of numerous functions, including data transfer.
- The choice to use mobile phones was necessitated by the lack of electricity in many of the rural schools, where desktop computers can not be used.
- Safaricom, Airtel and Orange mobile service providers have improved mobile phone networks in rural areas in Kenya.

3.6 Summary : Lessons learned

The literature highlighted a number of lessons that motivated this research. These lessons are outlined below.

1. There is need to motivate the initiation of strategic intervention for primary school students. This is because:
   (a) In the rural schools of the developing world where many students attain poor final examination marks drop out of school after primary level.
   (b) there is need to identify those students that need high intervention early enough, one or two years before they sit for the final examination to allow time for strategic intervention.
   (c) there is need to increase the number of those who will proceed to secondary and tertiary institutions in order to spur development.
2. Most academic performance prediction models have been built for the developed world; they use PCs, and data from: educational software, e-learning, MOOCs. Such models may not be usable in the developing world rural areas because of the high cost of buying and maintaining PCs.

3. Integration of mobile phones into the academic performance prediction models is not exploited, in the reviewed literature we did not come across any study that has been conducted in the developing world, especially Sub-Saharan Africa where the mobile phone is a technology that has been embraced.

4. The existing academic performance prediction models have mostly been built for university students; mature students. They have been built for distance learning students such as those that use datasets from MOOCs, e-learning systems, and university database systems.

These gaps implied a need to provide an academic performance prediction model that was applicable in the rural areas, one that integrated the available technology of mobile phones. The questions that had to be addressed to achieve such a system are:

1. Which is the best classifier model among the six common classifiers selected for the type of data used in this study?

2. What is the optimal subset of features from the total number of features from the two datasets used in this study?

3. What is the predictive performance of the Mobile Academic Performance Prediction System in classifying the high intervention class?

Therefore, this study was conducted to address the above three questions.

The literature also provided opportunities that were exploited while designing the academic performance prediction model. These are presented next:

1. When designing the academic performance prediction model, focus should be on the data mining process:

   (a) understanding the problem domain, namely the causes of poor academic performance, both from literature and through surveys.

   (b) Data understanding, which entails collecting the data and creating datasets that are meaningful

   (c) Data preparation, converting the data into a format suitable for the data mining techniques, this includes feature selection.

   (d) Data mining, building and comparing several data mining techniques to find the best.

   (e) Evaluation, use of k-fold cross validation evaluation criteria to determine the model performance by using different metrics. Also the use of confusion matrix to determine classifier performance on new data.

   (f) Using the discovered knowledge, entailed analysed the results and presentation of reports on findings of the performance of classifier models.

2. Design the academic performance prediction model for the purpose of motivating the initiation of early intervention of the at risk students, those who may fail a high stakes examination.

3. Design a binary classification model that categorises the students into two groups.

4. Classifier model evaluation should consider the following:

   (a) Use several metrics to validate the results: ROC area, F-Measure, Kappa value, Root mean square error, sensitivity, and specificity.

   (b) Use10-fold or 5-fold cross validation when finding the best classifier model

The next chapter discusses the data mining CRISP-DM process that was followed in building the academic performance prediction classifier model, indicating how the opportunities highlighted in this section were utilised in the design process.

Chapter 4

Methodology for Model Development

4.1 Introduction

The proposition of the present study is that a student academic performance prediction model could be developed to be useful and usable in rural areas of developing countries. The first step was to design the academic performance prediction model. To achieve this a data mining process called the Cross-Industry Standard Process (CRISP-DM) (Shearer, 2000) was followed to build six binary classifier models. These models were compared to determine that which is the best suited for the type of data used in the present study.

In addition to finding the best model, finding the optimal feature subset was also considered. In this chapter, five phases of the six CRISP-DM phases are discussed: (i) domain understanding; (ii) data understanding; (iii) data preparation; (iv) data mining, and (v) evaluation (Kurgan and Musilek, 2006). The chapter discusses how data from rural primary schools in developing countries could be successfully used to build the classifier models. The chapter concludes with a discussion of the metrics that were used in the evaluation of the performance of the classifier models and, further, points to the sixth phase, indicating that it would make use of the discovered knowledge to design and implement the mobile academic performance prediction system.

4.2 Educational Data Mining Process

Cross-Industry Standard Process for Data Mining (CRISP-DM) is a data mining process that has been used in the field of Educational Data Mining (EDM) (Cios et al., 2000). A revised version of CRISP-DM, slightly different from the original version, was proposed which is specific for educational research (Kurgan and Musilek, 2006) . Figure 4.1 shows the six phases process of the educational version of CRISP-DM.

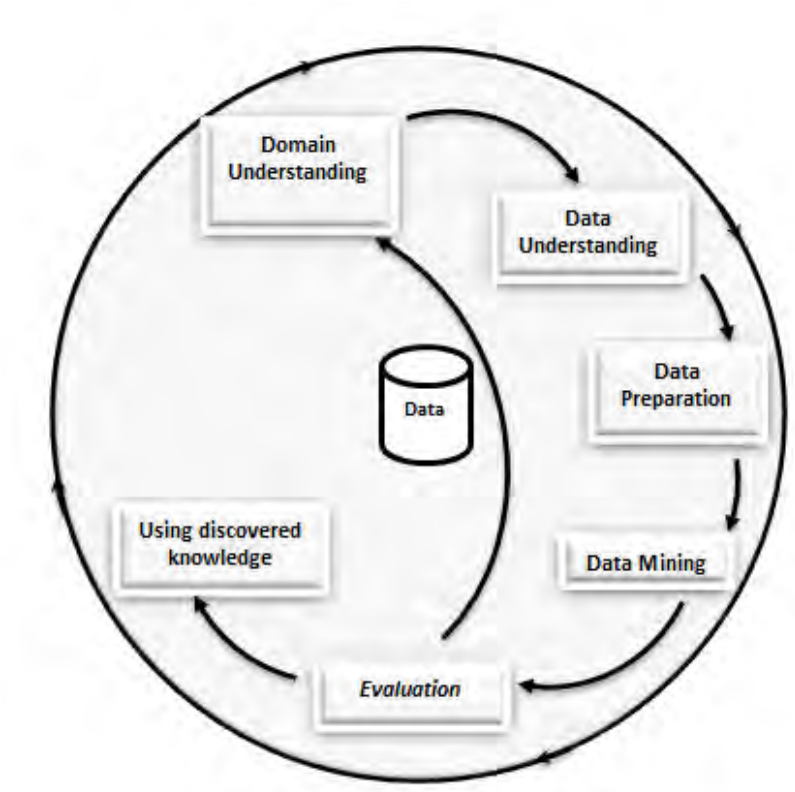


FIGURE 4.1: CRISP-DM Process Life Cycle (Kurgan and Musilek, 2006)

As shown in Figure 4.1 the six phases are briefly described next (Kurgan and Musilek, 2006):

Domain understanding - Entails identifying the key stakeholders in the research and looking for clarity and understanding of any useful knowledge that may be required. It is at this point that the goals are established.

Data understanding - Begins with data collection. This is followed by verifying the data for completeness, redundancy, and missing data. At this point data usefulness in terms of meeting the desired goal is also confirmed.

Data Preparation - Among other things, this step entails data cleaning and selecting the relevant feature subset. The goal in this step is to achieve a dataset that is suitable for selected methods of data mining.

Data Mining - Entails selecting the methods (classification, regression, or clustering) to be used for knowledge generation and applying those methods to the data. The generated data is also tested.

Evaluation - Entails interpretation of the results from the previous step of data mining. Interpretation includes looking out for novelty and interesting patterns that have been discovered. It may also entail revising the previous steps to identify possible alternative actions for improving the results.

Using the discovered knowledge - Entails putting the discovered knowledge to use by incorporating it into a performance system. It could also involve just documenting the knowledge and passing it to the interested stakeholders.

The aforementioned steps have been adopted in our study to generate a general research framework as discussed next.

4.3 Framework

This study adopted the six CRISP-DM steps in order to analyse the problem of poor academic performance in a rural area a of developing country. These steps are illustrated in Figure 4.2. The first five of these six steps are discussed in line with the problem domain in the preceding subsections.

4.3.1 Domain understanding: poor academic performance

Understanding the domain, namely poor academic performance, was the initial stage of the Education Data Mining process. At this point the goal of the study was put into focus: to develop an intervention-level classifier model that predicts whether a student will require high or low intervention to achieve passing marks in a primary school exit examination. Clearly this is a classification problem; classification techniques were employed.

The key stakeholders that confirmed the existence of the problem of poor academic performance in Kwale County, Kenya, included the County Director of Education (CDE), District Education Officers (DEOs), District Quality Assurance and Standards Officers (DQASOs), Area Education Officers (AEOs), and Head Teachers (HTs).

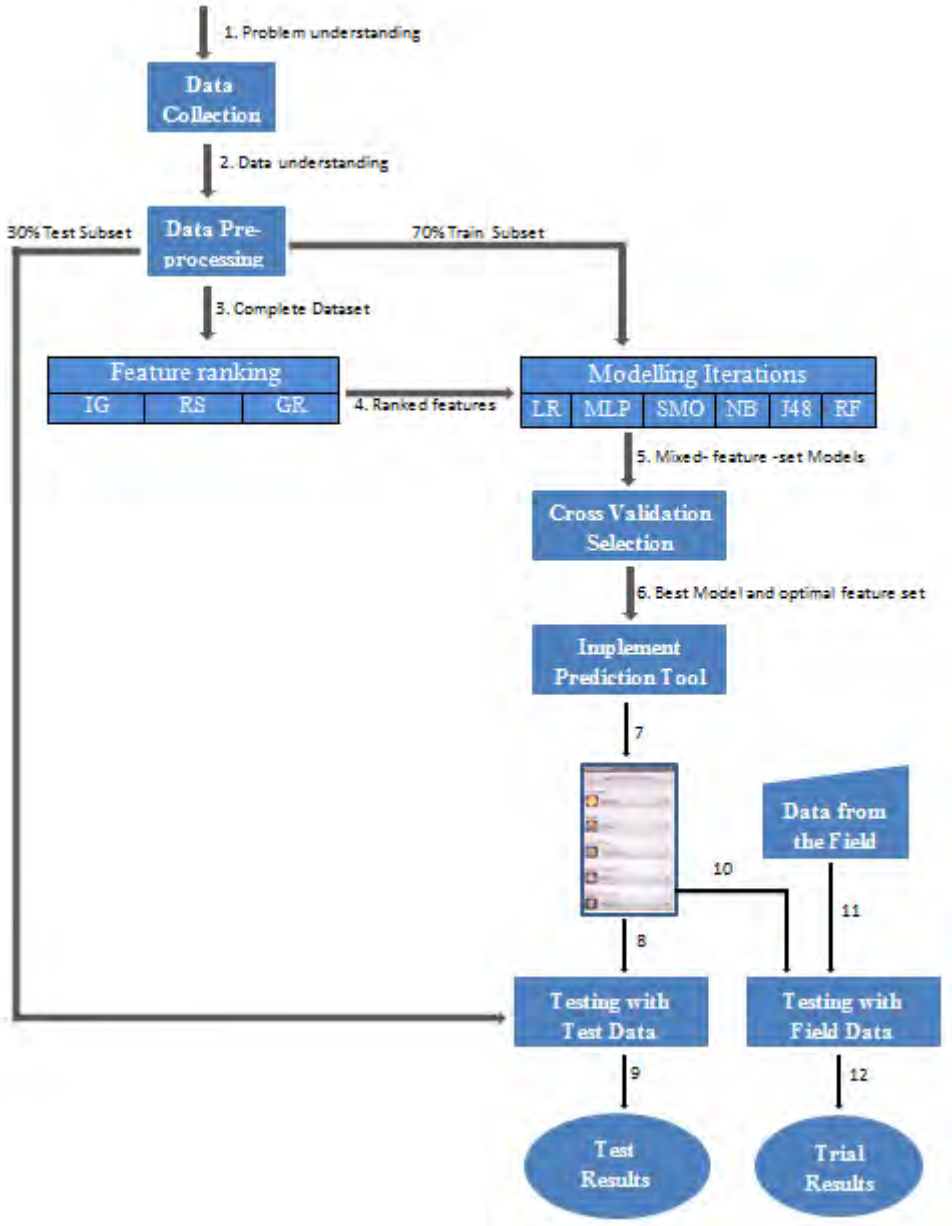


FIGURE 4.2: The Educational Data Mining Framework

A survey conducted with these senior staff revealed the problem of poor academic performance in Kwale County has has existed for over 5 years. One head teacher said, "...this problem has existed as long as I started teaching over 15 years ago....". Once the problem was understood, the study sought to collect data and understand it.

4.3.2 Data Understanding

Understanding the data entailed becoming familiar with the data types in each attribute and identifying data that would need conversion to make it suitable for building the

prediction models. The source of data was students in rural primary schools in Kwale County, Kenya. A second dataset collected for comparison was collected from Peri-urban schools (Mombasa County, Kenya). According to the County Director of Education in Kwale County, there are a total of 328 primary schools and an average of 15,000 students sit the KCPE exams each year. For this study a total of 65 schools, 54 from rural and 11 from peri-urban schools, were selected using stratified sampling (Levy and Lemeshow, 2013). A relatively high failure rate has been a concern in most schools. Therefore, if a big proportion of these students drop out every year, it is a concern not only to education stakeholders but a motivation to conduct this research.

To understand the data, it was first collected. The data collection process is discussed next.

4.3.2.1 Data Collection

The complete data collection process is shown in Figure 4.3.

As shown in Figure 4.3 a number of techniques were employed in the data collection process. The first activity involved compiling the attributes that could direct the data collection process. This was achieved through literature review and surveys, indicated by the number 1a. Activities and 1b. Findings in the figure. These are discussed next.

Activity 1a: Literature Review
The literature review was conducted on studies mostly from developed countries to determine the causes of school dropout. A structured approach was used, where, using the research boundaries, key words were identified and a search in the main journal databases was conducted. This was followed by a backward and a forward review (Webster and Watson, 2002). As discussed in Chapter 3 section 2, the causes of school dropout are similar to the causes of students' poor academic performance. The semi structured interviews and questionnaires discussed next were conducted to establish the causes that are more relevant to Kwale County.

Activity 1b: Semi-Structured Interview
Interviews were conducted with 7 education officers: the County Director of Education (CDE), the District Education Officers (DEOs), the District Quality Assurance and Standards Officers (DQASOs) and Area Education Officers (AEOs). Also interviewed were 14 head teachers during the period of June - July 2013. The interviews were conducted in Kenya, Kwale County. The questions for the interview are presented in Appendix F. The purpose of the interviews was to establish a clear understanding of the problem, mainly the causes of poor academic performance.

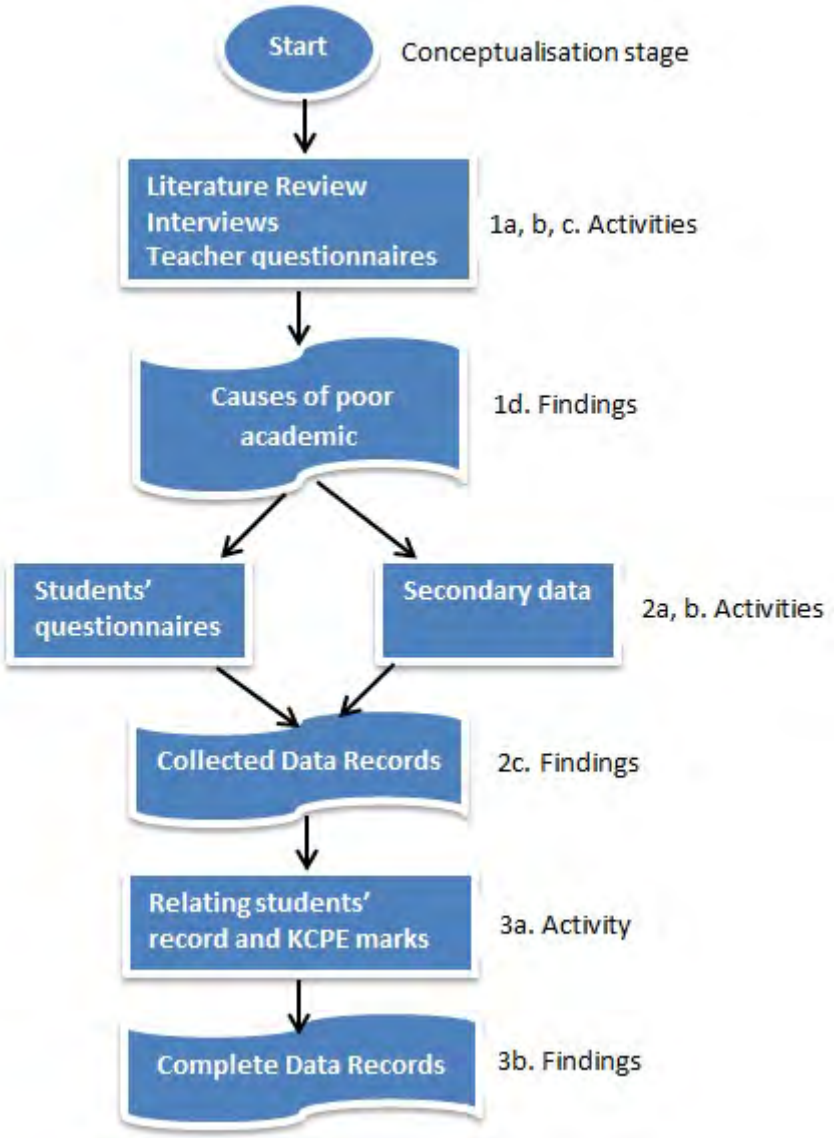


FIGURE 4.3: Systematic steps followed during data collection process

Activity 1c: Questionnaires for Teachers

Questionnaires were given to 124 teachers in 13 primary schools as part of the initial survey. This was carried out during June - July 2013. The teacher questionnaire is presented in Appendix G. As stakeholders who interact with the students on a daily basis, they provided vital insight to the causes of poor academic performance in their schools.

Activity 2a: Questionnaires for Students

Questionnaires were given to students in 54 primary schools during the period of October to December 2013. The purpose of these questionnaires was to collect data related to students' personal factors, parent and home factors, community factors, teacher and

school factors. Some of the students captured during the process are shown in the picture in Figure 4.4. The questionnaire is presented in Appendix H.

FIGURE 4.4: Students filling questionnaires during a data collection session

Activity 2b: Secondary Data Collection

Previous test marks formed the secondary data that was collected from the schools. The test marks comprised of three previous tests: Class Six year end test marks, Class Seven year end test marks, and Class Eight end of first term test marks. The test marks were collected from teachers in charge of examinations in each of the 54 schools. In some schools the test marks were obtained from the head teachers or the deputy head teachers. Most of this data was either in hand written papers, or typed hard copies. None of the schools had electronic copies. Most schools in Kwale County have no electricity. Even the ones that had electricity either did not have the computers or the manpower to create electronic mark sheets.

Activity 3a: Target marks

Classification tasks require target marks to complete the records and be able to carry out supervised learning. In this study, these marks are the Kenya Certificate of Primary Education(KCPE) exam results for the cohort of students who filled the questionnaires. These results were collected in January 2014 from the County education offices.

Completing the dataset entailed assigning digits to all the options for each attribute. Table 4.1 presents the complete list of all the features and their numeric codes.

From Table 4.1, the 22 attributes are the independent variables labeled X_1 to X_{22}.

The attributes are mainly of two types main types: categorical and numerical. The categorical types are such attributes as gender (male or female), age (normal or overage),

Name	**Var**	**Source**	**Description**	**Domain**
T MARKS	x_1	Secondary data	Test marks	350-500:1, 300-349:2, 250-299:3, 200-249:4, 0-199:5
AGE	x_2	Education officer	Student's age	normal:1, over age:2
GENDER	x_3	Education officer	Student's gender	female:1, male:2
DIST	x_4	teachers	Distance to school	1km:1, 2kms:2, 3kms:3, 4kms:4, >5kms:5
ABS	x_5	Teachers	Days absent from school per week	never:0, once:1, twice:2, >twice:3
STUD T	x_6	Teachers	Time to study at home	no time:1, little time:2, enough time:3
DISPL	x_7	Teachers	Student's discipline	never:1,once:2, twice:3, ⩾thrice:3, very often:4
COM ENG	x_8	Teachers	Student's command of English	speak local language:1, uncertain:2, speak English always:3
PUP M	x_9	Education officer	Student motivation towards education	motivated:1, neutral:2, not motivated:3
P ENC	x_{10}	All	Parent encouragement	they encourage:1, neutral:2, don't encourage:3
P ATT	x_{11}	All	Student education attitude	positive:1, neutral:2, negative:3
F HARM	x_{12}	All	Parents' state of harmony	live in harmony:1, neutral:2, don't live in harmony:3
F INCOME	x_{13}	All	Parents' income	needs not met:1, needs partly met:2, need fully met:3
PQ	x_{14}	Education officer	Parents' education level	Degree:4, diploma:3, secondary:2, primary:1, illiterate:0
F SIZ	x_{15}	Education officer	Family size in numbers	<6 persons:1, 6-10persons:2, >10persons:3
P PART	x_{16}	Teacher	Parents' participation level	they participate:1, neutral:2, they don't:3
C PART	x_{17}	Teacher	Community's level of participation	they participate:1, neutral:2, they don't:3
T ATT	x_{18}	Education officer	Teachers' attitude toward pupil	positive:1, neutral:2, negative:3
T COM	x_{19}	Education officer	Teachers' commitment to teaching	they are committed:1, neutral:2, they are not:3
T ABS	x_{20}	Education officer	Teacher absenteeism	never absent:1, neutral:2, always absent:3
S FAC	x_{21}	Head teacher	Availability of school facilities	facilities are inadequate:1, neutral:2, they are sufficient:3
L TEAC	x_{22}	Head teacher	Shortage of trained teachers	very adequate:1, adequate:2, not adequate:3

Table 4.1: The independent attributes and their numeric codes

and study time (no time, little time, or enough time). These attributes were digitised as shown in the Table 4.1. The numerical attributes were subdivided into discrete count and continuous types. An example of the continuous type is test marks, where the lowest mark is 0 and the highest is 500. These were also digitised so that the best grade was assigned 1 and the worst is assigned 5.

The target marks are the results for Class Eight exams Kenya Certificate of Primary Education (KCPE) exams. The target was labeled variable Y it takes two values: 1 for students who score below 250 marks and 0 for a student who score 250 marks and above. Therefore, the high intervention was assigned 1 and low intervention assigned 0.

The completed dataset is shown in Figure 4.5.

1	testmarks	sex	age	distance	absentsm	study_time	s_indiscipline	int_in_English	s_attitude	s_motivn	t_ attitude	t_commitnt	t_ absentsm	sh_facilities	sh _teachers	Target
2	2	2	1	1	0	2	4	1	1	1	1	1	1	1	3	0
3	2	2	1	5	0	2	3	3	1	1	1	1	3	1	1	0
4	3	2	1	2	0	3	3	1	1	1	1	1	1	3	1	0
5	3	2	2	1	0	3	2	1	1	1	1	1	3	1	3	0
6	4	2	2	1	0	2	2	1	1	1	1	1	1	1	1	0
7	4	1	2	4	0	3	2	1	1	1	1	1	1	1	1	0
8	3	2	2	2	0	2	3	3	1	3	1	1	3	1	3	0
9	4	1	2	2	0	2	4	1	1	1	1	1	1	1	3	0
10	4	2	2	2	2	3	3	1	1	1	1	1	1	3	3	1
11	4	2	2	4	0	1	3	1	1	1	1	1	1	1	1	0
12	4	1	2	2	1	2	1	3	1	1	1	1	1	3	3	1
13	4	2	2	1	0	3	3	1	1	1	1	1	1	1	3	1
14	4	1	1	2	1	3	4	3	1	1	1	1	1	3	3	1
15	4	1	2	4	1	3	3	1	1	2	1	1	1	1	3	1
16	4	2	2	3	1	2	4	1	1	1	1	1	1	1	1	1
17	4	2	2	2	0	3	2	3	1	1	1	1	3	1	1	1
18	1	2	2	1	0	3	2	3	1	1	1	1	1	1	1	0
19	4	1	2	3	1	3	2	1	1	1	1	1	1	3	3	1
20	2	2	2	1	1	3	2	3	1	1	1	1	2	1	1	0
21	4	1	2	1	2	1	3	1	1	1	1	1	3	1	1	1
22	4	2	2	1	1	2	2	2	1	1	1	1	1	1	1	1
23	4	1	2	2	1	0	2	1	1	3	1	1	1	1	1	1
24	3	1	2	1	0	1	2	3	1	1	1	1	1	3	1	1
25	4	2	2	2	1	1	2	3	1	3	1	1	3	3	3	1
26	4	1	2	3	1	0	2	3	1	3	1	1	3	1	1	1
27	4	1	2	2	1	0	2	3	1	3	2	1	1	3	3	1
28	3	1	2	1	0	1	2	3	1	1	1	1	1	3	3	0
29	3	2	2	2	0	2	2	3	1	1	1	1	3	3	3	0
30	4	2	2	2	0	2	3	3	1	1	1	1	1	1	1	1

FIGURE 4.5: The dataset to illustrate data understanding

Figure 4.5 shows the first row, the header which contains the attribute names. There are 16 attribute names shown. The total number of attributes is 23, including the target. Each raw is a complete record with an associated target mark. There are only 29 rows shown; however, there are a total of 2426 student records.

4.3.3 Data Preparation Process

Data preparation entailed digitising and enforcing validity, discretising, replacing missing values, deleting records with missing target marks, and feature subset selection. These are described next.

4.3.3.1 Digitising and Enforcing Validity

Digitising

Initially, the data records (Y, X_1, X_2,.....,X_{22}) were in different formats. For example, the X_1(test marks) attribute was of the numerical continuous type, having marks from 0 to 500 for the five tests. The independent variables, X_2 - X_{22}, were both categorical and discrete, having options from 2 to 5. Before being input to the machine learning classifiers, this raw data was preprocessed.

First, a common format was established for all attributes to make it possible for the classifier models to be built. All the worded options in attributes X_2 - X_{22} were converted into the corresponding numerical values as seen in Table 4.1.

Second, the continuous numerical values of test marks were discretised as in Table 4.2.

Total test marks	350 - 500	300 - 349	250 - 299	200 - 249	0 - 199
Letter Grade	A	B	C	D	E
Numerical Value	1	2	3	4	5

TABLE 4.2: Digitizing continuous test marks

The target marks (Y) were binarised so that they create a binary class. The marks from 0 to 249 were assigned digit 1, which stands for students requiring high intervention. The marks from 250 to 500 were assigned digit 0, which stand for students requiring low intervention.

The data was then entered in to an Excel worksheet and validated as discussed next.

Enforcing Validity

Enforcing validity was done during data entry into Excel worksheets. The data entry was done by two secretaries at the Technical University of Mombasa.

During data entry, functions in Excel were used to ensure data validity by employing 3 techniques: 'restricting response' , 'preventing missing data' and proof reading, a method used by Kupzyk and Cohen (2014).

Restricting Response Options

This method was used to ensure data validity by allowing only valid items to be entered. For items with a range of values, after determining the minimum and the maximum items, any entry outside this range would result in a warning box indicating"invalid entry". To illustrate this, if the lowest mark is 130 and the highest 450, any entry outside this range would be prevented by the spreadsheet application.

To enforce data validity of the response options from the questionnaires, the response options were restricted to only valid responses so that out-of-range values were not entered.

The data capturer could then select the option to enter in each cell from a drop down list. Only valid entries could be entered.

Preventing Missing Data

The Excel count function was used to validate the total number of items to be entered. Items that were accidentally left out were traced and reentered. The function was used to count the items entered and the missing items verified against the source document.

Proof Reading

After the secretaries entered the data, the researcher carefully proof read it to identify and correct typing errors. This was achieved by comparing the Excel worksheets, record by record, with the existing source documents.

4.3.3.2 Cleaning the Data

It was found to be necessary to clean the data and hence make it more suitable for the classification task. The process of data cleaning was achieved through replacing of missing values, and deleting records that did not have target attributes. These tasks are discussed next.

Replacing Missing Test Marks (X_1)

A record for each student consisted of test marks for three tests: Class Six and Class Seven end of year examination marks, and Class Eight term one examination marks. As expected, some records were missing test marks, because some students did not sit for the tests. They were either absent during the tests due to sickness or were transfer students. All the missing test marks for students who had the target mark (Y) and filled the questionnaires were replaced with the 'mean value' of the column as done in (Acuna and Rodriguez, 2004). The filled records were 551, which is 23% of the total number of records. Leaving them out would have reduced the number of records available for the machine learning.

Replacing Missing Values (X_2 - X_{22})

Some of the records among the X_2 - X_{22} attributes also had missing values because some students were absent during the researcher's visit to their school. Such values were replaced, as is the common practice in DM, with the most frequently occurring value in the column (García et al., 2015). The replacement was done only for those students'

records that had the test marks (X_1) and the target marks Y. Nearly 29% of the records had one or more missing values that were replaced.

Deleted Records

The records that did not have the target (Y) marks were considered incomplete and unsuitable to be part of the training data or test data; they were deleted to clean the data (Han and Xia, 2014). These records were deleted even though they had the test marks X_1 and the attributes X_2 - X_{22} because they are not useful for supervised learning. However, an insignificant number of records were deleted. This can be explained by the importance the final examination is given. It is on very rare occasions that students fail to sit for the examination.

4.3.3.3 Feature subset selection

Feature Selection is considered part of data preprocessing. The process determines a smaller subset with nearly equal predictive ability. The other features are eliminated because they are considered irrelevant or noisy (De Stefano et al., 2014). An optimal feature subset is known to increases processing speed and improve prediction accuracy (Bratu et al., 2008). The reduced feature subset was found suitable for use on the small screen mobile phone interface that formed part of MAPPS in this study. The optimal feature subset could also be of interest to education stakeholders. Knowing the subset of features that are prominently indicative of the academic performance can be helpful in coming up with focused interventions.

The filter algorithm (see chapter 3 section 3.3.2) was selected to rank the features in order of importance, from the most predictive to the least predictive. This was similar to a previous study that ranked features with an aim of discarding the features lower in the rank (Lazar et al., 2012). Our study ranked and selected the first 16 features in each rank, and the rest were discarded.

The three filter algorithms we used are: Information Gain, ReliefF (Hall and Holmes, 2003) , and Gain Ratio (Karegowda et al., 2010). The experimentation for all the three algorithms was done in the Waikato Environment for Knowledge Analysis (WEKA) machine learning environment (Yadav et al., 2014). The details of the three algorithms and how they achieve feature ranking is presented next.

Information Gain

Information Gain (IG) is a measure of the change in entropy due to the presence or absence of an attribute (Greven et al., 2014). It is a popular ranking method because it is fast, efficient and quite simple to interpret. It measures the dependence that exists

between the attributes and the labels. This is achieved by computing the information gain between the *ith* attribute A_i and the class labels C as illustrated in Equation 4.1:

$$IG(A_i, C) = H(A_i) - H(A_i|C) \tag{4.1}$$

where $H(A_i)$ is the entropy of A_i and $H(A_i|C)$ is the entropy of A_i after observing C. Both these entropies before and after observing the attribute are presented in Equation 4.2 and 4.3:

$$H(A_i) = -\sum_j P(x_j) log_2 P(x_j)) \tag{4.2}$$

$$H(A_i|C) = -\sum_k P(C_k) \sum_j P(x_j|C_k) log_2 P(x_j|C_k) \tag{4.3}$$

where C is the class and A_i is an attribute.

The amount by which the entropy of the class decreases reflects the additional information about the class provided by the attribute and is called information gain. Each attribute A_i is assigned a score based on the information gain between itself and the class.

An attribute is relevant if it has a high value of information gain and less relevant if it has a low value in the rank (Tang et al., 2014).

ReliefF

Relief is the original feature selection algorithm (Kira and Rendell, 1992). ReliefF is an improvement of Relief algorithm. Even though the original Relief algorithm is capable of determining the attributes that are most suitable, it is limited when dealing with data that is incomplete, and may not work where there are more that two classes (Kononenko et al., 1997).

Since ReliefF is an extension of Refief, an illustration of Relief algorithm is presented in Figure 4.6.

The objective of Relief is to find out those attributes whose values differentiate among the instances close to each other (Robnik-Šikonja and Kononenko, 1997). Therefore, Relief finds two nearest neighbors: from a similar class - 'nearest hit'; and from another class - 'nearest miss'. The function *diff*$(Attribute, Instance1, Instance2)$ computes the difference between the attribute values for the instances, it also computes the distance between the two instances to determine the nearest neighbour. The weights $W[A]$ define

```
1.  set all weights W[A] := 0.0;
2.  for i := 1 to m do begin
3.        randomly select an instance R;
4.        find nearest hit H and nearest miss M;
5.        for A := 1 to #all_attributes do
6.              W[A] := W[A] - diff(A,R,H)/m + diff(A,R,M)/m;
7.  end;
```

FIGURE 4.6: Basic Relief Algorithm; Adopted from (Robnik-Šikonja and Kononenko, 1997)

the quality of the attributes. As presented in the algorithm, the weights are updated to indicate the same value for instances from the same class (subtracting the difference $diff(A, R, H)$) and should differentiate between instances from different classes (adding the difference $diff(A, R, M)$)

ReliefF improves the Relief algorithm by ensuring a more reliable instance probability estimation, extends Relief to work with incomplete data and be usable in multi-classes (Kononenko et al., 1997). The present study used ReliefF because of the aforementioned improvements and also because it is readily available in the WEKA machine learning environment.

Gain Ratio

Gain ratio is a refinement of information gain: it reduces the bias of information gain towards multi-valued attributes; and considers the number and size of branches of a tree while selecting the attributes (Jantawan and Tsai, 2014). Gain ratio uses the C4.5 decision tree that is also known as the J48 algorithm in the WEKA environment (Karegowda et al., 2010).

The Gain Ratio of an attribute A is defined as the information gain of A divided by the intrinsic information - also known as the splitting information; information that is generated by splitting the training data into a defined number of partitions that correspond to the number of outcomes of a test on attribute A (Karegowda et al., 2010) .

It is expressed by the Equation 4.4:

$$GainRatio(A) = Gain(A)/SplitInfo_A(S) \tag{4.4}$$

where $Gain(A)$ is the information gain, and $SplitInfo_A$ is the information generated by splitting the training dataset S into a defined number of partitions corresponding to a similar number of outcomes of a test on a attribute A. The attribute with the highest gain ratio is taken as the splitting attribute. The non-leaf node of the decision tree generated are considered the relevant attributes (Karegowda et al., 2010).

4.3.4 Determining The Best Classifier Model and the Optimal Feature Subset

In this subsection, a description of the process of finding the best classifier model from a set of six selected classifier techniques is presented. The reason for selecting the six classifiers is that they are the more commonly used binary classifiers. Additionally, all classifiers use different classification methods, and none is known to perform better than the others in all situations (Asif et al., 2014) , therefore it was necessary to select the common classifiers and determine which one gives the best performance. In the present study, an investigation was conducted to find out which classifier model achieves the best prediction performance using the dataset in this study. Further, only one best classifier model was required for the implementation of MAPPS.

Similarly, it was important to determine an optimal feature subset because of the small screen limitation (Shudong and Higgins, 2005) of the mobile phone that was used as an interface for MAPPS. The task of optimal feature subset selection was carried out by first ranking the features using three filter algorithms, these features were then used to build successive classification models beginning with the top best common features in each rank. The classifier models were built using 70% of the dataset.

In order to determine the best classifier model, the following classifiers were built: logistic regression (LR), Artificial Neural Network Multilayer Perceptron (ANN MLP), Sequential Minimal Optimisation (SMO)- a version of Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (J48) and Random Forests (RF). Each of the classifier algorithms is discussed next.

4.3.4.1 Logistic Regression

Logistic regression is used to build models that perform binary classification (Liao and Chin, 2007). A function or relationship between the categorical target and the independent variables is determined by estimating the probabilities using a logistic function (Domínguez-Almendros et al., 2011). The present study seeks to classify students into a binary class of “high intervention” taking the digit 1, and “low intervention” taking the

digit 0. Using probabilities, a student who needs high intervention will have a probability $P(.5 < P < 1)$. A total of 22 variables are used, expressed as $(X_1, X_2, ..., X_{22})$.

Logistic regression has an output which is always 0 and 1, making it suitable for the classification case in this study.

The logistic regression hypothesis in Equation 4.9 that satisfies this condition is expressed as (Ng, 2011a):

$$h_\theta(x) = g(\theta^T x) = 1/(1 - e^{-(\theta^T x)}) \tag{4.5}$$

where $h_\theta(x)$ is the probability that the output is 1 on input x, θ are the parameters that need to be fitted to the data, and $g(z)$ is a sigmoid function that asymptotes at 1 and at 0, it is expressed as in Equation 4.6

$$g(z) = 1/(1 - e^{-z}) \tag{4.6}$$

The logistic regression hypothesis using all the 22 features used in this study is expressed as in Equation 4.7:

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_{22} x_{22}) \tag{4.7}$$

where g is the sigmoid function, θ are the parameters chosen from the training set, and $(x_1.....x_{22})$ are the 22 features used in the present study. The Equation 4.7 is called the decision boundary that divides between the high intervention students and the low intervention students

The values of the parameters θ are determined using the cost function for logistic regression which is expressed in Equation 4.8:

$$J(\Theta) = \frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)}) + (1 + y^{(i)}) log(1 - h_\theta(x^{(i)})] + \frac{y}{2m} \sum_{j=1}^{n} \Theta_j^2 \tag{4.8}$$

where $J(\Theta)$ is the cost and Θ is the vector of the parameters θ. The training set is expressed as $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ...(x^{(m)}, y^{(m)})\}$ where m is the number of records. The term on the right, known as a regularisation term, guards against overfitting. The term λ is the regularisation parameter that acts as a control on the fitting parameters. As the sizes of the fitting parameters increase, the penalty on the cost function increases;

the penalty depends on both the square of the parameters Θ and the size of λ (Lee et al., 2006).

To find the parameters θ, we determine those parameters that minimise $J(\Theta)$ in Equation 4.8. This is achieved by using gradient descent, which tunes the parameters in order to achieve a reasonable logistic regression model from the given input - output data (Wong and Chen, 1999).

4.3.4.2 The Multilayer Perceptron (MLP)

The MLP is a type of Artificial Neural Network (ANN). It is made up of interconnected process units that make use of learning algorithms to create models of knowledge, which are saved as weighted connections similar to the way the human brain functions (Zare et al., 2013). An ANNs uses the feed-forward approach where information is only allowed to move in one direction from the input to the output, and training is achieved using the back propagation algorithm (Kruse et al., 2013). It is a type of supervised learning algorithm that requires independent attributes and the target classes. MLPs are so named because they contain at least one hidden layer in addition to the input and output layers (Panchal et al., 2011).

A typical MLP is created by interconnecting artificial neurons. An artificial neuron is illustrated in Figure 4.7 (Ng, 2011b):

The output wire represents the hypothesis that takes the form of the activation function as shown in Equation 4.9; the logistic unit.

$$h_\theta(x) = g(\theta^T x) = 1/(1 - e^{-(\theta^T x)}) \tag{4.9}$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ x_{22} \end{bmatrix}$$

are the input attributes, and

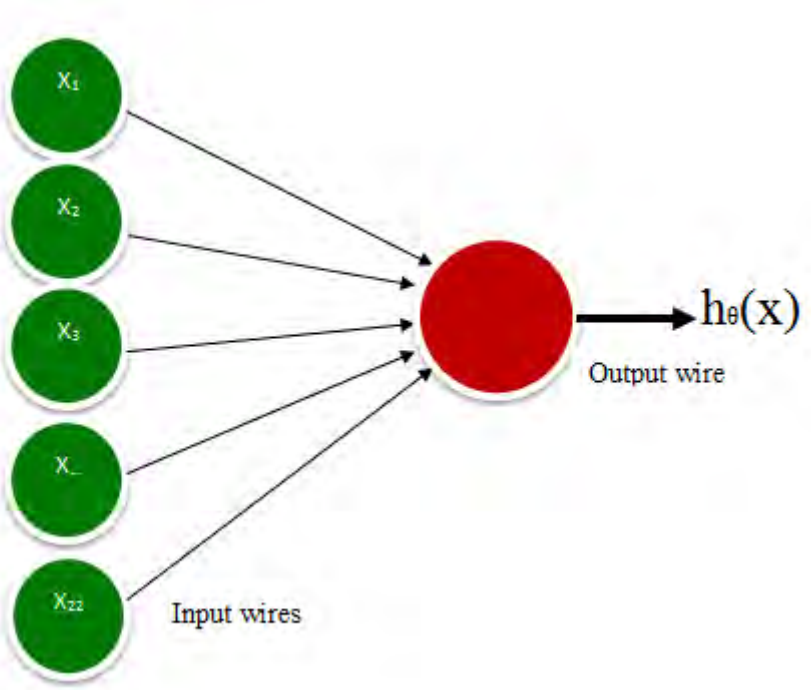


FIGURE 4.7: An artificial neuron network illustrating 22 input wires representing the attributes, and the output wire that represents the hypothesis learnt

$$\Theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ . \\ . \\ \theta_{22} \end{bmatrix}$$

is a matrix of weights that control the hypothesis mapping from one layer to the next layer.

A typical MLP is created by interconnecting artificial neurons. A typical neural network with one hidden layer is illustarused in Figure 4.8

The 'activation' values a_1.....a_{22} are computed as shown in equations 4.10 - 4.13 (Ng, 2011b):

$$a_0^{(2)} = g(\theta_{00}x_0 + \theta_{01}x_1 + \theta_{02}x_2 + ... + \theta_{022}x_{22}) \tag{4.10}$$

$$a_1^{(2)} = g(\theta_{10}x_0 + \theta_{11}x_1 + \theta_{12}x_2 + ... + \theta_{122}x_{22}) \tag{4.11}$$

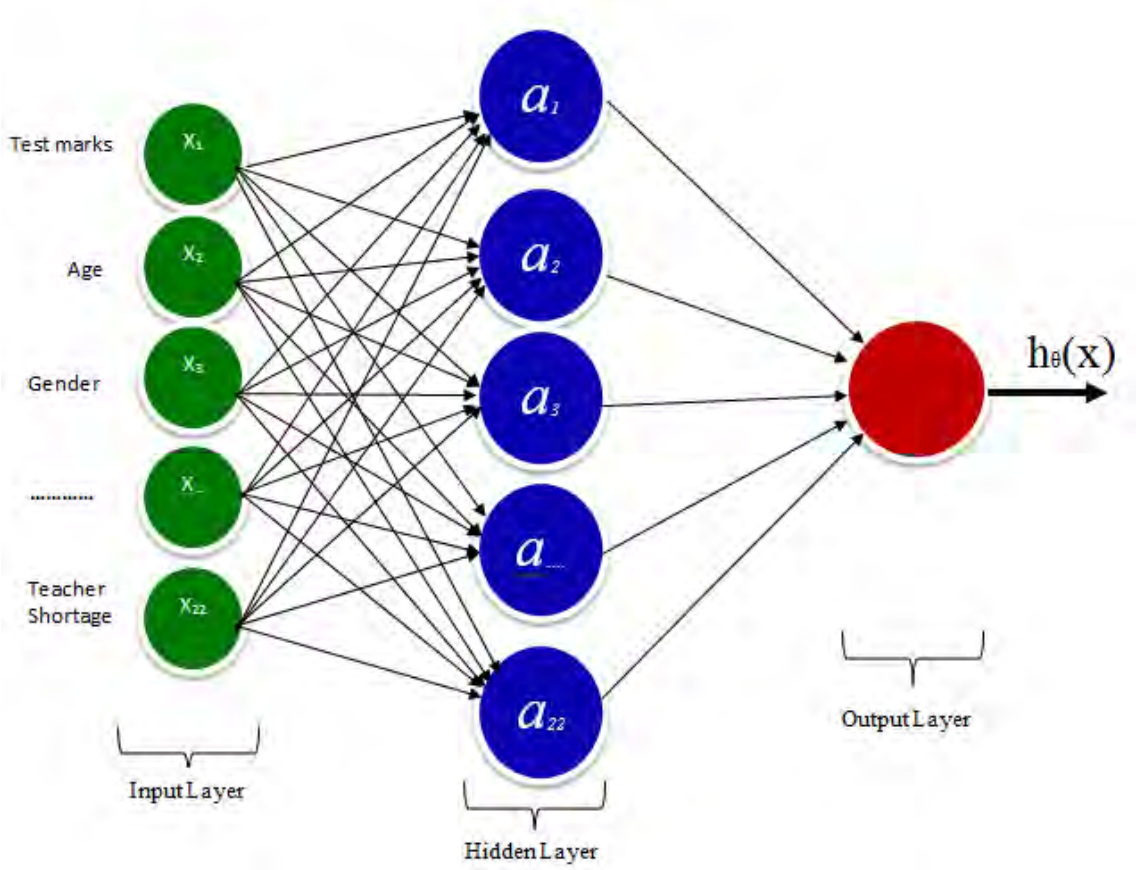


FIGURE 4.8: A neural network with one hidden layer that illustrates a multilayer perceptron that could have sevel hidden layers; the input layer is an attribute vector of the 22 attributes, the hidden layer shows the 'activation' of each of the units, and the output layer consists of the binary output hypothesis learnt to classify high intervention or low intervention students

$$a_2^{(2)} = g(\theta_{20}x_0 + \theta_{21}x_1 + \theta_{22}x_2 + ... + \theta_{222}x_{22}) \tag{4.12}$$

.....

$$a_{22}^{(2)} = g(\theta_{220}x_0 + \theta_{221}x_1 + \theta_{222}x_2 + ... + \theta_{2222}x_{22}) \tag{4.13}$$

Equation 4.10 computes a_0, the bias unit which is always equal to 1 and makes the activation function lift either to the left or to the right (Ng, 2011b). The other equations show the computation of the activation values for each of the 22 inputs. A combination of the activation values and the weights achieves the hypothesis as shown in Equation 4.14

$$h_\theta(x) = g(\theta_{10}a_0^{(2)} + \theta_{11}a_1^{(2)} + \theta_{12}a_2^{(2)} + ... + \theta_{122}a_{22}^{(2)}) \tag{4.14}$$

where in the activation values, $a_i^{(j)}$ i is the unit and j is the layer; g is the sigmoid function

The parameters θ are estimated by minimising the ANN cost function that is similar to the cost function of logistic regression Equation 4.8 with some modification as expressed in Equation 4.15:

$$J(\Theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} log(h_\theta(x^{(i)})_k + (1 - y_k^{(i)}) log(1 - h_\theta(x^{(i)})_k] + \frac{y}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\Theta_j^2 \tag{4.15}$$

The major differences between the logistic regression and ANN cost functions are: the summation of output units from 1 to k - this study uses one output unit and therefore this term may be eliminated; the summation of the number of layers from $(l-1)$ to $(L-1)$; the summation of the number of units in a layer l; and the summation of the square of parameters Θ. The last three summations are in the regularisation term; a term that improves classifier generalisation by eliminating over fitting of the learnt hypothesis (Lee et al., 2006). The cost function is optimised by first applying forward propagation that calculates the actication for a layer in order to find the hypothesis (Kalchbrenner et al., 2014). This followed by the use of back propagation algorithm for finding the parameters Θ that minimise $J(\Theta)$. The back propagation algorithm finds the gradients of the hypothesis, a partial derivative of the cost function that is used with gradient descent to find the suitable parameter values for the hypothesis (Maas et al., 2012). MLP has high accuracy in many applications, but has a drawback in that its inner working are not human understandable (Asif et al., 2014).

4.3.4.3 Sequential minimal optimization (SMO)

Support Vector Machine was designed to carry out binary classification (Tang, 2013). This makes it suitable for use in the present study that has training data and target, expressed as (x_i, y_i) where $(i = 1....m)$ and $y_i \in \{1, 0\}$; 1 represents a student requiring high intervention and 0 represents a student requiring low intervention. The discussion of SVM and the equations used were adopted from lecture notes by Ng (2012):

SVM decision boundary is derived from the logistic regression cost function as expressed in Equation 4.16 (Ng, 2012):

$$\min_{\Theta} C\sum_{i=1}^{m}[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)})] + \frac{1}{2}\sum_{i=1}^{n}\Theta_j^2 \tag{4.16}$$

where C is the regularisation parameter which when adjusted determines the position of the decision boundary or the line that separates between positive and negative samples. And $cost_1(\theta^T x^{(i)})$ is the cost function when $y = 1$ when $(\theta^T x^{(i)} >> 0)$, similarly, $cost_0(\theta^T x^{(i)})$ is the cost function when $y = 0$ when $(\theta^T x^{(i)} << 0)$. This is a special case where the selected parameters make the first part of Equation 4.16 equal to 0, the equation becomes Equation 4.17:

$$\min_{\Theta} C \times 0 + \frac{1}{2} \sum_{i=1}^{n} \Theta_j^2 \tag{4.17}$$

The equation is then subjected to the constrait so that: $(\theta^T x^{(i)} \geq 1)$ if $y^{(i)} = 1$ and $(\theta^T x^{(i)} \leq -1)$ if $y^{(i)} = 0$

When Equation 4.17 is minimised as a function of parameters Θ, a linearly separable decision boundary for the set of high intevention samples and low intervention samples as in the present study is obtained as illustrated in Figure 4.9

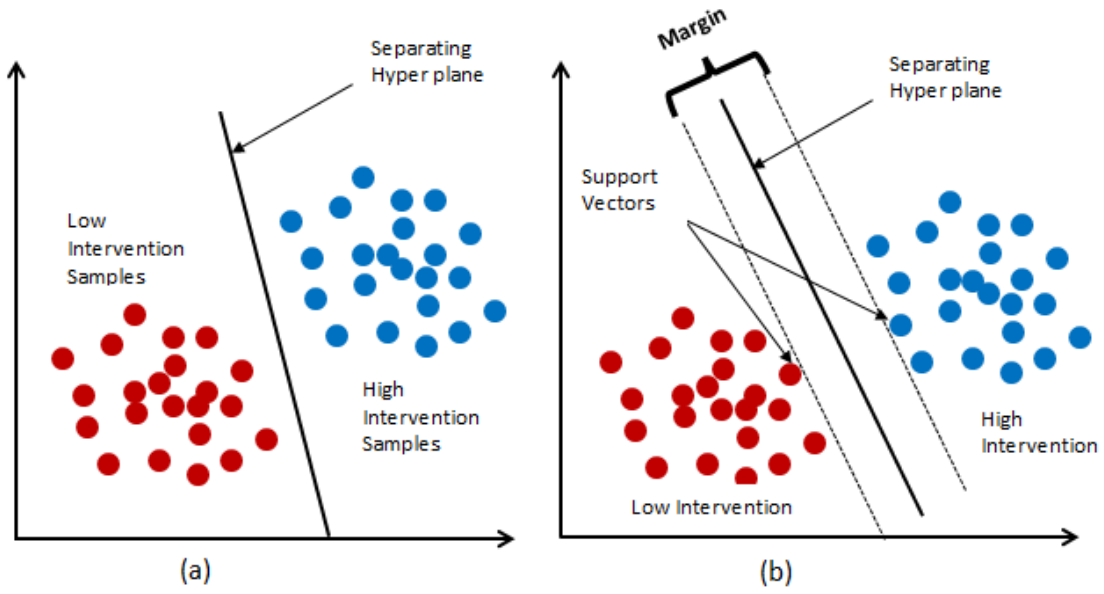


FIGURE 4.9: Support Vector Machine separating the high intervention class from the low intervention class with a hyperplane or the decision boundary as in figure (a) and at the point where the margin is greatest as in figure (b)

SVM chooses the decision boundary line that achieves the widest distance between the high intervention and low intervention samples. It is this capability of SVM to separate samples with as large a margin as possible that gives it its robustness.

Sequential minimal optimization was proposed by Platt (1999) as an algorithm that simplifies the training of Support Vector Machines (SVM) (Cao et al., 2006). This

simplification is achieved when SMO breaks the large quadratic problems normally solved by SVM into smaller problems that are then solved analytically, an approach that is time saving and faster to implement (Platt et al., 1999).

4.3.4.4 Naïve Bayes Classifier

The Naïve Bayes classifier is a type of Bayesian classifier, which assigns a given sample as described by an attribute vector to the most likely class that it belongs to (Leung, 2007).

Based on Bayes' theorem, naïve Bayes classifiers have a unique characteristic in that each attribute is assumed to be independent (Murphy, 2006).

The classification process with naïve Bayes proceeds as follows (Leung, 2007):

Letting $X = x_1, x_2, ..., x_{22}$ be a possible sample with components made of a set of the 22 attributes in our study. The term X is known as "evidence" when discussing Bayesian classifiers.

Next, assigning H to be a hypothesis that assigns the evidence X to one of the target classes, say C. The goal is therefore to determine the probability of the hypothesis H given the sample X, expressed as $P(H|X)$. This is also known as the "posterior probability" of the hypothesis H given X. Posterior probability is the probability of a student with a record X and knowing that the student will require high intervention - the probability when we have information. On the other hand, $P(H)$ is the prior probability of the hypothesis - the probability that the student will need high intervention even before taking into account the student record X.

By adopting Bayes' theorem, the required probability is expressed as in Equation 4.18.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{4.18}$$

where all the probabilities are obtained from the dataset.

Equation 4.18 can be rewritten to a naïve Bayes classifiers as follows: In the place of the hypothesis H we determine a class that maximises the probability - the highest probability that X is predicted to belong to a class C_i. The probability of this class is expressed as $P(C_i|X)$, the maximum posterior hypothesis. Expressed similarly as Equation 4.18 from Bayes theorem in Equation 4.19:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{4.19}$$

where P(X) is the prior probability that is the same for all classes. It is therefore eliminated from Equation 4.19, leaving us with the numerator to maximise the probability of a sample X appearing in a class C_i given the sample i.e. $P(C_i|X)$

To reduce the cost of computing $P(X|C_i)$ in Equation 4.19 and therefore reduce the analysis for $P(X|C_i)P(C_i)$, the assumption that the attributes are independent is made, which earns the classifier the name "naïve". This assumption allows for the mathematical expression as in Equation 4.20

$$P(X|C_i) \simeq \prod_{k=1}^{n} P(x_k|C_i) \tag{4.20}$$

where n is the number of attributes. The product of the probabilities $P(x_k|C_i)$ is obtained from the product of the probabilities $P(x_1|C_i)$, $P(x_2|C_i)$,....., $P(x_n|C_i)$.

To predict a class label for a new sample X_1, $P(X|C_i)P(C_i)$ is computed for our two classes: C_1 for high intervention and C_2 for low intervention. The sample X_1 is predicted to be in C_1 only if C_1 is the class that maximises $P(X|C_i)P(C_i)$

4.3.4.5 Decision Tree Classifiers (J48)

A popular decision tree classifier is C4.5, which is named J48 in the WEKA machine learning environment. A decision tree performs the classification task by repeatedly separating the attributes in branches like a tree (Şen et al., 2012). Mathematical algorithms such as information gain are used to determine one attribute and its threshold to split the attributes into two subgroups. This first node is known as the root node; the next nodes generated are known as leaf nodes. The above process is iterated at each leaf node until the tree is fully built. The last node is the end node. The J48 classifier is described as a powerful and popular tool for classification (Kabakchieva, 2013).

Decision trees have the known advantage that they represent rules that can easily be understood and interpreted by users, and they do not require complex data preparation. Also, they perform well for numerical and categorical variables.

4.3.4.6 Random Forests

Random forests are in the class of statistical classifiers, which have been known to achieve high performance on many classification tasks (Cutler et al., 2007). Their unique characteristic is that many classification trees are built from a given dataset. These trees are then combined to determine one common high prediction. Although this approach may result in difficulty in interpretation, the combination of many trees has been known to achieve improved results.

Random forests use the "bagging" procedure with some slight improvement to prevent the built trees from correlation in an effort to reduce variance. In the bagging procedure, also called "bootstrapping", samples are repeatedly taken from the training data. The idea is to create many bootstrapped training datasets. Each of these datasets are used to train a decision tree. Test data is then used to determine the predictive performance of each tree. For classification, the class that has been predicted by a majority of the trees for a test sample is voted the correct one (James et al., 2013). Random forests are an ensemble method which uses recursive partitioning to generate many trees and then aggregate the results (Kandaswamy et al., 2011) .

4.3.5 Evaluation Process

Evaluation is an important task in the building of classifier models and feature selection. 10-fold cross-validation was used in this study to evaluate the models. K-fold is a common machine learning experiment design that combines both training and testing. In the 10-fold design, 10 different subsets of equal size were created by randomly splitting the dataset. The procedure for model building involved training and testing the model 10 times. Each iteration involved training on nine portions and testing on one of the portions. Results from the 10 experiments were put together in one confusion matrix. Selected metrics measures were then generated to determine the performance of the classifier models (Şen et al., 2012).

4.3.5.1 Metrics

This section discusses all the metrics that were used throughout this study. The first evaluation process entails the cross-validation performance of classifier models, and the second is the evaluation of the mobile academic performance prediction system that is described in Chapter seven. The most commonly used metrics in the literature are: sensitivity, specificity, precision, F-Measure (Shaikh et al., 2015), Receiver Operating

Characteristics (ROC) area (Sarlis and Christopoulos, 2014), Cohen's Kappa (Romero and Ventura, 2010), and Root Mean Squared Error (RMSE) (Pardos et al., 2012).

The first four metrics are directly generated from the confusion metrics in Figure 4.10. In the confusion matrix, the performance of a classifier model is evaluated on the basis of counting the cross validation instances that are correctly and those that are incorrectly predicted by the model (Asif et al., 2014). This technique is common in prediction model studies (Hempel et al., 2012, Márquez-Vera et al., 2013).

ACTUAL / PREDICTED	**High Intervention(HI)**	**Low Intervention(LI)**
High Intervention(HI)	True High Intervention(TH)	False High Intervention(FH)
Low Intervention(LI)	False Low Intervention(FL)	True Low Intervention(TL)

FIGURE 4.10: confusion matrix (Márquez-Vera et al., 2013)

The confusion matrix is a tool for analysing the classification performance of classifiers, it accumulates the results so that they can be used as a basis for the accuracy analysis (Şen et al., 2012). Originally, it was popularised in the field of Machine Learning through Kohavi and Provost (Ron and Foster, 1998). The confusion matrix was adapted in this study to analyse the prediction performance of MAPPS. As shown in the figure, the columns present the actual high intervention and low intervention students. While, the rows represent what MAPPS predicted as high intervention and low intervention students. The correctly classified students appear in the true high intervention box and the true low intervention box. The false low intervention are the students that are actually high intervention cases, according to their final examination marks used for training, but the model has predicted them as belonging to the low intervention class. Similarly, the False low intervention are those students that actually belong to the low intervention class but the model has predicted them to be in the high intervention class. The accuracy of a classifier is therefore determined by the proportion of misclassified students; the smaller the proportion, the more accurate the classifier. A number of metrics are derived from the confusion matrix to analyse the classifier performance; these are discussed next.

Prevalence - This is the most basic measure derived from the confusion matrix - defined as the measure of the proportion of actual high intervention students to all the students. Mathematically, it is defined in Equation 4.21. Prevalence is an important measure; it indicates the target class distribution of the dataset under study.

$$\textbf{Prevalence} = \frac{TrueHigh + FalseLow}{TrueHigh + FalseHigh + FalseLow + TrueLow} \tag{4.21}$$

Sensitivity

Sensitivity, or recall, is the proportion of the number of high intervention records that have been correctly identified from among the actual high intervention records as identified by the target class. Sensitivity is a measure of the proportion of actual positives which are correctly identified as shown in Equation 4.22 (Sokolova and Lapalme, 2009).

$$Recall = \frac{TrueHigh}{TrueHigh + FalseLow} \tag{4.22}$$

Specificity

Specificity, on the other hand, relates to the proportion of the low intervention records identified from among the actual low intervention records. Specificity and sensitivity metrics have the advantage that they indicate the classifier's ability to classify the true positives and the true negatives.

Specificity measures the proportion of negatives that are correctly identified as such. It is defined in Equation 4.23

$$Specificity = \frac{TrueLow}{TrueLow + FalseHigh} \tag{4.23}$$

A perfect classifier would be 100% Sensitive, which means all high interventions are identified as high intervention, and 100% Specific, which means none of the low interventions is identified as high intervention.

Precision

Precision is a measure of the proportion of the actual correctly predicted high intervention students to all the students predicted as high intervention students. It has been defined as the positive predictive value, as given in (Equation 4.24) (Thai-Nghe et al., 2009).

$$Precision = \frac{TrueHigh}{TrueHigh + FalseHigh} \tag{4.24}$$

Accuracy

Accuracy is a common metric though not a preferred measure for imbalanced classes. It has been used in a number of studies (Doan et al., 2011, Zafra et al., 2009). It is also known as the overall accuracy, or the proportion of correctly predicted high and low intervention to all the students. It gives the overall performance of the classifier as shown in Equation 4.25 .

$$Accuracy = \frac{TrueHigh + TrueLow}{TrueHigh + FalseHigh + FalseLow + TrueLow} \tag{4.25}$$

In the five Equations, True High is the number of actual high intervention students correctly predicted. False High is the number of actual low intervention students predicted as high intervention students. True Low is the number of actual Low intervention students correctly predicted. False Low is the number of actual high intervention students predicted as low intervention students.

F-Measure

F-Measure determines the effectiveness of the classifier in classifying high intervention students or true positives, by combining both precision and recall to attain an average value that is balanced (Shaikh et al., 2015). It is preferred for imbalanced datasets because the classes are handled independently (Thai-Nghe et al., 2009).

F-Measure is expressed in Equation 4.26.

$$F\text{-}Measure = \frac{2Precision * Recall}{Precision + Recall} \tag{4.26}$$

Receiver Operating Characteristics (ROC) Area

ROC is a metric curve obtained by plotting sensitivity against specificity, it measures the ability of a classifier to categorise instances into different classes (Jiménez-Valverde, 2012). It is a common metric used to compare classifier performance (Sarlis and Christopoulos, 2014). The ROC metric is specifically preferred because of its advantage to achieve a trade-off between sensitivity and specificity. It is a reliable measure because it is not affected by imbalanced classes (Brown and Davis, 2006). A classifier that attains high values of sensitivity at low specificity ends up with a bigger ROC area, which means the classifier is accurate in detecting the 'True Positives'. The metric can be used to measure performance of classifiers even on skewed datasets, and has been known to achieve good results in comparing classifier models' performance (Tamhane et al., 2014). The ROC Area metric was suitable in this study because the training dataset used from rural schools has 1184 high intervention records and 512 low intervention records and is therefore not balanced.

Cohen's Kappa

Cohen's Kappa works in a similar manner to the statistic measure's correlation coefficient, which gives a correlation value between -1.0 and 1.0 (Wood, 2007). A Kappa value of 1.0 implies the rating by a pair of raters perfectly agree; a value of -1 implies they perfectly disagree, and a value of 0 implies a guess by the two raters (McHugh, 2012). There is a consensus by researchers that an acceptable agreement value between any two raters should be at least 0.60 (Wood, 2007). Cohen's Kappa has gained researchers' approval of being less error prone (Guyon and Elisseeff, 2003). It has therefore been used in studies to compare the performance of classifiers (Romero and Ventura, 2010). It is expressed in Equation 4.27 (Vieira et al., 2010):

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{4.27}$$

where K is Cohen's Kappa value, $Pr(a)$ is the total agreement probability, and $Pr(e)$ is the hypothetical value of probability of agreement among the raters. A perfect agreement between raters will achieve a value of $K = 1$ while in a case where the raters' agreement is only by chance, $K \leq 0$.

Root Mean Squared Error (RMSE)

RMSE has been used to rate classifier performance in terms of the size of errors the classifiers make (Pardos et al., 2012). RMSE has been found to be a more a appropriate metric for model performance compared to the popular mean absolute error metric (Chai and Draxler, 2014). The expression for RMSE is shown in Equation 4.28 (Chai and Draxler, 2014):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.28}$$

where n is the number of samples and $(y_i - \hat{y}_i)$ is the error calculated for $i = 1, 2, ..., n$

The six metrics measures discussed above, namely sensitivity, specificity, ROC area, F-Measure, Cohen's Kappa, and RMSE, were used to rate the performance of the classifier models in determining the best classifier and the optimal feature subset. These measures were found useful in other studies, although no single surveyed study has used all of them.

4.4 Chapter Summary

This chapter has shown how five out of the six-step data mining process framework was followed to achieve the mobile academic performance prediction system. The CRISP-DM's methodology of systematic data mining tasks were followed. The tasks and methods described are: domain understanding; data understanding - data collection and creating databases; data preparation - digitisation and discretisation, including ranking of features as a step to feature selection; data mining; and evaluation.

The building of six classifiers was discussed, these include logistic regression (LR), Artificial Neural Network Multilayer Perceptron (ANN MLP), Sequential Minimal Optimisation (SMO)- a version of Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree(J48) and Random Forests (RF). Evaluation was conducted using six metric measures: sensitivity, specificity, Receiver Operating Characteristics (ROC), F-Measure, Cohen's Kappa, and Root Mean Squared Error (RMSE).

Therefore, the chapter presented the methodology for the machine learning phase of this study. The machine learning process, with a focus on Educational Data Mining (EDM) is the backbone of this study. The sixth step of the CRISP-DM process is discussed in chapters six and seven, as it forms the second phase of the study, where the discovered knowledge was implemented to complete the design of the Mobile Academic Academic Prediction System.

The next chapter presents the results and a discussion of the five steps of the CRISP-DM process.

Chapter 5

Classifier Model and Optimal Feature Subset

5.1 Introduction

The purpose of this study was to investigate an academic performance prediction model that would classify rural primary school students into two categories: those that need high intervention, and those that need low intervention to pass the final exit examination. This chapter presents results and discussions of an educational data mining process, also known as the CRISP-DM process, followed in this study. This process was used to determine the best classifier model and an optimal subset of features. The experimental results of the classifier built, prediction performance, are obtained using 10-fold cross-validation. In the first phase of the experiments, the rural schools' dataset was used. All the 22 features were used in the experiments to determine the best classifier model. This was followed by experiments to determine the optimal feature subset. Features were ranked using three ranking algorithms: information gain; ReliefF; and Gain Ratio. Each ranked list was used to build successive models to determine the least number of features that give the highest prediction performance. The experiments were repeated for all the six classifiers. In the second phase, a peri-urban dataset was used, and similar experiments to those conducted with rural schools' dataset were conducted. The findings obtained provided important insights and guided the research to the next phase of designing and developing MAPPS.

5.2 Finding the Best Classifier Model Using Rural Schools' Dataset

The best classifier model is one that will attain best values of the selected metrics of performance measure. In this study six metrics were used including recall, specificity, ROC area, F-Measure, Cohen's Kappa, and RMSE. The performance measures were obtained using 10-fold cross-validation performance evaluation. Classifiers were built using a training set of 70% of the rural dataset. This dataset had a total of 22 features and 1696 records. Results and discussion for each of the six classifier models, including Logistic Regression, Multilayer Perceptron, SVM-SMO, Naïve Bayes, J48, and Random Forest are presented next.

5.2.1 Comparing the Performance of Six Classifier Models

5.2.1.1 Logistic Regression Model

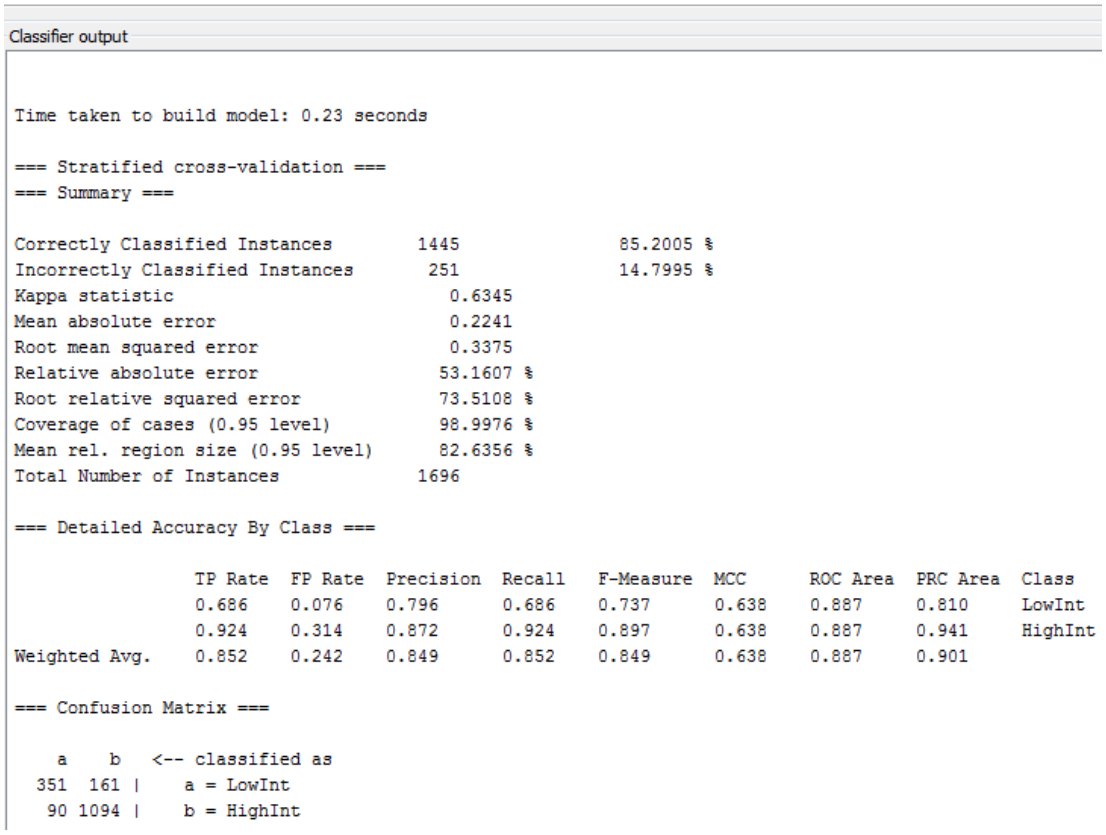
The first model to be built was Logistic Regression, it is a common classification algorithm that has been used in educational data mining (Baker and Inventado, 2014). Logistic regression is a binary classifier model that estimates the probability that a student requires high intervention based on a set of the features identified as causes of poor academic performance.

The model was built in WEKA, a machine learning environment for data mining (Hall et al., 2009). The logistic regression model performance details are shown in Figure 5.1.

The model performance for the six metrics identified from Figure 5.1 are summarised for clarity in Table 5.1

Metric	**Performance**
Recall value	0.924
Specificity value	0.686
ROC Area	0.887
F-Measure	0.897
Cohen's Kappa value	0.6345
RMSE value	0.3375

TABLE 5.1: A summary of results obtained from training and testing the Logistic Regression classifier model



```
Classifier output


Time taken to build model: 0.23 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1445               85.2005 %
Incorrectly Classified Instances       251               14.7995 %
Kappa statistic                          0.6345
Mean absolute error                      0.2241
Root mean squared error                  0.3375
Relative absolute error                 53.1607 %
Root relative squared error             73.5108 %
Coverage of cases (0.95 level)          98.9976 %
Mean rel. region size (0.95 level)      82.6356 %
Total Number of Instances             1696

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.686    0.076    0.796      0.686    0.737      0.638    0.887     0.810     LowInt
                 0.924    0.314    0.872      0.924    0.897      0.638    0.887     0.941     HighInt
Weighted Avg.    0.852    0.242    0.849      0.852    0.849      0.638    0.887     0.901

=== Confusion Matrix ===

    a    b   <-- classified as
  351  161 |    a = LowInt
   90 1094 |    b = HighInt
```

FIGURE 5.1: Logistic regression model showing the prediction performance with rural dataset

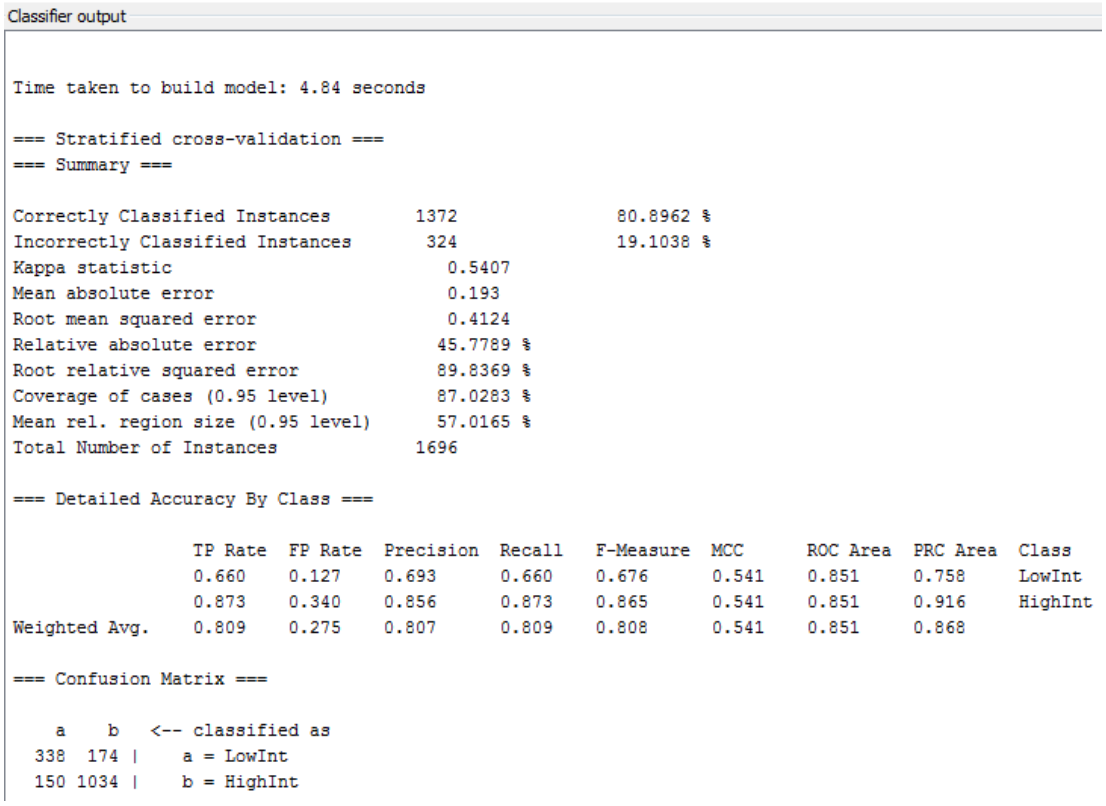
5.2.1.2 Multilayer Perceptron Model

The second classifier model built is Multilayer Perceptron (MLP). It is a type of Neural Network that has several layers of nodes: the input layer, the hidden layer(s), and the output layer (Ribeiro et al., 2012). MLP has the following features: the capacity to model non-linear functions; being a feed forward model that maps input features to an appropriate output; and utilising back propagation to facilitate supervised learning (Zare et al., 2013).

The results for the MLP model implemented in WEKA are presented in Figure 5.2. Table 5.2

5.2.1.3 Sequential Minimal Optimisation (SMO) Model

Sequential Minimal Optimisation (SMO) algorithm is an improved version of the Support Vector Machine (SVM). The concept of SVM is to obtain as large a margin as possible between a positive and negative sample (Huang et al., 2015). SMO, however is preferred because it is a simple and fast algorithm that is capable of solving the SVM's large



```
Classifier output

Time taken to build model: 4.84 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1372               80.8962 %
Incorrectly Classified Instances       324               19.1038 %
Kappa statistic                          0.5407
Mean absolute error                      0.193
Root mean squared error                  0.4124
Relative absolute error                 45.7789 %
Root relative squared error             89.8369 %
Coverage of cases (0.95 level)          87.0283 %
Mean rel. region size (0.95 level)      57.0165 %
Total Number of Instances             1696

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.660    0.127    0.693      0.660    0.676      0.541    0.851     0.758     LowInt
                 0.873    0.340    0.856      0.873    0.865      0.541    0.851     0.916     HighInt
Weighted Avg.    0.809    0.275    0.807      0.809    0.808      0.541    0.851     0.868

=== Confusion Matrix ===

    a    b   <-- classified as
  338  174 |    a = LowInt
  150 1034 |    b = HighInt
```

FIGURE 5.2: Multilayer Perceptron classifier model showing the prediction performance with rural dataset

Metric	**Performance**
Recall value	0.873
Specificity value	0.660
ROC Area	0.851
F-Measure	0.865
Cohen's Kappa value	0.5407
RMSE value	0.4124

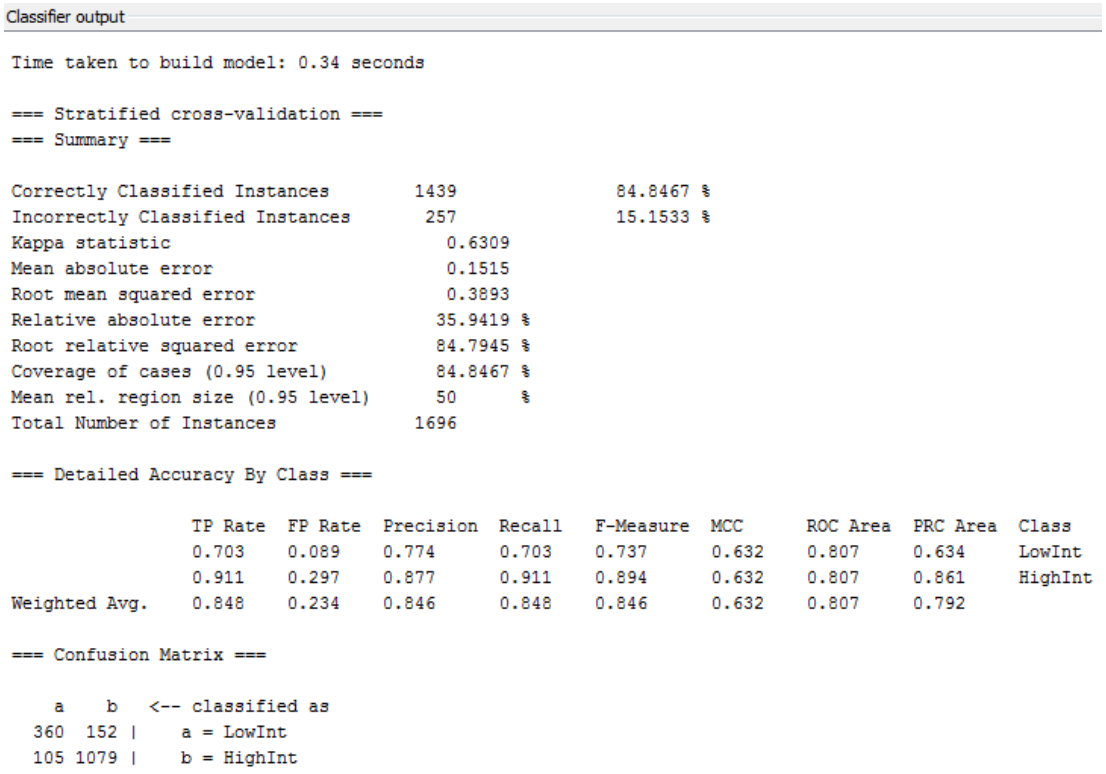
TABLE 5.2: A summary of results obtained from training and testing the Multilayer Perceptron classifier model

quadratic programming problems by splitting them into manageable problems that are solved analytically (Shevade et al., 2000).

The SMO model was conducted in WEKA. The results are presented in Figure 5.3. The summary of performance using the six metrics is presented in Table 5.3

5.2.1.4 Naïve Bayes Model

Naïve Bayes classifier is a simplified version of a group of Bayesian Classifiers, it uses Bayes' Theorem (Patil and Sherekar, 2013). It is described as naïve because it considers



```
Classifier output

Time taken to build model: 0.34 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1439               84.8467 %
Incorrectly Classified Instances       257               15.1533 %
Kappa statistic                          0.6309
Mean absolute error                      0.1515
Root mean squared error                  0.3893
Relative absolute error                 35.9419 %
Root relative squared error             84.7945 %
Coverage of cases (0.95 level)          84.8467 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances             1696

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.703    0.089    0.774      0.703    0.737      0.632    0.807     0.634     LowInt
                 0.911    0.297    0.877      0.911    0.894      0.632    0.807     0.861     HighInt
Weighted Avg.    0.848    0.234    0.846      0.848    0.846      0.632    0.807     0.792

=== Confusion Matrix ===

    a    b   <-- classified as
  360  152 |    a = LowInt
  105 1079 |    b = HighInt
```

FIGURE 5.3: Sequential Minimal Optimisation classifier model showing prediction performance with rural dataset

Metric	**Performance**
Recall value	0.911
Specificity value	0.703
ROC Area	0.807
F-Measure	0.894
Cohen's Kappa value	0.6309
RMSE value	0.3893

TABLE 5.3: A summary of results obtained from training and testing the Sequential Minimal Optimisation classifier model

every attribute as being independent - the probability of one attribute does not affect the other (Patil and Sherekar, 2013). Naïve Bayes has historically been shown to achieve high rates of correct classification results despite being simplistic (Cichosz, 2015). The results obtained through implementing naïve Bayes in WEKA are presented next in Figure 5.4. Summarised performance results for the six metrics are shown in Table 5.4

5.2.1.5 J48 Model

J48 is an enhanced version of the $C4.5$ decision tree classifier, which builds a tree model to achieve classification (Patil and Sherekar, 2013). The enhancement in J48 is in taking

```
Classifier output

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1240               73.1132 %
Incorrectly Classified Instances       456               26.8868 %
Kappa statistic                          0.4403
Mean absolute error                      0.2984
Root mean squared error                  0.4264
Relative absolute error                 70.7809 %
Root relative squared error             92.8744 %
Coverage of cases (0.95 level)          96.7571 %
Mean rel. region size (0.95 level)      82.1639 %
Total Number of Instances             1696

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.801    0.299    0.537      0.801    0.643      0.463    0.846     0.778     LowInt
                 0.701    0.199    0.891      0.701    0.784      0.463    0.846     0.898     HighInt
Weighted Avg.    0.731    0.229    0.784      0.731    0.742      0.463    0.846     0.862

=== Confusion Matrix ===

   a   b   <-- classified as
 410 102 |   a = LowInt
 354 830 |   b = HighInt
```

FIGURE 5.4: Naïve Bayes classifier model showing the prediction performance with rural schools' dataset

Metric	**Performance**
Recall value	0.701
Specificity value	0.801
ROC Area	0.846
F-Measure	0.784
Cohen's Kappa value	0.4403
RMSE value	0.4264

TABLE 5.4: A summary of results obtained from training and testing the Naïve Bayes classifier model

care of missing values, and pruning of the tree, among other advantages (Kaur and Chhabra, 2014).

The implementation was conducted in WEKA and results are presented next in Figure 5.5. Similarly, a summary of the results are presented in Table 5.5

5.2.1.6 Random Forests Model

Random Forest classifiers are created by putting together several tree predictors. These tree predictors are arranged so that each predictor tree relies on independently sampled

```
Classifier output

Time taken to build model: 0.25 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1415               83.4316 %
Incorrectly Classified Instances       281               16.5684 %
Kappa statistic                          0.5941
Mean absolute error                      0.2208
Root mean squared error                  0.372
Relative absolute error                 52.3754 %
Root relative squared error             81.0341 %
Coverage of cases (0.95 level)          96.0495 %
Mean rel. region size (0.95 level)      79.2748 %
Total Number of Instances             1696

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.670    0.095    0.754      0.670    0.709      0.596    0.822     0.669     LowInt
               0.905    0.330    0.864      0.905    0.884      0.596    0.822     0.861     HighInt
Weighted Avg.  0.834    0.259    0.831      0.834    0.831      0.596    0.822     0.803

=== Confusion Matrix ===

   a    b   <-- classified as
 343  169 |    a = LowInt
 112 1072 |    b = HighInt
```

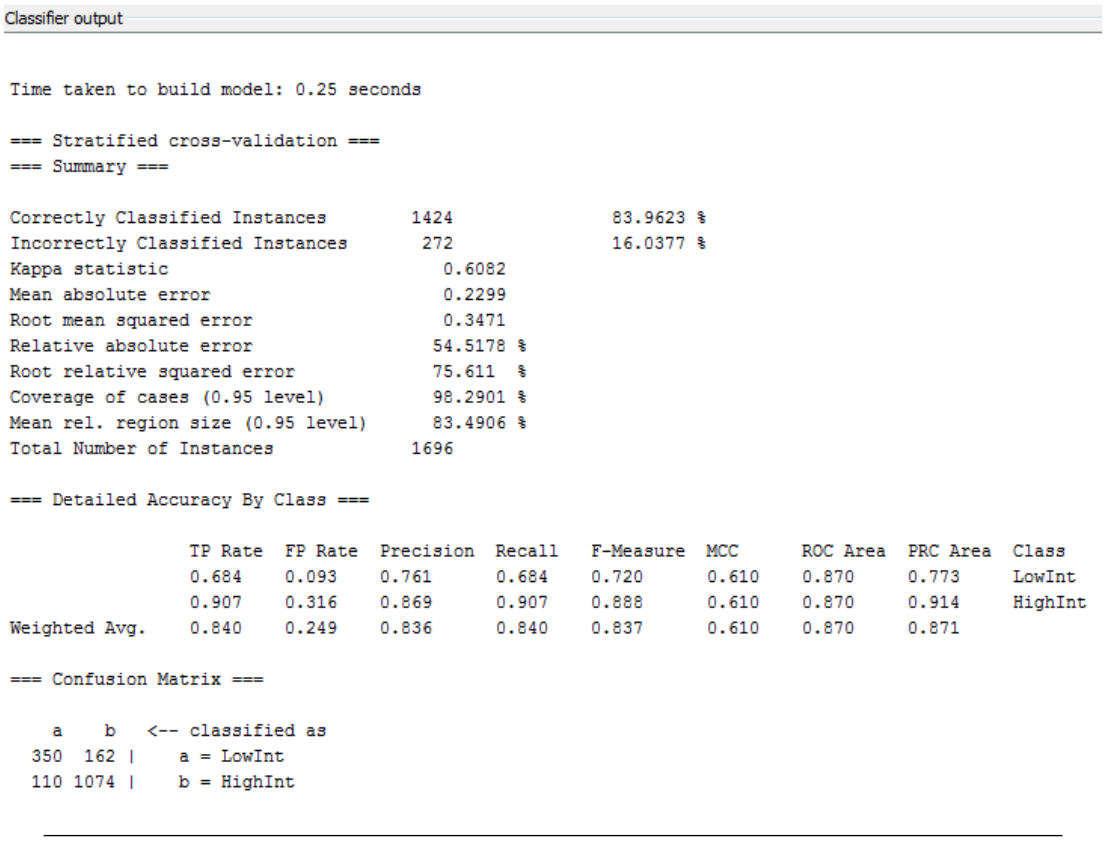
FIGURE 5.5: J48 classifier model showing prediction performance with rural schools' dataset

Metric	**Performance**
Recall value	0.905
Specificity value	0.670
ROC Area	0.822
F-Measure	0.884
Cohen's Kappa value	0.5941
RMSE value	0.3720

TABLE 5.5: A summary of results obtained from training and testing the J48 Decision tree classifier model

random vectors with similar distribution. It is therefore a classifier that is made up of other tree-structured classifiers (Breiman, 2001).

The implementation of Random Forest was also conducted in WEKA and the results are presented in Figure 5.6. A summary of the performance results are in Table 5.6



```
Classifier output

Time taken to build model: 0.25 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1424               83.9623 %
Incorrectly Classified Instances       272               16.0377 %
Kappa statistic                          0.6082
Mean absolute error                      0.2299
Root mean squared error                  0.3471
Relative absolute error                 54.5178 %
Root relative squared error             75.611  %
Coverage of cases (0.95 level)          98.2901 %
Mean rel. region size (0.95 level)      83.4906 %
Total Number of Instances             1696

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.684    0.093    0.761      0.684    0.720      0.610    0.870     0.773     LowInt
               0.907    0.316    0.869      0.907    0.888      0.610    0.870     0.914     HighInt
Weighted Avg.  0.840    0.249    0.836      0.840    0.837      0.610    0.870     0.871

=== Confusion Matrix ===

   a    b   <-- classified as
 350  162 |    a = LowInt
 110 1074 |    b = HighInt
```

FIGURE 5.6: Random Forest classifier model showing prediction performance with rural schools' dataset

Metric	**Performance**
Recall value	0.907
Specificity value	0.684
ROC Area	0.870
F-Measure	0.888
Cohen's Kappa value	0.6082
RMSE value	0.3471

TABLE 5.6: A summary of results obtained from training and testing the Random Forests classifier model

5.2.2 Discussion of Performance Findings

The results of experiments conducted to determine the best classifier model for the type of data used in this study were compared and discussed in this subsection. Table 5.7 presents the correctly and incorrectly classified actual numbers of student records.

The table shows a comparison of the classifiers' performance in terms of the actual numbers of student records that were correctly and incorrectly classified. As shown logistic regression correctly classified the highest number of students, and also misclassified the lowest number. SMO had the second highest number of correctly classified students

Selected Item	LR	MLP	SMO	NB	J48	RF
Students correctly classified	1445	1372	1439	1249	1415	1424
Students incorrectly classified	251	324	257	456	281	272
correctly classified high intervention students	1094	1034	1079	830	1072	1074
incorrectly classified high intervention students	90	150	105	354	112	110
correctly classified low intervention students	351	338	360	410	343	350
incorrectly classified low intervention students	161	174	152	102	169	162

TABLE 5.7: A comparison of the six classifier models performance in terms of numbers of students that were classified correctly and incorrectly

with a difference of only six students less that those for logistic regression. Random Forest forest was third with 21 records less than logistic regression. The same pattern is repeated for the number of students that were correctly classified as being in the high intervention class. This number is the focus of this study; a high number of correctly classified high intervention records means the classifier is suitable for the type of data used. Therefore, logistic regression could be considered the most suitable classifier for the type of data used in this study. A further analysis of the results obtained using the selected six metrics is presented next.

The results of the classifiers' performance using six selected metrics are shown in Table 5.8

Model	Recall	Specificity	ROC	F-Measure	Kappa	RMSE
LR	0.924	0.686	0.887	0.897	0.6345	0.3375
MLP	0.873	0.660	0.851	0.865	0.5407	0.4124
SMO	0.911	0.703	0.807	0.894	0.6309	0.3893
NB	0.701	0.801	0.846	0.784	0.4403	0.4264
J48	0.905	0.670	0.822	0.884	0.5941	0.3720
RF	0.907	0.684	0.870	0.888	0.6082	0.3471

TABLE 5.8: A comparison of the classifiers' performance using the six selected metrics

Table 5.8 presents the comparison of classification performance of all the six classifier models, using six metrics. The first metric is the recall value, also known as sensitivity, it is the proportion of all the high intervention records correctly identified from all the existing high intervention records. The results show that a majority of the classifiers are highly sensitive in classifying the high intervention class, with four out of the six classifiers achieving a recall probability of over 90%. Logistic regression achieved the highest recall probability, while naïve Bayes classifier has the lowest recall probability. The high intervention class is the class of interest, the objective being to identify the

students that require high intervention early enough so that intervention measures could be initiated for the students to improve in their final examination marks.

Specificity, the proportion of the low intervention records identified from among the actual low intervention records, ranges from 80% to 66%. The majority of the classifiers had specificity probabilities below 70%. This low specificity probability is attributed to the class imbalance, the dataset had nearly 50% as many low intervention samples as there were high intervention samples. Prevalence, the proportion of a given class, directly affects the mean square error which is correlated to classifier performance (Mazurowski et al., 2008). Therefore, because the low intervention class is underrepresented the classifiers tend to assign student records to the high intervention class.

The Receiver Operation Characteristic (ROC) Area is known to be a reliable measure of classifier performance, remaining stable even for imbalanced classes (Brown and Davis, 2006). The performance values therefore are a reasonable representation of the classifiers' performance. As seen in all the six figures, the ROC values remain constant for the two classes in each classifier. The highest performance was 88.7%, attained by logistic regression, and the lowest was 80.7%, attained by naïve Bayes.

The ROC curves for the six classifiers are illustrated in Figure 5.7.

As shown in the figure, the shape of each curve determines the area under it, a curve close to the 45 degrees diagonal line means the classifier makes a random guess, while a curve that shows a rapid increase creates a steep curve. Meaning, there are more true positives and less false positives. The classifier, therefore creates a bigger area under the curve. The bigger the area under the curve the better the classifier performance (Akobeng, 2007). As seen, logistic regression achieved the highest ROC value making it the most suitable of the six classifiers.

Another important measure is F-Measure. It is capable of determining the performance of the different classes separately. The F-Measure metric is a harmonic average of precision and the recall rates, considering only the high intervention class (Shaikh et al., 2015). This study focuses on the high intervention class, therefore, the table presents the F-Measure values for the high intervention class. As seen, the values were reasonable for all the classifiers except for naïve Bayes. Logistic regression attained the highest value of 89.7%. The high values could be explained by the fact that the F-Measure is about the high intervention class that also attained high recall probabilities.

Cohen's Kappa measures the extent of agreement between the classifiers on how they classify the student records into the classes of high and low intervention (Mackinnon, 2000). The Kappa value between any two raters should be at least 0.60 (Wood, 2007). This is, therefore, the Kappa value that a classifier needs to attain to be considered

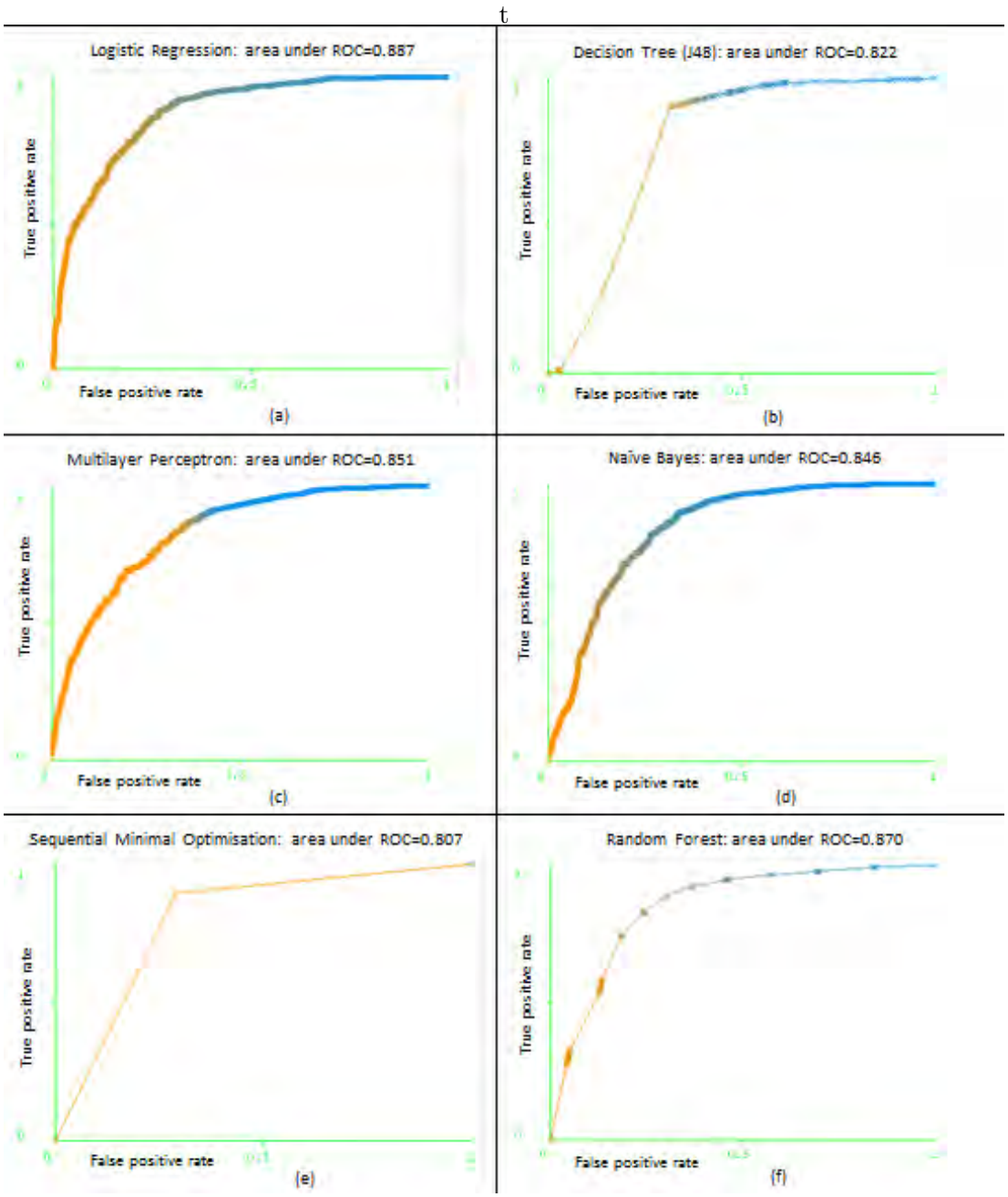


FIGURE 5.7: The ROC curves; showing the area under the curve drawn by varying sensitivity $x-axis$ and specificity $y-axis$ for: (a) logistic regression; (b) Decision Tree (J48); (c) Multilayer Perceptron; (d) Naive Bayes; (e) Sequential Minimal Optimisation; and (f) Random Forest.
The colors show the threshold values.

suitable for use. The Kappa values for logistic regression, SMO, and Random Forest reached the threshold values with logistic regression attaining the highest value of 0.6345.

Finally, root mean square error (RMSE) is the average of the error calculated by finding the difference between predicted values and the actual values (Chai and Draxler, 2014). A classifier is rated to perform well when it has a low RMSE value (Pardos et al., 2012).

The experimental results show that logistic regression has the lowest value of RMSE of 0.3375.

These experimental results, as obtained from using the six metrics, confirm that logistic regression is the most suitable classifier model for the data used in this study. Out of the six classifiers, logistic regression got the best metric value, making it the best classifier for the type of data used in this study.

The bar graph in Figure 5.8 presented summarises the classification performance of the six classifier models. The graph shows that logistic regression performance, represented by the blue bar is the best in all the metrics except for specificity. In the RMSE, the blue bar is the shortest, meaning that logistic regression has the lowest error and hence the best classifier. Logistic regression performed poorly with the specificity metric measure because, specificity is concerned with low intervention records, which that were fewer that the high intervention records; this class imbalance affected the specificity measure. However, the focus of this study is on the high intervention students so that intervention measures could be put in place early enough for the students to improve in their final examination scores.

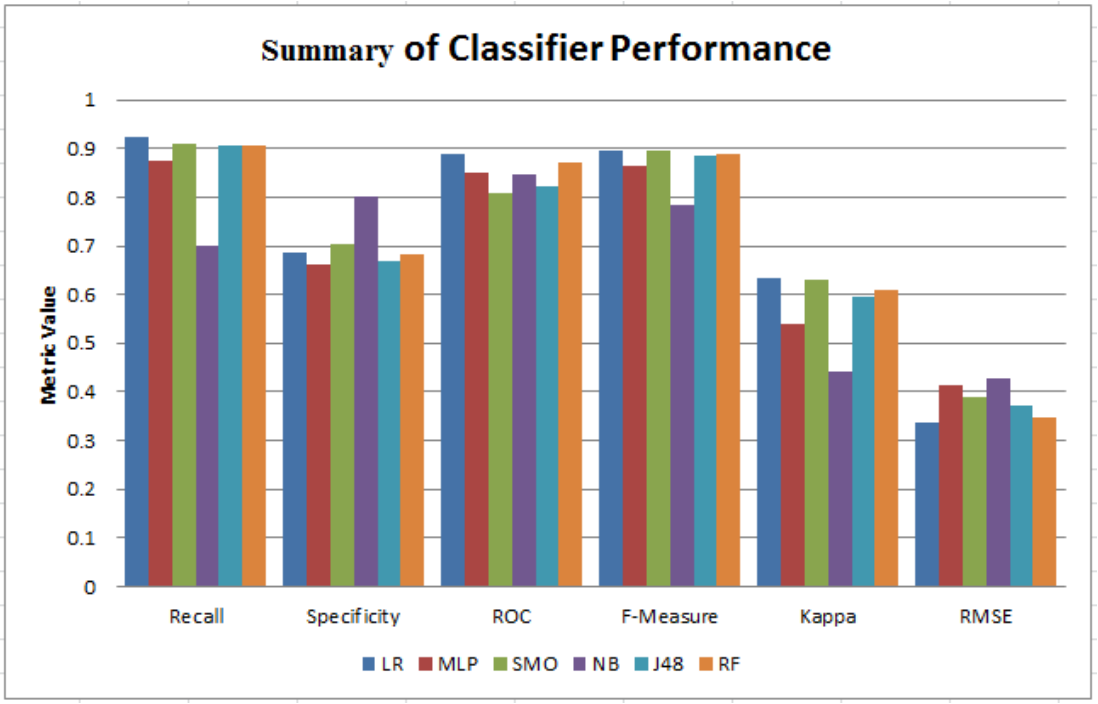


FIGURE 5.8: Summary of the six classifiers' prediction performance Using the Six Selected Metrics

5.3 Finding the Optimal Feature Subset Using Rural Schools' Dataset

This section presents the results of finding the optimal feature subset for a the dataset that was collected from rural schools. Determining the optimal feature subset is important because a reduced dataset will enhance effectiveness in modelling in terms of computer resource saving and suitability for use on a mobile phone. In this study, the optimal feature subset will be more suitable to implement on the Mobile Academic Performance Prediction System (MAPPS) since the mobile phone has a small screen.

This study adopted an approach of finding the optimal feature subset, where the optimal features are searched by successively modelling from a minimum determined subset to the maximum number of the features (Ramaswami and Bhaskaran, 2009). First, a complete dataset of 2426 student records with all the 22 features from rural schools was used to rank the features. This was achieved using three filter algorithms: ReliefF (RF) algorithm, Information Gain (IG), and Gain Ratio (GR).

5.3.1 Ranking of Features using ReliefF Algorithm

The ReliefF filter algorithm ranked the features as shown in Figure 5.9. The experiments were conducted in WEKA. As shown, the order of the features from the most important feature to the least is indicated in the third column.

5.3.2 Ranking of Features using Information Gain Algorithm

The Information Gain algorithm ranked features are presented in Figure 5.10. The ranking is presented in the third column. A notable observation is that the first three features in the information gain ranking are similar to the ranking of the first three features in the reliefF algorithm. This means, the three features are highly indicative of the target class and shows that these features could be considered as the least possible features. This method was used by Ramaswami and Bhaskaran (2009).

5.3.3 Ranking of Features using Gain Ratio Algorithm

The ranking of the features using Gain Ratio is presented in Figure 5.11. As shown in the third column, the first feature is the same as for reliefF and Information Gain, however, the others are different.

Attribute selection output

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1

average merit      average rank  attribute
 0.119 +- 0.003      1   +- 0        1 tot_mks
 0.025 +- 0.002      2.1 +- 0.3      2 sex
 0.02  +- 0.003      3.4 +- 0.49    22 teach_sh
 0.019 +- 0.002      3.7 +- 0.78    21 faclty_sh
 0.015 +- 0.002      5.1 +- 0.94    17 comm_inv
 0.013 +- 0.002      5.9 +- 0.54    20 teach_abs
 0.011 +- 0.001      7.5 +- 0.92    14 parnt_edc
 0.01  +- 0.002      8.7 +- 1.79    12 parnt_stb
 0.01  +- 0.001      9.2 +- 1.72    10 p_motvn
 0.009 +- 0.001     10.6 +- 1.43     7 displn
 0.008 +- 0.001     12.7 +- 2.37    13 famly_inc
 0.008 +- 0.002     12.8 +- 3.97    15 famly_s
 0.008 +- 0.001     12.9 +- 1.92     6 study_t
 0.008 +- 0.001     13   +- 1.26    11 parnt_enc
 0.007 +- 0.001     14.4 +- 1.96     4 dist
 0.006 +- 0.001     15.5 +- 2.25    16 parnt_inv
 0.006 +- 0.002     16.7 +- 2.87     8 inter_eng
 0.006 +- 0.001     17.3 +- 1.42    18 teach_att
 0.005 +- 0.001     18.2 +- 0.6     19 teach_com
 0.004 +- 0         19.9 +- 0.54     9 educ_att
 0     +- 0.002     21.2 +- 2.09     5 abst
 0.001 +- 0.002     21.2 +- 0.4      3 age
```

FIGURE 5.9: ReliefF ranked features from the most important to the least important

As observed in the three sets of ranked features, feature $A1(tot\_mks)$, or test marks, is the most important as it appears in position one in all the three subsets. The first tree features:$tot\_mks(testmarks), sex(gender), teach\_sh(teachershortage)(A1, A2, A22)$ are shared by the ranking of Information Gain and ReliefF algorithms. It can also be observed that $A10$(pupil motivation) and $A3$ (student age) are shared among the top 6 features by Information Gain and Gain Ratio. This shared features could mean they are indicative of the target class. Therefore, it is proposed that, the first three features, $A1$, $A2$, and $A22$ are the core features. A similar approach was used by Harb and Moustafa (2012). The results of the models built starting with the first three features are presented next.

Attribute selection output

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1

average merit      average rank  attribute
 0.321 +- 0.007      1    +- 0        1 tot_mks
 0.023 +- 0.002      2    +- 0        2 sex
 0.018 +- 0.002      3.2 +- 0.4     22 teach_sh
 0.016 +- 0.001      3.9 +- 0.54    10 p_motvn
 0.013 +- 0.001      5.1 +- 0.54    13 famly_inc
 0.012 +- 0.002      6.2 +- 0.87     3 age
 0.011 +- 0.002      7.6 +- 1.02     6 study_t
 0.01  +- 0.001      7.8 +- 0.87    18 teach_att
 0.008 +- 0.001      9.7 +- 1        8 inter_eng
 0.008 +- 0.002     10.3 +- 2.33    21 faclty_sh
 0.007 +- 0.001     10.6 +- 1.74     5 abst
 0.006 +- 0.001     12.7 +- 1.55    19 teach_com
 0.006 +- 0.001     13   +- 1.79    11 parnt_enc
 0.006 +- 0.001     13.4 +- 1.85     4 dist
 0.005 +- 0.001     14.1 +- 1.45     7 displn
 0.005 +- 0.001     16.7 +- 1        9 educ_att
 0.004 +- 0.001     16.8 +- 1.47    12 parnt_stb
 0.004 +- 0.001     18   +- 1.61    15 famly_s
 0.004 +- 0.001     18.1 +- 0.94    20 teach_abs
 0.002 +- 0.001     20.2 +- 0.75    16 parnt_inv
 0      +- 0        21.2 +- 0.75    17 comm_inv
 0      +- 0        21.4 +- 0.66    14 parnt_edc
```

FIGURE 5.10: Information Gain ranked features from the most important to the least important

5.3.4 Selecting the Optimal Feature Subset by Successive Modelling

In this subsection, the results of finding the optimal subset are presented. The same approach by Ramaswami and Bhaskaran (2009) was followed, but, in their study, they used metrics of accuracy and time. We used the ROC area and Root Mean Square Error (RMSE) metrics. These two metrics were selected out of the six selected metrics discussed earlier because of their ability to work well with unbalanced datasets as is the case with the dataset used in this study.

The optimal feature selection process started by first selecting the features that could be considered the top most predictive features of the target class from the three sets of ranked features. As discussed in the previous subsection above, three features $A1$, $A2$, and $A22$ were picked. These were then used to build the first six models. Results of performance are recorded before the next models are built. Other classifiers were then

Attribute selection output

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1

average merit      average rank  attribute
 0.176 +- 0.003      1   +- 0        1 tot_mks
 0.047 +- 0.004      2.2 +- 0.4     10 p_motvn
 0.037 +- 0.006      2.8 +- 0.4     18 teach_att
 0.024 +- 0.004      5.2 +- 1.25    19 teach_com
 0.023 +- 0.002      5.5 +- 1.2      2 sex
 0.021 +- 0.003      5.8 +- 1.25     3 age
 0.02  +- 0.004      6.9 +- 2.17     5 abst
 0.018 +- 0.003      7.8 +- 1.72    11 parnt_enc
 0.016 +- 0.002      8.7 +- 1.19     9 educ_att
 0.014 +- 0.001     10.3 +- 1       13 famly_inc
 0.014 +- 0.002     10.6 +- 1.36     6 study_t
 0.013 +- 0.001     11.2 +- 0.75    22 teach_sh
 0.008 +- 0.001     14.1 +- 1.14    21 faclty_sh
 0.008 +- 0.001     14.2 +- 0.6      8 inter_eng
 0.008 +- 0.001     15.2 +- 1.47     7 displn
 0.007 +- 0.001     16   +- 1.79    12 parnt_stb
 0.007 +- 0.001     16.8 +- 1.89     4 dist
 0.006 +- 0.001     17.7 +- 1.49    20 teach_abs
 0.004 +- 0.003     18.8 +- 1.72    16 parnt_inv
 0.004 +- 0.001     19.8 +- 0.87    15 famly_s
 0      +- 0        21   +- 0.77    17 comm_inv
 0      +- 0        21.4 +- 0.66    14 parnt_edc
```

FIGURE 5.11: Gain Ratio ranked features from the most important to the least important

successively built by adding one feature in every classifier building iteration until all the 22 features are used. In this study, however, it was decided that the feature subset search space be the first16 features in each rank. This is because the effect of adding a feature after the first 16 features did not have a significant effect on the accuracy.

The results of each of the successive sets of models are discussed next.

5.3.4.1 ReliefF Algorithm Ranked Features

Table 5.9 shows the ROC area values and Root Mean Square Error (RMSE) values, renamed (RE) in the table. The highest performance values are coloured red.

As seen from the table, the highest performance for the six classifiers appear in the range of $3 - 11$ features. Random Forest attains the highest ROC value of 0.889 and RMSE value of 0.3369 with the top three features. Logistic Regression attains the second

#F	LR		MLP		SMO		NB		J48		RF	
	ROC	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**
3	.874	.342	.871	.346	.812	.386	.879	.343	.802	.353	.889	.337
4	.874	.342	.886	.340	.812	.386	.875	.345	.802	.353	.887	.338
5	.874	.342	.875	.345	.812	.386	.875	.345	.802	.353	.882	.341
6	.873	.343	.877	.349	.812	.386	.875	.346	.802	.353	.876	.344
7	.873	.343	.876	.346	.812	.386	.876	.342	.805	.353	.858	.365
8	.880	.340	.867	.362	.812	.386	.878	.343	.806	.350	.853	.370
9	.882	.340	.868	.357	.809	.389	.873	.349	.804	.351	.848	.373
10	.884	.338	.864	.361	.809	.389	.875	.349	.832	.359	.851	.367
11	.886	.338	.868	.364	.809	.389	.876	.349	.838	.358	.857	.361
12	.886	.338	.858	.374	.809	.389	.875	.350	.832	.362	.861	.354
13	.886	.338	.841	.388	.810	.388	.876	.350	.820	.378	.871	.345
14	.886	.339	.846	.398	.809	.389	.866	.369	.827	.365	.864	.350
15	.885	.339	.855	.406	.809	.389	.864	.365	.818	.373	.866	.350
16	.884	.339	.843	.411	.809	.389	.859	.370	.823	.371	.864	.352
17	.885	.339	.857	.405	.809	.389	.861	.368	.819	.368	.860	.384
18	.884	.339	.850	.403	.808	.389	.858	.385	.8172	.369	.867	.349
19	.886	.339	.854	.405	.807	.389	.853	.407	.820	.369	.853	.356
20	.885	.339	.853	.40	.807	.389	.847	..425	.816	..368	.867	.349
21	.885	.339	.860	.405	.807	.389	.844	.432	.814	.374	.859	.354
22	.887	.338	.851	.412	.807	.389	.846	.426	.822	.372	.870	.347
max	.886	.338	.886	.340	.812	.386	.879	.342	.838	.350	.889	.337

TABLE 5.9: Performance of six classifiers on ReliefF ranked attributes

highest values with 11 features. It should also be noted that the two selected metrics agree. Except in the case of NB where the highest ROC value was at 3 features and the lowest RMSE at 7 features, and J48 where the highest ROC value was at 11 features and the lowest RMSE value at 8 features, the other four models have the two metric values at the same level. Since the concentration of highest ROC values and lowest RMSE values for all the models lies within the $3 - 11$ features range, it can be concluded that the first 11 form the suboptimal subset.

5.3.4.2 Information Gain Ranked Features

Table 5.10 shows the ROC area and the RMSE values obtained from the models built using the Information Gain (IG) ranked features.

The results show that the highest ROC area value of 0.893 was attained by MLP with 6 features and the lowest RMSE of 0.3378 rounded to 0.338 was attained by Logistic Regression with 7 features. As observed in the table, the highest ROC values and the lowest RMSE values for all the models lie in the range of 3 to 7 features. Because of the

#F	LR		MLP		SMO		NB		J48		RF	
	ROC	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**
3	.874	.342	.871	.346	.812	.386	.879	.343	.802	.353	.889	.337
4	.878	.340	.889	.337	.809	.389	.875	.345	.802	.353	.888	.337
5	.879	.340	.889	.338	.809	.389	.874	.346	.802	.353	.877	.345
6	.886	.338	.893	.338	.809	.389	.881	.342	.800	.354	.883	.345
7	.885	.338	.886	.341	.808	.389	.879	.342	.798	.355	.876	.353
8	.884	.339	.886	.341	.808	.389	.875	.352	.798	.355	.871	.355
9	.884	.338	.877	.347	.808	.389	.875	.351	.797	.360	.853	.359
10	.884	.338	.887	.347	.808	.389	.875	351	.797	.360	.853	.359
11	.853	.359	.878	.354	.808	.389	.875	.352	.798	.360	.849	.366
12	.883	.339	.869	.366	.808	.389	.362	.373	.801	.358	.860	.356
13	.884	.339	.867	.371	.807	.389	.857	.393	.801	.358	.862	.356
14	.883	.339	.857	.375	.807	.390	.857	.396	.818	.360	.854	.365
15	.884	.339	.872	.384	.808	.389	.867	.396	.856	.396	.862	.355
16	.883	.339	.864	.385	.807	.389	.851	.414	.828	.364	.861	.355
17	.887	.337	.863	.392	.808	.389	.808	.389	.855	.339	.871	.350
18	.887	.337	.853	.393	.807	.389	.851	.417	.823	.367	.864	.354
19	.887	.337	.856	.399	.807	.389	.850	.421	.825	.365	.871	.348
20	.886	.337	.862	.397	.807	.389	.846	.427	.834	.359	.863	.354
21	.886	.338	.849	.412	.807	.389	.846	.427	.832	.361	.863	.352
22	.887	.338	.851	.412	.807	.389	.846	.426	.822	.372	.870	.347
max	.886	.338	.893	.338	.812	.386	.881	.342	.802	.353	.889	.337

TABLE 5.10: Performance of six classifiers on Information Gain ranked attributes

decision to limit the search space to the first 16 features, the optimal subset using the Information Gain feature ranking algorithm therefore includes seven features.

5.3.4.3 Gain Ratio Ranked Features

Table 5.11 shows the ROC area and the RMSE values obtained from the models built using the Gain Ratio (GR) ranked features.

As shown, the highest ROC area value of0.888 and the lowest RMSE value was attained by Logistic Regression with 16 features. The second highest ROC area value was 0.881 with 12 features using MLP. As observed from the results, it is clearly difficult to make a decision of the optimal feature subset. Feature selection is the process of selecting a minimal number of features that achieves the best possible classifier performance (Kohavi and John, 1997). In the case of Gain Ratio ranked features, the highest values for all the models are scattered in the range of 3 to 16 features. This does not fit in the definition of a small number of features compared to the total of number of 22 features. It may therefore be concluded that the ranking by Gain Ratio does not attain an optimal

#F	LR		MLP		SMO		NB		J48		RF	
	ROC	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**
3	.868	.342	.865	.346	.808	.389	.857	.355	.802	.353	.861	.345
4	.870	.342	.868	.345	.808	.389	.852	.373	.802	.353	.859	.345
5	.873	.341	.868	.345	.809	.389	.857	.374	.802	.353	.866	.346
6	.881	.338	.873	.346	.808	.389	.863	.369	.801	.354	.874	.343
7	.880	.339	.869	.348	.809	.388	.860	.375	.801	.354	.869	.349
8	.879	.339	.867	.349	.809	.389	.854	.395	.801	.354	.870	.347
9	.879	.339	.868	.350	.807	.389	.848	.413	.802	.354	.869	.348
10	.883	.339	.869	.351	.807	.389	.851	.408	.802	.354	.867	.351
11	.883	.339	.861	.354	.807	.389	.850	.411	.802	.354	.856	.356
12	.883	.339	.881	.349	.808	.389	.852	.411	.800	.354	.880	.351
13	.882	.340	.882	.359	.807	.389	.851	.413	.800	.356	.869	.356
14	.883	.339	.862	.374	.808	.389	.852	.411	.803	.357	.866	.354
15	.884	.339	.869	.373	.808	.389	.853	.412	.811	.359	.867	.351
16	.888	.337	.861	.389	.809	.389	.853	.415	.798	.359	.863	.353
17	.887	.337	.863	.392	.808	.389	.851	.417	.820	.365	.871	.350
18	.887	.337	.857	.395	.808	.389	.850	.420	.822	.366	.870	.351
19	.870	.351	.852	.405	.808	.389	.808	.389	.833	.360	.866	.349
20	.886	.337	.862	.397	.807	.389	.846	.427	.834	.359	.863	.354
21	.886	.338	.849	.412	.807	.389	.846	.427	.832	.361	.863	.352
22	.887	.338	.851	.412	.807	.389	.846	.426	.822	.372	.870	.347
max	.888	.337	.881	.346	.809	.388	.863	.355	.834	.353	.871	.343

TABLE 5.11: Performance of six classifiers on Gain Ratio ranked attributes

feature subset that is useful for the purpose of reducing the number of features in order to make the subset suitable for use on a small screen mobile device.

5.3.4.4 A discussion and summary of the results

A summary of results from the search of an optimal feature subset are presented in Table 5.12. The three algorithms used to rank the features include ReliefF, Information Gain, and Gain Ratio. Iterations of successive modelling were carried out in order to discover the smallest number of features that achieve a high performance comparable to the complete set of features.

Algorithm	Highest value of ROC area	**Lowest value of RMSE**	**Range of Optimal subset**
ReliefF	0.889	0.338	3 - 11
Information Gain	0.893	0.338	3 - 7
Gain Ratio	0.888	0.337	3 - 16

TABLE 5.12: A summary of results obtained from successive modelling of classifiers to determine an optimal subset of features using three sets of ranked features with three selected algorithms: ReliefF, Information Gain, and Gain Ratio

Results from the experiments indicate that the Information Gain algorithm's ranked features provide the minimum number of features that also attains the highest value of ROC area and a low value of the Root Mean Square Error (RMSE). These experiments have demonstrated that a feature subset can be selected to reduce the dataset from 22 features to only 7 features with minimal loss in performance. The highest ROC area values for all 22 features using Logistic Regression is 0.887, while that of the 7 features is 0.886. The lowest error value is 0.3375 for all the 22 features while that of the 7 features using Logistic Regression is 0.3378. This proves there is an insignificant change in performance. The reduced dataset therefore saved resources without compromising the performance of classifiers. In this study, the 7 features as ranked using the Information Gain ranking algorithm were adopted as the optimal feature subset. These features are: test marks, gender, teacher shortage, student motivation, family income, student age, and study time. These features were used to build the Mobile Academic Performance Prediction System.

5.3.5 Classification Performance of Logistic Regression Using the Optimal Feature Subset

Having determined logistic regression as the best classifier model for the type of data used in this study, and the optimal feature subset containing seven features, this subsection presents results specifically for the classification performance of the model built using the determined optimal feature subset

A dataset of the 7 optimal features was extracted from the complete rural schools' dataset and used to build the logistic regression model. For consistency with the previous experiments, WEKA was used, and the results are presented in Figure 5.12

Table 5.13 presents a summary of the experimental results.

Selected Item	**Result**	**Metric**	**Value**
Students correctly classified	1444	Recall Value	0.916
Students incorrectly classified	252	Specificity	0.703
correctly classified high intervention students	1084	ROC area	0.885
incorrectly classified high intervention students	100	F-Measure	0.896
correctly classified low intervention students	360	Kappa Value	0.637
incorrectly classified low intervention students	152	RMSE value	0.3378

TABLE 5.13: The results of logistic regression model using the optimal dataset from rural schools

```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1444               85.1415 %
Incorrectly Classified Instances       252               14.8585 %
Kappa statistic                          0.637
Mean absolute error                      0.2282
Root mean squared error                  0.3378
Relative absolute error                 54.1339 %
Root relative squared error             73.5896 %
Coverage of cases (0.95 level)          99.2335 %
Mean rel. region size (0.95 level)      82.8715 %
Total Number of Instances             1696

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.703    0.084    0.783      0.703    0.741      0.639    0.885     0.801     LowInt
                 0.916    0.297    0.877      0.916    0.896      0.639    0.885     0.939     HighInt
Weighted Avg.    0.851    0.233    0.849      0.851    0.849      0.639    0.885     0.898

=== Confusion Matrix ===

    a    b   <-- classified as
  360  152 |    a = LowInt
  100 1084 |    b = HighInt
```

FIGURE 5.12: Model built on optimal features using rural data

The table shows the classification performance of logistic regression model built with the optimal feature subset. The performance measures, as seen, are good enough to motivate its adoption for implementation in the mobile academic performance prediction system. The recall value is 91.6%, which indicates the ability of the classifier to correctly identify high intervention students. Similarly, the F-Measure metric is 89.6%, this is the ability of the classifier to classify the high intervention class. The ROC area is 88.5% which indicates a reasonable overall prediction performance of the model. Further, the Kappa value of 0.637, indicates an above average classification ability compared to the other five classifiers. The low RMSE of 0.3378 indicates that logistic regression makes the smallest compared to the other classifiers. It can be concluded that the logistic regression model built on the optimal feature subset would be useful developing MAPPS.

5.4 Finding the best Classifier Model Using Peri-Urban Dataset

This section presents results for the second phase of experiments that used the peri-urban dataset. Because of limitation of resources, the study collected 1105 records from 11 public primary schools in the outskirts of Mombasa city, Kenya. Similar features to those used in the rural dataset were used to collect data. However, three features were

not included. These were: distance to school, command of English, and community involvement. For the first feature, it was noticed that most schools were built near residential places hence distance was constant for most students; for command of English, most students learn to speak English because of the metropolitan environment they live in, and for community involvement, in peri-urban areas, the city council takes the role of maintaining schools.

Similar experiments to those conducted with the rural schools' dataset to determine the best classifier model were conducted with each of the six classifiers. The classification performance results were analysed and results are presented in the next subsection. These experiments provided important insights in to the characteristics of students from the peri-urban region that could be contrasted with the student characteristics from rural schools.

5.4.1 Discussion of Performance Findings

Classifier models were built using the peri-urban data, and 10-fold cross validation evaluation conducted. The classification performance results are compared in Table 5.14.

Selected Item	**LR**	**MLP**	**SMO**	**NB**	**J48**	**RF**
Students correctly classified	559	523	566	525	550	546
Students incorrectly classified	104	140	97	138	113	117
correctly classified high intervention students	189	175	193	184	173	181
incorrectly classified high intervention students	54	68	50	59	70	62
correctly classified low intervention students	370	348	373	341	377	365
incorrectly classified low intervention students	50	72	47	79	43	55

TABLE 5.14: A comparison of the six classifier models in terms of the actual numbers of students that were classified correctly and incorrectly

The results compare the six classifiers' performance in terms of actual numbers of student records that were correctly and incorrectly classified. The results show that SMO performed better: SMO correctly classified 566 records, 7 records more than logistic regression which was the second best classifier. SMO also misclassified the least number of records, 97, 7 records less that than the misclassified records by logistic regression. A similar pattern is reflected in the number of correctly classified high intervention students and correctly classified low intervention students; SMO does marginally better than logistic regression by a small number of records.

Next, an analysis of the results obtained using the six metrics is presented in Table 5.15.

Model	Recall	Specificity	ROC	F-Measure	Kappa	RMSE
LR	0.778	0.881	0.897	0.784	0.661	0.3463
MLP	0.720	0.829	0.856	0.714	0.5468	0.4145
SMO	0.794	0.888	0.841	0.799	0.6841	0.3825
NB	0.757	0.812	0.850	0.727	0.5594	0.3953
J48	0.712	0.898	0.837	0.754	0.6342	0.3649
RF	0.745	0.869	0.874	0.756	0.6177	0.3618

TABLE 5.15: A comparison of the six classifiers' performance using the six selected metrics

The results show that SMO attained better performance in four measures: recall, specificity, F-Measure, and Kappa value. Logistic regression had the highest values in two: ROC area, and RMSE.

For this dataset, therefore, SMO is the best classifier model even though it only marginally did better than logistic regression. In a case where simplicity counts, either of the two could be selected as the best classifier, considering the simplicity of logistic regression. Therefore, although, the peri-urban data was very similar to the rural schools' data in terms of the attributes used, results have shown that there are differences in the attribute values which have a major effect in terms of the most suitable classifier model for the type of data. These results are in agreement with previous results, that no single classifier performs best in all situations (Asif et al., 2014).

5.5 Finding the Optimal Feature Subset Using Peri-Urban Schools' Dataset

In this section, the study seeks to investigate the features in the peri-urban dataset to discover the optimal feature subset. The objective is to compare the optimal feature subsets to get an understanding of the differences between the peri-urban student characteristics and the rural schools' student characteristics.

A similar procedure as that used for finding the optimal feature subset for the rural dataset was used. The features were first ranked using three ranking algorithms as discussed next.

5.5.1 Ranking the Features

5.5.1.1 Ranking of Features for Peri-Urban Data with ReliefF

The features ranked by ReliefF filter are as shown in Figure 5.13. The experiments were conducted in WEKA. As shown, the order of the features from the most important feature to the least is indicated in the third column.

```
=== Attribute selection 10 fold cross-validation (stratified)

average merit      average rank  attribute
 0.152 +- 0.006     1    +- 0       1 testmarks
 0.024 +- 0.002     2.6 +- 0.49   12 p_education level
 0.026 +- 0.004     2.7 +- 0.9    17 t_ absenteeism
 0.019 +- 0.002     4.8 +- 1.08    6 s_indiscipline
 0.017 +- 0.002     5.2 +- 0.87   19 f_l _teachers
 0.016 +- 0.004     5.4 +- 1.5     3 sex
 0.015 +- 0.002     6.4 +- 1.02   18 s_l_facilities
 0.009 +- 0.002     8.6 +- 0.49   10 p_stability
 0.009 +- 0.002     8.7 +- 0.78   13 familiy_size
 0.007 +- 0.002     9.6 +- 0.92   11 family_inc
 0.002 +- 0.001    12.2 +- 1.25   16 t_commitment
 0.003 +- 0.002    12.3 +- 1.73   15 t_ attitude
 0.002 +- 0.001    13    +- 1.48    5 study_time
 0.001 +- 0.001    14.5 +- 1.43    7 s_attitude
 0.001 +- 0.001    15.2 +- 1.47    8 s_motivation
 0.001 +- 0.002    15.3 +- 1.62    4 absenteesm
 0      +- 0.002    15.9 +- 1.76   14 p_ involvement
-0.001 +- 0.001    17.6 +- 0.49    9  p_encouragement
-0.005 +- 0.002    19    +- 0       2 age
```

FIGURE 5.13: ReliefF ranked features from the most indicative of the target class to the least least indicative using peri-Urban dataset

The ranking in the figure shows that *test-marks* is the highest indicator of the target class followed by *parent-educational-level*, and *teacher-absenteeism*, the last attribute is *student-age*. This ranking is indicated by the average merit where *test-marks* has the highest value and therefore is considered the most important. The average rank is obtained from the average attribute position throughout the 10-fold cross validation. The positions of each attribute appear in groups to signify the closeness of the predictive ability of the attributes. The attribute column shows the position of the attribute as determined by the average merit and the average rank.

5.5.1.2 Ranking of Features for Peri-Urban Data with Information Gain

The Information Gain algorithm ranked features are presented in Figure 5.14.

```
=== Attribute selection 10 fold cross-validation (stratified)

average merit      average rank  attribute
 0.418 +- 0.01      1   +- 0       1 testmarks
 0.041 +- 0.004     2   +- 0      12 p_education level
 0.024 +- 0.004     3.5 +- 0.5     2 age
 0.025 +- 0.002     3.5 +- 0.5    17 t_ absenteeism
 0.014 +- 0.006     5.6 +- 0.8     6 s_indiscipline
 0.013 +- 0.002     6.3 +- 0.9     3 sex
 0.011 +- 0.004     6.5 +- 1.12   11 family_inc
 0     +- 0         9.6 +- 1.74    4 absenteesm
 0     +- 0         9.8 +- 1.4     5 study_time
 0.002 +- 0.004    11.1 +- 1.81   19 f_l _teachers
 0.001 +- 0.004    11.5 +- 2.77    7 s_attitude
 0     +- 0        12.1 +- 2.95    8 s_motivation
 0     +- 0        13.6 +- 1.2    15 t_ attitude
 0.003 +- 0.005    13.9 +- 4.37   18 s_l_facilities
 0     +- 0        14.3 +- 1.42   16 t_commitment
 0     +- 0        14.5 +- 1.86    9  p_encouragement
 0     +- 0        15.8 +- 0.6    14 p_ involvement
 0     +- 0        16.4 +- 3.23   13 familiy_size
 0     +- 0        19   +- 0      10 p_stability
```

FIGURE 5.14: Information Gain ranked features using peri-Urban Dataset

A noticeable observation in the ranking of information gain is that the first two features, *test marks*, and *parent-education-level*, hold the same top places as in the reliefF ranked features. These two features could therefore be seen as the most important. The two attributes also achieved the highest average merit, making them the most important features for predicting the target class.

5.5.1.3 Ranking of Features for Peri-Urban Data with Gain Ratio

The ranking of the features using Gain Ratio is presented in Figure 5.15.

As shown in the first features are *test-marks*, *parent-education-level*, and *teacher-absenteeism*. These three features are similar to those ranked by reliefF, and the first two features are similar to those ranked by Information Gain.

Therefore, it was proposed that the two features, *test-marks*, and *parent-education-level*, that are at the top in all the three ranks: by reliefF, Gain Ratio, and Information gain,

```
=== Attribute selection 10 fold cross-validation (stratified)

average merit      average rank  attribute
 0.228 +- 0.006     1    +- 0       1 testmarks
 0.041 +- 0.004     2    +- 0      12 p_education level
 0.025 +- 0.002     4    +- 1      17 t_ absenteeism
 0.024 +- 0.004     4.1 +- 0.83     2 age
 0.022 +- 0.009     4.3 +- 1.19     6 s_indiscipline
 0.013 +- 0.002     6.9 +- 0.83     3 sex
 0.014 +- 0.005     7.3 +- 2.97    11 family_inc
 0     +- 0         9.2 +- 1.33     4 absenteesm
 0     +- 0        10.1 +- 1.51     5 study_time
 0.003 +- 0.008    10.6 +- 2.84     7 s_attitude
 0.003 +- 0.006    10.9 +- 2.21    19 f_l _teachers
 0     +- 0        12.1 +- 3.21     8 s_motivation
 0     +- 0        13.7 +- 0.9     15 t_ attitude
 0.004 +- 0.006    14.1 +- 4.23    18 s_l_facilities
 0     +- 0        14.4 +- 1.2     16 t_commitment
 0     +- 0        14.6 +- 1.96     9  p_encouragement
 0     +- 0        15.2 +- 2.09    14 p_ involvement
 0     +- 0        16.5 +- 3.01    13 familiy_size
 0     +- 0        19   +- 0       10 p_stability
```

FIGURE 5.15: Gain Ratio ranked features using peri-Urban Dataset

be considered the core features. A similar approach was followed in (Harb and Moustafa, 2012, Ramaswami and Bhaskaran, 2009). The two features were used to build the first model and successive models were then built using the different lists of features in the three ranked lists. The results of the models built starting with the first two features are presented in the next subsection.

5.5.2 Selecting the Peri-Urban Optimal Feature Subset by Successive Modelling

5.5.2.1 ReliefF Algorithm Ranked Features

Table 5.16 shows the values of ROC area and Root Mean Square Error (RMSE) values, renamed (RE) in the table. The best performance for each classifier is coloured in red for easy identification.

As shown in the table, the highest performance for the six classifiers appear in the range of *2-11* features. Logistic regression attains the highest ROC value of 0.905 and a RMSE value of 0.338 with the top eleven features. Multilayer perceptron (MLP) and Random Forest attain the their highest values with the first two features, while naïve Bayes and

#F	LR		MLP		SMO		NB		J48		RF	
	ROC	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**
2	.882	.341	.899	.342	.841	.382	.884	.351	.813	.352	.892	.342
3	.894	.339	.891	.346	.841	.382	.892	.331	.814	.353	.890	.344
4	.898	.339	.889	.352	.841	.382	.894	.349	.813	.356	.866	.370
5	.897	.340	.889	.355	.841	.382	.888	.356	.818	.356	.869	.370
6	.898	.340	.884	.357	.841	.382	.890	.356	.811	.359	.863	.372
7	.899	.341	.879	.363	.841	.382	.885	.361	.819	.357	.876	.361
8	.902	.339	.882	.377	.841	.382	.884	.362	.821	.355	.864	.372
9	.903	.339	.879	.379	.841	.382	.883	.363	.821	.355	.859	.375
10	.905	.338	.868	.389	.841	.382	.881	.366	.816	.360	.873	.362
11	.905	.339	.888	.383	.841	.382	.880	.367	.832	.366	.875	.360
12	.904	.339	.876	.399	.841	.382	.876	.369	.825	.369	.875	.362
13	.904	.341	.870	.390	.841	.382	.876	.377	.825	.369	.871	.364
14	.904	.341	.871	.400	.841	.382	.863	.381	.825	.370	.881	.358
15	.901	.344	.861	.421	.841	.382	.857	.387	.825	.370	.868	.369
16	.901	.344	.865	.415	.841	.384	.855	.388	.832	.367	.881	.357
17	.900	.344	.858	.413	.841	.382	.855	.389	.832	.367	.872	.367
18	.897	.346	.858	.432	.841	.382	.852	.393	.837	.365	.878	.358
19	.897	.346	.856	.414	.841	.382	.850	.395	.837	.365	.874	.362
max	.905	.338	.899	.342	.841	.382	.894	.331	.832	.352	.892	.342

TABLE 5.16: Performance of six classifiers on ReliefF ranked attributes using peri-urban dataset

J48 attain the highest values with four and eleven features respectively. SMO is not affected by the change in the number and type of attributes.

The results show that the range of features giving the maximum performance are between 2 and 11. Therefore, these eleven features could be considered the set of features that are most predictive of the target class. More experiments using information gain ranked features are discussed next.

5.5.2.2 Information Gain Algorithm Ranked Features

Table 5.17 shows the experimental results using ROC area and RMSE metric values.

The results show that logistic regression attained the highest value of ROC area of 0.902 and the lowest error, RMSE of 0.340 with seven features. Similarly, NB attained the highest ROC area value with six features. The other four classifiers obtained the highest values with the top two to three features. Clearly, the optimal set of features lie in the range of 2 to 7 features.

#F	LR		MLP		SMO		NB		J48		RF	
	ROC	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**
2	.882	.341	.899	.342	.841	.382	.884	.351	.813	.352	.892	.342
3	.880	.342	.894	.343	.841	.382	.885	.356	.885	.356	.879	.349
4	.893	.341	.884	.351	.841	.382	.888	.357	.814	.353	.884	.354
5	.896	.341	.887	.359	.841	.382	.890	.356	.809	.358	.855	.383
6	.897	.341	.885	.359	.841	.382	.891	.356	.811	.359	.859	.377
7	.902	.340	.880	.368	.841	.382	.891	.355	.814	.355	.854	.379
8	.899	.342	.871	.379	.841	.382	.884	.362	.814	.355	.866	.369
9	.898	.342	.885	.387	.841	.382	.882	.365	.817	.356	.863	.374
10	.897	.343	.871	.397	.841	.382	.881	.367	.817	.356	.865	.371
11	.896	.344	.873	.396	.841	.382	.876	.371	.817	.356	.864	.372
12	.895	.345	.859	.409	.841	.382	.868	.379	.817	.365	.869	.371
13	.894	.345	.873	.403	.841	.382	.864	.384	.817	.357	.870	.368
14	.893	.346	.848	.427	.841	.382	.862	.385	.832	.363	.872	.369
15	.893	.347	.857	.409	.841	.382	.858	.388	.832	.363	.869	.369
16	.893	.347	.873	.399	.841	.382	.857	.388	.832	.363	.877	.363
17	.893	.344	.862	.408	.841	.382	.854	.392	.832	.365	.874	.362
18	.897	.346	.859	.419	.841	.382	.853	.393	.837	.365	.867	.368
19	.897	.346	.856	.414	.841	.382	.850	.395	.837	.365	.874	.362
max	.902	.340	.899	.342	.841	.382	.891	.351	.885	.352	.892	.342

TABLE 5.17: Performance of six classifiers on Information Gain ranked attributes using peri-urban dataset

5.5.2.3 Gain Ratio Algorithm Ranked Features

Table 5.11 shows the metrics of performance for the features ranked by Gain Ratio (GR).

The results show that the highest ROC area value of 0.902 and the lowest RMSE value of 0.340 was attained by logistic regression with 7 features. This was followed by MLP with just two features, NB and RF with three and two features respectively. However, J48 attained a high value of ROC area with fourteen features.

Since an optimal feature set should have a minimal number of features (Kohavi and John, 1997), the fourteen features out of a total of 19 features may not be considered optimal as the features giving the highest values are scattered in the range of 2 to 14 features. Therefore, the features ranking by Gain Ratio does not attain an optimal feature subset that is useful for the purpose of reducing the number of features in order to make the subset suitable for use on a small screen mobile device.

#F	LR		MLP		SMO		NB		J48		RF	
	ROC	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**	**ROC**	**RE**
2	.882	.341	.899	.342	.841	.382	.884	.351	.813	.352	.892	.342
3	.894	.339	.891	.346	.841	.382	.892	.331	.814	.353	.890	.344
4	.893	.341	.884	.351	.841	.382	.888	.357	.814	.353	.884	.354
5	.896	.341	.887	.359	.841	.382	.890	.356	.809	.358	.855	.383
6	.900	.340	.880	.362	.841	.382	.889	.356	.814	.355	.857	.379
7	.902	.340	.880	.368	.841	.382	.891	.355	.814	.355	.854	.379
8	.899	.342	.871	.379	.841	.382	.884	.362	.814	.355	.866	.369
9	.898	.342	.885	.387	.841	.382	.882	.365	.817	.356	.863	.374
10	.897	.343	.871	.397	.841	.382	.881	.367	.817	.356	.865	.371
11	.896	.344	.873	.396	.841	.382	.876	.371	.817	.356	.864	.372
12	.895	.345	.859	.409	.841	.382	.868	.379	.817	.365	.869	.371
13	.894	.345	.873	.403	.841	.382	.864	.384	.817	.357	.870	.368
14	.893	.346	.848	.427	.841	.382	.862	.385	.832	.363	.872	.369
15	.893	.347	.857	.409	.841	.382	.858	.388	.832	.363	.869	.369
16	.893	.347	.873	.399	.841	.382	.857	.388	.832	.363	.877	.363
17	.893	.344	.862	.408	.841	.382	.854	.392	.832	.365	.874	.362
18	.897	.346	.859	.419	.841	.382	.853	.393	.837	.365	.867	.368
19	.897	.346	.856	.414	.841	.382	.850	.395	.837	.365	.874	.362
max	.902	.340	.899	.342	.841	.382	.892	.331	.832	.352	.892	.342

TABLE 5.18: Performance of six classifiers on Gain Ratio ranked attributes using peri-urban dataset

5.5.2.4 Discussion of The Results

The results for the search of an optimal feature subset from the peri-urban data is presented in Table 5.19. Successive modelling was carried out starting with the top two common features.

Algorithm	Highest value of ROC area	**Lowest value of RMSE**	**Range of Optimal subset**
ReliefF	0.905	0.338	2 - 11
Information Gain	0.902	0.340	2 - 7
Gain Ratio	0.902	0.340	2 - 14

TABLE 5.19: A summary of results obtained from successive modelling of classifiers to determine an optimal subset of features for peri-urban data using three sets of ranked features with three selected algorithms: ReliefF, Information Gain, and Gain Ratio

The experimental results indicate that the Information Gain algorithm's ranked features provide the least number of features, whose maximum ROC area and least RMSE values are the same as for the Gain Ratio ranked features. However, for Gain Ratio, the features that achieved the classifier maximum values lie in the range of up to fourteen features. The ReliefF's eleven features attained slightly better ROC area value and lower error value. Eleven features would mean a much bigger dataset for training and four more

features to be incorporated in the small mobile screen. Therefore, it was decided that, the information gain seven features are the optimal feature subset.

The optimal features identified are: *test-marks*, *parent-educational-level*, *student-age*, *teacher-absenteeism*, *student-discipline*, *gender*, and *family-income*. A comparison with the rural school data optimal features obtained earlier shows that four of the features are shared, including *test-marks*, *gender*, *family-income*, and *student-age*. The features unique to the peri-urban data are: *teacher-absenteeism*, *student discipline*, and *parent-educational-level*. Those that are unique to the rural schools' data are: *teacher-shortage*, *student-motivation*, and *study-time*.

These findings show important differences in student characteristics among the two groups of students. In rural schools, there is a much more serious problem of teacher shortage than in peri-urban schools because most rural areas are hardship areas and fewer teachers are willing to teach in these areas, those that are posted look for every opportunity to move to urban schools (Monk, 2007). Students in rural schools are at risk of low motivation and less likely to excel in education because there are fewer role models and less exposure to the benefits of acquiring education (Hardré et al., 2009). Study-time for the rural student, especially in developing countries, may be interfered with due to lack of electricity in their homes; many parents have financial challenges and may not be able to afford it (Bornstein and Bradley, 2014).

On the other hand, the peri-urban student may face challenges of teacher-absenteeism, teachers face health problems, and family problems that could cause them to engage in businesses to offset the high costs of living in peri-urban areas (Sezgin et al., 2014). Although teachers in rural areas face the same challenges, life is cheaper in the rural areas. Student-discipline is a problem because in peri-urban areas, where there are many opportunities that negatively influence them and lead them to deviant behaviour. Student behaviour is termed as deviant when they indulge in drug abuse, engage in romantic actions, and start families. This behaviour will affect their classwork and their examination scores. Parent education level plays a key role. The students with educated parents living in peri-urban areas benefit from having role models at home, getting financial support for buying reading materials and paying for extra tuition (Fall and Roberts, 2012). Clearly, the two categories of students: rural school students and peri-urban students are different if the three features unique to each group are considered.

5.6 Chapter Summary

This chapter focused on finding the best classifier model, the optimal feature subset, and determining the performance of the best classifier model with the optimal feature subset. A comparison of student characteristics was also conducted between students from rural schools and those those from peri-urban schools.

Six models were built, including logistic regression, Multilayer Perceptron, sequential minimal optimisation, na'ive Bayes, J48, and Random Forest. Their classification performance was compared using six metrics, including recall, specificity, ROC area, F-Measure, Cohen's Kappa, and Root Mean Squared Error (RMSE). The first phase of results show that logistic regression was identified as the best model when the rural dataset was used. However, it was the second best model after SMO when peri-urban data set was used. Logistic regression was selected for implementation in the prediction system, because it turned out the best for the rural dataset, the focus of this study.

A search for the optimal feature subset yielded 7 features out of the total 22 features from the rural dataset. Similarly, 7 features were identified as the optimal feature subset for the peri-urban dataset. The peri-urban features were 19 because three features were excluded from the list, namely distance to school, command of English, and community involvement. Therefore, the feature selection reduced the dataset to less than a third for the rural dataset without compromising the quality of classifier performance.

The process of feature selection involved, firstly, ranking the features using three filter algorithms: ReliefF, Gain Ratio, and Information Gain. Followed by building successive models. The range of features that attained the best performance were selected as the optimal subset. The selected features for the rural data are: *test-marks*, *gender*, *teacher-shortage*, *student-motivation*, *family-income*, *student-age*, and *study-time*. Those of the peri-urban data include *test-marks*, *parent-educational-level*, *student-age*, *teacher-absenteeism*, *student-discipline*, *gender*, and *family-income*. Four features were common in both lists while three features were unique to each list. This finding confirmed the difference in student characteristics from the two environments, and presented an opportunity for further research on the peri-urban students.

The identification of the best classifier model and the optimal feature subset was an important step towards implementing the Mobile Academic Performance Prediction System. The design and implementation process for the system is discussed in the next chapter.

Chapter 6

Design and Implementation of the Mobile Academic Performance Prediction System

6.1 Introduction

The proposition of this study is that academic performance prediction for students in rural regions could integrate a mobile phone interface to make the system affordable in developing regions that have scarce resources. The first task of the study was to find the best classifier algorithm and the optimal feature subset as discussed in Chapters four and five. The second task is the design and integration of the mobile application interface with the classifier algorithm. To achieve this second task, a User-Centered Design (UCD) approach was adopted. This approach puts a great emphasis on involving the users of the system being developed from the beginning of the design process to the end (Marsden et al., 2008). UCD was important in this study because the aim was to ensure the system built is usable. Additionally, it acts as a standard process of doing research, since it offers the standards for conducting and evaluating this type of research (Venable, 2010).

The mobile academic performance prediction system overview is first discussed followed by a discussion on the adopted UCD methodology's four phase interaction model. These phases are: gathering requirements; developing alternative designs; building interactive

designs; and evaluating the designs (Preece et al., 2015). The chapter concludes with the usability evaluation of the high fidelity prototype.

6.2 The Study Perspective

This study aimed at developing a technology that is useful for predicting those students that require high intervention early enough. To achieve this goal, the technology had to be both sustainable and accessible. The trend has been to develop artefacts using mobile phones since that is the technology that is most readily available in developing countries. The use of a technology that is available is in line with the concept that an ICT4D study needs to come up with local means for intervening in a local problem (Wicander, 2011).

It is with this understanding that this study followed the approach of involving users from the initial stage. The approach helped to eliminate complexities of the system, increase understanding and improve system acceptance.

6.2.1 Components of The Mobile Academic Performance Prediction System (MAPPS)

Although the focus in this Chapter is the design process for the mobile application interface for MAPPS, the other components are briefly discussed. Figure 6.1 illustrates the system.

As shown in the figure, the mobile application interface is linked to the server via the Internet. The mobile application on the client side allows the entry of a student's record, consisting of the seven features. The options have been binarised to reduce the data size and enhance speed of data transfer. The server contained the Logistic regression classifier model, previously identified as the best model. Logistic regression was implemented in Python using machine learning libraries such as Scikit-learn (Pedregosa et al., 2011).

The server interface used the Django Web framework (Widman, 2011). Django was preferred in this study for the following reasons: it is a framework that is popular and has proved to be effective; it significantly reduces the difficulty in developing efficient Web applications; and lastly, Django has a flexible open source license (Widman, 2011).

The system does the following: it transfers a student's record entered through the mobile interface via the Internet to the logistic regression classifier on the server. On the server, the classifier model predicts the result for this new record. The result is next transfered

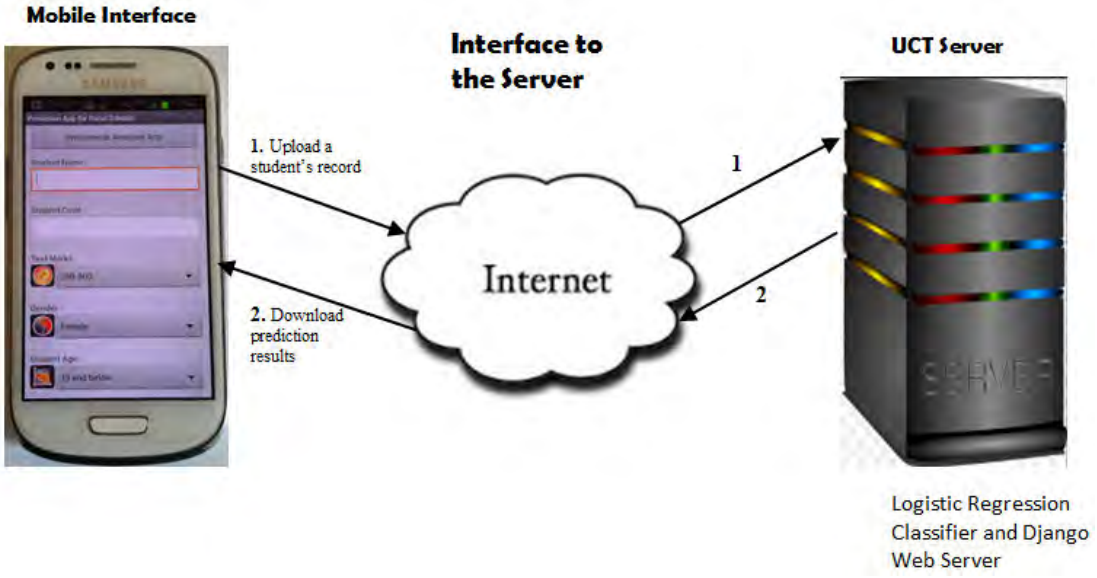


FIGURE 6.1: Mobile Academic Performance Prediction System Architecture

via the Internet to the mobile interface. Django Web server facilitated the record transfer to the server and the result from the server to the mobile interface. For this prototype, we used a computer science department server at The University of Cape Town.

6.2.2 Design Process

This section presents the design followed in developing the Mobile application interface for MAPPS. The design started from the Educational Data Mining process (see Chapter 4). This was followed by the process of finding the best classifier model and the optimal subset in Chapter 5.

From Chapters 4 and 5, the following functional requirements were derived for the design of the prediction system:

- Predict a student's intervention level as either high intervention or low intervention
- Use the optimal feature subset (test marks, age, gender, study time, student motivation, family income, and shortage of teachers)
- Use the best classifier model determined, which is the logistic regression classifier model.

6.2.3 Why Use Mobile Phones?

Mobile phones have extensively spread in developing countries. The mobile phone usage in sub-Saharan Africa as at 2012 included over 500 million people, which was about two thirds of the total population (Ojanen et al., 2015) . Mobile phones are said to be ubiquitous, referring to the fact that mobile services are found by the users anytime and everywhere, especially in those places where desktop computers cannot be used because of lack of electricity (Okazaki and Mendez, 2013). The invention of smart-phones has made it possible for many mobile applications to be developed and used (Okazaki and Mendez, 2013). In education, mobile phones have been used to improve learning among primary school children and in adult education with evidence of success (Rotberg and Aker, 2013). Mobile phones have been found useful in education (OḂannon and Thomas, 2015) . However, the closest use of mobile phones in prediction predicted exam results through the usage of a mobile phone learning system, where the usage of the learning system formed the independent variables (Boticki et al., 2015).

Our study proposed to use mobile phones as a component of a system that predicts a student's academic performance to make the system usable in rural areas . Most other studies of predicting academic performance have used desktop computers.

In the next subsection, a discussion of the design process for the mobile phone application interface is presented.

6.3 User-Centred Design

User-Centered Design puts a great emphasis on involving the users of the system being developed from the beginning of the design process to the end (Marsden et al., 2008). It is seen as an interactive approach of system development that emphasises developing a system that is usable, depending on the characteristics of the users, the environment they operate in and what they are to achieve. The aim of using the UCD process in this study is to achieve a system that will be both useful and usable by the users (Gitau, 2013).

The UCD methodology that was used in this study is the four phase interaction model as shown in Figure 6.2 (Preece et al., 2015).

The figure shows the overall UCD model adapted in this study. At the top is the study defined task, building the mobile academic performance prediction system. The first task toward achieving this goal was to determine the best classifier algorithm, logistic regression, followed by finding the optimal feature subset of seven features out of the

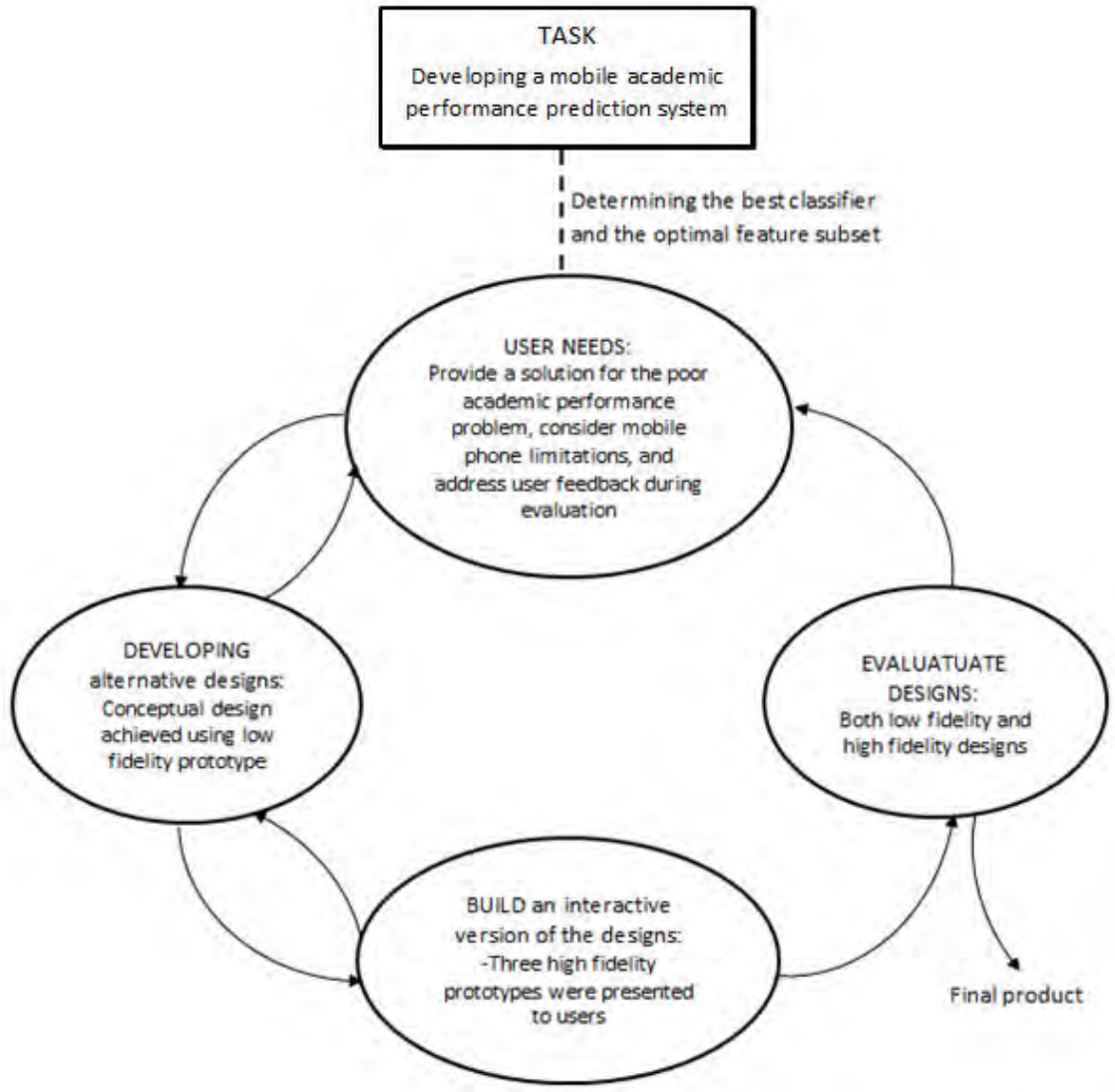


FIGURE 6.2: The four phase methodology of Interaction Design (Preece et al., 2015)

total of 22 of the rural school dataset. The four interactive task phases form the second phase of the overall process, and are discussed next.

The Process of Gathering Requirements

Requirements gathering is the first step towards the design process. It involves identifying the system users, what they do and want to achieve by using the system, in addition to knowing their environment. There are two types of requirements, including functional and non-functional requirements. The functional requirements refer to what the system will be able to achieve or perform (Glinz, 2007), while the non-functional requirements specify the attributes or constraint that the system must respect (Chung

and do Prado Leite, 2009). This definition was earlier expressed to separate the requirements that focus on how good the software is from what the software is capable of doing (Paech and Kerkow, 2004) .

Developing Alternative Designs

Phase two presents design alternatives, which are designs for the system, as generated from the requirements. The first phase of the design process is called conceptual design, followed by the physical design. The conceptual design represents and validates the requirements gathered. It is accomplished in collaboration between the designer and the users. One of the ways of achieving conceptual design is through low fidelity prototypes, a method that this study adopted.

Building Interactive Versions of the Designs

The third phase is to build prototypes of the system that will allow interaction with the users. These are built iteratively as reference is made to the requirements and the conceptual model. Evaluation is conducted in every iteration as discussed next.

Evaluating the Designs

Evaluation is carried out for any part of the system that has been built. The goal of evaluation is to make sure the final system is what was expected. Evaluation is achieved through a user-centered approach that involves users in every level of the design process.

Following the UCD methodology, the first phase was to understand the potential system users' needs as discussed next.

6.3.1 Requirements

6.3.1.1 Educational Stakeholders' Expectations

In order to understand the users' expectations, 19 teachers from different schools and 2 education officers were selected and engaged in interview sessions to extract requirements. These were selected from among the 54 schools earlier visited for data collection and the 7 education officers. The participants were selected because they showed interest, willingness to participate, and were also technology literate. The following list of user requirements was compiled:

1. Facilitate school inspection by the education officers and head teachers when they need to submit reports on quality of education for a given school.

2. Categorise those students who need high intervention early enough for strategic intervention to be put in place.

3. Provide results quickly and accurately.
4. The interface must be simple and easy to use.
5. The system must motivate the users to continue using it.
6. The system must not be expensive to maintain.
7. An error prevention mechanism, must prevent users from making wrong entries.

These requirements gave the researcher an insight into the design and development process. The following were the implications of the requirements:

1. The system has to be used in rural regions that have no electricity, which means, mobile phones would be the most suitable since they can be charged with solar power. This refers to requirement 1, where, school inspection and academic performance report is needed from rural schools.
2. Identify those students that need to be put in strategic intervention as accurately as possible and early enough before they sit for the final national standardised examination, as presented in requirement 2 and 3.
3. The student records need to be as light as possible to ensure a small data size and hence increase the speed as indicated in requirement 3.
4. The system interface needs to have clearly labeled icons and buttons that guide the user to complete a task, as expected by requirement 4.
5. The system interface needs to have icons that clearly differentiate the features, and there should be a minimum number of steps to achieve a task, as indicated in requirement 5.
6. The mobile component of the system has to be usable on a cheap smart phone to make it affordable. Refer to requirement 6.
7. The design of the mobile interface had to use forms to avoid wrong entries and hence prevent errors. Refer to requirement 7.

6.3.2 Conceptual Design: Low Fidelity Prototype

After compiling the user requirements, the next task was to conceptualise the system design. A computer printout paper prototype with the 7 optimal features was presented to the system users. The users were 19 teachers and 2 education officers. Each of them

was given the computer printed paper prototype as shown in Figure 6.3. Because of the wide rural area in which the users were situated, the researcher visited them at their stations. In some cases, two teachers from the same school were engaged in the design.

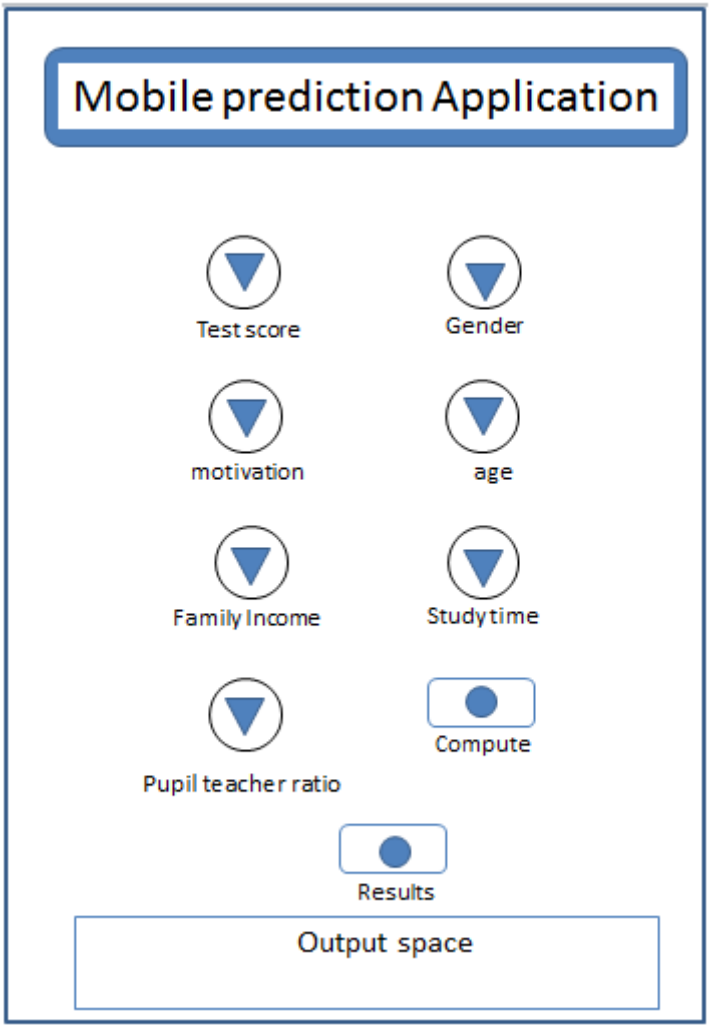


FIGURE 6.3: Computer printout prototype showing the features for the main system interface

The figure shows the 7 optimal features:*test-scores*, , *student-gender*, *motivation-to-learn*, *student-age*, *study-time*, *family-income*, and *teacher-pupil-ratio*, that were determined by the machine learning process (see Chapter 5). This printout was given to the users so that they can give the options for each of the features. The common options are presented in Figure 6.4, Figure 6.5, and Figure 6.6.

These Figures show the common options selected by the users. As seen on Figure 6.4, and Figure 6.5, there was a consensus on the test scores that the first range of marks should be 350 - 500, and the lower marks to reduce by 50 marks up to 200 marks. The focus for this study was to identify those students that obtain 250 marks and above,

FIGURE 6.4: Sample one prototype showing options for the seven features

and those who score below 250 marks. Therefore, Figure 6.4, and Figure 6.5 were found suitable for that purpose. Extra subdivision for the lower marks as in Figure 6.6 would not be of any benefit.

The *student-age* options were similar in the three figures. These options were adopted as they are in the system. However, the age options suggested by the participants varied between 14 years and 13 years. This study adopted the threshold of 13 years because we targeted Class Six students who still have two years to sit for the final examinations. These would benefit most from the strategic intervention.

FIGURE 6.5: Sample two prototype showing options for the seven features

Student-motivation-to-learn had a consensus of three options with different wording. The study adopted the three options with words that formed a compromise for the various suggestions: very good; good; and fair. *Study-time*had four options suggested, we opted for three general terms that would be acceptable to even those students that may not have clocks in their homes. The options including no-time, little-time, and enough-time. Similarly, the *family-income* feature was proposed to have three options with different wording, we come up with compromise wording for the three options: needs-not-met; needs-partly-met; and needs-fully-met. This wording wording relates more to the student since they are able to relate needs such as having no meals, no uniform, and not being able to pay the school levies.

Finally, there seemed to be a consensus on 40 as the number of students that are considered a normal class. An acceptable or average class was considered to be between 50 and 60, and a large class over 60 students. We adopted the three terms: standard-class, acceptable-class, and large-class.

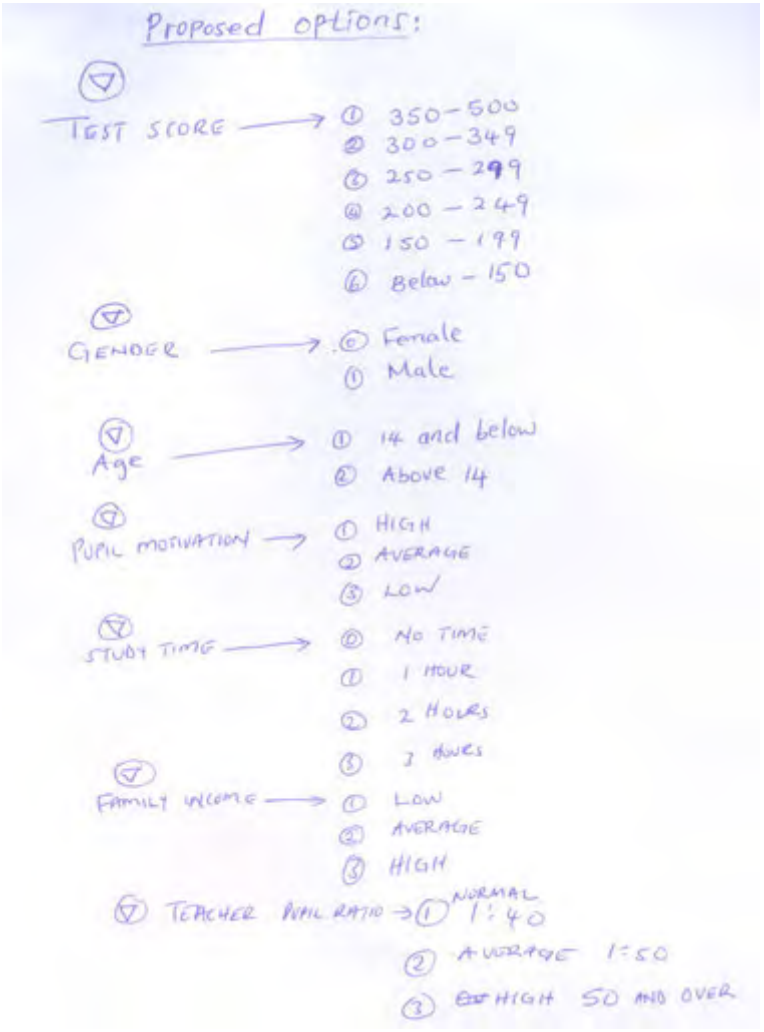


FIGURE 6.6: Sample three prototype showing options for the seven features

The next subsection discuses the building of the high-fidelity prototype using the options of the seven features obtained in this subsection.

6.3.3 Building the Interactive Versions: High-Fidelity Prototype

The features determined in the machine learning process (see Chapter 4 and 5) and the options for each of the seven features determined in the previous section were implemented on an Android platform. Django Web server was used to send the student record to the server and the results back. This subsection presents the design process for the mobile interface. Three versions of prototypes were built in this study. Figure 6.7 illustrates the first version prototype interface overview. The features are shown in blue. This first version interface prototype is discussed next.

6.3.3.1 First Version Prototype

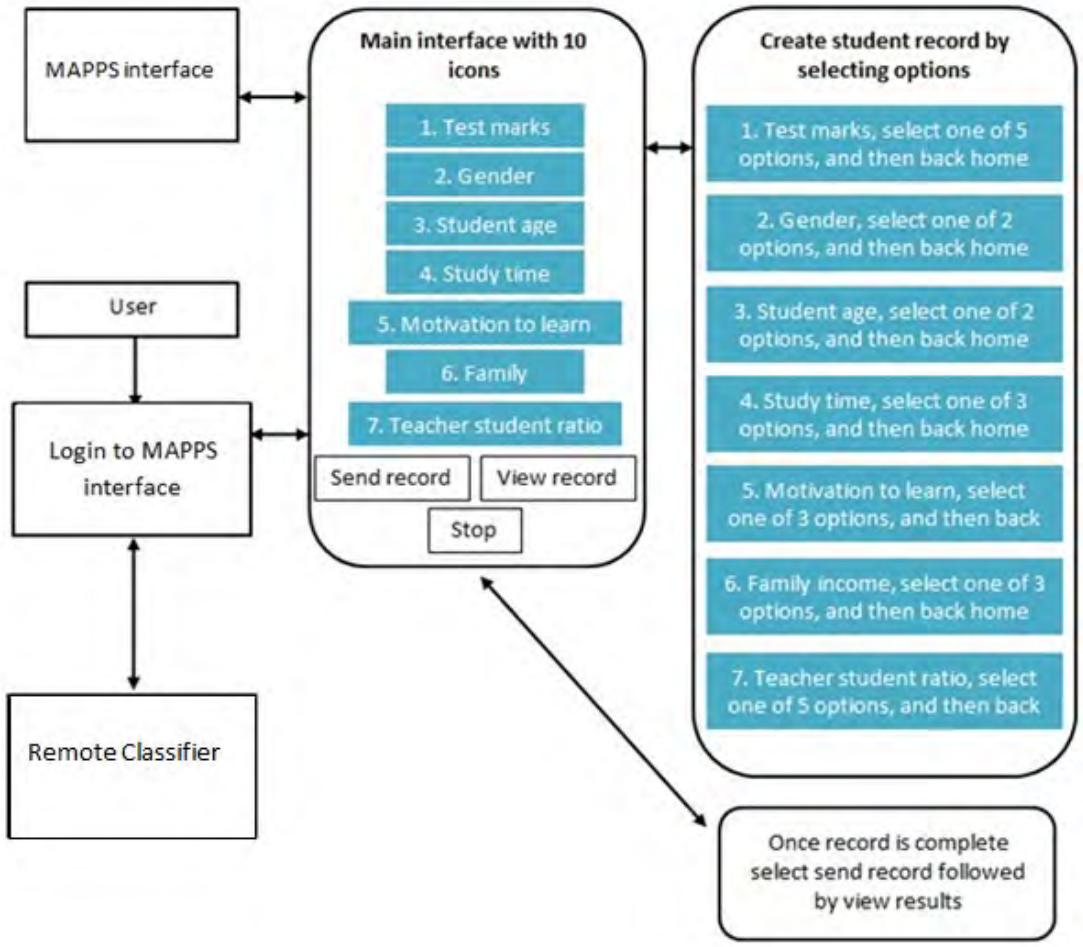


FIGURE 6.7: Mobile interface system prototype of the first version showing the seven features at the main interface and the create student record screen

Login to the System

The user logs in by identifying and selecting the prediction model icon on a smart phone that has been preloaded MAPPS. It was not found necessary to register users and log in with a password because the focus was not on the users but the students' records. Further, the system had to be as easy to use as possible.

Main Interface

This main interface shows the layout of the seven icons representing the seven features, that were previously determined as the optimal features.

This first version prototype icon layout is represented in Figure 6.8. The interface layout also shows: the welcome bar, send record, view results, and close icons. For new users, a welcome bar gives basic instructions on the process of creating a student record, sending the record to the server for the classification process, and getting the result of prediction.

FIGURE 6.8: Mobile phone main interface icons for prototype 1

In this first version, the user had to tap or select each icon and make the selection from the menu that appears, and then select the back home button to go back to the main interface for the next icon. It is only after all the feature options have been selected that the user can select the send record icon; otherwise, the system will report an error.

Create Student Record

A student's record is created in the screens that appear once a feature icon has been selected (see Figure 6.9, Figure 6.10, Figure 6.11, and Figure 6.12). These are the options for each of the seven features. To compile a student record, a user tapped an icon which opened the options. An option was selected by tapping a check box which would be marked as shown in the figures. Tapping the back home button returns to the main interface for the next feature icon selection. This process was repeated until a student record of the seven features is complete. Results were obtained by tapping the send record button followed by the view results button. To exit the system, the close button is tapped.

6.3.3.2 First Version Prototype Evaluation

To obtain users' feedback, this prototype was presented to 7 users, teachers in the primary schools previously visited during data collection. These schools were visited because of the teachers that had been identified as representatives of the system users;

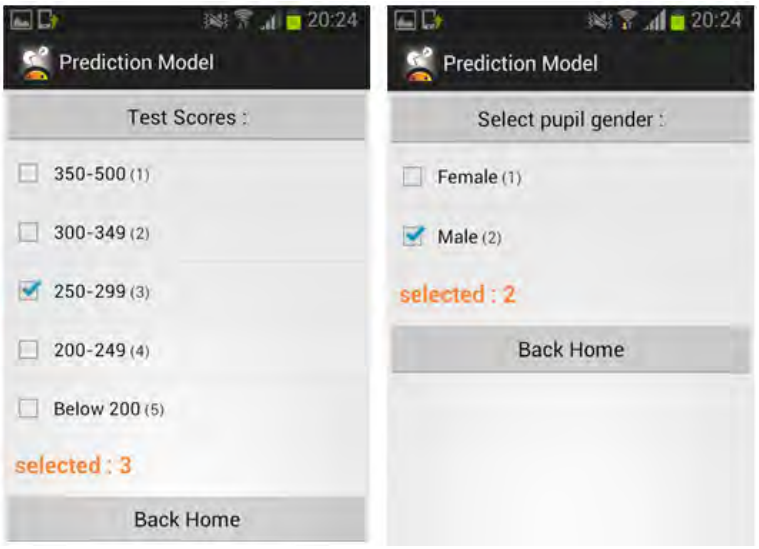


FIGURE 6.9: Test score options (left figure) showing the 5 options of test marks categorisation, where 350 - 500 are the top students and 0 - 249 are the students below average. Gender options are shown in the figure on the right

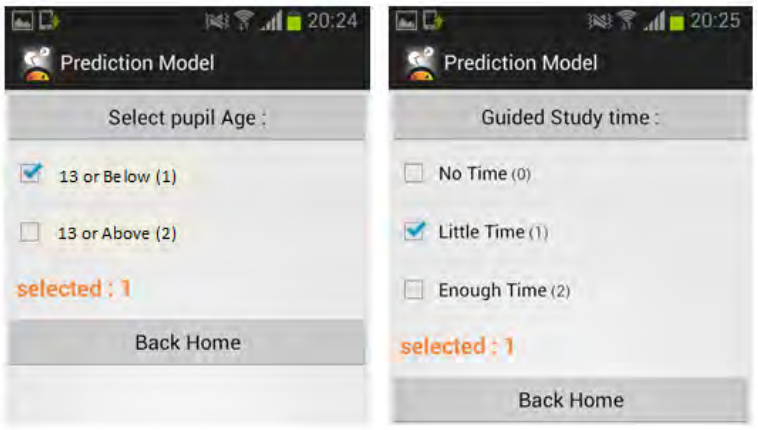


FIGURE 6.10: The age feature options indicates the average age when students are in Class Six (left figure). The study-time options (right figure) show the time in hours students spend studying after school: no time - 0 hours; little time - about 1 hour; and enough time-two hours or more

again working with them would ensure smooth progression in the research process. Their feedback was obtained in the following areas: their perception towards the system; and the challenges they faced while testing the system. This initial part of the prototype evaluation used an interview with each of the users. The reason for the open ended interview was to gather the users' feedback freely. The challenges faced by the users were identified as the users performed the following system tasks:

- Selecting a correct option from each of the seven features that make up a student record.

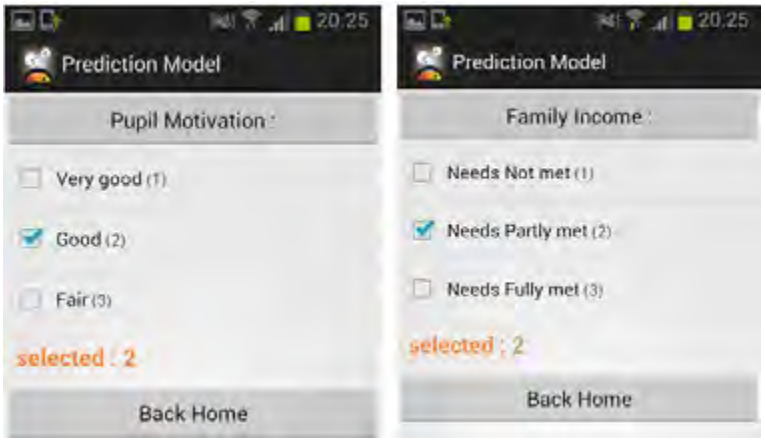


FIGURE 6.11: The motivation-to-learn options (left figure) shows the three motivation options; and family-income options (right figure) present the three income options

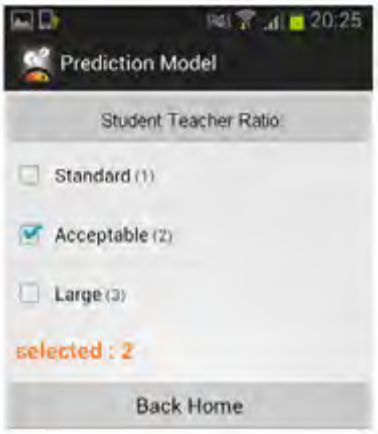


FIGURE 6.12: The student-teacher-ratio options: standard, for up to 40 students; acceptable, for up to 60 students; and large, for over 60 students in a given class

- Confirming that the student record is the correct one.
- Sending the record to the server.
- Obtaining the predicted intervention result for that particular record.
- Clearing the result screen in order to enter another record.

The objectives for this initial prototype evaluation were achieved: to identify the challenges the system users faced while using the system; and to find out from the users the modifications that could improve the system.

Most users were able to obtain the prediction result for a record entered into the system. However, they had difficulties while entering a student's record. They needed to remember the order of the features and their options because they were not visible once

selected. Additionally, users noted that there were a lot more buttons to tap to complete an intervention prediction task.

This feedback was useful in improving the mobile interface to use forms as discussed next.

6.3.3.3 Second Version Prototype

Figure 6.13 shows the overview of the second version prototype with the modifications from the first version prototype. The modifications were as a result of the feedback obtained during the evaluation of the first version prototype. The modifications are on the main interface and the function for creating a student record as described next.

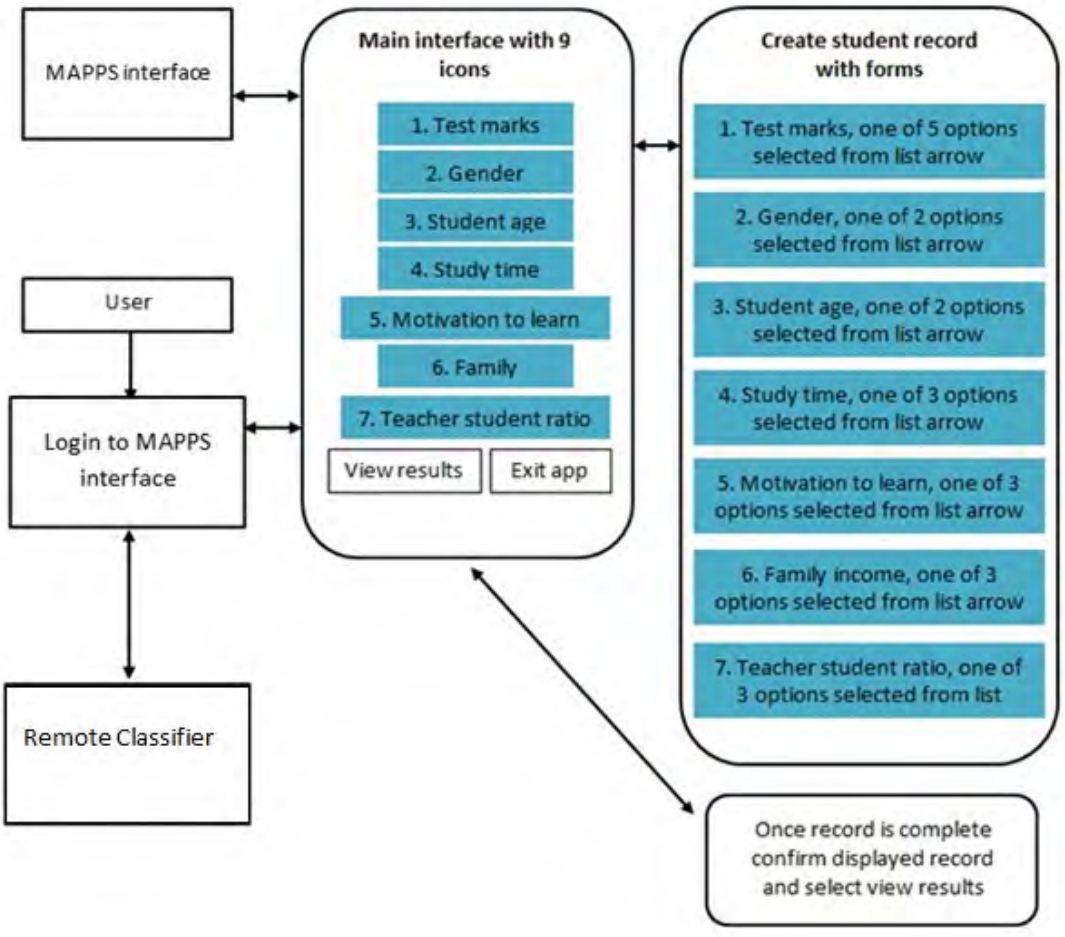


FIGURE 6.13: Mobile interface system prototype of the second version showing the seven features at the main interface and the create student record screen

Modifications on the Main Interface

The main interface of the second version prototype was modified to use forms, the feature icons appear on the left, and on the right is a drop down list arrow that when tapped

opens a list of the options. Between the icon and the list arrow appears the current option selection as shown in Figure 6.14.

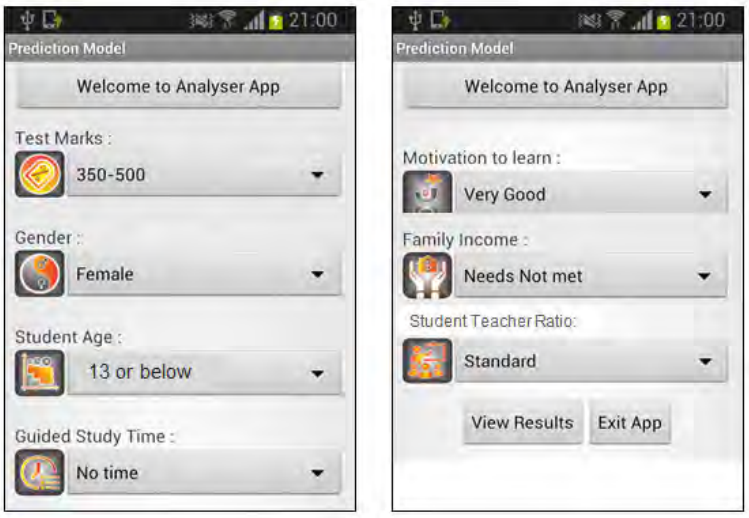


FIGURE 6.14: Mobile phone interface showing icons for prototype 2

Once an option is selected, it remains displayed beside the feature icon until another selection is made. A selected student record is verified by simply comparing the selection that appears beside each icon and the record on a source document. The process of sending the record and viewing results has been condensed into tapping the view results button. The exit app button closes the system.

Modifications on the Create Student Record

The modification of the screen for creating a student record was to reduce the number of buttons to tap. Once an option was selected, the option screen automatically closed to allow for next feature option selection from the main interface which remains visible throughout the record formation. Figure 6.15 - Figure 6.17 show the modifications in the Create Student Record screens.

The left Figure 6.15 illustrates the five test marks options. To select an option, a user taps one of the five test marks options. The marks categories refer to the average student test marks for the whole year for either Class Six or Seven. A student who scores above 350 marks is rated a top performer, 300 to 349 is a good student, 250 to 299 is an average student, 200 to 249 is below average student, and below 200 is a poor student.

The right Figure 6.15 illustrates the gender options. One taps either the female or the male options. All schools visited in this study were mixed, having female and male students.

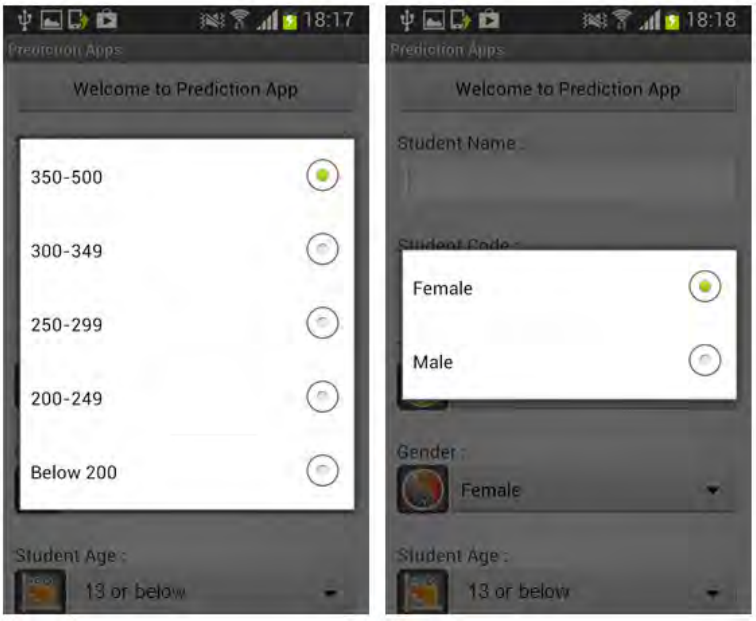


FIGURE 6.15: Test score options: 350-500, 300-349, 250-299, 200-249, below 200 (left figure) and gender options: female, and male (right figure)

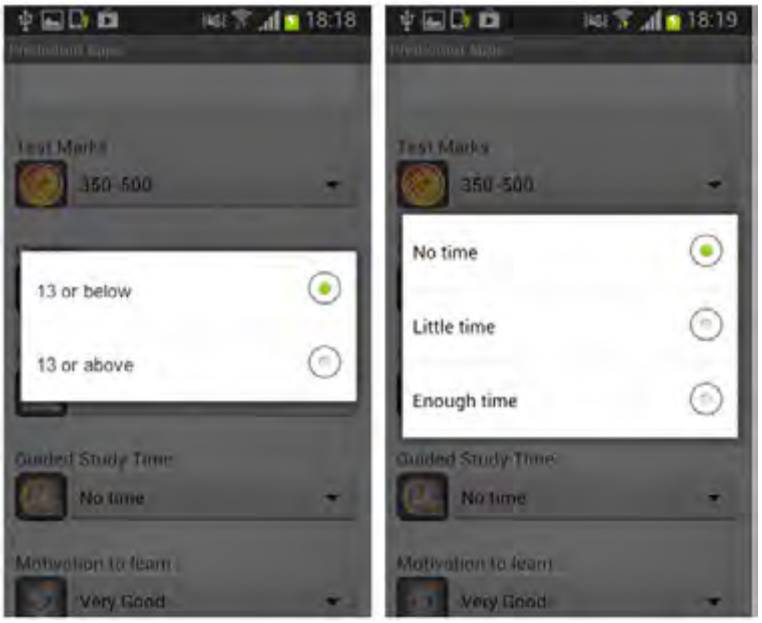


FIGURE 6.16: Student age options: 13 and below, 13 and above (left figure) indicating the student's age category when they are in Class Six; and the study time options: no time, little time, and enough time (right figure)

Figure 6.16 (left figure) shows the age options that are accessed by tap on the drop down list arrow beside the age icon, and another tap on one of the options, 13 and below, or 13 and above.

Figure 6.16 (right figure) presents the three options for study time. The user asks the student the time they spend doing their studies after school and taps one of the three options: no time, for zero hours of study; little time, for about 1 hour; and enough time, for two hours and above. In rural schools, it is possible that a student finds no time to study because of lack of electricity in their homes or house chores, especially for girls.

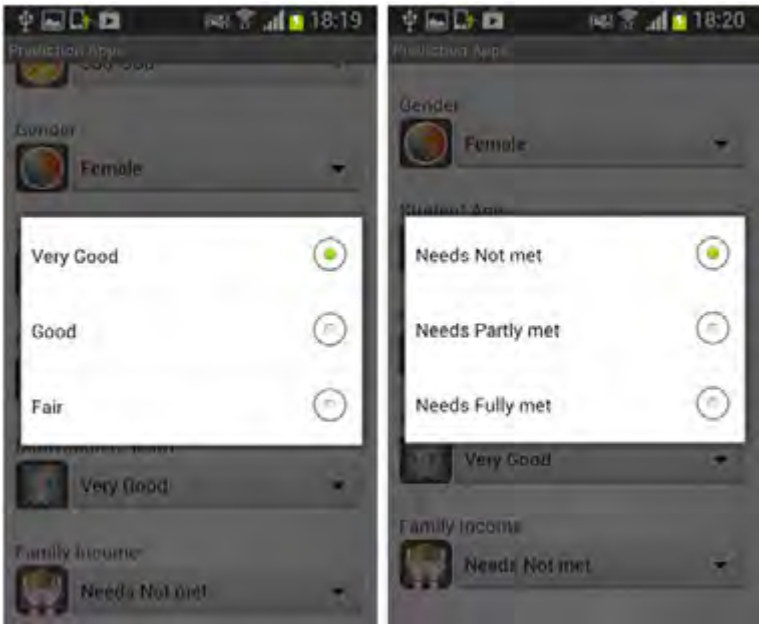


FIGURE 6.17: Student motivation-to-learn, showing the options:very good, good, and fair (left figure) and the family income options: needs-not-met, needs-partly-met, and needs-fully-met (right figure)

Figure 6.17 (left) shows student's motivation to learn options. A user, after deciding where the student's level of motivation falls taps one of the three options. For example, a student who finishes all the class work and homework, and is active in class could be classified as having very good motivation to learn.

Figure 6.17 (right) presents the three options of family income. A user makes the decision on which option to tap after asking the student some questions or make observations on the student. For example, a student whose needs are not met, could be a one who comes to school without breakfast, and has unpaid school levies among other conditions.

The Figure 6.18 shows the student-teacher-ratio's three options: standard, for up to 40 students; acceptable, for up to 60 students; and large, for more than 60 students. The user, taps on one of the options based on their knowledge if they are class teachers or

FIGURE 6.18: Student-teacher-ratio showing the options: standard, acceptable, and large

they could get get the information from class registers, or directly from a colleague who teaches the class.

6.3.3.4 Second Version User Evaluation

During the evaluation process, the same 7 users in 7 primary schools were given the system to perform similar tasks as those used in the evaluation process of the first prototype (see Section 6.3.3). A brief training was conducted for each of the users before using the system to perform the tasks. Thereafter, an interview was conducted with each of the users to determine their feedback.

The users gave the following feedback: that the design improved because one did not have to remember previous selections; it was also able to confirm the student's record entered before selecting the view results button for submitting the record to the server and displaying the results; and there were fewer buttons to touch before obtaining the intervention prediction.

Some users proposed that they would wish to identify the records stored in the server so that it could be possible to follow up the student progress in the proceeding years. This comment prompted slight modification of the prototype.

6.3.3.5 Third Version Prototype

Modifications on the main interface

Except for the student identifier text place holders on the main interface, the system overview for the third version prototype was similar to that of the second version. The modification is presented in Figure 6.19.

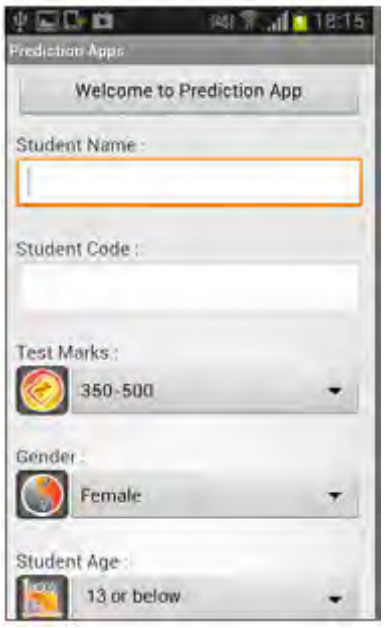


FIGURE 6.19: Main interface of the third version prototype showing the main modification of the second version prototype, with only three features out of the 7 features visible in this screenshot

The figure shows three of the seven features that could fit in the screen shot and the modification of the text place holders suggested during the evaluation of version 2 prototype. Everything else of the second version prototype is utilised in the third version. Similar to the second version, forms have been utilised to reduce the number of taps and to improve the usability of the system.

Sample results for Intervention Prediction

Figure 6.20 shows the results of intervention prediction for a student's record was sent to the server via the Django Web server. This results are displayed on the results panel once a user taps the view results button. As seen, the actual result is the student's intervention, either the student requires high intervention or the student requires low intervention. However, the student record is also displayed as a way of validating that the results belong to same record that was sent to the server.

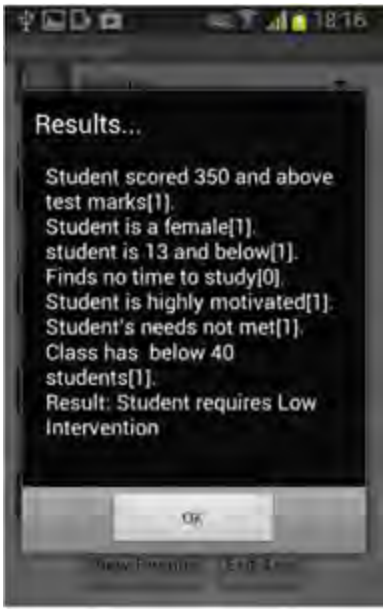


FIGURE 6.20: Example of intervention prediction results obtained using the third version prototype showing the student record and the predicted intervention

6.3.4 Usability evaluation of the third version Prototype

The evaluation of the final prototype was conducted with 15 schools where one lead teacher was identified and given the mobile phone application interface for a period of three weeks. The teachers were expected to perform the five tasks previously mentioned in Section 6.3.3. However, this time, they were to use the system to predict intervention for all their students in Class Six and Seven.

The system was preloaded on Samsung Galaxy pocket phones. Each phone was loaded with airtime worth USD 2.5. The period was limited to three weeks because it was seen to be enough time since the selected users were already familiar with the concept and the system.

After the three weeks of using the system, a questionnaire was given to the users to provide feedback on their interaction with the system. The questionnaire had a five-point Likert-scale with options: Strongly disagree, Disagree, Neutral, Agree, and Strongly agree (Munshi, 2014).

This evaluation process is called summative evaluation (Preece et al., 2015). Meaning, it is evaluation that is conducted at the final stage of the product. It is carried out after other evaluations in earlier stages of the design process have been conducted. This final evaluation of the design process tested the functionality of the system and the general usability.

Figures 6.21 illustrates a photograph taken during some of the sessions when teachers were using the prototype.

FIGURE 6.21: A picture showing some participants during the prototype evaluation sessions

As shown in the picture, each student's record is entered in the presence of the student, as some of the features, such as age and family income, require direct response from the student.

The results of the evaluation process are presented in Table 6.1.

Table 6.1 shows that 92.31% of all the participants would use the system frequently. Those who strongly agreed were 61.54% , while, 30.77% agreed. From the rest, 3.4% were neutral and another 3.4% disagreed. The observation shows that most participants have no reason to avoid the system.

A large number of participants, 88.46%, said the various functions of the system were well integrated. Those who strongly agreed were 69.23%, while 19.23% agreed. For the rest: 7.69% were neutral; and 3.4% disagreed. Therefore, participants liked the system.

A total of 85.61% of the participants said there are no inconsistencies in the system, 38.38% strongly disagreed and 47.15% disagreed. Of the remaining 14.39%, 7.59% were neutral and 6.8% agreed that inconsistencies existed.

Nearly all the participants (96.60%) felt confident to use the system; 50% strongly agreed, and 46.6% agreed. Only 3.4% of the participants were not confident to use the system. This feedback was expected because of the UCD approach used, it involved all the users in the design and evaluation process.

User feedback	**strongly disagree**	**disagree**	**neutral**	**agree**	**strongly agree**
I would like to use the system frequently	0%	3.4%	3.4%	30.77%	61.54%
This system is unnecessarily complex	61.54%	15.38%	3.4%	15.38%	3.4%
The system is easy to use	3.4%	0%	8.14%	50%	38.46%
I would need assistance to use the system	30.77%	3.4%	15.38%	26.92%	23.1%
Various functions of the system are well integrated	0%	3.4%	7.69%	19.23%	69.23%
There is too much inconsistency in the system	38.46%	47.15%	7.59%	3.4%	3.4%
Most people would learn to use this system quickly	3.4%	3.4%	15.38%	26.92%	50%
The system is cumbersome and awkward to use	92.31%	3.4%	0%	3.4%	0%
I had no confidence in using the system	0%	3.4%	0%	46.6%	50%
I had to learn a lot before using the system	19.23%	15.38%	23.1%	19.23%	23.1%

TABLE 6.1: Usability evaluation of the Mobile Phone Application Interface for MAPPS

On whether the tool is complex or not, the majority of the participants (76.92%) did not find it complex, possibly because the algorithm's complexities were hidden from the users. The choice to use smart phones also helped to eliminate the complexities as the operation became simply tapping buttons. An small number of participants (18.8%) found the system complex. It is possible that these are the participants that were not familiar with smart phones. The researcher noted that many of the participants did not own smart phones.

On whether the system was easy to use or not, a total of 88.46% of the participants accepted it was easy to use; 50% agreeing, and 38.46% strongly agreeing. The rest of the participants were distributed as follows: neutral group (8.14%); and strongly disagreed (3.4%). The high percentage of participants that agreed the system is easy may be attributed to the UCD approach where users were involved in testing the previous two versions of the prototype.

Half of the participants said they would need assistance to use the system, with 26.92% agreeing and 23.1% strongly agreeing. The rest, 34.17% said they would not need assistance, while 15.38% were neutral. It is possible that the participants may not have been clear as to what type of assistance is being referred to here, as revealed by the percentage of 88.46% of those who said the system is easy to use. It may be that the

users were considering the complexities of the intervention prediction and not the simple system usability.

When the participants were asked whether any person given the system would be able to quickly learn it, a total of 76.92% agreed; 50% strongly agreeing and 26.92% agreeing. For the rest, 15.38% were neutral; probably, they thought about understanding the complex classification process carried out in the background. Only 6.8%, thought they will not be able to quickly learn the system.

A large number of participants strongly disagree (92.31%) that the system was cumbersome or awkward to use. Those who disagreed were 3.4%. The total number of participants who disagreed that the system is cumbersome or awkward to use is 95.71%. This could be attributed to the UCD approach used. Users were involved from the initial stage of the study to the design stage of the mobile interface. A small percentage of participants (3.4%) thought it was cumbersome and awkward to use; these could be considered outliers.

Lastly, 42.33% of the participants thought one needed to learn a lot before using the system. While one third of the participants (34.61%) thought someone did not have to learn a lot to use the tool. The rest (23.1%) of the participants were neutral. During the researcher's interaction sessions with the participants, it became clear that they were referring to the concepts of classification. To most of the participants, the idea of predicting students' performance through classification techniques sounded advanced and abstract. To most of them, the idea of using a mobile phone for prediction of academic performance was completely new.

6.4 Chapter Summary

The chapter has illustrated how a mobile phone interface application could be integrated on an academic performance prediction system to predict the intervention of a student. The chapter followed the UCD four phase interaction model. The users' rural environment where there is lack of electricity and Internet determined the use of the mobile phone interface. Specifically, the chapter illustrated how the features determined using machine learning, and their options, determined by the users, have been implemented on a mobile phone. Therefore, this chapter has shown the possibility of integrating a classifier model with an interface on a mobile phone.

The chapter has also presented an overview of three prototype versions. Using an example of intervention prediction, it has been shown that the intervention prediction is

possible. The chapter presents user evaluation for the first prototype version that motivated modification to realise the second version, and similarly for the second version to realise the third version. Finally, a usability evaluation was conducted for the third version prototype.

The next chapter presents quantitative evaluation of the mobile academic performance prediction system on real data as gathered from actual students in Class Six and Class Seven from 15 rural primary schools.

Chapter 7

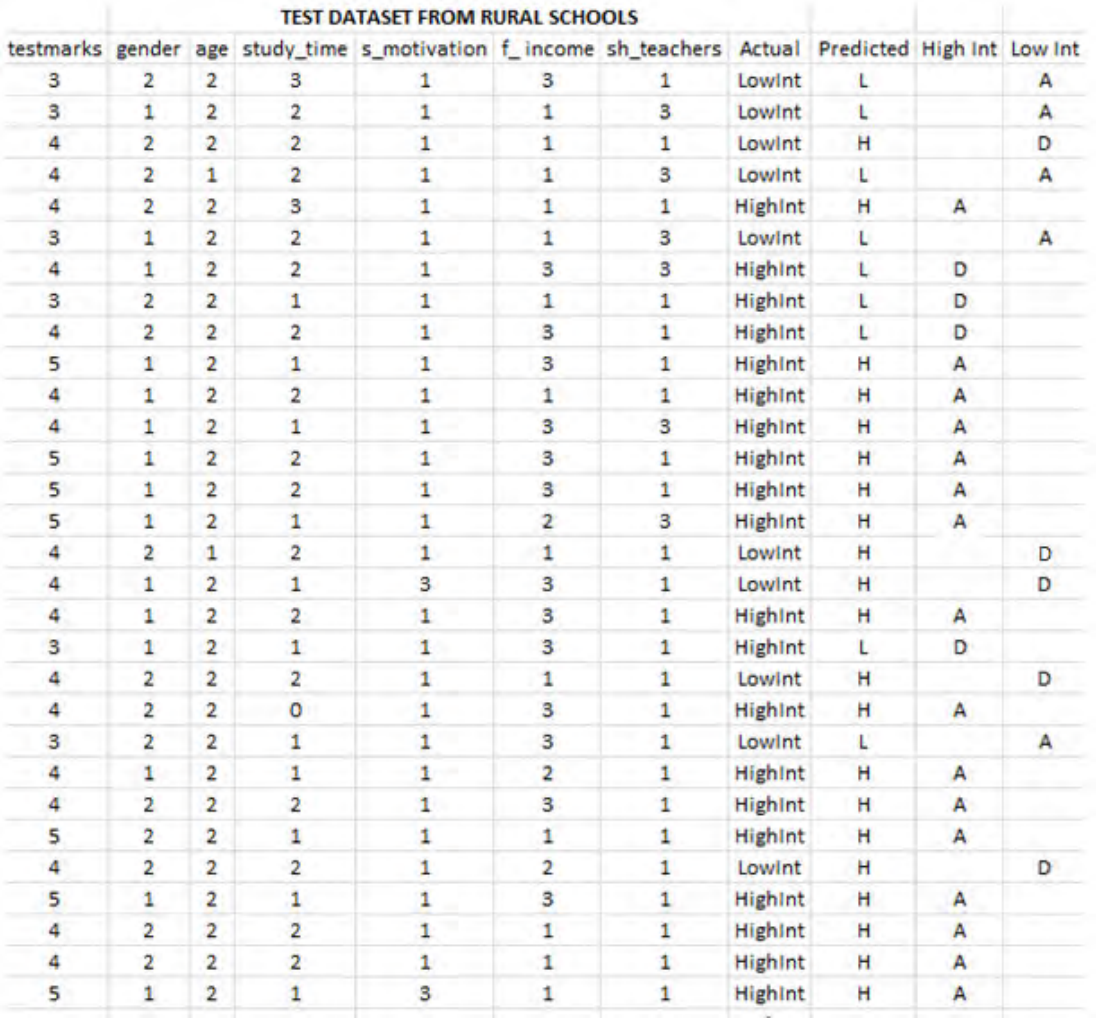
Results and Discussion of the Mobile Academic Performance Prediction System

7.1 Introduction

Data for the experiments conducted to determine the performance of the mobile academic performance prediction system came from three sources: 30% of the original dataset that was set aside as test data, while the rest, 70%, was used for training the models; 40% of a dataset obtained from 11 peri-urban schools; and 1839 student records from 15 rural primary schools made up of Class Six and Seven students. This chapter discusses the results and analysis of the three datasets as per the metrics generated from the confusion matrix discussed earlier (see Chapter 4). The metrics used are: prevalence, sensitivity, specificity, precision, F-Measure, and accuracy. Results and discussion are presented.

7.2 First Experiment: Rural Schools' Test Dataset

From the original complete dataset of 2426 student records collected from 54 rural schools in Kwale County, 695 records, 30% of the total, was randomly selected and set aside as the test data . A sample of the test data is presented in Figure 7.1. The data collection process was discussed in Chapter 4.



TEST DATASET FROM RURAL SCHOOLS

testmarks	gender	age	study_time	s_motivation	f_ income	sh_teachers	Actual	Predicted	High Int	Low Int
3	2	2	3	1	3	1	LowInt	L		A
3	1	2	2	1	1	3	LowInt	L		A
4	2	2	2	1	1	1	LowInt	H		D
4	2	1	2	1	1	3	LowInt	L		A
4	2	2	3	1	1	1	HighInt	H	A	
3	1	2	2	1	1	3	LowInt	L		A
4	1	2	2	1	3	3	HighInt	L	D	
3	2	2	1	1	1	1	HighInt	L	D	
4	2	2	2	1	3	1	HighInt	L	D	
5	1	2	1	1	3	1	HighInt	H	A	
4	1	2	2	1	1	1	HighInt	H	A	
4	1	2	1	1	3	3	HighInt	H	A	
5	1	2	2	1	3	1	HighInt	H	A	
5	1	2	2	1	3	1	HighInt	H	A	
5	1	2	1	1	2	3	HighInt	H	A	
4	2	1	2	1	1	1	LowInt	H		D
4	1	2	1	3	3	1	LowInt	H		D
4	1	2	2	1	3	1	HighInt	H	A	
3	1	2	1	1	3	1	HighInt	L	D	
4	2	2	2	1	1	1	LowInt	H		D
4	2	2	0	1	3	1	HighInt	H	A	
3	2	2	1	1	3	1	LowInt	L		A
4	1	2	1	1	2	1	HighInt	H	A	
4	2	2	2	1	3	1	HighInt	H	A	
5	2	2	1	1	1	1	HighInt	H	A	
4	2	2	2	1	2	1	LowInt	H		D
5	1	2	1	1	3	1	HighInt	H	A	
4	2	2	2	1	1	1	HighInt	H	A	
4	2	2	2	1	1	1	HighInt	H	A	
5	1	2	1	3	1	1	HighInt	H	A	

FIGURE 7.1: A screen shot of a section of the rural test data showing the student records, the actual intervention, and the predicted intervention as obtained using MAPPS

Each of the 695 records of seven feature was entered into the mobile interface of the MAPPS to obtain a predicted intervention. The predicted intervention was placed in the Predicted column. A comparison was them made between the actual and predicted, 'A' stands for agree, and 'D' stands for disagree. The letter 'A' or'D' was placed on the High Int or Low Int column depending on the Actual Intervention of the record. Four counts were generated to complete the confusion matrix in Table 7.1: the number of high intervention students that wrere correctly identified, made up of all the 'As' in the High Int column; the number of low intervention students that were correctly identified, made up of all the 'As' in the Low Int column; the students that were incorrectly identified as requiring low intervention, made up of all the 'Ds' in the High Int. column; and the students that were incorrectly identified as requiring high intervention, made up of all the 'Ds' in the Low Int. column.

Data Set		Actual HI	Actual LI
Test Data	Predicted HI	337	71
	Predicted LI	113	174

TABLE 7.1: Confusion matrix to determine the correctness of prediction for MAPPS on rural test data

The results show that 337 high intervention students were correctly recognised, and 174 low intervention students were correctly recognised. There were 113 high intervention students that were incorrectly identified as low intervention students, and 71 low intervention students that were incorrectly identified as high intervention. The metrics presented next give the analysis of the prediction performance of MAPPS.

Prevalence

The proportion of the actual high intervention students (450) to the total number of students (695) gives the high intervention prevalence which is 64.7%. This is compared to the proportion of the low intervention students (245) to the total number of students (695) that gives the low intervention prevalence of 35.3%. This suggests that in rural schools, the proportion requiring high intervention is about twice that requiring low intervention. This agrees with other research findings that students' academic performance is affected by socioeconomic status to a large extent (Sirin, 2005). Most of the students in rural areas of Kwale County come from low socioeconomic status homes.

Sensitivity

The ratio of correctly predicted high intervention students (337) to the actual number of students requiring high intervention (450) is 75%. Meaning, MAPPS is 75% sensitive in identifying the high intervention students. There are 113 students that have been identified as low intervention students when they actually need high intervention. A good number of students were seen to scores marginal marks, these are the students who lie on the boarder line of the high and low intervention classes. However, this is a concern that future research needs to address to reduce the error. It is also worth noting that because the prevalence of the high intervention class is 64.7%, the obtained sensitivity is better than random guessing.

Specificity

The ratio of correctly predicted low intervention students (174) to the actual number of students in the low intervention class (245) is 71%. Considering the prevalence of the low intervention class is only 35%, the specificity ratio reflects the ability of the MAPPS to reasonably identify the students who require low intervention.

Precision Rate

This is the ratio of correctly identified high intervention students (337) to the total number of the students predicted to be in the high intervention class (408). A value of 82.6% was attained. Only 71 students out of a total of 245 students in the low intervention class were misclassified and placed in the high intervention class. However, this is not a serious error because such students could benefit from the strategic intervention given to the high intervention students.

Accuracy

It is the ratio of both correctly identified low and high intervention (511) to the total number of students in the test data (695). It is the overall, or the average performance of the classifier in correctly classifying both classes. The attained value is 73.5%. This implies MAPPS misclassified 184 students and correctly classified 511 students out of a total of 695 students. As noted, the number of misclassified students is high and future efforts need to be put into reducing it. However, in this study, the focus is on the high intervention students, for which a better metric is the F-Measure discussed next.

F-Measure

This focuses on the accuracy of predicting the high intervention class. It is a combination of both sensitivity and precision. Notably, both metrics relate to the high intervention class, which makes this measure more useful. When computed, value of 79% was attained. Meaning, MAPPS identified correctly up to 79% of the high intervention students and misclassified up to 21% of the high intervention students. This is the total misclassification, but those students that have been placed erroneously in the high intervention class may not be a serious error, since they will benefit from the strategic intervention.

7.3 Second Experiment: Peri-urban Schools

The peri-urban dataset consisted of 1105 student records collected from 11 public primary Peri-urban schools in Mombasa County. The data was divided into 60% (663 records) training data and 40% (442 records) test data. Peri-urban schools share the qualities of both urban and rural schools (Tao, 2013). This study sought to test MAPPS with peri-urban data because of some of the shared characteristics with rural school students. The test data from the peri-urban schools had similar features as the rural school test dataset as shown in Figure 7.2.

Each record of the seven features was entered into MAPPS and the results of prediction noted and compared with the actual intervention as was done in the case of the rural test

	TEST DATASET FROM PERI -URBAN SCHOOLS									
test marks	Gender	Age	study_time	s_motivation	f_income	sh _teachers	Actual	Predicted	High Int	Low Int
4	2	1	2	1	2	3	LowInt	L		A
3	2	2	2	1	2	3	LowInt	L		A
1	2	2	2	1	2	3	LowInt	L		A
1	1	1	2	1	2	3	LowInt	L		A
3	1	1	0	1	2	3	LowInt	L		A
3	2	1	1	1	2	3	LowInt	L		A
2	2	2	2	1	1	3	LowInt	L		A
2	1	2	0	1	2	3	LowInt	L		A
2	2	1	0	1	1	3	LowInt	L		A
1	2	1	2	1	1	3	LowInt	L		A
3	2	1	2	1	1	3	LowInt	L		A
4	1	1	1	1	1	3	HighInt	H	A	
3	2	1	2	1	1	3	LowInt	L		A
2	1	2	1	1	1	3	LowInt	L		A
3	2	2	2	1	2	3	HighInt	L	D	
3	1	1	1	1	1	3	LowInt	L		A
2	1	1	2	1	1	3	LowInt	L		A
2	1	2	2	1	2	3	LowInt	L		A
3	1	2	0	2	2	3	LowInt	L		A
3	1	1	2	1	1	3	LowInt	L		A
2	1	1	0	1	3	3	LowInt	L		A
2	2	1	2	1	1	3	LowInt	L		A
3	1	1	0	1	3	3	LowInt	L		A
3	1	2	2	1	1	3	LowInt	L		A
3	1	2	1	1	3	3	LowInt	L		A
5	1	2	2	1	2	3	HighInt	H	A	
2	1	1	2	1	1	3	LowInt	L		A
3	1	1	2	1	1	3	LowInt	L		A
4	2	1	2	1	2	3	LowInt	L		A
2	2	1	1	1	1	3	LowInt	L		A
3	1	1	0	1	3	3	LowInt	L		A

FIGURE 7.2: A screen-shot of a section of the peri-urban test data showing the seven features, the actual intervention and the predicted intervention for both high and low intervention as determined using MAPPS

dataset. The four counts were also generated in a similar manner as for the rural test data, these are: true high intervention; true low intervention; false high intervention; and false low intervention, are shown in the confusion matrix in Table 7.2.

Data Set		**Actual HI**	**Actual LI**
Peri-Urban	Predicted HI	111	18
	Predicted LI	41	271

TABLE 7.2: Confusion Matrix showing correctness of prediction for MAPPS for Peri-Urban Data

The counts from the confusion metrics are: correctly identified high intervention students, 111; correctly identified low intervention students, 271; high intervention students that were incorrectly identified as low intervention, 41; and the low intervention students

that were incorrectly identified as high intervention were 18. The metrics discussed next presented the analysis of the prediction performance of MAPPS.

Prevalence

The number of students requiring high intervention were 152 out of the total of 441 which computes to 34.5%, while, the low intervention students were 289 out of the total of 441, which computes to 65.5%. The results indicate a smaller proportion of students require high intervention compared to those that require low intervention in peri-urban schools. This situation could be attributed to the higher socioeconomic state in peri-urban schools compared to rural schools (Tao, 2013). Students who experience low socioeconomic factors have lower chances of performing well academically (Okaya et al., 2013). The low prevalence of high intervention obtained in this study seems to agree with the previous studies that higher socioeconomic status positively contributes to improve academic performance.

Sensitivity

The ratio of correctly predicted high intervention students (111) to that of the actual number of students requiring high intervention (152) is 73%. Therefore, 27% of the students were incorrectly predicted as requiring low intervention. Compared to the sensitivity obtained for rural students, which was 75%, the reduction in sensitivity could be attributed to the low prevalence of the high intervention class.

Specificity

The ratio of the correctly predicted low intervention students (271) to that of the total number of students requiring low intervention (289) is 93.8%. Only 6.2% of the students belonging to the low intervention class were incorrectly placed in the high intervention class. This high specificity may imply a relationship between the prevalence and specificity. Previously for rural schools, a low intervention prevalence of 35%, had a corresponding specificity of 71%. In the current case, the low intervention prevalence is 65.5% and the corresponding sensitivity is 93.8%

Precision

The ratio of correctly predicted high intervention students to all the students predicted to be in the high intervention class is 86%. Of the 129 students predicted to be in the high intervention class, 111 students were correctly predicted and only 18 were wrongly predicted from the low intervention class. This is a high predictive value with a reduced error rate. It compares well with the precision of 83% obtained using rural data.

Accuracy

The overall performance of MAPPS in the peri-urban schools is 86.6%. This is a much

higher value compared to the accuracy of 74% obtained using the rural data. A possible explanation for this variation could be the unreliability of the accuracy metric for unbalanced data (Thai-Nghe et al., 2009).

F-Measure

The ability of the model to predict the high intervention class, given by the harmonic average between precision and sensitivity, was 79%. Incidentally, this value is the same as that obtained when using rural data. These two values are the same despite differences in the prevalence of the high intervention classes and the amount of test data, which implies that F-Measure is a stable and reliable metric measure that is not affected by imbalanced datasets (Rahman and Devanbu, 2013). The result also points to a possibility that MAPPS could be generalised for use in different environments such as the peri-urban environment with a different student data distribution.

7.4 Third Experiment: Rural Schools' Field Data

The third experiment used data obtained in schools as teachers used MAPPS for a period of three weeks. The goals of the experiment were to determine the usefulness of MAPPS, and the intervention level one or two years before students sit for the final Class Eight examinations. Early determination of the level of intervention would allow time for strategic intervention to be put in place (Scharf, 2013). A good example was the prediction of Grade Eight students' academic performance as early as in Grade Five in the US (Tamhane et al., 2014). MAPPS aimed to classify students in Class Six and Class Seven. These are considered upper primary, and hence their academic performance records are usually kept. They were made available to the researcher.

Teachers in 15 primary schools used MAPPS for three weeks between January 6, 2015 and January 30, 2015. The school term started on January 5. The teachers used MAPPS to predict the intervention levels of students entering Class Seven and those entering Class Eight. Class Six and Class Seven end of year test marks respectively were used. The experiments were conducted early in the year because this is when their end of year results from the previous year would be available. This allowed us to predict the students' performance one and two years before the students sit for the standardised exit examination at the end of Class Eight. Data was collected as the teachers asked the students questions to supply inputs into the MAPPS. To obtain the predicted intervention of a record, it was sent to the server to be classified, the output was either the student requires high intervention or the student requires low intervention. To determine the performance of prediction, standardised County exams were used in place of the actual intervention for the students in Class Seven, and Sub-county exams

for students in Class Six. County exams were used because students sit for only one standardised national examination at the end of Class Eight. However, studies show that such standardised exams correlate with the standardised national exams (Beleche et al., 2012) .

7.4.1 Experiment with Class Seven Data

Figure 7.3 shows a teacher's compiled test data of the seven features in the first seven columns, while the next two columns show the intervention prediction obtained from MAPPS.

Code No.	Test marks	Gender	Student age	Study time	Student motivation	Family income	Teacher/ student ratio	High Int.	Low Int.
01									X
02									X
03									X
04								X	
05									X
06								X	
07								X	
08								X	
09								X	
010									X
011									X
012								X	
013									X
014									X
015									X
016									X
017									X
018									X
019									X
020								X	
021								X	
022									X
023									X
024								X	
025									X
026									X
027									X
028								X	
029									X
030									X
031									X
032								X	
033									X
034									X
035									X

KEY								
	350-500=1	Female= 1	13 and below = 1	No time=0	Very good=1	Needs not met =1	Standard=1	X	
	300-349=2	Male = 2	Above 13 = 2	Little time=1	Good=2	Partially met=2	Acceptable= 2		X
	250-299=3			Enough time = 2	Fair=3	fully met=3	Large =3		
	200-249=4								
	<200 = 5								

KINANGO
P.O.BOX 3-
KINANGO

FIGURE 7.3: Sample test data obtained directly from students in Class Seven and the MAPPS intervention prediction for one of the primary schools

As shown in the figure, each student's record was predicted to be either in the high intervention or low intervention. A student code was included so that they can be used to identify each student's record to be associated with the actual intervention as obtained from the County standardised examinations. MAPPS intervention was compared with the associated actual prediction. The prediction performance results were compiled using the four counts of a confusion matrix as conducted previous in Sections 7.2 and 7.3.

Experimental results for Class Seven students in all the 15 schools are presented in the combined confusion matrix in Table 7.3.

School		**Actual HI**	**Actual LI**
School 1	Predicted HI	7	1
	Predicted LI	14	25
School 2	Predicted HI	67	12
	Predicted LI	0	10
School 3	Predicted HI	21	3
	Predicted LI	0	2
School 4	Predicted HI	38	10
	Predicted LI	20	34
School 5	Predicted HI	26	17
	Predicted LI	4	15
School 6	Predicted HI	13	2
	Predicted LI	7	34
School 7	Predicted HI	8	23
	Predicted LI	8	49
School 8	Predicted HI	17	5
	Predicted LI	0	3
School 9	Predicted HI	12	9
	Predicted LI	3	6
School 10	Predicted HI	18	2
	Predicted LI	18	16
School 11	Predicted HI	12	13
	Predicted LI	3	23
School 12	Predicted HI	40	8
	Predicted LI	12	40
School 13	Predicted HI	25	7
	Predicted LI	11	3
School 14	Predicted HI	6	4
	Predicted LI	1	34
School 15	Predicted HI	27	14
	Predicted LI	10	9

TABLE 7.3: Confusion Matrix for Class Seven Students in each of the 15 selected schools

The table shows the obtained classification of students in each of the 15 schools. As can be noticed from the confusion matrix, 7 schools (schools 2, 3, 4, 8,10, 13,15) had a higher

prevalence of students requiring high intervention. Five schools (school 1, 6, 7, 11, 14) had a higher prevalence of students requiring low intervention. Three schools had an average prevalence (school 5, 9, 12). In summary, there are seven schools with a high prevalence of high intervention, five schools with a high prevalence of low intervention, and three schools that have average prevalence. There are 448 actual high intervention students, and 413 actual low intervention students. This gives a uniform distribution that helps in overcoming the problem of unbalanced databases (Dal Pozzolo et al., 2014).

A combined confusion matrix for Class Seven students is presented in Table 7.4

School		**Actual HI**	**Actual LI**
Combined	Predicted HI	337	110
	Predicted LI	111	303

TABLE 7.4: Combined Confusion Matrix for Class Seven Students

With the help of the metrics as discussed in the previous two experiments, the intervention prediction performance analysis is presented next

Prevalence

As indicated the actual number of students requiring high intervention are 448 out of the total number of 861 students, prevalence is 52%. Implying, the prevalence of students requiring low intervention is 48%. Although this is good for the classifier, as the dataset is balanced, it shows a significant difference with the first set of test data. The two sets of data were obtained from the same rural area and are expected to show the same distribution. This difference could be attributed to the fact that, in the first experiment, a national standardised exam was used that may be more accurate.

Sensitivity

The true high intervention records were 337 and the actual number of students requiring high intervention were 448, giving a sensitivity of 75.2%. This sensitivity rate compares with that of the first experiment which was 75%. The data used in the first experiment was collected from a rural region, similar to the data used in this experiment.

Specificity

The true low intervention records were 303, and the actual low intervention records were 413, giving a specificity of 73.4%. The specificity obtained in the first experiment was 71%. The specificity value in this experiment is higher compared to that of the first experiment, which could be explained by the higher prevalence (48%) of low intervention, compared to 35.3% for the first experiment.

Precision

The value of precision was computed as a fraction of the true high intervention (337

records) and the total records predicted as high intervention (447). The precision value is 75.4%. This value is lower than both of the previous experiments. A possible explanation is the large number of misclassified low intervention students that have been identified as high intervention students. However, as mentioned previously, this type of misclassification is not negative because those students who may have passed marginally will be given the strategic intervention.

Accuracy
This metric was computed by getting a fraction of the total correctly classified records (337 + 303) and the total number of student records (866). The accuracy is 73.9%. This is the overall performance of the MAPPS. It is similar to the accuracy of 73.5% which was obtained in the first experiment.

F-Measure The harmonic mean between sensitivity and precision was computed and is 75.3%. Because both sensitivity and precision contribute to this value, the low values of the two measures generate a similarly lower F-Measure value.

7.4.2 Experiment with Class Six Data

Data was obtained in a similar way as for the previous experiment that used Class Seven data. Class Six data was compiled by the teachers as shown in Figure 7.3 for the previous experiment.

Table 7.5 shows the combined confusion matrix for all the 15 schools. As can be observed, nearly two thirds of the schools had more students requiring high intervention that those requiring low intervention. One third of the schools had fairly balanced students' records. No school recorded more students requiring low intervention. Overall therefore, there were more students requiring high intervention (766) compared to (262) who required low intervention. This could be the actual situation on the ground. It could also be as a result of a problem in the zonal standardised tests that were used as the target marks. Unlike the county exams, these are sub-county exams that may not undergo high level quality control. However, this was beyond the researcher's ability to control.

As indicated in Table 7.6 out of a total of 766 students identified to be in the high intervention class, 585 were correctly predicted to be in this class, while 131 students were wrongly placed in the low intervention class. Similarly, out of 262 students in the low intervention class, 181 students were correctly predicted and 81 students were erroneously predicted to be in the high intervention class. Failing to identify students who need high intervention is a more critical error than failing to identify students who

School		Actual HI	Actual LI
School 1	Predicted HI	20	0
	Predicted LI	16	13
School 2	Predicted HI	79	0
	Predicted LI	3	1
School 3	Predicted HI	67	1
	Predicted LI	4	1
School 4	Predicted HI	29	13
	Predicted LI	9	34
School 5	Predicted HI	34	4
	Predicted LI	6	10
School 6	Predicted HI	29	4
	Predicted LI	18	19
School 7	Predicted HI	29	14
	Predicted LI	11	31
School 8	Predicted HI	27	2
	Predicted LI	0	1
School 9	Predicted HI	26	11
	Predicted LI	3	5
School 10	Predicted HI	48	9
	Predicted LI	8	13
School 11	Predicted HI	29	3
	Predicted LI	17	6
School 12	Predicted HI	35	4
	Predicted LI	12	23
School 13	Predicted HI	57	4
	Predicted LI	9	1
School 14	Predicted HI	13	7
	Predicted LI	2	12
School 15	Predicted HI	62	5
	Predicted LI	13	11

TABLE 7.5: Confusion Matrix for Class Six Students in each of the 15 selected schools

require low intervention (Aguiar et al., 2014). Hence the 81 students erroneously placed in the high intervention class could actually benefit from strategic intervention measures.

School		Actual HI	Actual LI
Combined	Predicted HI	585	81
	Predicted LI	131	181

TABLE 7.6: Combined Confusion Matrix for Class Six Students

An analysis of the four confusion matrix counts is presented next.

Prevalence

The actual number of students requiring high intervention are 716 out of the total number of 978 students, and the high intervention prevalence is 73.2%. On the other hand, the low intervention prevalence is 26.8%. This prevalence could be a reflection

of the situation in rural areas as seen in the first experiment with 64.7% prevalence. However, the prevalence is on the higher side as it presents an extreme imbalance in the dataset.

Sensitivity

The true high intervention records were 585 and the actual number of students requiring high intervention were 716, this computes to a sensitivity of 81.7%. This sensitivity is higher than all the previous experiments, and could perhaps be explained by the higher prevalence of the high intervention class.

Specificity

The true low intervention records were 181, and the actual low intervention records 262, giving a specificity of 69%. Again, this is the lowest specificity probably due to the low prevalence of the low intervention class.

Precision

The value of precision was computed as a fraction of 585 true high intervention records and the total of 666 records predicted as high intervention. The precision is 87.8%. This value was also influenced by the high prevalence of the high intervention class.

Accuracy

The total number of correctly classified records is 766, and the total number of student records is 978 which computes to an accuracy of 78.3%. Being overall performance, it is a contribution of correctness to identify the high intervention students and the low intervention students. The high value of precision, therefore, explains why the accuracy has improved.

F-Measure

The F-Measure was computed as in the previous experiments, and a result of 84.6% was obtained. As earlier described, F-Measure is a harmonic mean between sensitivity and precision, whose values are 81.7% and 87.8%, which explains the high average performance.

7.4.2.1 Summary of Experimental Findings

This section presents a summary of the experimental results obtained using the three data sets and the user feedback. The summary is subdivided into three sections: predictive performance of the system with the test data; testing generalisability of the methodology; and testing the suitability and usefulness of the system.

In total, four test datasets were used to test the performance of MAPPS as described in the experiments. The purpose of these experiments was to analyse the ability of

MAPPS to correctly classify the high intervention students. A summary of the results is presented in Table 7.7

Metric	**Exp. 1**	**Exp. 2**	**Exp.3-Class 7**	**Exp.3-Class 6**
Prevalence	64.7%	34.5%	52%	73.2%
Sensitivity	75%	73%	75.2%	81.7%
Specificity	71%	93.8%	73.4%	69%
Precision	82.6%	86%	75.4%	87.8%
Accuracy	73.5%	86.6%	73.9%	78.3%
F-Measure	79%	79%	75.3%	84.6%

TABLE 7.7: An analysis of the four experiments using six metrics

The table shows the performance of MAPPS on the three test datasets. The prevalence metric shows that the rural dataset has a higher prevalence of the high intervention category (experiments 1 and 3) than the Peri-urban dataset. There could be fewer challenges affecting student academic performance in Peri-urban areas compared to the rural areas. This is indicated by the higher prevalence of the students that require low intervention in peri-urban areas.

The sensitivity of the system range from a low value of 73% to a high value of 82%. The lowest value was obtained with the Peri-urban data. Sensitivity is affected by prevalence; a lower prevalence in the peri-urban dataset resulted in the low sensitivity. Similarly, a high value (as seen in the case of Class 6 dataset) resulted in a higher sensitivity. The study adopts the sensitivity of 75% because it was achieved with balanced datasets (Experiment 3 Class 7). Further research requires to be conducted to improve this value.

Specificity was not so uniform, the highest was 93.8%, and the lowest was 69%. Specificity also seems to be affected by prevalence of the low intervention. As seen, the prevalence of low intervention was highest for the dataset in experiment 2 at 65.5%, and lowest in experiment 3 for Class 6 at 26.8%. A reasonable value from the experiments is therefore about 73%. Again, this value requires further improvement.

The precision value was highest at 87.8% and lowest at 75.4%, both in the third experiment. This reasonable precision means MAPPS misclassified fewer low intervention records as high intervention records. However, this could be as a results of the low intervention records being fewer compared to the high intervention records. From these reasonable results, it can be concluded that MAPPS is able to correctly identify the high intervention records,bearing in mind that those students that are incorrectly placed in this category are likely to benefit from the strategic intervention.

Accuracy gives the overall performance of a classifier. The highest accuracy was 86.6% with peri-urban data, and the lowest 73.5% with rural test dataset in experiment 1.

Accuracy shows the average ability of the classifier to classify both the high intervention and low intervention classes. As explained earlier, accuracy is affected by unbalanced datasets, hence the variation.

F-Measure is a more relevant metric in the sense that it determines the accuracy of a class separately. The high intervention F-Measure had the lowest value of 75.3% and the highest value of 84.6%. The fact that it is a harmonic mean between sensitivity and precision makes it a stable metric. An F-Measure value of nearly 80% attained by MAPPS is reasonably accurate in determining the high intervention class. It means for a given dataset, about 80% of high intervention records will be correctly classified by MAPPS. Only about 20% will be misclassified to belong to the low intervention class or some low intervention records will be classified to belong to the high intervention class. As stated earlier, the bad error is to place high intervention students in the low intervention class. However, since the 20% is shared between the two errors, fewer students are likely to be affected. This indicates the effectiveness of MAPPS.

Lastly, the third experiment helps in determining whether MAPPS can be used one or two years before students sit for KCPE. As seen, the results of the metrics are comparable with the rest, and in some cases, even better. For example, the highest F-Measure was attained by the Class Six dataset. The findings therefore, suggest that MAPPS could be used as early as two years before students sit for the final examination.

7.4.3 User Feedback from Rural Schools

In this section the feedback from the teachers who used MAPPS for three weeks is reported. Feedback is important because it assists researchers to improve the features and functionality of a software tool used for research, and it could also be a means of determining the relevance of the software and its impact on users (Pagano and Brügge, 2013). In this study, a total of 17 teachers gave their feedback after using the Mobile Academic Performance Prediction System (MAPPS) for three weeks. Teachers were given a set of 36 cards with words describing the possible feedback. The words were a mixture of both positive words (60%) and negative words (40%). Users were asked to select five top words to describe what they think of the system. Using these words, users were engaged in an interview to give the feedback for MAPPS. This approach has been found useful getting feedback from software users (Benedek and Miner, 2002, Kalz et al., 2014, van der Weegen et al., 2014).

The picture in Figure 7.4 shows the researcher engaging one of the users in getting his feedback using the cards and recorded the response on a piece of paper.

FIGURE 7.4: A picture showing the researcher engaging one of the users to collect feedback

Table 7.8 lists the words selected, and the number of teachers who selected each word.

Selected Word	**Number of Users**
Useful	9 teachers
Motivating	8 teachers
Relevant	7 teachers
Reliable	6 teachers
Applicable	5 teachers
Organised	5 teachers
Efficient	5 teachers
Adaptable	4 teachers
Easy to use	4 teachers
Valuable	4 teachers
Time consuming	2 teachers
Sophisticated	1 teacher

TABLE 7.8: Summary of Feedback from Teachers

Useful

Nine teachers selected this word. They were then asked to explain why they thought the system was useful, and had the following to say: the system can bring change of academic performance in the schools if put in use; education stakeholders can be made aware of the students' situation early enough; teachers can know the weaknesses of learners; teachers are able to discuss challenges affecting their students; and useful information is obtained which can lead to motivating strategic intervention.

Motivating

Eight teachers selected the word motivating, stating the following: it is exciting to see that the system is able to give a feedback response according to the challenges the students face that affect their academic performance; the system motivates us to ask the students to talk to their parents concerning the issues that affect them; it motivates us to follow up on the students who need to work hard in order to improve their academic performance; It provokes us to speak to the students about the situation affecting them; and finally, it motivates the user to want to continue using the system as it produces results quickly.

Relevant

Seven teachers selected this word and said the following statements: it is about the students' education; prediction given is accurate to the learners' intervention level; reveals the challenges faced by the student; once teachers know the learners situation, they will assist them; and since it deals with teachers and students, it is relevant to education.

Reliable

Six teachers selected the word and had the following to say: the exact information given by the student is displayed and the result given; the program correctly predicts the category of the student; the program does not cheat, it gives the holistic picture of the students' situation; answers obtained by the program agree with what is known about the students.

Applicable

Five teachers had the following to say: the program is easy to understand and relate the results to the students; it helps in gauging the level of understanding of the student; the information fed correctly brings out the results, hence it can be applied; the program can be adapted for use in secondary schools.

Organised

Five teachers said: the items are well structured to assist the student speak out their challenges; the students are asked one at a time, this ensures privacy.

Efficient

Five teachers said: it produces results immediately after feeding in the information; and results are given correctly, easily, and fast.

Adaptable

Four teachers said: students' weaknesses will be known so that the right intervention measures can be put in place; and it can be implemented in schools in order to identify students with specific needs.

Easy to use

Four teachers said the following: the program does not require much knowledge for one to start using it; the user only feeds the data, the program is not difficult to learn; it is self explanatory, no need to be guided; and can be used with little training.

Time consuming

Two teachers gave the following comments: the process of interviewing students requires set aside time, a role which should be assigned to a specific teacher; and a teacher has to create extra time to conduct the interviews with students.

Sophisticated

One teachers said: this is innovative technology in the education environment.

7.4.3.1 Summary of User Feedback

The experimental analysis is reflected in the user feedback as presented in the feedback analysis. The words that were selected by all the users are displayed in Figure 7.5

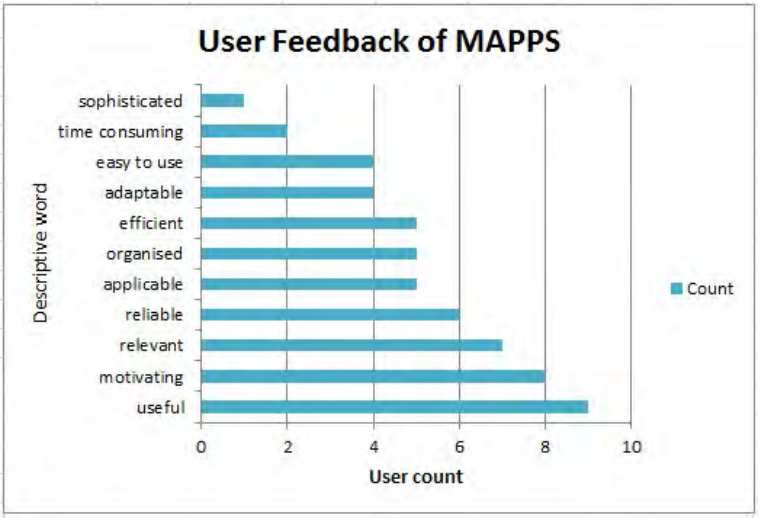


FIGURE 7.5: User feedback rating in terms of the selected descriptive words

The figure shows that the majority of the users agree that the system is useful in the sense that it could improve the students academic performance. When stakeholders are made aware of the students' status early enough, they are likely to think of ways and means to assist them. Besides, as teachers get data from the students, it helps to understand them so that they are able to discuss the students' status with their parents and other stakeholders who could come up with strategic intervention.

Motivated was also a popular word, which means the system was accepted by both the students and teachers. Students were excited as they identified the challenges they face

while giving their data. Similarly, teachers were excited by the fact that the system produced the result of intervention almost immediately. The speed of processing was made possible by Safaricom mobile provider network which covers most of the rural areas where the study was conducted. Therefore, MAPPS succeeded in performing the predictions.

The word *relevant* was also popular because users looked at MAPPS as a device that is student centered. The fact that it was about how to improve the students' academic performance by facilitating interaction between stakeholders concerning the student. The fact that users saw the relevance of MAPPS implies it was accepted and was successful in predicting the high intervention students. Further, users agreed with the results of prediction for their students and concluded that it is a reliable system.

The four words: *useful*, *motivating*, *relevant*, and *reliable*, dominated the user feedback. However other terms such as: *applicable*, *organised*, *efficient*, *adaptable*, and *easy of use*, were also selected as positive words to confirm the success of MAPPS from the users' point of view. Notably, two users thought the system is time consuming and one user thought the system is sophisticated. As stated previously, for time consuming, the users meant a person has to be assigned the role of using the system. The good thing is that, the system does not require to be used daily. It could be used at the beginning of the term or year when teachers are not busy and students have not started learning. On sophistication, the user was actually positive in the sense that MAPPS is a new innovation in the education field.

7.5 Chapter Summary

Three types of test datasets were used to conduct four experiments. The first two experiments used test data that was separated from the training data, and the third experiment used data that was collected while teachers used MAPPS for three weeks with their students.

The first experiment was conducted with 30% test dataset that was set aside from the rural dataset. The records for this dataset were complete with their target marks that were used to determine the accuracy of prediction. Results from this test dataset indicate that MAPPS enabled the classification of the high intervention student records with an acceptable accuracy. The performance could be improved with further research.

The second experiment was conducted with peri-urban test data which was 40% of the total data collected from peri-urban schools. This data set was also complete with target marks. Similar features were used for the test records as discussed previously. A key

difference with this dataset was that the number of low intervention records were more than those of the high intervention records. Similarly, results indicate that MAPPS predicted the high intervention class with reasonable accuracy.

The third experiment used new records collected when teachers used MAPPS for a period of three weeks. They used the system to predict individual Class Six and Seven students' intervention level. To determine MAPPS performance, County standardised exams given at the end of the year were adopted to generate the actual intervention for each student. Results indicate that MAPPS was able to achieve reasonable performance even when it was used with students who had two and one year left before completing primary school education. Therefore, the results indicate that MAPPS can be used two years before students sit for KCPE. This will allow strategic intervention to be put in place in order to help the students improve their marks in KCPE. Further, results from Class Six test data indicate that MAPPS successfully classified the high intervention students with the highest F-Measure. This is a further confirmation that MAPPS depends purely on data to do classification, hence given any set of data with similar features, it is able to predict correctly.

The results from all the experiments indicate the success of MAPPS in classifying the high intervention students. This is further confirmed by the user feedback. The users who used the system for three weeks were able to successfully predict the intervention levels of their students as discussed in the third experiment. They also provided their feedback that the system was useful, motivating, relevant, and reliable.

The next chapter presents an analysis of how these results have addressed the research questions, and a concludes the thesis.

Chapter 8

Conclusions

8.1 Introduction

This study proposed that an academic performance prediction model designed based on the EDM CRISP-DM process could be integrated with a mobile phone interface and used in resource-constrained areas of developing countries and that the integrated system, named MAPPS, could predict the high intervention students in rural primary schools so that strategic intervention could be put in place early enough. The system design was conducted following the user-centered design method. The design was carried out in collaboration with users who were technology literate.

To achieve this goal, the study was guided by the following three research questions:

1. Which is the best classifier model among the six common classifiers selected for the type of data used in this study?
2. What is the optimal subset of features from the total number of features from the two datasets used in this study?
3. What is the predictive performance of the Mobile Academic Performance Prediction System in classifying the high intervention class?

The chapter begins with a synthesis of how the findings from the educational data mining process and how the experimental findings of testing MAPPS addressed the research questions. This is followed by a discussion on the implications of the study. Finally, the chapter discusses the limitations of the study and recommendations for future work.

8.2 Synthesis of EDM Process Findings

8.2.1 Which is the best classifier model among the six common classifiers selected for the type of data used in this study?

The findings for this question are presented as per the following objectives that operationalise the research question.

To compare the prediction performance of the six selected classifier models on the rural dataset in terms of numbers of correctly classified and incorrectly classified students

Table 8.1 shows the predictive performance of the six classifier models were compared in terms of how well they correctly classified both the high intervention and low intervention student records. Logistic regression turned out to be the best classifier model for the rural dataset by correctly classifying the highest number of records from the training data of 1696 students records. Since the smaller the number of misclassified students, the better the model, logistic regression proved to be the most suitable for the rural dataset.

Algorithm	**Correctly classified records**	**Misclassified records**
logistic regression	1445	251
SMO	1439	257
Random Forest	1424	272
J48	1415	281
MLP	1372	324
Naïve Bayes	1240	456

TABLE 8.1: Comparison of classifier performance in terms of how well they correctly classified both the high intervention and low intervention student records

To compare the prediction performance of the six classifier models using the six selected metrics

A comparison of the classifier models' performance was carried out using six metrics (recall, specificity, ROC area, F-Measure, Kappa, and RMSE). Results showed that logistic regression led the other classifiers in attaining the best predictive rates in 5 metrics out of the six. The results show that logistic regression achieved the highest sensitivity of 90% in classifying the high intervention class, the best ROC area of 88.7%, the best F-Measure of 89.7%, the best Kappa of 0.6345, and the lowest error as measured by RMSE of 0.3375.

The only metric where its value was low is specificity, a metric that measures how well the classifier identifies the low intervention records. It could be that its rate was low

because of the low prevalence of students who belong to the low intervention class.

To determine the best classifier model according to the classifier performance results obtained using peri-urban data

In terms of which classifier model correctly classified the highest number of records, SMO turned out to be marginally better than logistic regression. SMO correctly classified 566 records - 7 records more than logistic regression, which was the second best classifier. SMO also misclassified the least number of records(97) - 7 records less than the records misclassified by logistic regression.

Further, an analysis of the classifier performance using the six metrics show SMO attained better performance in four measures: recall, specificity, F-Measure, and Kappa value, while logistic regression had the highest values in two: ROC area, and RMSE.

The fact that SMO turned out to be the best classifier model with peri-urban data, and not logistic regression, which was the best with the rural dataset, agree with previous findings that no single classifier performs best in all situations (Asif et al., 2014). Our results show that, although the datasets were similar in terms of the features used, they were collected from different environments, which had a major effect on the most suitable classifier.

8.2.2 What is the optimal subset of features from the total number of features for both rural and peri-urban datasets?

The reported findings in this question are presented in terms of the subquestions that were answered to address this second question.

To determine the most predictive features from the three lists that have been ranked using ranking algorithms.

Features were first ranked using three ranking algorithms - reliefF, information gain, and gain ratio. For the rural dataset, three features were identified as the most predictive: test-marks, student gender, and teacher shortage. They were ranked the best in two of the three lists of the ranking algorithms. Test-marks was ranked top in all the three lists. These features were considered to be the core features. Successive classifier models were built starting with these three features.

Identifying the top features could be helpful to education stakeholders as indicators of where to put strategic intervention. Test-marks, which topped the list, is the most

important indicator of academic performance in final examinations. Figure 8.1 shows the correlation. The figure shows that test marks correlate with final examination marks as the scatter plots are seen to have a direct relationship. Students who score low test marks should therefore be closely monitored to understand the possible causes.

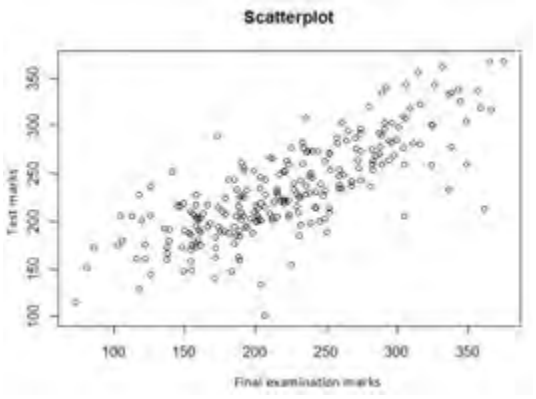


FIGURE 8.1: Test marks shown as a strong indicator of final examination

Student gender was the second most predictive feature. In most rural areas, girls are likely to have less study time than boys because they are expected to help with household chores. Figure 8.2 shows that girls, represented by the yellow boxplot, have a lower median mark than the boys, represented by the cyan boxplot, in final examinations in Kwale County.

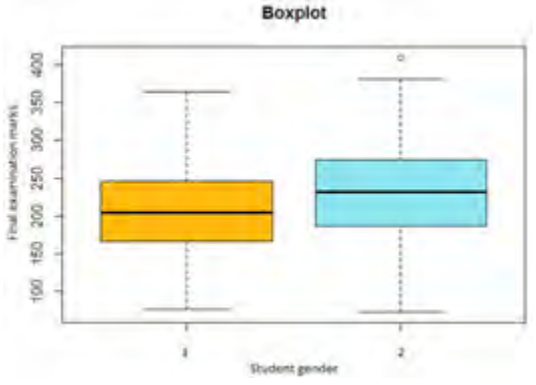


FIGURE 8.2: Gender shown to affect student performance in final examination

Teacher shortage also appeared as one of the key features. Whenever teachers are posted to rural schools they always look for an opportunity to transfer to urban and peri-urban areas because of the poor working conditions.

In the case of peri-urban data, two features appeared top in all of the three lists. These are test marks and parent education level. Hence, these were taken to be the core features; the successive modelling was done starting with them.

The findings show that test marks is the top feature in the two regions. That is, the test marks that students score during their learning process are highly indicative of the final examination academic performance. Therefore, a student who gets low test marks should be a concern to the education stakeholders. However, the low marks obtained in test marks are caused by other factors. This study has identified six, namely gender, teacher shortage, student motivation, family income, student age, and study time such factors using data mining methods.

To determine the optimal number of features that achieve the highest predictive performance of the selected classifier models in the two datasets

The search for the optimal features was achieved through successive modelling of the six classifiers starting from the three identified top features down to the last feature in each of the three ranked lists. The metric values attained with Information Gain ranked features were the highest with the least number of features. Seven features were identified as the optimal number of features using the rural dataset. It was concluded that the 7 features are the optimal feature subset. The features are: test marks, gender, teacher shortage, student motivation, family income, student age, and study time.

Similarly, in the case of the peri-urban dataset, seven features were identified as the optimal feature subset. These are: test marks, parent educational level, student age, teacher absenteeism, student discipline, gender, and family income.

These two sets of optimal features are different, with at least four features shared: test marks, gender, family income, and student age. Three features are unique to the rural dataset: teacher shortage, student motivation, and study time. Another three are unique to the peri-urban dataset, including teacher absenteeism, student discipline, and parent educational level. The finding confirms that students from these two regions have different features that affect their academic performance.

8.2.3 What is the predictive performance of the Mobile Academic Performance Prediction System in classifying the high intervention class?

Similarly, the findings to this question are reported by answering the following subquestions.

To compare the MAPPS classification performance between the rural schools dataset and the peri-urban dataset

The test data for rural schools, consisting of 695 records, was used in the experiments. This data was set aside to be used only for testing purposes. Results show that the high intervention prevalence was 64.7%. And, out of the 450 actual high intervention student records, 337 records were correctly classified. This gives a sensitivity rate of 75% in determining the high intervention category of students. The other metrics used are: precision - out of 408 student records predicted as being in the high intervention class, 71 students were misclassified to be in the low intervention class, giving a precision rate of 82.6%; accuracy, which measures the overall performance of MAPPS, attained a rate of 73.5%; and F-Measure, the harmonic mean of sensitivity and precision, attained a rate of 78.8%. The F-Measure is a rate that measures the predictive ability of MAPPS in classifying the high intervention class. These results show that MAPPS is capable of identifying the high intervention students with a performance which is nearly 80%. Meaning, there is an error of about 20% that combines both high and low intervention misclassification. The high intervention misclassification is not an inferior error, since such students will benefit from the strategic intervention.

The performance results using the peri-urban test dataset are: prevalence for the high intervention is 34.5%; sensitivity is 73%; precision is 86%; accuracy is 86.6%; and F-Measure, or a harmonic mean between sensitivity and precision, is 79%.

Comparing the two sets of results shows that MAPPS is capable of being used to classify students and, specifically, identify the high intervention class with reasonable accuracy. Although the values of most of the metrics vary between the two test datasets, the F-Measure values are similar. These similar values of F-Measure suggest that MAPPS could be used with different student datasets. In this study, the dataset records had similar features but had different prevalence for the high interention class.

To compare the prediction performance of MAPPS in the two student datasets; one for students two years and the other for students one year before they sit for the final examination

The MAPPS classification performance on Class Seven student records was compared to that of Class Six student records, and these two were also compared with the results obtained when MAPPS was used with the rural and peri-urban test data.

Classification performance of MAPPS for Class Seven records was calculated by comparing the predicted outcome from the student records and marks of a standardised County examination that students sit for at the end of the year. The results were: sensitivity rate of 75.2%; precision rate of 75.4%; accuracy of 73.5%; and a harmonic mean value - F-Measure- of 75.3%.

Likewise, for Class Six, performance of MAPPS was calculated by comparing the predictive performance of MAPPS and the marks in standardised exams offered at the end of Class Six. The results were: sensitivity rate of 82%; precision of 88%; accuracy of 78.3%; and an F-Measure value of 84.6%.

The performance values for Class Six were unexpectedly higher than those of Class Seven. A possible explanation for these unexpected results may have to do with the target examination used. The level of standardisation at Class Six may not be as thorough as that for Class Seven. As seen from the results, the values for Class Seven are comparable with the test data values. Therefore, Class Seven values give a better reflection of the reality.

However, the experiments with the field data show that MAPPS could be used one or two years before students sit for KCPE. As seen, the results of the metrics are comparable with the test data results.

8.3 Summary of the Conclusions

8.3.1 Educational Data Mining Framework

Educational Data Mining (EDM) formed the underlying theoretical framework for this study since it embodies the concept that support the building of academic performance prediction, and the concepts of motivating strategic intervention for students who may fail in final examinations. EDM was chosen for the following reasons (Siemens and Baker, 2012):

EDM research is focused on conducting academic performance prediction. There are many studies related to our work in EDM. Most of the studies reviewed in literature (see Chapter 3) that gave this study a strong base are in EDM. EDM studies are scarce in developing countries of Africa; this study seeks to fill this gap.

EDM follows a standard framework where features are studied as a complete set of features and in subsets. This study followed the framework and conducted feature selection. Experiments were conducted beginning with all the features in order to eliminate redundant or irrelevant features. This process reduced the database to one third, without affecting the classifier performance, and improved the speed of the system. Further, the small feature subset was more convenient for implementing on a small mobile screen.

EDM focuses on automated discovery of knowledge as opposed to human judgment, which is the case in learning analytics. In this study, training data was used to build

six classifier models. The prediction performance of each of the six selected models was compared to identify the best, which was used to build the academic performance prediction system. This system was then used to automatically discover the intervention levels required by new students. This possibility of automating knowledge discovery is a strength for EDM.

Finally, EDM is focused on automation to empower education stakeholders. This was the focus in this study: to develop an academic performance prediction system with a mobile phone interface. The mobile phone interface would make the system usable in rural environments with limited infrastructure where PCs may not be usable or affordable. Although teachers may have an idea of their students' level of intervention required, the prediction system would empower other education stakeholders such as parents and education officers to come together and initiate strategic interventions.

8.3.2 The CRISP-DM Process

Chapter 4 presents the detailed EDM CRISP-DM design process followed in this study. This is the process that was used to find the best classifier model- logistic regression - and the optimal subset of features. To complete the design of MAPPS, Chapter 6 presents the design process for the mobile phone interface.

The complete design process was guided by the limitations of the mobile phone, the rural environment of the users and the standard six-steps of CRISP-DM that were adopted for the design process. The limitations of the mobile phones - small memory and low processing power - are standard limitations for the type of mobile phones that are widely available and affordable to users in rural areas of developing countries. Further, users in most rural areas lack infrastructure that could allow the use of PCs. These challenges motivated this study. Therefore, we proposed a process framework that incorporates a machine learning process and a mobile interface design process for developing the Mobile Academic Performance Prediction System. This system could be replicated for other mobile academic performance prediction systems.

The last step of the CRISP-DM process makes use of the discovered knowledge. In this study, using this discovered knowledge entailed designing prototypes and deploying them for evaluation by the users.

8.3.3 Contribution to Knowledge

The study contributes knowledge to ICTs that are involved in enhancing the socio-economic development known as Information and Communication Technologies for Development or ICT4D (Donner and Toyama, 2009). It is a wide area of research, with research methods such as the experimental intervention that this study used.

The 'T' in ICT4D represents technology in the form of the Educational Data Mining classifier model that was integrated with the Mobile phone interface to develop MAPPS. The 'D' for development component is a theory contribution that this study contributes towards reducing the number of undereducated youth in rural areas of developing countries. Increasing the number of educated youth is possible if more primary school graduates transit to secondary and tertiary education; this is one of the ways to promote development in a community (UN, 2015).

In most of the rural areas of the developing nations, especially in Sub-Saharan Africa, desktop computers and laptops may not be affordable. Mobile phone technology, therefore, has been embraced, and has become available and usable everywhere because phones can be recharged with solar power. Additionally, mobile service providers have enhanced the mobile network in rural areas. The solution, therefore is to develop an academic performance prediction model that integrates a mobile phone on the user end. This study has shown that this can be achieved.

The prototype developed in this study could be adapted in future studies that seek to identify the students who require strategic intervention in other levels of the education cycle, and in other areas of developing regions. Studies have shown that predicting academic performance of students early enough is helpful in motivating strategic intervention (Tamhane et al., 2014).

8.4 Limitations of this Study

The focus of this research was to predict academic performance for primary school students and not secondary or tertiary institutions. Further, the primary school students were from rural schools; some peri-urban school students were used for the purpose of comparing the results. Urban primary school students were not used; neither were private school students. Additionally, the system only categorised the students into two classes: high intervention and low intervention. The choice of the two classes was motivated by the goal of identifying the students who are likely to fail (high intervention class) so that appropriate strategic intervention could be motivated for and initiated by education stakeholders.

The choice of smart phones for the system interface means that the more popular feature phones cannot be used. However, cheap smart phones are increasing in the market. Other mobile phone limitations include limited memory to contain large datasets required for training the models, and limited processing power to be able to do the classifier model training. This forced us to place the classifier model on the university server. It was linked to the mobile interface via the Web server. The mobile phone also had the limitation of a small screen that this study made provision for by searching for an optimal feature subset.

This study was also not involved in proposing or initiating intervention. The focus was on the design and development of MAPPS. The classification performance of MAPPS was the only way to know whether or not users would be motivated to initiate strategic intervention. The performance evaluation was conducted by testing the system's performance on three types of test datasets, to determine whether it achieved reasonably accurate prediction that could motivate intervention.

Finally, the study did not conduct evaluation on the long-term impact for the students who were classified as requiring high intervention; this will be part of the future work.

8.5 Further Research

8.5.1 Extending the capabilities of the system

As a proof-of-concept prototype, the system built in this study has a lot of room to be extended. Firstly, the two components of this system: the academic performance prediction classifier model and the mobile phone interface could be merged to be resident in the mobile phone. This could be possible as the memory size and processing power of smart phones improve, especially the smart phones affordable to the rural communities of developing countries. This will eliminate the need for Internet and make the system more affordable to the users in rural schools.

Further, the possibility of allowing for batch processing of students being predicted is another possible extension. New students whose records have been compiled beforehand could be entered in the system in batches. This could reduce the time spent using the system.

In addition to the prediction of high and low intervention, the system could be given the ability to suggest a strategic intervention for a student or a group of students. The system could point out the possible activities and who among the stakeholders

should undertake those activities to best assist the student so that they improve in their academic performance and in the final examination.

Additionally, further research could look into the possibility of increasing the training dataset as new student records that are predicted are added to the training dataset. As the database grows and hence the training data, it would allow the system to improve in prediction accuracy for any new student records.

With these extensions, the system will be more effective, efficient and more useful.

8.5.2 Longitudinal study

The addition of some of the above-mentioned extensions could also reduce the cost of airtime required to send student records to the server and get feedback. It may then be possible to conduct a longitudinal study and carry out many more experiments with primary school students. Further, with some modifications, the system could be used with secondary school students. For the primary school students, experiments could be started with Grade Four students all the way to Grade Eight. Likewise, for the secondary school students, experiments could be started from year one to the final fourth year. More users could also be involved among the rural and peri-urban schools for a thorough comparison and evaluation of the system.

8.5.3 Use of the system with other education stakeholders

In this study, the researcher only worked with teachers in primary schools because of the limitation of resources. Teachers used the system to predict Class Six and Seven student required intervention. However, it is possible to conduct a long-term study with other education stakeholders.

Parents in collaboration with teachers could use the system with their children so that they put in place specific strategic intervention over an extended period of time. Similarly, it is possible for education officers to use the system during school inspections. The education department would then be in a position to put in place relevant strategic intervention over an extended period of time, such as increasing the number of teachers in the school with high numbers of students requiring high intervention. Such experiments would measure the long term impact of MAPPS.

Appendix A

University of Cape Town Ethical Clearence

Department of Environmental and Geographical Science
University of Cape Town
RONDEBOSCH 7701
South Africa

e-mail: Michael.meadows@uct.ac.za
phone : + 27 21 650 2873
fax : +27 21 650 3791

14th June 2013

Mr Mvurya Mgala
Department of Computer Science
University of Cape Town
mmvurya@yahoo.com

Dear Mr Mgala

Investigating prediction modelling of academic performance for primary school pupils: a data mining approach

I am pleased to inform you that, having scrutinized the details of your above-named application for research ethics clearance, the Faculty of Science Research Ethics Committee has approved your proposal in terms of its attention to ethical principles.

Your approval code for the project is: SFREC 018_2013

I wish you success in the work involved.

Yours sincerely

Signed

Michael E Meadows
Professor and Head of Department
Chair: Science Faculty Research Ethics Committee

FIGURE A.1: UCT Research Ethics committee proposal approval, code SFREC 018_2013

Appendix B

Certicate of Ethical Approval in Kenya-by Pwanu University as Agent of NACOSTI

ERC/PhD/002/2013

PWANI UNIVERSITY CHAIRMAN 04 JUL 2013 ETHICS REVIEW COMMITTEE P. O. Box 195-80108, KILIFI

Pwani UNIVERSITY

P.O. Box 195-80108 Kilifi, Kenya | Tel: +254-41-7522498/59 | MOBILE: +254-706-727898/9 | FAX: +254-41-7522128 | E-Mail: info@pwaniuniversity.ac.ke

ETHICS REVIEW COMMITTEE

CERTIFICATE OF ETHICAL APPROVAL

THIS IS TO CERTIFY THAT THE PROPOSAL SUBMITTED BY:
Mr. Mvurya Mgala

REFERENCE NO:
ERC/PhD/002/2013

ENTITLED:
Investigating Mobile based Prediction Modelling of Academic Performance for primary School Pupils: a Data Mining Approach

TO BE UNDERTAKEN AT:
Kwale County in Kenya

FOR THE PROPOSED PERIOD OF RESEARCH

HAS BEEN **APPROVED** BY THE ETHICS REVIEW COMMITTEE

AT ITS SITTING HELD AT PWANI UNIVERSITY, KENYA

ON THE 11th DAY OF APRIL 2013

CHAIRMAN SECRETARY LAY MEMBER

Signed

decision or on account of misuse or error appearing on the face

Pwani

FIGURE B.1: Semi-structured Interview

Appendix C

Permission of Entry in Kwale County, Kenya

MINISTRY OF EDUCATION

Telegrams: "EDUCATION", KWALE
Telephone: Kwale 040/2104010
Email Address:
kwalecde@gmail.com
Please when replying quote
REF: KWL/CDE/R.4/1

The County Director of Education
P. O. BOX 20
KWALE

9/07/2013

RE: MVURYA MGALA REF: ERC/PhD/002/2013 A PhD STUDENT

The above PhD student of Pwani University is hereby authorized to conduct research in the County.

Mr. Mgala is out to investigate the problems of low performance among Primary school pupils in the County.

He will use semi structured interview questions as part of his methodology in the County.

The County believes that, with this kind of research, the County will not what so ever perform dismally in the Primary Section.

Signed

JUMA MWATENGAR
COUNTY DIRECTOR OF EDUCATION
KWALE

All DEO's
KWALE COUNTY

FIGURE C.1: Permission to conduct survey in Kwale County, Kenya

Appendix D

Permission of Entry in Mombasa County, Kenya

REPUBLIC OF KENYA
MINISTRY OF EDUCATION, SCIENCE AND TECHNOLOGY
STATE DEPARTMENT OF EDUCATION

Telegrams: "SCHOOLING",
Mombasa
Telephone: Mombasa 2315327 /
2230052
When replying please quote
Email: pdecoast@yahoo.com

COUNTY DIRECTOR OF EDUCATION
MOMBASA COUNTY,
P. O. BOX 90204 – 80100,
MOMBASA.

Ref. No.MC/ED/GEN/23/5

16th July, 2014

All Heads of Primary Schools
MOMBASA COUNTY

RE: RESEARCH CONDUCTION
MR. MVURYA MGALA-A PHD STUDENT

The above named PHD student of Pwani University is hereby authorized to conduct research in the County.

Mr. Mgala is out to investigate the problems of how performance among primary school pupils in the County.

He will use semi structured interview questions as part of his methodology of conducting the research.

Any assistance accorded to him will be appreciated.

Signed

Abdikadir M. Kike
COUNTY DIRECTOR OF EDUCATION
MOMBASA COUNTY

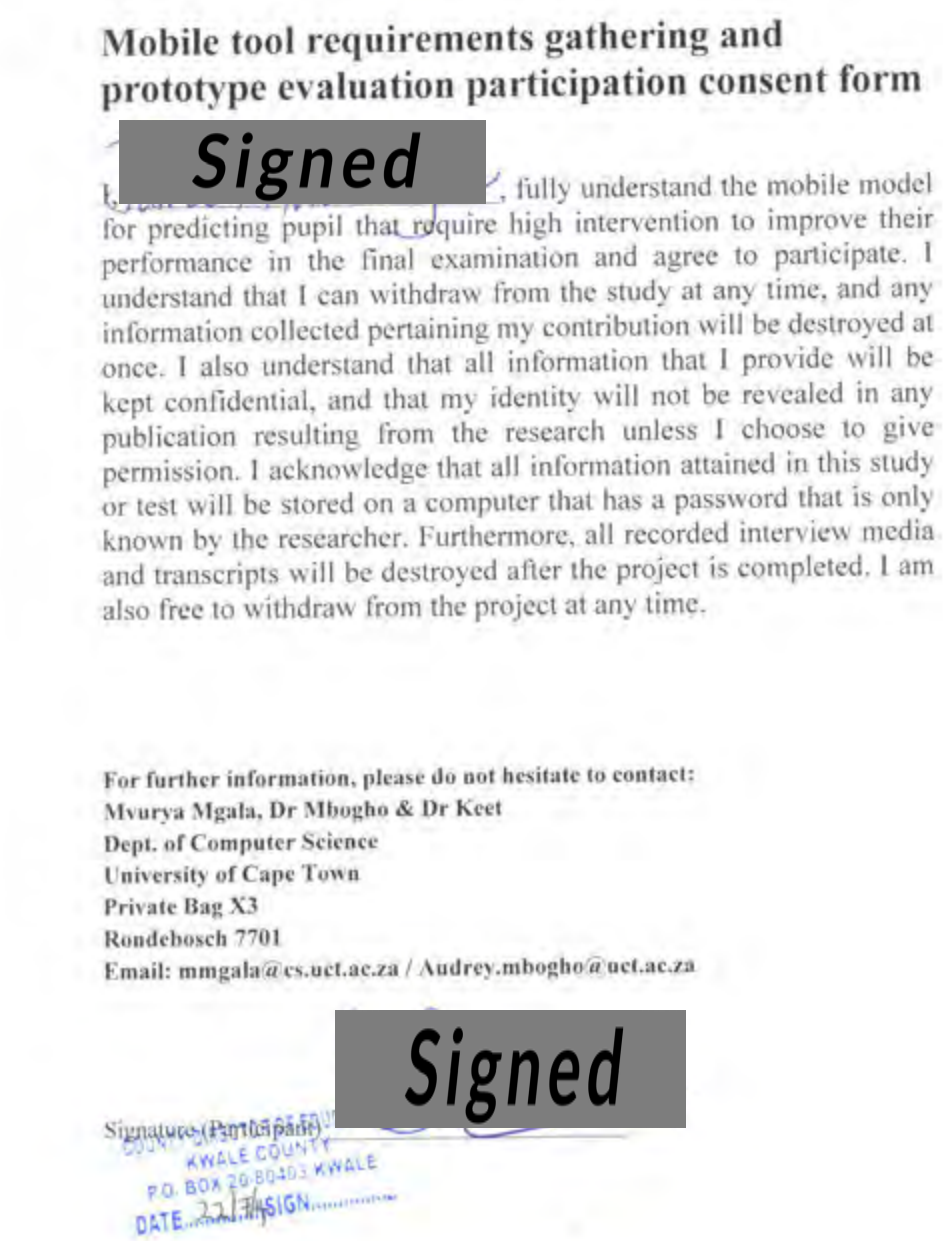
Copy to:

All Sub-County Director of Education
MOMBASA COUNTY

FIGURE D.1: Permission to conduct survey in Mombasa County, Kenya

Appendix E

Consent Form for Participants



Mobile tool requirements gathering and prototype evaluation participation consent form

Signed

, fully understand the mobile model for predicting pupil that require high intervention to improve their performance in the final examination and agree to participate. I understand that I can withdraw from the study at any time, and any information collected pertaining my contribution will be destroyed at once. I also understand that all information that I provide will be kept confidential, and that my identity will not be revealed in any publication resulting from the research unless I choose to give permission. I acknowledge that all information attained in this study or test will be stored on a computer that has a password that is only known by the researcher. Furthermore, all recorded interview media and transcripts will be destroyed after the project is completed. I am also free to withdraw from the project at any time.

For further information, please do not hesitate to contact:
Mvurya Mgala, Dr Mbogho & Dr Keet
Dept. of Computer Science
University of Cape Town
Private Bag X3
Rondebosch 7701
Email: mmgala@cs.uct.ac.za / Audrey.mbogho@uct.ac.za

Signed

Signature (Participant)

KWALE COUNTY
P.O. BOX 20-80403 KWALE
DATE..... SIGN.....

FIGURE E.1: Participants consent form for protoytype evaluation

Appendix F

Semi-structured Interview

Semi-structured Interview questions

Why has Kwale County been performing poorly in the Kenya Certificate of primary education examinations?

[Respondents are County director of education, District education officer, District quality assurance and standards officers, Area education officers and Head teachers]

Thank you for taking the time to participate in this interview.

There are 13 questions in this survey

1. What is your position?
   - ○ County director of education
   - ○ DQASO
   - ○ DEO
   - ○ AEO
   - ○ Head teacher

2. How many years have you worked in the current position in this County?
   - ○ Less than 1 year
   - ○ Between 1 and 3 years
   - ○ Between 3 and 5 years
   - ○ Over 5 years

3. Please state at least one of your main responsibilities as an education officer in the County.

4. For how many years has the trend of poor performance existed in the County?
   - ○ The last two years
   - ○ The last 5 years
   - ○ For more than 5 years

5. What pupil personal characteristics cause poor performance in this County?

6. What teacher characteristics contribute negatively to academic performance in this County?

7. What local community characteristics contribute to low academic performance in the County?

8. What parents' characteristics affect their children's academic performance?

9. What school characteristics cause low academic performance?

10. State any other causes of poor academic performance in the County.

11. Please rank the characteristics from highest impact to lowest impact. Please number each line in order of preference from 1 to 6

    Pupils' personal characteristics
    Teacher characteristics
    Community characteristics
    School characteristics
    Parents' characteristics
    Others

12. Please give your opinion on what can be done to end the trend of low academic performance in the County.

13. What information will be relevant for the decision makers to solve the problem of low academic performance in the County?

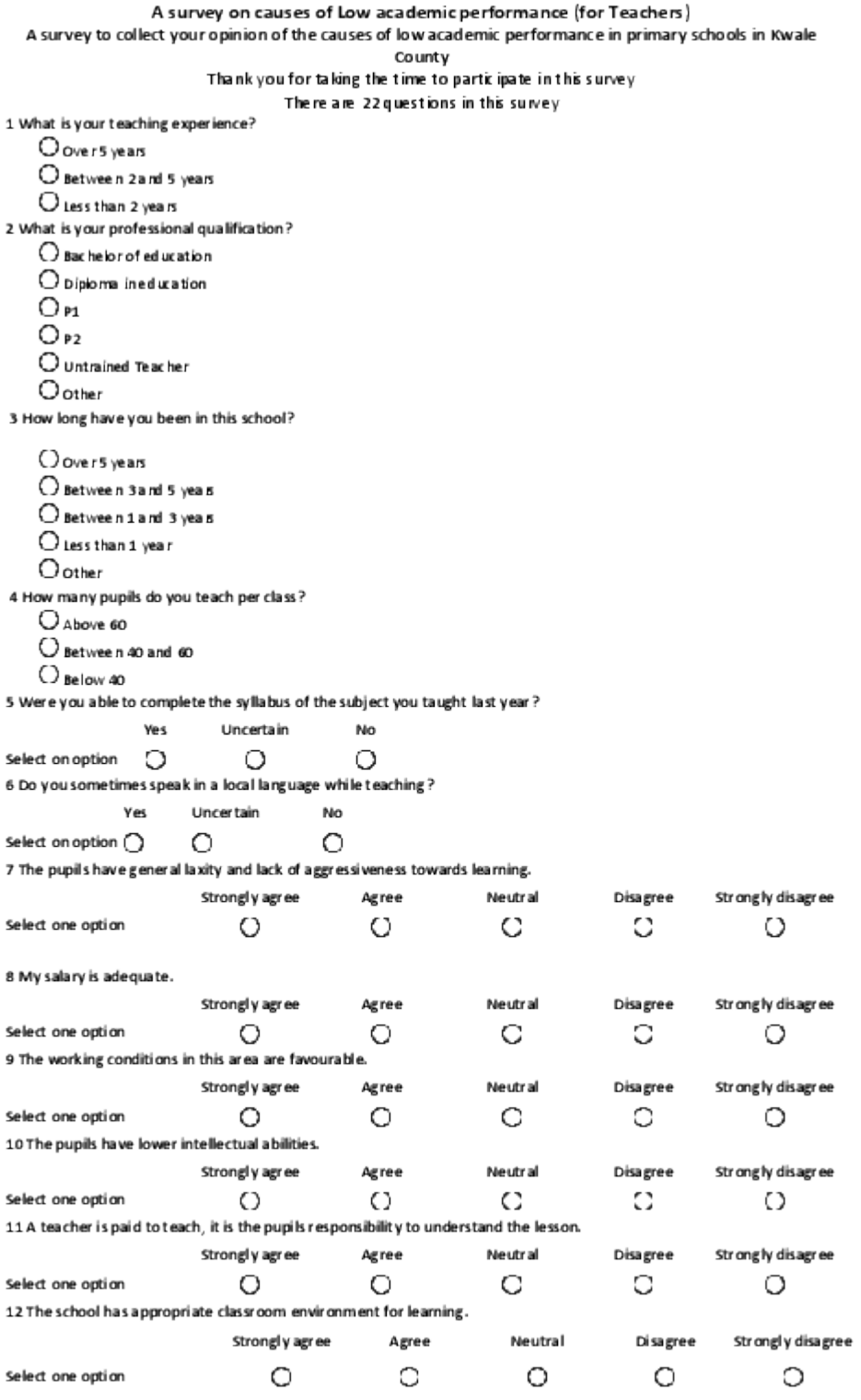
Thank you for completing this interview, we hope the results of this research will help to change the trend of low academic performance in this County

.

FIGURE F.1: Semi-structured Interview for education officers and head teachers

Appendix G

Teachers' Questionnaire



A survey on causes of Low academic performance (for Teachers)

A survey to collect your opinion of the causes of low academic performance in primary schools in Kwale County

Thank you for taking the time to participate in this survey

There are 22 questions in this survey

1 What is your teaching experience?

- ○ Over 5 years
- ○ Between 2 and 5 years
- ○ Less than 2 years

2 What is your professional qualification?

- ○ Bachelor of education
- ○ Diploma in education
- ○ P1
- ○ P2
- ○ Untrained Teacher
- ○ Other

3 How long have you been in this school?

- ○ Over 5 years
- ○ Between 3 and 5 years
- ○ Between 1 and 3 years
- ○ Less than 1 year
- ○ Other

4 How many pupils do you teach per class?

- ○ Above 60
- ○ Between 40 and 60
- ○ Below 40

5 Were you able to complete the syllabus of the subject you taught last year?

	Yes	Uncertain	No
Select on option	○	○	○

6 Do you sometimes speak in a local language while teaching?

	Yes	Uncertain	No
Select on option	○	○	○

7 The pupils have general laxity and lack of aggressiveness towards learning.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

8 My salary is adequate.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

9 The working conditions in this area are favourable.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

10 The pupils have lower intellectual abilities.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

11 A teacher is paid to teach, it is the pupils responsibility to understand the lesson.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

12 The school has appropriate classroom environment for learning.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

FIGURE G.1: Teachers' Questionnaire pg1

13 The school has enough textbooks and exercise books.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select on option	◯	◯	◯	◯	◯

14 There is a secure working relationship between the head teacher and the teachers.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	◯	◯	◯	◯	◯

15 There is a close supervision from the school administration on academic matters.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	◯	◯	◯	◯	◯

16 There is a close supervision of the teaching and learning activities by the District quality assurance and standards officers, they visit the school at least once a term.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
select one option	◯	◯	◯	◯	◯

17 The administration in this school creates a conducive teaching and learning environment.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	◯	◯	◯	◯	◯

18 There is a serious shortage of teachers in this school.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	◯	◯	◯	◯	◯

19 Parents are very keen with their children progress; they at least come once a term to see the teachers to discuss their children's academic performance.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	◯	◯	◯	◯	◯

20 The children are generally undisciplined.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	◯	◯	◯	◯	◯

21 Please give your opinion why pupils in Kwale County perform poorly in KCPE?

22 Please give your suggestions on what can be done to improve performance in the County.

Thank you for completing this survey. We hope the results will help in changing the trend of low academic performance in the County.

FIGURE G.2: Teachers' Questionnaire pg2

Appendix H

Students' Questionnaire

Last name:.....First name.....Index No:.....

A survey on causes of Low academic performance (for students)

A survey to collect your opinion of the causes of low academic performance in primary schools in Kwale County

Thank you for taking the time to participate in this survey.
There are 22 questions in this survey
These questions require you to give personal information

1. How old are you?
Please choose **only one** of the following:
- ○ 13 years and below
- ○ 14 years and above
- ○ I don't know

2. What is your gender?
Please choose **only one** of the following:
- ○ Male
- ○ Female

3. How far is your home from school?
Please choose **only one** of the following:
- ○ About 1 km away
- ○ About 2 kms away
- ○ About 3 kms away
- ○ About 4 kms away
- ○ Over 5 kms away

4. How often are you absent from school?
Please choose **only one** of the following:
- ○ Never
- ○ Once a week
- ○ Twice a week
- ○ More than twice a week

5. Do you find time to do your studies at home?
Please choose **only one** of the following:
- ○ I don't find time
- ○ About 1 hour
- ○ About 2 hours
- ○ About 3 hours

6. How many times have you been punished since the beginning of the year 2013?
Please choose **only one** of the following:
- ○ I have never been punished
- ○ About twice
- ○ More than three times
- ○ Very often

7. Do you sometimes speak your local language with your fellow pupils in class?
Please choose only one of the following:
- ○ **Yes**
- ○ **Uncertain**
- ○ **No**

8. Education will help me to get a good job in future.

	Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

9. I need to work very hard to go to a national school.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

FIGURE H.1: Students' Questionnaire pg1

10. My parents encourage me to work hard in my studies.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select on option	○	○	○	○	○

11. My father and mother live harmoniously

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select on option	○	○	○	○	○

12. My parents are able to pay for my further education.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

13. My parent education levels are.

	Up to university	Up to Diploma	Up to secondary school	Up to primary	Did not go to school
Select one/two options	○	○	○	○	○

14. The total number of my family members is.

	Between 3 and 5	Between 6 and 10	More than 10
Select one option	○	○	○

15. My parents attend all meetings whenever they are called by the school.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

16. The community around supports building of classrooms, library, toilets etc.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

17. Our teachers encourage us to work hard.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

18. Our teachers are available and willing to assist us in our studies.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

19. Our teachers are never absent without a good reason.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select one option	○	○	○	○	○

20. Lack of learning facilities in school and at home is what makes pupils perform poorly.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select on option	○	○	○	○	○

21. Lack of enough teachers is what makes pupils perform poorly.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Select on option	○	○	○	○	○

22. Please state why you think pupils in Kwale County perform poorly in KCPE.

.....

.....

.....

.....

.....

.....

Thank you for completing this survey. We hope the results will help in changing the low academic performance trend in the County.

FIGURE H.2: Students' Questionnaire pg2

Appendix I

Prototype Evaluation Questionnaire

PROTOTYPE EVALUATION SCALE

	Strongly disagree	disagree	Neutral	agree	Strongly agree
1.Interest to use the tool frequently					
2.Tool is unnecessary complex					
3. Tool is easy to use.					
4.I would need assistance to use the tool					
5.Various functions of the tool are well integration					
6.There are inconsistencies in the tool					
7.People would learn to use the tool quickly					
8.Found the tool cumbersome/ awkward to use					
9.Felt confident to use the tool					
10.One needs to learn a lot of things before using the tool					

FIGURE I.1: Prototype Evaluation questionnaire

Bibliography

Abuya, B. a., Admassu, K., Ngware, M., Onsomu, E. O., and Oketch, M. (2015). Free Primary Education and Implementation in Kenya: The Role of Primary School Teachers in Addressing the Policy Gap. *SAGE Open*, 5(1).

Acuna, E. and Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications*, pages 639–647. Springer.

Agnihotri, L. and Ott, A. (2014). Building a student at-risk model: An end-to-end perspective from user to data scientist. In *Educational Data Mining 2014*.

Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., and Goodrich, V. (2014). Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, pages 103–112. ACM.

Aker, J. C. (2008). Does digital divide or provide? the impact of cell phones on grain markets in niger. *Center for Global Development Working Paper*, (154).

Akobeng, A. K. (2007). Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta paediatrica*, 96(5):644–647.

Alexander, K., Entwisle, D., and Kabbani, N. (2001). The dropout process in life course perspective: Early risk factors at home and school. *The Teachers College Record*, 103(5):760–822.

Archana, S. and Elangovan, K. (2014). Survey of classification techniques in data mining. *International Journal of Computer Science and Mobile Applications*, 2(2):65–71.

Asif, R., Merceron, A., and Pathan, M. K. (2014). Predicting Student Academic Performance at Degree Level: A Case Study. *International Journal of Intelligent Systems and Applications*, 7(1):49–61.

Bae, D. and Wickrama, K. (2014). Family socioeconomic status and academic achievement among korean adolescents linking mechanisms of family processes and adolescents' time use. *The Journal of Early Adolescence*, page 0272431614549627.

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7:112–118.

Baker, R. S. and Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning Analytics*, pages 61–75. Springer.

Baker, R. S., Lindrum, D., Lindrum, M. J., and Perkowski, D. (2015). Analyzing early at-risk factors in higher education e-learning courses. In *Proceedings of the 8th International Conference on Educational Data Mining*.

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1):3–17.

Balfanz, R., Herzog, L., and Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, 42(4):223–235.

Banfield, R. E., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. (2007). A comparison of decision tree ensemble creation techniques. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(1):173–180.

Barnett, W. S. and Belfield, C. R. (2006). Early childhood development and social mobility. *The future of children*, 16(2):73–98.

Basch, C. E. (2011). Teen pregnancy and the achievement gap among urban minority youth. *Journal of School Health*, 81(10):614–618.

Battin-Pearson, S., Newcomb, M. D., Abbott, R. D., Hill, K. G., Catalano, R. F., and Hawkins, J. D. (2000). Predictors of early high school dropout: A test of five theories. *Journal of educational psychology*, 92(3):568.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T., and Popelinsky, L. (2012). Predicting drop-out from social behaviour of students. *International Educational Data Mining Society*.

Beleche, T., Fairris, D., and Marks, M. (2012). Do course evaluations truly reflect student learning? evidence from an objectively graded post-test. *Economics of Education Review*, 31(5):709–719.

Benedek, J. and Miner, T. (2002). Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association*, 2003:8–12.

Bhardwaj, B. K. and Pal, S. (2012). Data mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.

Bhargava, N., Sharma, G., Bhargava, R., and Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6).

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3):483–519.

Bornstein, M. H. and Bradley, R. H. (2014). *Socioeconomic status, parenting, and child development*. Routledge.

Boticki, I., Baksa, J., Seow, P., and Looi, C.-K. (2015). Usage of a mobile social learning platform with virtual badges in a primary school. *Computers & Education*, 86:120–136.

Bowen, G. L., Hopson, L. M., Rose, R. A., and Glennie, E. J. (2012). Students' perceived parental school behavior expectations and their academic performance: A longitudinal analysis. *Family Relations*, 61(2):175–191.

Bowen, N. K., Wegmann, K. M., and Webber, K. C. (2013). Enhancing a brief writing intervention to combat stereotype threat among middle-school students. *Journal of Educational Psychology*, 105(2):427.

Braa, K. and Vidgen, R. (1999). Interpretation, intervention, and reduction in the organizational laboratory: a framework for in-context information system research. *Accounting, Management and Information Technologies*, 9(1):25–47.

Bratu, C. V., Muresan, T., and Potolea, R. (2008). Improving classification accuracy through feature selection. In *Intelligent Computer Communication and Processing, 2008. ICCP 2008. 4th International Conference on*, pages 25–32. IEEE.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brewer, E., Demmer, M., Du, B., Ho, M., Kam, M., Nedevschi, S., Pal, J., Patra, R., Surana, S., and Fall, K. (2005). The case for technology in developing regions. *Computer*, 38(6):25–38.

Brown, C. D. and Davis, H. T. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80(1):24–38.

Brundrett, M. (2011). The global challenge for primary schools: education in a world of 7 billion people. *Education 3-13*, 39(5):451–453.

Brundrett, M. (2014). Education for all: the challenges of achieving universal early childhood care and primary education. *Education 3-13*, 42(3):233–236.

Bruns, B. and Rakotomalala, R. (2003). *Achieving universal primary education by 2015: A chance for every child*, volume 1. World Bank Publications.

Bühlmann, P. (2012). Bagging, boosting and ensemble methods. In *Handbook of Computational Statistics*, pages 985–1022. Springer.

Buku, M. W. and Meredith, M. W. (2012). Safaricom and m-pesa in kenya: financial inclusion and financial integrity. *Wash. Jl tech. & arts*, 8:375.

Cao, L. J., Keerthi, S. S., Ong, C.-J., Zhang, J. Q., and Lee, H. P. (2006). Parallel sequential minimal optimization for the training of support vector machines. *Neural Networks, IEEE Transactions on*, 17(4):1039–1049.

Castro, F., Vellido, A., Nebot, À., and Mugica, F. (2007). Applying data mining techniques to e-learning problems. In *Evolution of teaching and learning paradigms in intelligent environment*, pages 183–221. Springer.

Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107.

Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

Chattopadhay, T. (2014). School as a site of student social capital: An exploratory study from brazil. *International Journal of Educational Development*, 34:67–76.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357.

Chen, C.-c. (2013). Joyce."opportunities and challenges of moocs: Perspectives from asia.". *IFLA WLIC 2013*.

Chen, W.-B. and Gregory, A. (2009). Parental involvement as a protective factor during the transition to high school. *The Journal of Educational Research*, 103(1):53–62.

Chepken, C. (2012). Telecommuting in the developing world: a case of the day-labour market.

Chung, H. L., Mulvey, E. P., and Steinberg, L. (2011). Understanding the school outcomes of juvenile offenders: An exploration of neighborhood influences and motivational resources. *Journal of youth and adolescence*, 40(8):1025–1038.

Chung, L. and do Prado Leite, J. C. S. (2009). On non-functional requirements in software engineering. In *Conceptual modeling: Foundations and applications*, pages 363–379. Springer.

Churchill, D. and Hedberg, J. (2008). Learning object design considerations for small-screen handheld devices. *Computers & Education*, 50(3):881–893.

Cichosz, P. (2015). Naïve bayes classifier. *Data Mining Algorithms: Explained Using R*, pages 118–133.

Cios, K. J., Teresinska, A., Konieczna, S., Potocka, J., and Sharma, S. (2000). A knowledge discovery approach to diagnosing myocardial perfusion. *Engineering in Medicine and Biology Magazine, IEEE*, 19(4):17–25.

Cohen, J. et al. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Collins, R. T., Liu, Y., and Leordeanu, M. (2005). Online selection of discriminative tracking features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1631–1643.

Collis, B. and Moonen, J. (2001). *Flexible learning in a digital world: Experiences and expectations*. Psychology Press.

Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology*, 104(1):32.

Corbett, A. T., Koedinger, K. R., and Anderson, J. R. (1997). Intelligent tutoring systems. *Handbook of humancomputer interaction*, pages 849–874.

Crossley, S., Danielle, S., Baker, R., Wang, Y., Paquette, L., Barnes, T., and Bergner, Y. (2015). Language to completion: Success in an educational data mining massive open online class.

Crowder, K. and South, S. J. (2003). Neighborhood distress and school dropout: The variable significance of community context. *Social Science Research*, 32(4):659–698.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792.

Dahl, G. B. and Lochner, L. (2012). The impact of family income on child achievement: Evidence from the earned income tax credit. *The American Economic Review*, 102(5):1927–1956.

Dal Pozzolo, A., Johnson, R., Caelen, O., Waterschoot, S., Chawla, N. V., and Bontempi, G. (2014). Using hddt to avoid instances propagation in unbalanced and evolving data streams. In *Neural Networks (IJCNN), 2014 International Joint Conference on*, pages 588–594. IEEE.

Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., and Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of learning disabilities*, 39(6):507–514.

De Stefano, C., Fontanella, F., Marrocco, C., and di Freca, A. S. (2014). A ga-based feature selection approach with an application to handwritten character recognition. *Pattern Recognition Letters*, 35:130–141.

Dekker, G., Pechenizkiy, M., and Vleeshouwers, J. (2009). Predicting students' drop out: a case study. In *Educational Data Mining 2009*.

Doan, T.-A., Zhang, J., TJHI, W. C., and LEE, B. S. (2011). Analyzing students' usage of e-learning systems in the cloud for course management. In *Proceedings of the 19th International Conference on Computers in Education (ICCE'11)*, page 297.

Domínguez-Almendros, S., Benítez-Parejo, N., and Gonzalez-Ramirez, A. (2011). Logistic regression models. *Allergologia et immunopathologia*, 39(5):295–305.

Dong, S., Fabian, E., and Luecking, R. G. (2015). Impacts of school structural factors and student factors on employment outcomes for youth with disabilities in transition a secondary data analysis. *Rehabilitation Counseling Bulletin*, page 0034355215595515.

Donner, J. (2008). Research approaches to mobile use in the developing world: A review of the literature. *The information society*, 24(3):140–159.

Donner, J. (2009). Mobile-based livelihood services in africa: pilots and early deployments. *Communication technologies in Latin America and Africa: A multidisciplinary perspective*, pages 37–58.

Donner, J. and Toyama, K. (2009). Persistent themes in ict4d research: priorities for inter-methodological exchange. *57th Session of the International Statistics Institute, Durban, South Africa*, pages 17–21.

Dotenco, S., Götzelmann, T., and Gallwitz, F. (2014). Smartphone input using its integrated projector and built-in camera. In *Human-Computer Interaction. Applications and Services*, pages 124–133. Springer.

Draper, N. R. and Smith, H. (2014). *Applied regression analysis*. John Wiley & Sons.

Du, K.-L. and Swamy, M. (2014). Perceptrons. In *Neural Networks and Statistical Learning*, pages 67–81. Springer.

Duflo, E., Dupas, P., and Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from kenyan primary schools. *Journal of Public Economics*, 123:92–110.

Duflo, E., Hanna, R., and Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *The American Economic Review*, pages 1241–1278.

Duncombe, R. (2012). Understanding mobile phone impact on livelihoods in developing countries: A new research framework.

Durrant, B., FRANK, E., HUNT, L., HOLMES, G., MAYO, M., PFAHRINGER, B., SMITH, T., and WITTEN, I. (2014). Weka 3: Data mining software in java. *Machine Learning Group at the University of Waikato*.

Easterly, W. (2009). How the millennium development goals are unfair to africa. *World Development*, 37(1):26–35.

Elmadani, M., Mitrovic, A., Weerasinghe, A., and Neshatian, K. (2015). Investigating student interactions with tutorial dialogues in eer-tutor. *Research and Practice in Technology Enhanced Learning*, 10(1):1–21.

Erturk, E. (2012). A case study in open source software security and privacy: Android adware. In *Internet Security (WorldCIS), 2012 World Congress on*, pages 189–191. IEEE.

Espejo, P. G., Ventura, S., and Herrera, F. (2010). A survey on the application of genetic programming to classification. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(2):121–144.

Etsey, K. (2005). Causes of low academic performance of primary school pupils in the shama sub-metro of shama ahanta east metropolitan assembly (saema) in ghana. In *Regional Conference on Education in West Africa Dakar, Senegal*, pages 1–34.

Evjemo, B., Akselsen, S., Slettemeås, D., Munch-Ellingsen, A., Andersen, A., and Karlsen, R. (2014). I expect smart services!": User feedback on nfc based services addressing everyday routines. *Mobility and Smart Cities, Mobility IoT*.

Fall, A.-M. and Roberts, G. (2012). High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. *Journal of adolescence*, 35(4):787–798.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

Feldmann, V. (2003). Mobile overtakes fixed: Implications for policy and regulation. *Geneva: ITU*.

Feng, P.-M., Ding, H., Chen, W., and Lin, H. (2013). Naive bayes classifier with feature selection to identify phage virion proteins. *Computational and mathematical methods in medicine*, 2013.

Finn, J. D., Gerber, S. B., and Boyd-Zaharias, J. (2005). Small classes in the early grades, academic achievement, and graduating from high school. *Journal of Educational Psychology*, 97(2):214.

Finn, J. D. and Zimmer, K. S. (2012). Student engagement: What is it? why does it matter? In *Handbook of research on student engagement*, pages 97–131. Springer.

Ford, M. and Batchelor, J. (2007). From zero to hero–is the mobile phone a viable learning tool for africa?

Gakure, R. W., Mukuria, P., and Kithae, P. P. (2013). An evaluation of factors that affect performance of primary schools in kenya: A case study of gatanga district. *African Journal of Agricultural Science*, 1(2):26–35.

Gakuru, M., Winters, K., and Stepman, F. (2009). Innovative farmer advisory services using ict. *documento presentado en el taller de W3C "Africa perspective on the role of movile technologies in fostering social development", Maputo*, 1.

Garcia, S., Derrac, J., Cano, J. R., and Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):417–435.

García, S., Luengo, J., and Herrera, F. (2015). Dealing with missing values. In *Data Preprocessing in Data Mining*, pages 59–105. Springer.

Gitau, S. W. (2013). Designing ummeli a case for mediated design, a participatory approach to designing interactive systems for semi-literate users.

Glinz, M. (2007). On non-functional requirements. In *Requirements Engineering Conference, 2007. RE'07. 15th IEEE International*, pages 21–26. IEEE.

Golding, P. and Donaldson, O. (2006). Predicting academic performance. In *Frontiers in education conference, 36th Annual*, pages 21–26. IEEE.

Goodman, D. (2005). Linking mobile phone ownership and use to social capital in rural south africa and tanzania. *INTERMEDIA-LONDON-*, 33(4):26.

Gorey, K. M. (2001). Early childhood education: A meta-analytic affirmation of the short-and long-term benefits of educational opportunity. *School Psychology Quarterly*, 16(1):9.

Goundar, S. (2011). What is the potential impact of using mobile devices in education. In *Proceedings of SIG GlobDev Fourth Annual Workshop*.

Graesser, A. C., Conley, M. W., and Olney, A. (2012). Intelligent tutoring systems.

Gray, G., McGuinness, C., and Owende, P. (2013). An investigation of psychometric measures for modelling academic performance in tertiary education. In *Educational Data Mining 2013*.

Greer, J. and Mark, M. (2015). Evaluation methods for intelligent tutoring systems revisited. *International Journal of Artificial Intelligence in Education*, pages 1–6.

Greven, A., Keller, G., and Warnecke, G. (2014). *Entropy*. Princeton university press.

Grönlund, Å., Andersson, A., and Hatakka, M. (2008). Mobile technologies for development–a comparative study on challenges. In *Sig GlobDev Workshop, pre-conference ICIS*. Citeseer.

Guruler, H., Istanbullu, A., and Karahasan, M. (2010). A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education*, 55(1):247–254.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato.

Hall, M. A. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 15(6):1437–1447.

Hämäläinen, W. and Vinni, M. (2006). Comparison of machine learning methods for intelligent tutoring systems. In *Intelligent Tutoring Systems*, pages 525–534. Springer.

Han, J., Kamber, M., and Pei, J. (2011). *Data mining: concepts and techniques: concepts and techniques*. Elsevier.

Han, Y. and Xia, K. (2014). Data preprocessing method based on user characteristic of interests for web log mining. In *Instrumentation and Measurement, Computer, Communication and Control (IMCCC), 2014 Fourth International Conference on*, pages 867–872. IEEE.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Hanna, M. (2004). Data mining in the e-learning domain. *Campus-wide information systems*, 21(1):29–34.

Harb, H. M. and Moustafa, M. A. (2012). Selecting optimal subset of features for student performance model. *Int J Comput Sci*, (9):253–262.

Hardré, P. L., Sullivan, D. W., and Crowson, H. M. (2009). Student characteristics and motivation in rural high schools. *Journal of Research in Rural Education (Online)*, 24(16):1.

Harichandan, S. (2009). Role of mobile technology in learning and teaching. In *Collected Conference Papers and Abstracts September 2009*, page 209.

Harris, D. N. and Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of public economics*, 95(7):798–812.

Hashemi, M., Azizinezhad, M., Najafi, V., and Nesari, A. J. (2011). What is mobile learning? challenges and capabilities. *Procedia-Social and Behavioral Sciences*, 30:2477–2481.

Hassanien, A. E., Tolba, M., and Azar, A. T. (2014). *Advanced Machine Learning Technologies and Applications: Second International Conference, AMLTA 2014, Cairo, Egypt, November 28-30, 2014. Proceedings*, volume 488. Springer.

Heck, R. H. and Mahoe, R. (2006). Student transition to high school and persistence: Highlighting the influences of social divisions and school contingencies. *American Journal of Education*, 112(3):418–446.

Hempel, S., Shetty, K. D., Shekelle, P. G., Rubenstein, L. V., Danz, M. S., Johnsen, B., Dalal, S. R., et al. (2012). Machine learning confusion matrix, text terms distinguishing relevant and irrelevant citations, and reviewer disagreements.

Hill, C., Corbett, C., and St Rose, A. (2010). *Why so few? Women in Science, Technology, Engineering, and Mathematics*. ERIC.

Hoy, W. K., Tarter, C. J., and Hoy, A. W. (2006). Academic optimism of schools: A force for student achievement. *American educational research journal*, 43(3):425–446.

Huang, S. H. (2015). Supervised feature selection: A tutorial. *Artificial Intelligence Research*, 4(2):p22.

Huang, X., Shi, L., and Suykens, J. A. (2015). Sequential minimal optimization for svm with pinball loss. *Neurocomputing*, 149:1596–1603.

Hughes, N. and Lonie, S. (2007). M-pesa: mobile money for the "unbanked" turning cellphones into 24-hour tellers in kenya. *Innovations*, 2(1-2):63–81.

Ibrahim, Z. and Rusli, D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In *21st Annual SAS Malaysia Forum, 5th September*.

İşcan, T. B., Rosenblum, D., and Tinker, K. (2015). School fees and access to primary education: Assessing four decades of policy in sub-saharan africa. *Journal of African Economies*, page ejv007.

Islam, M. S. and Grönlund, Å. (2011). Bangladesh calling: farmers' technology use practices as a driver for development. *Information Technology for Development*, 17(2):95–111.

Jacob, B. A. (2001). Getting tough? the impact of high school graduation exams. *Educational evaluation and policy analysis*, 23(2):99–121.

Jain, A. and Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Jantawan, B. and Tsai, C.-F. (2014). A comparison of filter and wrapper approaches with data mining techniques for categorical variables selection. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(6):4501–4508.

Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the south indian fisheries sector. *The quarterly journal of economics*, pages 879–924.

Jiang, H., Bai, J., Zhang, S., and Xu, B. (2005). Svm-based audio scene classification. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, pages 131–136. IEEE.

Jiang, S., Williams, A., Schenke, K., Warschauer, M., and O'dowd, D. (2014). Predicting mooc performance with week 1 behavior. In *Educational Data Mining 2014*.

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (auc) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, 21(4):498–507.

Johnson, S. D., Aragon, S. R., Shaik, N., and Palma-Rivas, N. (2000). Comparative analysis of learner satisfaction and learning outcomes in online and face-to-face learning environments. *Journal of interactive learning research*, 11(1):29.

Jurečková, J. and Picek, J. (2005). Two-step regression quantiles. *Sankhyā: The Indian Journal of Statistics*, pages 227–252.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, 13(1):61–72.

Kafyulilo, A. (2014). Access, use and perceptions of teachers and students towards mobile phones as a tool for teaching and learning in tanzania. *Education and Information Technologies*, 19(1):115–127.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Kalz, M., Lenssen, N., Felzen, M., Rossaint, R., Tabuenca, B., Specht, M., and Skorning, M. (2014). Smartphone apps for cardiopulmonary resuscitation training and real incident support: a mixed-methods evaluation study. *Journal of medical Internet research*, 16(3).

Kandaswamy, K. K., Chou, K.-C., Martinetz, T., Möller, S., Suganthan, P., Sridharan, S., and Pugalenthi, G. (2011). Afp-pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *Journal of Theoretical Biology*, 270(1):56–62.

Karbach, J., Gottschling, J., Spengler, M., Hegewald, K., and Spinath, F. M. (2013). Parental involvement and general cognitive ability as predictors of domain-specific academic achievement in early adolescence. *Learning and Instruction*, 23:43–51.

Karegowda, A. G., Manjunath, A., and Jayaram, M. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277.

Kaur, G. and Chhabra, A. (2014). Improved j48 classification algorithm for the prediction of diabetes. *International Journal of Computer Applications*, 98(22).

Kelly, K., Arroyo, I., and Heffernan, N. (2013). Using its generated data to predict standardized test scores. In *Educational Data Mining 2013*.

Kimosop, P. K., Otiso, K. M., and Ye, X. (2015). Spatial and gender inequality in the kenya certificate of primary education examination results. *Applied Geography*, 62:44–61.

King'oina, O. J. (2011). *Thye role of quality assurance and standards officers in enhancing primary school teachers'effectiveness in Marani Division, Marani District Kenya.* PhD thesis, KENYATTA UNIVERSITY.

Kira, K. and Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134.

Koedinger, K., Cunningham, K., Skogsholm, A., and Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. In *Educational Data Mining 2008*.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324.

Kononenko, I., Šimec, E., and Robnik-Šikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, 7(1):39–55.

Koprinska, I., Stretton, J., and Yacef, K. (2015). Students at risk: Detection and remediation.

Kotsiantis, S., Pierrakeas, C., and Pintelas, P. (2002). Efficiency of machine learning techniques in predicting students' performance in distance learning systems. *Educational Software Development Laboratory Department of Mathematics, University of Patras, Greece*.

Kotsiantis, S. B., Pierrakeas, C., and Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 267–274. Springer.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Krishnaiah, V., Narsimha, G., and Chandra, N. S. (2014). Survey of classification techniques in data mining.

Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., and Held, P. (2013). Multi-layer perceptrons. In *Computational Intelligence*, pages 47–81. Springer.

Kumar, G. and Kumar, K. (2011). A novel evaluation function for feature selection based upon information theory. In *Electrical and Computer Engineering (CCECE), 2011 24th Canadian Conference on*, pages 000395–000399. IEEE.

Kumar, M., Kamal, K. K., Varyani, B., and Barwad, S. (2015). Paradigm shift in mobile communication carriers. In *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, pages 1753–1756. IEEE.

Kupzyk, K. A. and Cohen, M. Z. (2014). Data validation and other strategies for data entry. *Western journal of nursing research*, page 0193945914532550.

Kurgan, L. A. and Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(01):1–24.

Laird, J., Kienzl, G., DeBell, M., and Chapman, C. (2007). Dropout rates in the united states: 2005. compendium report. nces 2007-059. *National Center for Education Statistics*.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., De Schaetzen, V., Duque, R., Bersini, H., and Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4):1106–1119.

Lee, M. S. and Moore, A. (2014). Efficient algorithms for minimizing cross validation error. In *Machine Learning Proceedings 1994: Proceedings of the Eighth International Conference*, page 190. Morgan Kaufmann.

Lee, S.-I., Lee, H., Abbeel, P., and Ng, A. Y. (2006). Efficient l˜ 1 regularized logistic regression. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 401. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Leung, K. M. (2007). Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*.

Leventhal, T. and Brooks-Gunn, J. (2000). The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. *Psychological bulletin*, 126(2):309.

Levy, P. S. and Lemeshow, S. (2013). *Sampling of populations: methods and applications*. John Wiley & Sons.

Lewin, T. (2013). Universities abroad join partnerships on the web. *The New York Times*, 21.

Lewis, R. J. (2000). An introduction to classification and regression tree (cart) analysis. In *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*, pages 1–14.

Liao, J. and Chin, K.-V. (2007). Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945–1951.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502.

Lucas, A. M. and Mbiti, I. M. (2012). Access, sorting, and achievement: The short-run effects of free primary education in kenya. *American Economic Journal: Applied Economics*, 4(4):226–253.

Lucas, A. M. and Mbiti, I. M. (2014). Effects of school quality on student achievement: Discontinuity evidence from kenya. *American Economic Journal: Applied Economics*, 6(3):234–263.

Luo, L., Koprinska, I., and Liu, W. (2015). Discrimination-aware classifiers for student performance prediction.

Ma, Y., Liu, B., Wong, C. K., Yu, P. S., and Lee, S. M. (2000). Targeting the right students using data mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–464. ACM.

Maas, A. L., Le, Q. V., O'Neil, T. M., Vinyals, O., Nguyen, P., and Ng, A. Y. (2012). Recurrent neural networks for noise reduction in robust asr. In *INTERSPEECH*, pages 22–25.

Macfadyen, L. P. and Dawson, S. (2010). Mining lms data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, 54(2):588–599.

MacKenzie, G. and Peng, D. (2014). *Introduction*. Springer.

Mackinnon, A. (2000). A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. *Computers in biology and medicine*, 30(3):127–134.

Malviya, R. and Umrao, B. K. (2014). Comparison of nbtree and vfi machine learning algorithms for network intrusion detection using feature selection. *International Journal of Computer Applications*, 108(2).

Mangels, J. A., Good, C., Whiteman, R. C., Maniscalco, B., and Dweck, C. S. (2012). Emotion blocks the path to learning under stereotype threat. *Social cognitive and affective neuroscience*, 7(2):230–241.

Márquez-Vera, C., Cano, A., Romero, C., and Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, 38(3):315–330.

Marquez-Vera, C., Romero, C., and Ventura, S. (2010). Predicting school failure using data mining. In *Educational Data Mining 2011*.

Marsden, G., Maunder, A., and Parker, M. (2008). People are people, but technology is not technology. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 366(1881):3795–3804.

Martin, M. A. (2012). Family structure and the intergenerational transmission of educational advantage. *Social science research*, 41(1):33–47.

Masino, S. and Niño-Zarazúa, M. (2015). What works to improve the quality of student learning in developing countries?

Masters, K. (2005). Low-key m-learning: a realistic introduction of m-learning to developing countries. In *Sixth Conference on Communications in the 21st Century: Seeing, Understanding, Learning in the Mobile Age, Budapest*.

Matyila, P., Botha, A., Alberts, R., and Sibiya, G. (2013). The design of accessible and usable mobile services for low literate users. In *2013 International Conference on Adaptive Science and Technology*, pages 1–6. IEEE.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., and Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, 21(2):427–436.

Mbogo, C. C. (2015). *Scaffolding Java programming on a mobile phone for novice learners*. PhD thesis, Department of Computer Science, Faculty of Science, University of Cape Town.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Mehana, M. and Reynolds, A. J. (2004). School mobility and achievement: A meta-analysis. *Children and Youth Services Review*, 26(1):93–119.

Merceron, A. and Yacef, K. (2005). Educational data mining: a case study. In *AIED*, pages 467–474.

Mgala, M. and Mbogho, A. (2015). Data-driven intervention-level prediction modeling for academic performance. In *ICTD*, page 2.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., and Punch, W. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. In *Frontiers in education, 2003. FIE 2003 33rd annual*, volume 1, pages T2A–13. IEEE.

Monahan, K. C., Lee, J. M., and Steinberg, L. (2011). Revisiting the impact of part-time work on adolescent adjustment: Distinguishing between selection and socialization using propensity score matching. *Child development*, 82(1):96–112.

Monk, D. H. (2007). Recruiting and retaining high-quality teachers in rural areas. *The Future of Children*, 17(1):155–174.

Muhammad, A., Ahamd, F., and Shah, A. (2015). Resolving ethical dilemma in technology enhanced education through smart mobile devices. *International Arab Journal of e-Technology*, 4(1):25–31.

Munshi, J. (2014). A method for constructing likert scales. *Available at SSRN 2419366*.

Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*.

Nakhanu, S. B. (2012). Effect of syllabus coverage on secondary school students' performance in mathematics in kenya. *International Journal of Education Science*, 4(1):31–34.

Neupane, B., McDonald, S. D., and Beyene, J. (2015). Identifying determinants and estimating the risk of inadequate and excess gestational weight gain using a multinomial logistic regression model. *Open Access Medical Statistics*, 5.

Ng, A. (2011a). Cs229 lecture notes on machine learning. Technical report, Technical report, Stanford University, Department of Computer Science, Stanford, USA, 2011. 39, 116, 140.

Ng, A. (2011b). Sparse autoencoder. *CS294A Lecture notes*, 72:1–19.

Ng, A. (2012). Cs 229 lecture notes: Support vector machines. *online] cs229. stanford. edu/notes*.

Nsofor, C., Bello, A., Umeh, A. E., Oboh, C., et al. (2015). The future of educational technology in the 21st century nigeria: Changing educational landscape through emerging technologies. *JOURNAL OF EDUCATIONAL POLICY AND ENTREPRENEURIAL RESEARCH*, 2(3):28–37.

OḂannon, B. W. and Thomas, K. M. (2015). Mobile phones in the classroom: Preservice teachers answer the call. *Computers & Education*, 85:110–122.

Ojanen, E., Ronimus, M., Ahonen, T., Chansa-Kabali, T., February, P., Jere-Folotiya, J., Kauppinen, K.-P., Ketonen, R., Pitkänen, M., Ngorosho, D., et al. (2015). Graphogame-a catalyst for multi-level promotion of literacy in diverse contexts. *Frontiers in Psychology*, 6:671.

Okaya, T. M., Horne, M., Laming, M., and Smith, K. H. (2013). Measuring inviting school climate: A case study of a public primary school in an urban low socioeconomic setting in kenya. *Journal of Invitational Theory and Practice*, 19:15.

Okazaki, S. and Mendez, F. (2013). Perceived ubiquity in mobile services. *Journal of Interactive Marketing*, 27(2):98–111.

Oliver, J. J. and Hand, D. J. (2014). On pruning and averaging decision trees. In *Machine Learning: Proceedings of the Twelfth International Conference*, pages 430–437.

Orodho, J. A., Waweru, P. N., Ndichu, M., and Nthinguri, R. (2014). Home-based challenges to effective implementation of curriculum under free primary education system in nomadic kenya. *Journal of Education and Practice*, 5(26):134–144.

Ou, S.-R., Mersky, J. P., Reynolds, A. J., and Kohler, K. M. (2007). Alterable predictors of educational attainment, income, and crime: Findings from an inner-city cohort. *Social Service Review*, 81(1):85–128.

Pachler, N., Bachmair, B., and Cook, J. (2009). *Mobile learning: structures, agency, practices*. Springer Science & Business Media.

Paech, B. and Kerkow, D. (2004). Non-functional requirements engineering-quality is essential. In *10th International Workshop on Requirments Engineering Foundation for Software Quality*.

Pagano, D. and Brügge, B. (2013). User involvement in software evolution practice: a case study. In *Proceedings of the 2013 international conference on Software engineering*, pages 953–962. IEEE Press.

Page, T. (2014). Application-based mobile devices in design education. *International Journal of Mobile Learning and Organisation*, 8(2):96–111.

Pampaka, M. (2011). Receiver operating characteristic. *The SAGE Dictionary of Quantitative Management Research*, page 267.

Panchal, G., Ganatra, A., Kosta, Y., and Panchal, D. (2011). Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. *International Journal of Computer Theory and Engineering*, 3(2):332–337.

Panda, M., Abraham, A., and Patra, M. R. (2010). Discriminative multinomial naive bayes for network intrusion detection. In *Information Assurance and Security (IAS), 2010 Sixth International Conference on*, pages 5–10. IEEE.

Pardos, Z. A., Wang, Q. Y., and Trivedi, S. (2012). The real world significance of performance prediction. *International Educational Data Mining Society*.

Patil, T. R. and Sherekar, S. (2013). Performance analysis of naive bayes and j48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, 6(2):256–261.

Patton, D. U., Woolley, M. E., and Hong, J. S. (2012). Exposure to violence, student fear, and low academic achievement: African american males in the critical transition to high school. *Children and Youth Services Review*, 34(2):388–395.

Pears, K. C., Kim, H. K., and Fisher, P. A. (2008). Psychosocial and cognitive functioning of children with specific profiles of maltreatment. *Child abuse & neglect*, 32(10):958–971.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Pedro, M. O., Baker, R., Bowers, A., and Heffernan, N. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Educational Data Mining 2013*.

Pedro, M. O., Ocumpaugh, J., Baker, R., and Heffernan, N. (2014). Predicting stem and non-stem college major enrollment from middle school interaction with mathematics educational software. In *Educational Data Mining 2014*.

Perreira, K. M., Harris, K. M., and Lee, D. (2006). Making it in america: High school completion by immigrant and native youth. *Demography*, 43(3):511–536.

Peteiro-Barral, D., Bolón-Canedo, V., Alonso-Betanzos, A., Guijarro-Berdiñas, B., and Sánchez-Maroño, N. (2013). Toward the scalability of neural networks through feature selection. *Expert Systems with Applications*, 40(8):2807–2816.

Philip, T. M. and Garcia, A. (2015). Schooling mobile phones assumptions about proximal benefits, the challenges of shifting meanings, and the politics of teaching. *Educational Policy*, 29(4):676–707.

Platt, J. et al. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods—support vector learning*, 3.

Platt, J. C. (1999). 12 fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, pages 185–208.

Pomerantz, E. M., Moorman, E. A., and Litwack, S. D. (2007). The how, whom, and why of parents' involvement in children's academic lives: More is not always better. *Review of educational research*, 77(3):373–410.

Pong, S.-L. and Ju, D.-B. (2000). The effects of change in family structure and income on dropping out of middle and high school. *Journal of Family Issues*, 21(2):147–169.

Porteous, D. (2011). The enabling environment for mobile banking in africa, bankablefrontier.

Poverty, E. (2015). Millennium development goals. *United Nations. Available online: http://www. un. org/millenniumgoals/(accessed on 23 August 2011)*.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Preece, J., Sharp, H., and Rogers, Y. (2015). *Interaction Design-beyond human-computer interaction*. John Wiley & Sons.

Pudil, P., Novovičová, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125.

Raghavan, N. S. (2005). Data mining in e-commerce: A survey. *Sadhana*, 30(2-3):275–289.

Rahman, F. and Devanbu, P. (2013). How, and why, process metrics are better. In *Proceedings of the 2013 International Conference on Software Engineering*, pages 432–441. IEEE Press.

Raisamo, R. (2014). Investigating students' behavioural intention to adopt and use mobile learning in higher education in east africa joel s. mtebe university of dar es salaam, tanzania. *International Journal of Education and Development using Information and Communication Technology*, 10(3):4–20.

Ramaswami, M. and Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. *arXiv preprint arXiv:0912.3924*.

Rätsch, G., Onoda, T., and Müller, K.-R. (2001). Soft margins for adaboost. *Machine learning*, 42(3):287–320.

Ream, R. K. (2005). Toward understanding how social capital mediates the impact of mobility on mexican american achievement. *Social forces*, 84(1):201–224.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems*, pages 532–538. Springer.

Reschly, A. L. and Christenson, S. L. (2006). Prediction of dropout among students with mild disabilities a case for the inclusion of student engagement variables. *Remedial and Special Education*, 27(5):276–292.

Ribeiro, G., Duivesteijn, W., Soares, C., and Knobbe, A. (2012). Multilayer perceptron for label ranking. In *Artificial Neural Networks and Machine Learning–ICANN 2012*, pages 25–32. Springer.

Robnik-Šikonja, M. and Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression. In *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)*, pages 296–304.

Roebuck, M. C., French, M. T., and Dennis, M. L. (2004). Adolescent marijuana use and school attendance. *Economics of Education Review*, 23(2):133–141.

Rogers, A. (2014). Pisa, power and policy: the emergence of global educational governance. *International Review of Education*, 60(4):591–596.

Romero, C., López, M.-I., Luna, J.-M., and Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68:458–472.

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618.

Romero, C., Ventura, S., and De Bra, P. (2004). Knowledge discovery with genetic programming for providing feedback to courseware authors. *User Modeling and User-Adapted Interaction*, 14(5):425–464.

Romero, C., Ventura, S., Espejo, P. G., and Hervás, C. (2008). Data mining algorithms to classify students. In *Educational Data Mining 2008*.

Romero, C., Ventura, S., Pechenizkiy, M., and Baker, R. S. (2010). *Handbook of educational data mining*. CRC Press.

Ron, K. and Foster, P. (1998). Special issue on applications of machine learning and the knowledge discovery process. *Journal of Machine Learning*, 30:271–274.

Roscigno, V. J. and Crowle, M. L. (2001). Rurality, institutional disadvantage, and achievement/attainment*. *Rural Sociology*, 66(2):268–292.

Rotberg, R. I. and Aker, J. C. (2013). Mobile phones: uplifting weak and failed states. *The Washington Quarterly*, 36(1):111–125.

Rumberger, R. and Lim, S. A. (2008). Why students drop out of school: A review of 25 years of research.

Rumberger, R. W. and Palardy, G. J. (2005). Test scores, dropout rates, and transfer rates as alternative indicators of high school performance. *American educational research journal*, 42(1):3–42.

Rumberger, R. W. and Thomas, S. L. (2000). The distribution of dropout and turnover rates among urban and suburban high schools. *Sociology of Education*, pages 39–67.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517.

Samuel, J., Shah, N., and Hadingham, W. (2005). Mobile communications in south africa, tanzania, and egypt: Results from community and business surveys. *Africa: the impact of mobile phones*, 2:44–52.

Sarlis, N. V. and Christopoulos, S.-R. G. (2014). Visualization of the significance of receiver operating characteristics based on confidence ellipses. *Computer Physics Communications*, 185(3):1172–1176.

Scharf, M. (2013). Children's social competence within close friendship: The role of self-perception and attachment orientations. *School Psychology International*, page 0143034312474377.

Scheuer, O. and McLaren, B. M. (2012). Educational data mining. In *Encyclopedia of the Sciences of Learning*, pages 1075–1079. Springer.

Şen, B., Uçar, E., and Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications*, 39(10):9468–9476.

Sezgin, F., Koşar, S., Kılınç, A. Ç., and Öğdem, Z. (2014). Teacher absenteeism in turkish primary schools: A qualitative perspective from school principals. *International Online Journal of Educational Sciences*, 6(3).

Shaikh, A., MAHOTO, N., KHUHAWAR, F., and MEMON, M. (2015). Performance evaluation of classification methods for heart disease dataset. *Sindh University Research Journal-SURJ (Science Series)*, 47(3).

Shanks, D. R. and John, M. F. S. (1994). Characteristics of dissociable human learning systems. *Behavioral and brain sciences*, 17(03):367–395.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., and Murthy, K. R. K. (2000). Improvements to the smo algorithm for svm regression. *Neural Networks, IEEE Transactions on*, 11(5):1188–1193.

Shudong, W. and Higgins, M. (2005). Limitations of mobile phone learning. In *Wireless and Mobile Technologies in Education, 2005. WMTE 2005. IEEE International Workshop on*, pages 3–pp. IEEE.

Siemens, G. and Baker, R. S. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254. ACM.

Sife, A. S., Kiondo, E., and Lyimo-Macha, J. G. (2010). Contribution of mobile phones to rural livelihoods and poverty reduction in morogoro region, tanzania. *The Electronic Journal of Information Systems in Developing Countries*, 42.

Sifuna, D. N. (1992). Prevocational subjects in primary schools in the 8-4-4 education system in kenya. *International Journal of Educational Development*, 12(2):133–145.

Silver, D., Saunders, M., and Zarate, E. (2008). What factors predict high school graduation in the los angeles unified school district. *Policy Brief*, 14.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, 75(3):417–453.

Sisungo, Z. W. (2012). *A study of the role of the District Education Officers in the management and supervision of primary school education programmes in the three districts of Western province in Kenya*. PhD thesis.

Site, S. and Mishra, D. S. K. (2013). A review of ensemble technique for improving majority voting for classifier. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(1).

Sivaraj, R. and Ravichandran, T. (2011). A review of selection methods in genetic algorithm. *International journal of engineering science and technology*, 3(5).

So, H.-J., Seow, P., and Looi, C. K. (2009). Location matters: Leveraging knowledge building with mobile devices and web 2.0 technology. *Interactive Learning Environments*, 17(4):367–382.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Somerset, A. (2010). *Consortium for Research on Educational Access , Transitions and Equity Free Primary Education and After in Kenya : Enrolment impact , quality effects , and the transition to secondary school Moses Oketch*. Number 37.

Srithar, U. and Selvaraj, D. (2015). Learning at your own pace: M-learning solution for school students. *International Journal of Information and Electronics Engineering*, 5(3):216.

Ssembatya, R. (2014). Designing an architecture for secure sharing of personal health records: a case of developing countries.

Stearns, E., Moller, S., Blau, J., and Potochnick, S. (2007). Staying back and dropping out: The relationship between grade retention and school dropout. *Sociology of Education*, 80(3):210–240.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., and Abreu, R. (2015). A comparative study of classification and regression algorithms for modelling students' academic performance.

Suganya, P. and Sumathi, C. (2014). Classifier rules in data mining—a survey. In *Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on*, pages 1–3. IEEE.

Sullivan, K., Perry, L. B., and McConney, A. (2013). How do school resources and academic performance differ across australia's rural, regional and metropolitan communities? *The Australian Educational Researcher*, 40(3):353–372.

Suryadarma, D., Suryahadi, A., Sumarto, S., and Rogers, F. H. (2006). Improving student performance in public primary schools in developing countries: Evidence from indonesia. *Education Economics*, 14(4):401–429.

Swaffield, S., Jull, S., and Ampah-Mensah, A. (2013). Using mobile phone texting to support the capacity of school leaders in ghana to practise leadership for learning. *Procedia-Social and Behavioral Sciences*, 103:1295–1302.

Sweeten, G. (2006). Who will graduate? disruption of high school education by arrest and court involvement. *Justice Quarterly*, 23(4):462–480.

Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M., and Appleton, J. (2014). Predicting student risks through longitudinal analysis. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1544–1552. ACM.

Tan, M. and Shao, P. (2015). Prediction of student dropout in e-learning program through the use of machine learning method. *International Journal of Emerging Technologies in Learning (iJET)*, 10(1):pp–11.

Tang, J., Alelyani, S., and Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications. Editor: Charu Aggarwal, CRC Press In Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.*

Tang, Y. (2013). Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*.

Tao, S. (2013). Why are teachers absent? utilising the capability approach and critical realism to explain teacher performance in tanzania. *International Journal of Educational Development*, 33(1):2–14.

Thai-Nghe, N., Busche, A., and Schmidt-Thieme, L. (2009). Improving academic performance prediction by dealing with class imbalance. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, pages 878–883. IEEE.

Thao, T. T. P. and Nam, N. D. (2014). A model for using mobile phones in teaching and learning mathematics. In *Proceedings of the 7th International Conference on Educational Reform.*

Tong, S. and Koller, D. (2002). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

Traxler, J. and Kukulska-Julme, A. (2005). Mobile learning in developing countries.

Traxler, J. and Leach, J. (2006). Innovative and sustainable mobile learning in africa. In *Wireless, Mobile and Ubiquitous Technology in Education, 2006. WMUTE'06. Fourth IEEE International Workshop on*, pages 98–102. IEEE.

Turner, A. and Thompson, A. (2015). School mobility and educational outcomes of off-reserve first nations students.

UN (2015). Transforming our world: the 2030 agenda for sustainable development. *see, related, The Global Initiative Against Transnational Organized Crime,"Organized Crime: A Cross-Cutting Threat to Sustainable Development"(Geneva: January 2015)*.

Unesco (2001). *EFA global monitoring report: Education for All*. Unesco.

Van de Grift, W. and Houtveen, A. (2006). Underperformance in primary schools. *School Effectiveness and School Improvement*, 17(3):255–273.

van der Weegen, S., Verwey, R., Tange, H. J., Spreeuwenberg, M. D., and de Witte, L. P. (2014). Usability testing of a monitoring and feedback tool to stimulate physical activity. *Patient preference and adherence*, 8:311.

Vandamme, J.-P., Meskens, N., and Superby, J.-F. (2007). Predicting academic performance by data mining methods. *Education Economics*, 15(4):405–419.

Venable, J. R. (2010). Design science research post hevner et al.: criteria, standards, guidelines, and expectations. In *International Conference on Design Science Research in Information Systems*, pages 109–123. Springer.

Verweij, K. J., Huizink, A. C., Agrawal, A., Martin, N. G., and Lynskey, M. T. (2013). Is the relationship between early-onset cannabis use and educational attainment causal or due to common liability? *Drug and alcohol dependence*, 133(2):580–586.

Vieira, S. M., Kaymak, U., and Sousa, J. (2010). Cohen's kappa coefficient as a performance measure for feature selection. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pages 1–8. IEEE.

Vihavainen, A., Luukkainen, M., and Kurhila, J. (2013). Using students' programming behavior to predict success in an introductory mathematics course. In *Educational Data Mining 2013*.

Von Wachter, T., Song, J., and Manchester, J. (2011). Trends in employment and earnings of allowed and rejected applicants to the social security disability insurance program. *The American Economic Review*, 101(7):3308–3329.

Wade, R. H. (2002). Bridging the digital divide: new route to development or new form of dependency? *Global governance*, pages 443–466.

Warren, J. R., Jenkins, K. N., and Kulick, R. B. (2006). High school exit examinations and state-level completion and ged rates, 1975 through 2002. *Educational Evaluation and Policy Analysis*, 28(2):131–152.

Webster, J. and Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review.

Wegmann, K. and Bowen, G. L. (2010). Strengthening connections between schools and diverse families: A cultural capital perspective. *Prevention Researcher*, 17(3):7–10.

Wicander, G. (2011). Mobile supported e-government systems: Analysis of the education management information system (emis) in tanzania.

Widman, L. (2011). Design and implementation of a web-based time tracking system.

Wong, C.-C. and Chen, C.-C. (1999). A hybrid clustering and gradient descent approach for fuzzy modeling. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 29(6):686–693.

Wood, J. M. (2007). Understanding and computing cohen's kappa: A tutorial. *WebPsychEmpiricist. Web Journal at http://wpe. info/.*

Worrell, F. C. and Hale, R. L. (2001). The relationship of hope in the future and perceived school climate to school completion. *School Psychology Quarterly*, 16(4):370.

Wycliffe, A., Samson, G. O., and Ayuya, V. C. (2013). Can education system be repaired? ideological dearth in kenya's educational practice and its implications for reforms in the education sector. *Journal of Educational and Social Research*, 3(2):213.

Xiao, X., Xu, H., and Xu, S. (2015). Using ibm spss modeler to improve undergraduate mathematical modelling competence. *Computer Applications in Engineering Education.*

Yadav, A. K., Malik, H., and Chandel, S. (2014). Selection of most relevant input parameters using weka for artificial neural network based solar radiation prediction models. *Renewable and Sustainable Energy Reviews*, 31:509–519.

Yu, L. and Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *International Conference on Machine Learning (ICML)*, pages 1–8.

Yukselturk, E. and Ozekes, S. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open Distance and e-Learning*, 17(1):118–133.

Zafra, A., Romero, C., and Ventura, S. (2009). Predicting academic achievement using multiple instance genetic programming. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, pages 1120–1125. IEEE.

Zare, M., Pourghasemi, H. R., Vafakhah, M., and Pradhan, B. (2013). Landslide susceptibility mapping at vaz watershed (iran) using an artificial neural network model: a comparison between multilayer perceptron (mlp) and radial basic function (rbf) algorithms. *Arabian Journal of Geosciences*, 6(8):2873–2888.