

**MACHINE LEARNING MODEL FOR PREDICTION OF STUDENTS'  
ACADEMIC PERFORMANCE, KENYA**

**Obadiah Matolo Musau**

**A Thesis Submitted in Partial Fulfilment for the Requirements of the Award for the  
Degree of Doctor of Philosophy in Information Technology of Masinde Muliro  
University of Science and Technology**

**October, 2020**

## DECLARATION

This Thesis is my original work, except where otherwise stated and has not been presented for a degree in any other University or any other award.

.....

**Obadiah Matolo Musau**

**SIT/H/09/11**

.....

**Date**

## CERTIFICATION

The undersigned certify that they have read and hereby recommend for acceptance of Masinde Muliro University of Science and Technology a Thesis entitled “**Machine Learning Model for Prediction of Students’ Academic Performance, Kenya**”

.....

**Dr. Kelvin K. Omieno**

**Department of Information Technology and Informatics**

**School of Computing and Information Technology**

**Kaimosi Friends University College**

.....

**Date**

.....

**Dr. Raphael Angulu**

**Department of Computer Science**

**School of Computing & Informatics**

**Masinde Muliro University of Science and Technology**

.....

**Date**

## **DEDICATION**

I dedicate this work to my dear parents; Mom Susan Mwethya Musau and my late Dad Samuel Musau Muthoka who gave me a chance into this world and above all, a good foundation to hold on.

God bless you Mom. Daddy, may God rest your soul in eternal peace.

## ACKNOWLEDGEMENT

I am grateful to God for giving me the strength, hope and courage to undertake and accomplish this work. My sincere appreciation to my supervisors Dr. Kelvin Omieno and Dr. Raphael Angulu for their professional guidance throughout my research

I wish to acknowledge my friend Dr. Stephen Ikikii who psyched me up to complete this doctorate degree. Special acknowledgement to many other friends not possible to mention all of them for cheering me on to complete this doctorate degree.

God bless you all.

## ABSTRACT

Prediction of students' academic performance with high accuracy is useful in many ways in academic institutions. Institutions would like to know which students are likely to have low academic achievements or need assistance in order to finish their studies. Successful students' academic performance prediction at an early stage in learning depends on many factors. Machine learning techniques can be utilized to predict students' future academic performance. The primary objective of this study was to develop a machine learning model for prediction of students' academic performance. To achieve this objective, the study was guided by the following theoretical and empirical objectives: 1. To analyse existing studies on students' academic performance prediction, 2. To find out the most significant factors that affect students' academic performance, 3. To develop a model for students' academic performance prediction in Kenya and, 4. To validate the students' academic performance prediction model. Student data was collected from 1720 former secondary school students currently enrolled in tertiary institutions using questionnaires. The data included students' academic performance, demographic features, social features and school related features. Naïve Bayes, Decision Trees and Neural Networks were used to predict students' final examination grade. The performance of the prediction models was validated using 10-fold cross-validation method. J48 Decision Tree prediction model achieved 85.9 % prediction accuracy, Naïve Bayes prediction model achieved 78.96% prediction accuracy and Neural Networks Multi Perceptron prediction model achieved the lowest prediction accuracy of 73.73%. This work will help educational institutions, school managements, government ministries, parents, donors and other education stakeholders to predict students' performance and identify nonperforming student that need assistance to finish their studies.

## TABLE OF CONTENTS

<b>DECLARATION</b> .....	<b>ii</b>
<b>CERTIFICATION</b> .....	<b>ii</b>
<b>DEDICATION</b> .....	<b>iii</b>
<b>ACKNOWLEDGEMENT</b> .....	<b>iv</b>
<b>ABSTRACT</b> .....	<b>v</b>
<b>TABLE OF CONTENTS</b> .....	<b>vi</b>
<b>LIST OF TABLES</b> .....	<b>xiii</b>
<b>LIST OF FIGURES</b> .....	<b>xiv</b>
<b>ABBREVIATIONS</b> .....	<b>xvi</b>
<b>INTRODUCTION</b> .....	<b>1</b>
1.1 Background of the Study .....	1
1.2 Statement of the Problem.....	6
1.3 Overall Objective .....	7
1.4 Specific Objectives.....	8
1.5 Research Questions .....	8
1.6 Contribution of the Thesis .....	8
1.7 Justification of the study .....	9
1.8 Scope of the Study.....	9
1.9 Assumptions of the Study.....	10
1.10 Limitation of the Study.....	10
1.11 Structure of Thesis.....	11
<b>LITERATURE REVIEW</b> .....	<b>13</b>
2.1 Introduction.....	13
2.2 Machine Learning (ML) .....	13
2.2.1 Introduction to Machine Learning.....	13
2.2.2 Machine Learning and Data Mining.....	14

2.2.3 Theory of Machine Learning .....	14
2.2.3.1 Machine Learning Theory.....	15
2.2.3.2 Information Theory.....	15
2.2.4 Machine Learning in Education .....	16
2.2.5 Applications of Machine Learning.....	16
2.2.5.1 Application of Machine Learning in Education.....	17
2.2.5.2 Application of Machine Learning in Industry.....	17
2.2.6 Types of Machine Learning .....	18
2.2.6.1 Supervised Learning .....	18
2.2.6.2 Unsupervised Learning .....	19
2.2.6.3 Semi-Supervised Learning .....	19
2.2.6.4 Reinforced Learning .....	19
2.2.7 Classification Techniques .....	20
2.2.7.1 Decision Trees.....	21
2.2.7.1.1 Pruning.....	22
2.2.7.1.2 Algorithm.....	22
2.2.7.1.3 Entropy.....	24
2.2.7.1.4 Information Gain .....	24
2.2.7.2 Naïve Bayes .....	24
2.2.7.3 Support Vector Machines (SVM).....	25
2.2.7.4 Neural Networks (NN) .....	25
2.2.7.5 K-Nearest Neighbours (KNN) .....	26
2.2.7.6 Random Forest .....	27
2.2.8 Data Representation in Machine Learning .....	27
2.2.9 Evaluation of Machine Learning Predictive Models.....	28
2.2.9.1 Evaluation Metrics for Classification Models.....	28
2.2.9.1.1 Classification Accuracy .....	28

2.2.9.1.2 Confusion Matrix.....	28
2.2.9.1.3 Precision.....	29
2.2.9.1.4 Recall .....	29
2.2.9.1.5 F-score.....	30
2.2.9.1.6 Area Under the Curve (AUC) - Receiver Operating Characteristics (ROC) curve .....	30
2.2.9.2 Evaluation Techniques.....	31
2.2.9.2.1 Hold-out Method .....	31
2.2.9.2.2 Cross-Validation.....	31
2.2.9.2.3 Bootstrap Method .....	33
2.2.10 Feature Selection .....	34
2.2.10.1 Filters .....	35
2.2.10.1.1 Correlation-based Feature Selection (CFS) .....	36
2.2.10.1.2 Relief.....	37
2.2.10.1.3 Variance Thresholds .....	37
2.2.10.1.4 Information Gain .....	37
2.2.10.1.5 Gain Ratio .....	38
2.2.10.2 Wrappers.....	38
2.2.10.2.1 Step forward feature selection.....	39
2.2.10.2.2 Step Backwards Feature Selection.....	39
2.2.10.2.3 Exhaustive Feature Selection .....	40
2.2.10.3 Embedded Methods .....	40
2.3 Literature review of students’ performance related work .....	41
2.3.1 Factors used in predicting students’ academic performance .....	41
2.3.2 Prediction Methods used for Predicting Students’ Academic Performance.....	42
2.3.3 Gap Analysis.....	53
2.4 Theoretical Framework.....	55



2.4.1 Identifying the Factors Affecting Students’ Academic Performance .....	56
2.4.1.1 Theoretical Perspective on Factors Affecting Students’ Academic Performance	56
2.4.1.1.1 Tinto’s Integration Theory .....	56
2.4.1.1.2 Bean’s Longitudinal Student Attrition Model.....	58
2.4.1.1.3 Ogude, Kilfoil and Du Plessis student academic development and excellence model (SADEM).....	60
2.4.1.2 Summary on Theoretical Perspective .....	61
2.4.2 Conceptual Framework.....	61
2.5 Summary of Literature Review .....	63
<b>RESEARCH METHODOLOGY .....</b>	<b>64</b>
3.1 Introduction.....	64
3.2 Research Philosophy .....	64
3.3 Research design.....	66
3.4 Location of Study .....	68
3.5 Target Population .....	68
3.6 Sampling Techniques .....	68
3.7 Sampling Size .....	70
3.8 Instruments of Data Collection .....	71
3.9 Validity and Reliability of Research Instruments .....	72
3.9.1 Validity of Research Instrument .....	72
3.9.2 Reliability of Research Instrument.....	74
3.10 Data Collection for Prediction Model Development.....	75
3.10.1 Data Sources.....	75
3.10.2 Data Collection Procedure .....	76
3.11 Data Analysis .....	76
3.12 Machine Learning Methodology for Developing Prediction Model.....	78
3.13 Model Validation.....	81

3.14 Ethical Considerations .....	82
<b>RESULTS AND DISCUSSION .....</b>	<b>83</b>
4.1 Introduction.....	83
4.2 Student Data Set.....	83
4.3 Pre-Processing the Student Data Set .....	84
4.3.1 Digitization .....	84
4.3.2 Missing Data .....	88
4.3.3 Data Conversion.....	89
4.4. Feature Selection.....	89
4.4.1 Feature Selection Process .....	89
4.4.2 Using Feature Selection Techniques to Rank Features .....	91
4.4.2.1 Feature Selection Using Information-Gain Based Technique .....	91
4.4.2.2 Feature Selection Using Correlation-Based Technique.....	93
4.4.2.3 Feature Selection Using One Rule Technique .....	95
4.4.3 Discussion of Feature Selection .....	97
4.4.4 Finding the Optimal Feature Subset.....	103
4.5 Model Development.....	104
4.5.1 Training of the Predictive Models.....	104
4.5.1.1 Naïve Bayes Model .....	105
4.5.1.2 J48 Decision Tree Model.....	108
4.5.1.3 Multilayer Perceptron – Neural Networks Model.....	110
4.5.2 Using Successive Modelling to Select the Optimal Feature Subset.....	112
4.6 Discussion of Research Findings .....	115
4.6.1 Selection of the Optimal Feature Subset .....	115
4.6.2 Finding the Best Prediction Model.....	116
4.6.2.1 Using Performance Metrics .....	116
4.6.2.1.1 Precision (Confidence).....	118

4.6.2.1.2 Recall (Sensitivity) .....	119
4.6.2.1.3 F-Measure .....	120
4.6.2.1.4 Specificity .....	121
4.6.2.1.5 ROC Area.....	121
4.6.2.1.6 Accuracy .....	124
4.6.2.2 Using Data Augmentation to Compare Models' Performance .....	125
4.6.2.3 Using Voting Technique.....	132
4.6.2.3.1 Classification for Grade A .....	134
4.6.2.3.2 Classification for Grade B.....	134
4.6.2.3.3 Classification for Grade C.....	135
4.6.2.3.4 Classification for Grade D .....	136
4.6.2.3.5 Classification for Grade E.....	138
4.6.2.3.6 Comparing Precision for Different Classes.....	138
4.6.3 Selected Prediction Model.....	140
4.7 Summary of Results and Discussion.....	141
<b>FINAL MODEL FOR PREDICTION STUDENTS' ACADEMIC PERFORMANCE .....</b>	<b>144</b>
5.1 Introduction.....	144
5.2 Predictive Model Presentations.....	144
5.3 Predictive Model Structure .....	145
5.3.1 Model Description.....	145
5.3.2 Model Parameters.....	146
5.3.3 Model Attributes .....	146
5.3.4 Model Performance .....	149
5.3.5 Class Attribute.....	149
5.3.6 Confusion Matrix .....	150
5.4 Chapter Summary.....	151

<b>SUMMARY, CONCLUSION AND RECOMMENDATIONS .....</b>	<b>152</b>
6.1 Introduction.....	152
6.1.1 Objective 1: To analyse existing studies on students’ academic performance prediction.....	152
6.1.2 Objective 2: To find out significant factors that affect students’ academic performance .....	153
6.1.3 Objective 3: To develop a model for students’ academic performance prediction in Kenya .....	155
6.1.4 Objective 4: To validate the students’ academic performance prediction model..	156
6.2 Summary of the Conclusion .....	156
6.2.1 Most Significant Factors Affecting Students’ Academic Performance .....	156
6.2.2 Best Machine Learning Algorithms for Modelling Academic Performance Prediction Model.....	157
6.2.3 Academic Performance Prediction Model.....	157
6.3 Contribution of the Thesis .....	157
6.4 Limitations of the Study .....	158
6.5 Recommendations for Future Work.....	158
<b>REFERENCES.....</b>	<b>160</b>
<b>APPENDICES .....</b>	<b>175</b>

## LIST OF TABLES

Table 1.11. Structure of Thesis.....	11
Table 3.7 Target Population .....	71
Table 3.9.1: Validity of Research Instruments .....	73
Table 3.10: Reliability Statistics.....	75
Table 4.3.1 Attribute Description .....	84
Table 4.3.2 Class Attribute Grouping .....	88
Table 4.4.3 Summary of Feature selection.....	99
Table 4.5.1.1 Performance of the Naïve Bayes Classifier on the Top Ranked Twenty Features .....	108
Table 4.5.1.2 Performance of the J48 Classifier on the Top Ranked Twenty Features..	110
Table 4.5.1.3 Performance of the Multilayer Perceptron Classifier on the Top Ranked Twenty Features.....	112
Table 4.5.2 Comparing Performance of the Models.....	114
Table 4.6.2.1 Performance of Models Based on the optimal Feature Subset.....	117
Table 4.6.2.2 Performance Metrics of the Models.....	118
Table 4.6.2.1.6 Prediction Accuracy of Classifiers .....	124
Table 4.6.2.2.1 Class instances before and after data augmentation .....	126
Table 4.6.2.2.2 Performance of the three classifiers before and after data augmentation	129
Table 4.6.2.3.1 Classification of Grade A using Majority Voting.....	134
Table 4.6.2.3.2 Classification of Grade B using Majority Voting.....	135
Table 4.6.2.3.3 Classification of Grade C using Majority Voting.....	136
Table 4.6.2.3.4 Classification of Grade D using Majority Voting.....	137
Table 4.6.2.3.5 Classification of Grade E using Majority Voting.....	138
Table 4.6.2.3.6.1 Comparing Precision for Different Classes.....	139
Table 4.6.2.3.6.2 Comparing Recall for Different Classes .....	139
Table 4.6.2.3.6.3 Comparing F-measure for Different Classes.....	140

## LIST OF FIGURES

Figure 2.2.7 Workflow of supervised machine learning algorithm [45].....	21
Figure 2.2.7.1.2 Pseudo Code of ID3 Algorithm [38] .....	23
Figure 2.2.9.1.2 Confusion Matrix [52] .....	29
Figure 2.2.9.1.6 ROC Curve [53] .....	31
Table 2.3.2 Factors used in predicting students’ academic performance and prediction methods .....	49
Figure 2.4.1.1 Tinto Conceptual Schema for Dropout in College [39].....	57
Figure 2.4.1.2: Tinto’s Longitudinal Model of Institutional Departure [75].....	58
Figure 2.4.1.1.2: Bean Longitudinal Student Attrition Model [76] .....	59
Figure 2.4.1.1.3 Student academic development and excellence model [77] .....	60
Figure 2.4.2 Conceptual Framework.....	62
Figure 3.2: The Research Pyramid, 2010 .....	66
Figure 3.3: Pretest-posttest with control group design.....	68
Figure 3.11.1: Number of respondents per Institutions.....	77
Figure 3.11.2: Grade Distribution .....	78
Figure 3.11 Machine Learning Processes [98] .....	80
Figure 4.3.1 Student Data Set.....	88
Figure 4.4.2.1 WEKA Information Gain Based Feature Selection Technique Results ....	92
Figure 4.4.2.2 WEKA Correlation-Based Feature Selection Method Results .....	94
Figure 4.4.2.3 WEKA OneR Feature Selection Technique Results .....	96
Figure 4.4.3: Tree generated using J48 .....	103
Figure 4.5.1.1: Naïve Bayes Model showing the Optimal Prediction Performance.....	107
Figure 4.5.1.2 J48 Decision Tree Model showing the Optimal Prediction Performance	109
Figure 4.5.1.3: Multilayer Perceptron Model showing the Optimal Prediction Performance.....	111
Figure 4.6.2.1.5.1 Class a ROC Curve for J48 Classifier.....	121
Figure 4.6.2.1.5.2 Class b ROC Curve for J48 Classifier .....	122
Figure 4.6.2.1.5.3 Class c ROC Curve for J48 Classifier.....	122
Figure 4.6.2.1.5.4 Class d ROC Curve for J48 Classifier .....	123
Figure 4.6.2.1.5.5 Class e ROC Curve for J48 Classifier.....	123
Figure 4.6.2.1.6 Prediction Accuracy of Classifiers .....	124
Figure 4.6.2.2.2 Randomizing File data using Randomize Technique .....	128
Figure 4.6.2.2.3 Naïve Bayes classifier output after data augmentation.....	130

Figure 4.6.2.2.4 J48 classifier output after data augmentation.....	131
Figure 4.6.2.2.5 Multilayer Perceptron classifier output after data augmentation .....	132
Figure 4.6.2.3 Majority Voting Process [109].....	133
Figure 5.3.1 Performance Predictive Model Description.....	146
Figure 5.3.2 Predictive Model Parameters .....	146
Figure 5.3.3 Predictive Model Attributes.....	149
Figure 5.3.4 Predictive Model Performance.....	149
Figure 5.3.5 Class Attributes .....	150
Figure 5.3.6 Confusion Matrix .....	150

## ABBREVIATIONS

<b>AI</b>	Artificial intelligence
<b>ANN</b>	Artificial Neural Network
<b>ARFF</b>	Attribute-Relation File Format
<b>CART</b>	Classification and Regression Trees
<b>CHAID</b>	Chi-Square Automatic Interaction Detector
<b>CSV</b>	Comma-Separated Value
<b>EDM</b>	Educational Data Mining
<b>EFA</b>	Education For All
<b>FFNN</b>	Feed- Forward Neural Networks
<b>FPE</b>	Free Primary Education
<b>GA</b>	Genetic Algorithm
<b>GCE</b>	General Certificate of Education Examination
<b>IS</b>	Information Systems
<b>KCSE</b>	Kenya Certificate of Secondary Education
<b>KDD</b>	Knowledge Discovery in Databases
<b>KMTC</b>	Kenya Medical Training College
<b>LA</b>	Learning Analytics
<b>LP-ITS</b>	Linear Programming Intelligent Tutoring System
<b>ML</b>	Machine Learning
<b>MOOC</b>	Massive Open Online Course
<b>MSE</b>	Mean Square Error
<b>NARC</b>	National Rainbow Coalition
<b>PTML</b>	Predictive Toxicology Mark-up Language
<b>SADEM</b>	Student Academic Development And Excellence Model
<b>SMOTE</b>	Synthetic Minority Oversampling Technique
<b>SPSS</b>	Statistical Package for Social Sciences
<b>SSCE</b>	Senior Secondary Certificate Examination
<b>SVM</b>	Support Vector Machines
<b>TTC</b>	Teacher Training College
<b>TTI</b>	Technical Training Institute



**UNICEF**

United Nations Children's Fund

**WEKA**

Waikato Environment for Knowledge Analysis

**XML**

Extensible Mark-up Language

# CHAPTER ONE

## INTRODUCTION

### 1.1 Background of the Study

Making elementary education free and compulsory by the Government of Kenya in 2003 was foreseen to; accelerate progress towards quality education for all children, remove cost barriers to education and eventually improve the economy of the country [1]. However, increased enrolment led to many schools being overcrowded with students hence putting a lot of pressure on the existing infrastructural and personnel resources. This came with unique challenges of increased student failure rate, drop out prior to completing school and high rate of class repetition. The pupil-teacher ratio increased from 34:1 in 2002 to 40:1 in 2003 [1]. These among other factors affects the quality of education and performance of students [2]. Predicting students' academic performance at elementary levels in focus of these challenges will help the government and school managements when making educational policies and budget [1].

According to [3], prediction is a positivist theory that provide replicability and predictive power through its ability to control any interventions in an experiment such that only the experimental variables change. Predictive modelling is based on positivist research philosophy which assumes that phenomena can be observed objectively and rigorously; and that research must possess the virtues of reductionism, refutability, and repeatability [3]. Student performance prediction models work by identifying factors that are most influential in determining student performance, and developing prediction models based on such factors to predict future performance. Machine learning techniques are applied to develop the models that predict students' academic performance using educational data mining [4] [5] [6].

Successful implementation of student performance prediction models in educational processes can support poorly performing students through intervention programmes especially in the developing countries [7]. Therefore, to make the objectives of free and compulsory elementary education relevant in developing countries, it is crucial to ensure that students at these levels achieve better grades and indeed learn skills that will improve their wellbeing and that of the nation at large. This can be achieved through such programs like early intervention programs on students that perform poorly and are more likely to get lower grades in the final examinations. Critical move towards effective intervention is to build models that can continuously track and accurately predict students' future academic performance, such as what are they likely to get in final examination given current and previous performance [1] [8].

Educational Data Mining (EDM) as an emerging interdisciplinary field uses data mining and machine learning techniques to turn data from educational settings to useful knowledge. The objective of EDM is to find out predictions and patterns that best characterize student's behaviour and performance. According to [1], students' success in learning is linked to several factors that include experience, language and culture, practices, gifts, traits, the external and internal school environments, and interests. Studies on building prediction models to predict student academic performance in secondary school, especially in the developed countries, have been based on specific educational and academic backgrounds [9] [10] [11] [12] [13] [14]. Majority of these studies predict: progress of student performance in college programs [15]; performance of students in courses like engineering [16]; student performance improvement [17]; tracking student academic performance and prediction of university student performance

in various courses [18].

However, there are diverse factors that measure student academic performance and differ from one educational setting to another [19]. These include student demographics, educational background, psychological, student academic progress and other environmental variables [19]. Review of previous studies on prediction of students' academic performance shows student factors differ significantly from the current study in terms of the educational and academic settings [15] [16] [17] [18], and therefore such results cannot be generalized to the developing countries. Previous studies are based on: tertiary institutions [15] [16] [17] [18] where learners are mostly adults whereas learners in secondary schools are minors from diverse socio-economic backgrounds and still need to be guided and modelled; the learning environment in universities and colleges differ significantly from elementary schools such as secondary school in terms of teaching and learning styles; student final performance in the tertiary institutions such as universities is incremental from evolving students' academic performance whereas in secondary school, instance assessment is applied.

The modes of evaluation adopted in different studies differ significantly worldwide in terms of geographical location and levels of study. For example, in Kenya, the mode of evaluation is an end of cycle (exit) examination called Kenya Certificate of Secondary Education (KCSE) administered at the end of the four years' secondary school schooling. In developed countries such as Portugal, France, Venezuela and other European countries where most of the studies have been carried out, students are evaluated in three periods during their years of schooling and the last evaluation corresponds to the final grade [9] [10]. In Nigeria, students sit for the Senior Secondary Certificate Examination

(SSCE) also called General Certificate of Education Examination (GCE) at the end of the six years secondary school period [11].

Education systems in the developed world differ significantly from those in developing countries [9] [10]. Whereas in Kenya, the education system advocates for four years of schooling in secondary school education preceding eight years of primary education. In the developed countries such as Portugal and some other European countries, secondary education consists of three years of schooling preceding nine years of basic education [9] [10]. In Nigeria, secondary school education consists of six years preceding six years of primary education [11].

Another significant difference between our study and the previous study is in terms of the grading systems applied. Countries like Kenya where the study was conducted uses an expanded letter grade ranging from A to E as follows: A is expanded to A, A-; B is expanded to B+, B, B-; C is expanded to C+, C, C-; D is expanded to D+, D, D- and E which is not expanded. These grades are based on a numeric 12-point scale where A is equivalent to 12 points representing excellent and, E is equivalent to 1 point representing poorest. On the contrary, European countries such as Portugal, France or Venezuela use a 20-point grading scale where 0 represents the lowest score and 20 represents the highest score [9] [10]. The grading system in Nigeria consists of nine grades for each subject. They include distinction grades - A1, A2, A3; credit grades - C4, C5, C6; pass grades - P7, P8 and Failure grade - F9 [11].

According to Xu, et al. [15], predicting performance of students that have diverse factors such as student demographics, previous academic progress, and other environmental

variables require more tailored approach to address the diversity which has effect on performance of individual students. Use of existing model for prediction of students' academic performance without discovering all the underlying correlation among the influential factor can lead to incorrect predictions [20] [15] [18] [19]. Although there is a vast publications on prediction of students' academic performance, however, most of the studies do not constitute a conclusive list of students' attributes that may potentially influence their performance and the quality of the prediction model [9] such as students' social and cultural characteristics. In other studies, the predictions are based on performance in a single subject such as mathematics or local language course which might introduce biasness since some students may be good in certain subjects and poor in other subjects and hence the subjects may not carry equal weights [9]. Therefore, the findings of such studies are not easily generalizable.

Previous studies on predictive modelling are silent on how the predictive models handle subjects that are dropped by students after selecting their majors (specialization areas), given that student prediction is largely reliant on student's past performance [15] [18] [19] [17] [12]. Unlike for universities where most students join their areas of interest directly on admission, subject specialization in secondary schools occur in later stages of study. The question on whether subjects studied by a student in their early years of study and dropped later after specialization should be considered when predicting students' academic performance in later stages of study has not been fully investigated.

Therefore, due to these challenge, applying all the past performance records of a student in secondary school on the existing predictors may not give an accurate measure of student academic potential. Furthermore, most of these studies are done in first world

countries [9] [10] which have diverse educational policies and environmental settings. Such challenges support the need to develop a secondary school students' academic performance prediction model for developing countries such as Kenya. To the best of our knowledge, there was none of the previous studies that predict students' academic performance at elementary levels of study in Kenya which was the focus of the current study.

## **1.2 Statement of the Problem**

The application of machine learning techniques to predict students' academic performance, based on student's demographics, educational background, previous in-term performance and other environmental variables, has proven to be very useful in student grade prediction and for foreknowing nonperforming students in various levels of education. However, research in the area of student academic performance in secondary school still remains limited, and the few studies that exist have been carried out in the developed world [9] [10] [11]. Majority of studies on prediction of students' academic performance have been based on tertiary institutions [19] [18] [15] [16] [17] [12] [21] [13] [14] such as universities and colleges. The existing models for prediction of students' academic performance from previous studies differ from this study in several ways: First, unlike in developed countries, students, in secondary schools in developing countries like Kenya come from diverse socio-economic backgrounds [1] hence they face diverse challenges due to their geographical diversity especially those located in areas that are socio-economically disadvantaged. These factors differ among students

from different educational backgrounds. Predicting performance of students that have diverse factors such as student demographics, previous academic progress, and other environmental variables require more tailor made approach to address the diversity [15]. Livieris, et al [9] indicated that most of the previous studies on academic performance prediction models do not constitute a conclusive list of students' attributes in order to generalize the findings of such studies [19]. For example, previous studies are silent on how the prediction models handle subject specialization and the effect of specialization on academic performance in secondary schools [15] [18] [19] [17] [12]. Secondly, previous studies reveal that tertiary learning institutions such as universities apply incremental assessment [12] [19] [22] [13] unlike secondary school where instance assessment is used. Although study has shown that there is a direct influence of previous incremental assessments on future performance of the student, such influence has not been investigated in secondary schools in developing countries.

It is therefore evident that the existing prediction models learn from features based on specific academic settings [19] such as grading system, mode of evaluation, learning and assessment styles. These features differ from one education system to another as well as from one levels of study to another. Therefore, the existing models cannot be directly applied for prediction of secondary school students' academic performance in the developing countries. The current study was on prediction of secondary school students' academic performance in the developing countries. All the experiments were conducted in Kenya.

### **1.3 Overall Objective**

The overall objective of the study was to develop a machine learning model for



prediction of students' academic performance in Kenya

#### **1.4 Specific Objectives**

The study sought to achieve the following objectives

- i. To analyse existing studies on students' academic performance prediction
- ii. To find out significant factors that affect students' academic performance
- iii. To develop a model for students' academic performance prediction in Kenya
- iv. To validate the students' academic performance prediction model

#### **1.5 Research Questions**

The research questions of this study were:

- i. What are the algorithms used in prediction of students' academic performance?
- ii. How to find out the most significant factors for predicting students' academic performance?
- iii. How can we model student academic performance prediction based on significant factors?
- iv. How can we validate a students' academic performance prediction model?

#### **1.6 Contribution of the Thesis**

This study makes the following contributions:

- i. Provide a comprehensive and analytical review of student' academic performance prediction, factors affecting students' academic performance, algorithms for academic performance prediction, feature selection techniques for student data and evaluation techniques for predictive models.
- ii. Provide a model for prediction of students' academic performance. The model will help relevant government ministries, parents, school administrators, teacher

and students within the context of improving and reforming the learning environment of secondary school in Kenya.

### **1.7 Justification of the study**

According to [1], Kenya's need for universal primary and secondary education dates back to the post-independence era in 1964 when the first commission (called Ominde commission) to chart course for education was established. Since then, the government has placed education at the centre of development. The free primary education (FPE) and education for all (EFA) initiatives by the National Rainbow Coalition (NARC) government in 2003 resulted in massive enrolments in the elementary levels of study with increased transition rate from primary level to secondary level. However, this came with unique challenges of increased student failure rate, drop out prior to completing the primary cycle and high rate of class repetition.

The Nairobi workshop [1] dubbed "School Fees Abolition" organized by UNICEF and World Bank in April 5-7 2006 underscored the need for quality monitoring of learning process through a national system that effectively monitors learning achievements at all levels. According to [23], student performance has been a big concern to the policy makers and school administrators. Maximising student course completion rate require consultative efforts and innovations. This study will benefit the government, school managements, parents, donors and all education stakeholders, as the adoption of the model will help in monitoring students' performance at all levels to improve on completion rates and overall students' academic performance, and help in earlier identification of at-risk students [24].

### **1.8 Scope of the Study**

The study focused on identification of influential student factors for prediction of academic performance and development of academic performance prediction model. Academic performance is measured at various levels of study including primary school level, secondary school level and at tertiary institutions such as universities, colleges or technical training institutions. The focus of this study was on students' academic performance at secondary school level of study in Kenya. The features were extracted from a student dataset consisting of 1720 instances and 62 features (attributes). Machine learning techniques Naïve Bayes, Decision Trees and Neural Networks were used to learn the prediction model.

### **1.9 Assumptions of the Study**

The following assumption were taken:

- i. All the respondents would answer the questionnaire honestly and to the best of their knowledge
- ii. The respondents have studied and completed their secondary school education in Kenya
- iii. The results obtained in data analysis are a representative of the target population

### **1.10 Limitation of the Study**

One of the limitation was getting students that scored grade E in KCSE. The study targeted respondents ranging from those who scored grade A up to grade E in KCSE, respondents who scored grade E may have shied away or were unwilling to fill the questionnaires either due to the stigma associated with poor performance, or it could be possible that only a few students ended up scoring grade E. The study managed to get

only 4 respondents who scored grade E. Again, some respondents avoided to answer the question on disability but responded to the other questions.

### 1.11 Structure of Thesis

The structure of the thesis is summarized in Table 1.11

**Table 1.11. Structure of Thesis**

<b>Chapter</b>	<b>Description</b>
Chapter One	Chapter one introduces the study and provides the background for the study, statement of the problem, research objectives and research question. The contribution of the study, justification, scope, research assumptions and limitations of the study are also discussed in this chapter.
Chapter Two	Chapter two begins by providing an overview of machine learning and predictive modeling, review of common machine learning algorithms and techniques, data representation in machine learning and feature selection techniques, evaluation of predictive models, review of students' academic performance prediction literature, theoretical framework and conceptual frameworks.
Chapter Three	Chapter three describes the research methodology used in the study. This comprises of; research philosophy, research design, target population, sampling techniques and sample size, research instruments, validity and reliability of research instruments, data collection and data analysis, model validation and finally ethical considerations.
Chapter Four	Chapter four presents a description of study data, feature selection, model development and discussions of research findings.
Chapter Five	Chapter five presents a detailed description of the machine learning

	students' academic performance prediction model.
Chapter Six	Chapter six presents a summary of the study, conclusion, recommendations and suggestions for future work.

## **CHAPTER TWO**

### **LITERATURE REVIEW**

#### **2.1 Introduction**

In chapter one we reviewed literature on student academic performance predictive modelling and the use of machine learning techniques in building academic performance prediction models. This chapter will discuss this issues in depth. The chapter is divided into two sections: Section one will discuss the machine learning process, types of machine learning, classification techniques and their application in building academic performance prediction modelling. Machine learning data representation, feature selection and evaluation of predictive models are also discussed in this section. The second section focuses on detailed review of previous studies related to students' academic performance prediction modelling, theoretical framework and the conceptual framework.

#### **2.2 Machine Learning (ML)**

##### **2.2.1 Introduction to Machine Learning**

In the recent past, application of machine learning techniques in education has grown exponentially, spurred by the fact that educators can now uncover new, interesting and useful insights about students [9]. Machine learning has enabled the development of more sophisticated and efficient performance predictive models in the educational sector. This models have the ability to classify and identify weak students with low achievements than was previously possible [12] [13] [25] [26] [27] [28] [29] [30] [31] [32] [33]. Mitchell [26] defined machine learning as computer programs that have the ability to automatically learn and improve from experience without being explicitly programmed. Lernverfahren [34] also defined it as a discipline of computer science that focuses on methods and algorithms that use predictive models to generate information

about new unseen data. Machine learning can also be seen as natural outgrowth of the intersection between computer science and statistics [35]. According to Mitchell [35], a machine learns with respect to a particular task T, performance metric P, and type of experience E, if such a system improves its performance P at task T, following experience E.

### **2.2.2 Machine Learning and Data Mining**

Machine learning is often associated with data mining and predictive modelling. However, depending on how we specify the task T, performance metric P and type of experience E, the learning task might also be called by names such as data mining, autonomous discovery etc [35]. According to Danso [36], data mining is a machine learning discipline inspired by pattern recognitions. It works by applying machine learning techniques to historical data to improve future decisions. Predictive modeling on the other hand works by applying a machine learning algorithm to previously collected data to predict future outcomes. Machine learning and data mining often use the same methods and techniques, however, despite of the overlap, the two are different in terms of their roles. Machine learning main focus is on prediction based on known properties learned from the training data while in data mining, the main focus is discovery of patterns and trends from unknown properties in the data. In terms of methods, data mining uses machine learning methods to achieve its goals while machine learning uses data mining methods such as unsupervised learning to achieve its goals.

### **2.2.3 Theory of Machine Learning**

Machine learning (ML) is a subfield of artificial intelligence that deals with the study of algorithms and statistical models that are used by computer systems to make predictions or decisions without being explicitly programmed but relying on patterns and inference instead. The core objective of a learner in machine learning is to build a generalized

model from experience. Several theoretical foundations of traditional machine learning approaches have played an important role in the development of machine learning techniques. Majority of previous studies in machine learning are based on computational learning theory and statistical learning theory. These theories are described here.

### **2.2.3.1 Machine Learning Theory**

Machine learning theory, which is also known as computational learning theory in other literature, is a fundamental theory that helps to advance the state of the art in software [37]. It provides a mathematical framework for designing new machine learning algorithms. This theory deals with the study of the design and analysis of machine learning algorithms. The goal is to understand the fundamental principles of a computational process. Machine learning theory draws elements from the theory of computation and the fields of statistics. From Computation, the computational learning theory's objective is to develop algorithms that are able to learn quickly [37]. The data in a machine learning algorithm is represented in terms of features which are processed by the learning algorithm to make some prediction. The statistical learning theory on the other hand solves the problem of finding a predictive function based on study data. This provides a framework for machine learning that draws from the statistics fields and functional analysis.

### **2.2.3.2 Information Theory**

Most studies in machine learning are based on information theory [38]. Information theory was developed as a result of contributions made by several individuals from various backgrounds whose perspectives and interests shaped the direction of information theory [39]. Information theory forms the basis of the decision tree ID3 algorithm and by extension C4.5 [38].



#### **2.2.4 Machine Learning in Education**

Machine learning represents promising areas of research in the education field [40]. The demand for use of machine learning techniques and other educational data mining techniques has been in the rise driven by the abundance of educational datasets available. Several machine learning and data mining technologies have been successfully implemented in the business world for some time now. However, according to Osmanbegović, and Suljić [40], their use in the education sector is still relatively new. The goal for educational institutions is to improve the quality of education and human capital. According to Iqbal et al [32] , the success of developing quality human capital has always been a subject of a continuous analysis. Prediction of students' success is therefore crucial to attaining quality education and quality human capital [41]. Despite the sector having experienced rapid growth of educational data in the recent past, still more research needs to be done to earnest the benefits from these data [40]. Data mining has the potential to identify and extract new and potentially valuable knowledge from student data. By converting educational data into knowledge, machine learning techniques can be applied to develop models that can guide conclusions on students' academic success [4] [5] [6]. Successful implementation of student performance prediction models in educational processes can support the specific needs of each student and other educational stakeholders [7].

#### **2.2.5 Applications of Machine Learning**

Machine learning has become a highly successful discipline in the recent past. Literature on machine learning has suggested several applications of machine learning in the educational sector and the industry. In this section, some selected fields are discussed.

### **2.2.5.1 Application of Machine Learning in Education**

From the educational sector, several applications have been proposed that include: development of prediction models for early alert systems to address concerns regarding declining student retention outcomes in higher education [29] [42], prediction of first-year to second year retention rates [28], use of machine learning algorithm to predict student pass rates in online education [25], prediction of the final grade of a university student before graduation [13], prediction of student academic performance in several levels [14] [30] [24] and predicting who will drop out of courses [31] [43] [44].

### **2.2.5.2 Application of Machine Learning in Industry**

There is a broad range of significant real-world machine learning applications such as autonomous mobile robots that can learn to navigate from their own experience, to medical applications that can learn to predict which future patients will respond best to which treatments, to search engines that automatically customize to their user's interests [35]. Other recent successful machine learning applications include the self-driving vehicles and of late the smart cities. The applications can be grouped into several areas including: Robot control which use machine learning methods in a number of robot systems such as flight control; Speech recognition which use machine learning to train computer systems to recognize speech; Computer vision programs ranging from face recognition systems to systems that can automatically classify microscope images of cells using machine learning; Bio-surveillance systems for detection and tracking purposes using machine learning technologies; and for accelerating empirical sciences

where data-intensive sciences make use of machine learning methods to aid in the scientific discovery process [35].

Other uses of machine learning techniques include customer retention systems where businesses have turned to predictive modeling to get solutions on how to retain customers. This is achieved by applying machine learning techniques on historical data of customers to come up with predictive models that can flag customers who are exhibiting behaviours indicative of possible exit. Machine learning is also used in risk assessment by insurance companies and banking institutions, weather forecasting where scientists use data from many sources on weather conditions to analyse data for tracking and predicting storm paths [33], online advertising and marketing which involves use of predictive models for marketing and decisions-making based on those projections, Other successful areas of application include spam filters where predictive models are used to identify the probability that a given email message is spam, and fraud detection where predictive models are commonly used in banks to identify outliers in a dataset that point toward fraudulent activity.

### **2.2.6 Types of Machine Learning**

Mitchell [35] described learning in machine learning as a process that improves the performance  $P$  of a system in a particular task  $T$  using some type of experience  $E$ . Machine learning is divided into four basic types: supervised learning, unsupervised learning, semi-supervised learning and reinforced learning [5]. Each of these types of learning is described here.

#### **2.2.6.1 Supervised Learning**

In supervised learning, the objective is to build a prediction model for predicting the true labels of unseen future data [9]. The input dataset consists of labeled data [45].

Supervised learning is used to solve two types of problems: classification and regression problem. In classification problem, the output variable is a category (continuous) while in regression problems, the output variable is usually a real value (discrete) [5] [7].

#### **2.2.6.2 Unsupervised Learning**

In unsupervised learning, the objective is to infer the natural structure present within a given dataset by applying the learning algorithms directly to the dataset and letting the algorithms learn on their own the structure in the data [7]. The dataset in unsupervised learning consists of unlabelled data. Unsupervised learning problems are categorized into clustering or association problems. In clustering problem, the goal is to discover the inherent groupings in the data while in association problem the goal is to discover rules that describe large portions of data.

#### **2.2.6.3 Semi-Supervised Learning**

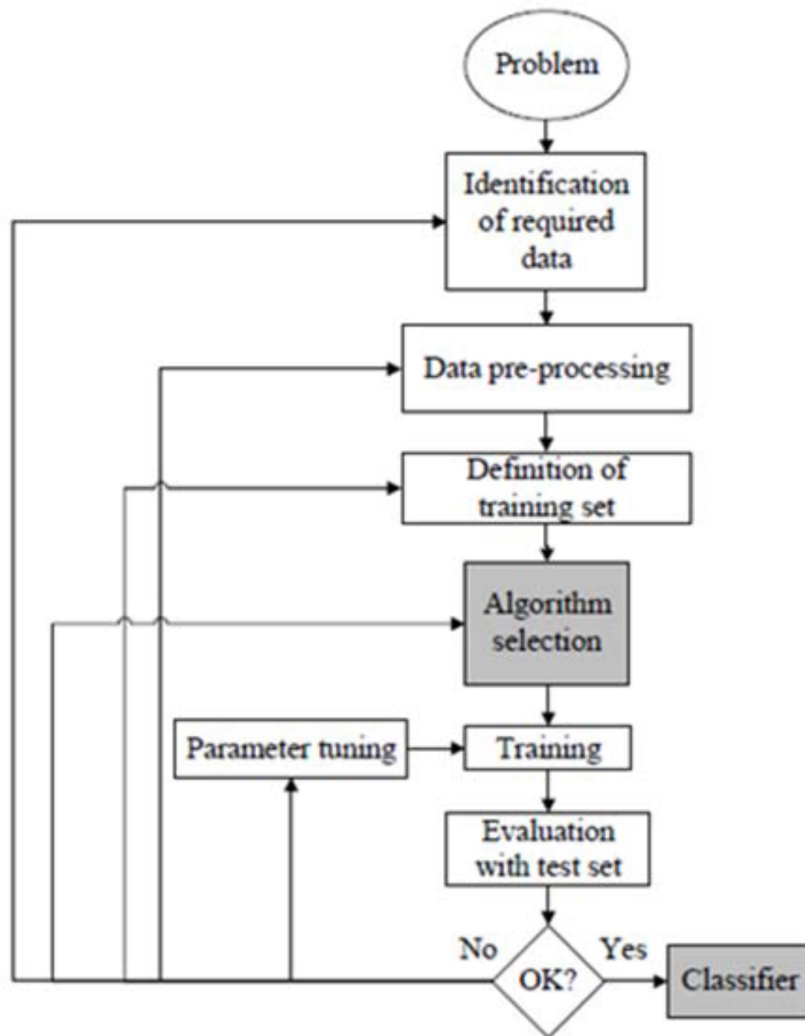
Semi-supervised learning deals with problems that can neither be categorized as supervised and unsupervised learning problems. Such problems are classified as semi-supervised machine learning. A combination of supervised and unsupervised techniques can be used to solve semi-supervised machine learning problems.

#### **2.2.6.4 Reinforced Learning**

In reinforced learning, the objective is to allow the learning algorithm train on their own continually through trial and error from past experience. In this learning, the algorithm learns based on feedback from the environment and keep on learning or adapting as time goes by. However, the challenge with this type of learning is that if the problems are very modular; similar learning reappear often, thus learning everything all over again. The bottleneck to this kind of learning is that it requires huge memory to store values of each state thus making it very expensive.

### **2.2.7 Classification Techniques**

Classification is one of the most frequently researched problem in supervised machine learning [12] [21]. It is used to predict the value of a target attribute (class attribute) based on the values of given predicting attributes or independent attributes [12]. Classification is the task of classifying or placing a target item to their correct target class using some supervised machine learning algorithm. Supervised machine learning algorithms use labelled input data with in advance familiar class to which data belong for building models, and then predict the class to which unknown data (unlabelled data) will belong using the constructed model [45] [6]. The process of classification in supervised machine learning is shown in Figure 2.3.2. The objective of supervised machine learning algorithms is to build a model which automates the process of classifying future unknown data in an easier way [46].



**Figure 2.2.7 Workflow of supervised machine learning algorithm [45]**

There are several classification algorithms used to create a prediction model [35]. Although they all perform essentially the same task, i.e., to predict a class variable (dependent variable) based on independent variables, they are however based on different mathematical methods [35]. This section will discuss the commonly used machine algorithms for classification problems.

### **2.2.7.1 Decision Trees**

Decision tree is a classification algorithm which uses a tree structure to build classification models [47]. The decision tree algorithm uses a recursive process to build the tree by breaking down the training data set into discrete groups as homogeneous as possible with respect to the class variable (predicted attribute) [38]. The final output of the recursive process is a tree which comprises of nodes and branches. The end nodes (leaf node) represent a decision and non-final node represent a test that the node can take [5] [45].

#### **2.2.7.1.1 Pruning**

Decision tree uses pruning technique to address overfitting problem. Overfitting problem occur when resultant decision function performs best only with a given of training data set. Overfitting problem affects prediction error rate [38]. Pruning technique works by reducing the size of decision trees by removing parts of the decision tree that provide little power to classify an instance. In the implementation of pruning technique using C4.5 decision tree algorithm, the sum of estimated errors of the branches of a sub tree are compared with the estimated error of expected leaf assuming that the sub-tree is exchanged with a leaf; if the approximated error given by a leaf is less than the approximated error of the branches, the entire sub-tree is pruned (replaced with a leaf) [38].

#### **2.2.7.1.2 Algorithm**

The core algorithm for building decision trees is ID3. It is a supervised learning algorithm that builds a decision tree from a given set of features(examples). C4.5 is an extension of ID3 algorithm. Information theory forms the basis of the ID3 algorithm and by extension C4.5 [38]. Decision tree C4.5 algorithm employs a top-down greedy search through possible branches and does not allow backtracking. This approach successively splits the training data set into distinct groups until no further subdivision is possible.

This is achieved using the divide-and-conquer strategy. It uses the “if-then” rules for model representation which is the most commonly used type for model representation especially in machine learning and data mining [48]. ID3 uses Entropy and Information Gain to construct a decision tree.

Given a set of attributes  $C_1, C_2, \dots, C_n$  and  $C$  as the target attribute, and a set  $S$  of recording learning [38], the pseudo code for ID3 algorithm is given shown in Fig 2.2.7.1.2.

```

Inputs:  $R$ : a set of non- target attributes,  $C$ : the target attribute,  $S$ : training data.
Output: returns a decision tree
Start
Initialize to empty tree;
    If  $S$  is empty then
        Return a single node failure value
    End If
    If  $S$  is made only for the values of the same target
then
        Return a single node of this value
    End if
    If  $R$  is empty then
        Return a single node with value as the most common value of the target attribute values found in  $S$ 
    End if
     $D \leftarrow$  the attribute that has the largest Gain ( $D, S$ ) among all the attributes of  $R$ 
     $\{d_j, j = 1, 2, \dots, m\} \leftarrow$  Attribute values of  $D$ 
     $\{S_j \text{ with } j = 1, 2, \dots, m\} \leftarrow$  The subsets of  $S$  respectively constituted of  $d_j$  records attribute value  $D$ 
        Return a tree whose root is  $D$  and the arcs are labeled by  $d_1, d_2, \dots, d_m$  and going to sub-trees  $ID3 (R-\{D\}, C, S_1), ID3 (R-\{D\} C, S_2), \dots, ID3 (R-\{D\}, C, S_m)$ 
End

```

Figure 2.2.7.1.2 Pseudo Code of ID3 Algorithm [38]



### 2.2.7.1.3 Entropy

Entropy is the degree of randomness of elements or the measure of impurity. Roughly speaking, entropy is a measure of how much variance the given data set has. Mathematically, it can be calculated with the help of probability of the items as:

$$Entropy(p) = \sum_{x=0}^n p(x) \log p(x)$$

Where  $p(x)$  is the probability of feature  $x$ .

### 2.2.7.1.4 Information Gain

Information gain is used to measure the amount of information an attribute gives about the class attribute. It is based on the reduction in entropy after a dataset is split on an attribute. Information gain is a metric used to measure the quality of a split. We use information gain to determine which attribute in a given set of training features gives the highest information (most significant attribute). While constructing a decision tree, the attribute that returns the highest information gain value becomes the root node. Information gain is calculated as:

$$Gain(p, T) = Entropy(p) - \sum_{x=0}^n (p(x) - Entropy p(x))$$

Where  $p(x)$  is the probability of feature  $x$

### 2.2.7.2 Naïve Bayes

Naive Bayes is a classification method which consists of a group of simple probabilistic classifiers. These classifiers are usually based on the Bayes' theorem with strong independence presumptions between the features. The assumptions are that: the predictive attributes are conditionally independent with familiar classification, and that there are no hidden attributes that could interfere with the process of prediction. Naive

Bayes is a very robust model which has quite often outperformed sophisticated models. It provides a very efficient algorithm for data classification [40].

### **2.2.7.3 Support Vector Machines (SVM)**

Support Vector Machine algorithms works on the principle of margin calculation between the classes [45]. SVM uses a hyperplane. This is the line that best splits the points in the input variable space by their class. The class can be either class 0 or class 1. The margin is the distance between the hyperplane and the closest data points. The optimal hyperplane that separates the two classes is the line that has the largest margin. SVM learning algorithm is used to find the coefficients that results in the best separation of the classes by the hyperplane. The points are called support vectors and are relevant in both defining the hyperplane and construction of the classifier. SVM classifier or model is used to predict whether a new example falls into one category or the other.

### **22.7.4 Neural Networks (NN)**

The Neural Network algorithm mimics the structure of the human brain [12]. They consist of a set of highly interconnected entities that mimic the human neurons referred to as processing unit (or artificial neuron). The processing units are interconnected (through synapses) to transmit signal from one neuron to another. The processing units have the ability to receive or accept a set of inputs (signals), process it and respond with an output to the neurons connected to it [13].

A neuron has two modes of operation: the first mode is called the “training mode” whose objective is to determine the input-output mapping. This is achieved through training the network using a set of paired data to allow the neuron learn when to fire and when not to

fire. The second mode is the “using mode” where the weights of the connections between neurons are then fixed and the network is used to determine the classifications of a new set of data [22]. At this point, the neuron will detect and fire the output associated to any input pattern. However, if the input pattern is not among the list of the taught input patterns, then the firing rule is applied to decide whether the neuron will fire or not fire. The signal is represented in form of a real number at any connection between the neurons.

The neurons and connections normally have a weight that keeps on adjusting itself as learning proceeds by either increasing or decreasing the strength of the signal at a connection. Typically, the neurons may be assigned some threshold, in such a case, the signal is fired only if the aggregate signal crosses the threshold. Neurons in a neural network are usually organized in layers. The input signal traverses through all the layers from the first layer (also called input layer) through the network layers to the last layer (also called output layer). Each layer is designed to perform certain kinds of transformations on the input signal or data. Where necessary the traversal may traverse iteratively. NN has the capability of self-learning and self-adapting which makes it to be more efficient and accurate than other classification techniques [5] [26] [48].

#### **2.2.7.5 K-Nearest Neighbours (KNN)**

K-Nearest Neighbors uses supervised learning algorithms used for classification and regression problem. To predict for a new input (also called point), KNN algorithm searches through the whole training set to get the K most similar instances (also called the neighbors). To determine the similarity between the data instances, KNN uses the Euclidean distance, a number which is calculated directly based on the differences

between each input variable [49]. However, KNN requires a lot of memory to store all of the data. The technique has been used in a number fields such as health.

#### **2.2.7.6 Random Forest**

Random forest is a set of decision trees with each built on random samples using a different policy for splitting a node. Random Forest belongs to the ensemble machine learning algorithm that is called bootstrap aggregation or bagging. Bootstrap and bagging are statistical method. Bootstrap works by taking multiple samples of data, calculating the mean of each sample and later average all of the mean values to get a better estimation of the true mean value. Bagging on the other side takes multiple samples of the training data and create models for each data sample. To predict for new data, each model makes a prediction, then an average of the predictions is taken to give a better estimate of the true output value.

#### **2.2.8 Data Representation in Machine Learning**

According to Hall [50], in supervised machine learning research, the raw data is usually represented inform of a table of instances with each instance representing a fixed number of features or attributes. The features are of one of the two data types: nominal data which is classified without a natural order or, ordinal data which has a predetermined or natural order. In machine learning, data is split into two sets: the training dataset which is used to construct the model and the test dataset which is used to evaluate the accuracy of the model. In an experiments, the test dataset consists of fewer instances compared to the training dataset.

## **2.2.9 Evaluation of Machine Learning Predictive Models**

Choosing the right evaluation techniques and the evaluation metrics for evaluating the performance of a model is paramount to the success of machine learning applications [51]. There are several methods used to estimate the accuracy (or expected prediction error) of the model. This section will discuss on the evaluation metrics and evaluation methods for classification models.

### **2.2.9.1 Evaluation Metrics for Classification Models**

Evaluation metric are used for measuring and judging the performance of a model. There are several evaluation metrics, the commonly used evaluation metrics in classification problems are discussed here.

#### **2.2.9.1.1 Classification Accuracy**

Accuracy is one metric for evaluating classification models. The accuracy of a given model is computed as the ratio of number of correct predictions to the total number of predictions made (or total number of input samples) multiplied by 100 to convert it into a percentage. Accuracy can also be calculated in terms of positives and negatives as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives. However, Classification accuracy alone is generally not enough information to make a judgement on the performance of the model.

#### **2.2.9.1.2 Confusion Matrix**

Confusion Matrix is a table (or matrix) that is used to describe the performance of a classification model. The basic terms used in confusion matrix are: True Positives (TP) which is an outcome where the model correctly predicts the positive class, True Negatives (TN) which is an outcome where the model correctly predicts the negative

class, False Positives (FP) which is an outcome where the model incorrectly predicts the positive class (also known as a Type I error), and False Negatives (FN) is an outcome where the model incorrectly predicts the negative class (or a Type II error).

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

**Figure 2.2.9.1.2 Confusion Matrix [52]**

Confusion Matrix forms the basis for the other types of evaluation metrics. It is used for measuring Recall, Precision, Specificity, Accuracy and AUC-ROC Curve.

### **2.2.9.1.3 Precision**

Precision is defined as the total number of correct positive outcomes divided by the total number of positive results predicted by the model. It measures the exactness of a classifier.

$$Precision = \frac{TP}{TP + FP}$$

### **2.2.9.1.4 Recall**

Recall is the total number of correct positive outcomes divided by the total number of all relevant samples which should have been identified as positive. It measures the completeness of a classifier.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

### 2.2.9.1.5 F-score

F-score is used to measure Recall and Precision at the same time making it easier to compare precision and recall of two models at the same time. It is calculated as follows:

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

### 2.2.9.1.6 Area Under the Curve (AUC) - Receiver Operating Characteristics (ROC) curve

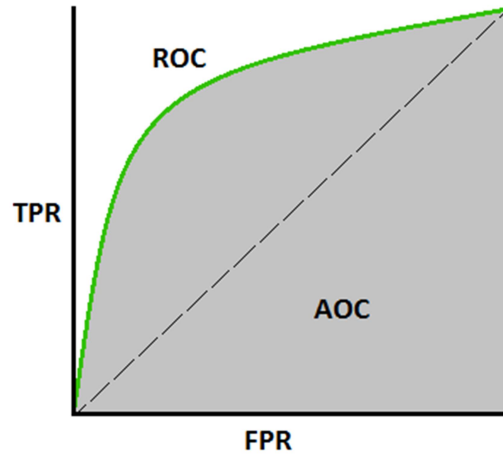
AUC - ROC curve is used to measure performance for classification problem at various thresholds settings. The ROC curve graph displays the performance of a classification model at all classification thresholds. The graph or curve is plotted against two parameters: True Positive Rate (TPR) which is a synonym for recall and is calculated as;

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

and False Positive Rate (FPR) given as;

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

If the AUC is near to the 1 it means an excellent model which is good measure of separability where as if AUC is near to the 0 means it is a poor model.



**Figure 2.2.9.1.6 ROC Curve [53]**

### **2.2.9.2 Evaluation Techniques**

In most applications, there exists only a limited amount of data available, hence in machine learning we split the original dataset into two, a training set for training the model and a test set for evaluating the performance of the trained model. There are several methods for accuracy estimation used to evaluate performance of machine learning models such as cross-validation, bootstrap methods, hold-out, etc. In this section we will discuss some of these model evaluation techniques.

#### **2.2.9.2.1 Hold-out Method**

Hold-out method splits the data into two mutually exclusive subsets, a training set and a test set. This method relies on a single split of data. Usually the training set contains two thirds of the entire dataset and the rest forms test set [54] [51]. However, the hold-out method is considered a pessimistic estimator because only a fraction of the data is used for training the algorithm leaving out more instances for training hence increasing the bias of the estimate.

#### **2.2.9.2.2 Cross-Validation**

Cross-validation is a technique for validating a model's performance by splitting the dataset into two subsets, the training subset and the test subset. The training subset is



used for training the model and the test subset is used for evaluating the model's performance. Cross-validation technique has been widely applied on model selection because of its simplicity and universality [54]. There are several cross-validation techniques that can be used in validating a model, however, choosing the best cross-validation technique will depend on the particular features of the problem in hand. Some of the common types of cross-validation are discussed below:

#### **2.2.9.2.2.1 Leave-one-out (LOO)**

This is an exhaustive cross-validation procedure in which the dataset is split into two parts, one part contains one observation which forms the test data and the other remaining part forms the training subset. The process is iterated successively leaving out each data point out from the dataset and using it as test data until the entire dataset is exhausted. This approach is less bias since the entire dataset is used for training. However, using a single observation can introduce variability especially if the data point is an outlier then the variability is higher. Again, LOO is computationally a very expensive method especially for large datasets.

#### **2.2.9.2.2.2 K-fold cross-validation**

K-fold cross-validation divides the dataset into K-blocks, then the  $K^{\text{th}}$  block is used to make the test block and the rest of the data makes the training data. The model is trained on k-1 folds. The process is repeated K-times. The commonly used value is  $k=10$  although there may be some variation of  $K=5$  or others use  $K=20$ . The advantage with this technique is that the execution time is greatly reduced, the bias and variance of the resulting estimate is also reduced. Another advantage is that because of systematic data partitioning, all available samples are eventually used for both training and testing the model. K-fold cross-validation is seen as a compromise between the holdout and leave-one-out estimator.

#### **2.2.9.2.2.3 Stratified cross-validation**

Stratified cross-validation reorganises the dataset into folds/strata such that each fold has a representative of the entire dataset. The advantage with Stratified cross-validation method is that it reduces the bias and variance.

#### **2.2.9.2.2.4 Leave-p-out (LPO)**

LPO is the exhaustive cross-validation procedure where every possible subset of  $p$  data is successively left out of the sample and used for validation.

#### **2.2.9.2.3 Bootstrap Method**

In bootstrap method, the dataset is uniformly sampled with replacement to generate a training set (bootstrap sample) with the same number of instances as the original data set, and with some instances repeated more than once in the training set. The rest of the instances not sampled are picked as the test set. Bootstrap usually fails or gives wrong results if the inducer or classifier used has a memorizer module such as in nearest neighbour or unpruned decision tree [55]. There are several variants of bootstrap methods, the common ones are discussed here.

##### **2.2.9.2.3.1 0.632 Bootstrap estimator**

Bootstrap estimator is defined by the bootstrap formula for computing the bootstrap sample. This is because the bootstrap sample or training set is created by uniformly sampling the input dataset with replacement, then the probability that a particular instance will be chosen to be in the bootstrap sample is 0.632. The remaining 0.368 instances of the dataset are picked for the test set.

##### **2.2.9.2.3.2 Out-of-bootstrap estimator**

The out-of-bootstrap method works similar to the repeated holdout validation approach where the data split ratio 63.2:36.8 and repeating the process severally.

### **2.2.10 Feature Selection**

Feature selection is a pre-process step in machine learning that is used to remove irrelevant features in a data set [56] [57]. According to Blum, et al [58] and Yu, et al [59], a feature is said to be relevant to the output variable if correlation between that feature and the output variable is high enough to make it predictive of the output variable. The objective of feature selection is to improve prediction accuracy by selecting input features which are highly influential and give high predictive information [56]. Feature selection indirectly helps reduce computational time and model constructional cost through elimination of irrelevant features in the training and classification phases [56].

Feature selection techniques use induction algorithms or heuristics to identify relevant features and remove features that are considered redundant and irrelevant with respect to what is being learned [50]. According to Blum, et al [58] and Yu, et al [59], a feature is said to be relevant to the output variable if correlation between that feature and the output variable is high enough to make it predictive of the output variable. The search algorithms are required to define four mandatory characteristics: start point, search organisation, evaluation strategy and stopping criterion. The start point indicates the search starting point which in turn determines the direction of search. For example, a search algorithm can start with an empty feature set and proceed successively adding the features in which the search proceeds forward through the entire search space or, the algorithm can start searching the entire feature set and then successively remove the features in which the search proceeds backward through the search space or, the search can start at somewhere in the middle of the feature set and move outwards. The search

organisation is used to select an initial subset of the features since an exhaustive search through the entire feature vector is highly costly. Use of heuristic search strategies yields better results when operating on a data set that consist of many features. Evaluation strategy defines how the features are evaluated. For example, depending on the feature selection approach selected, there are those methods that use heuristics to rank the features based on general characteristics of the data while others use a combination of an induction algorithm and a statistical resampling method to estimate final accuracy given by feature subsets. The stop criterion consists of a set of condition(s) that halt the search. For example, a stop criterion can be to halt search when neither of the remaining combination features gives better performance than current feature subset or, to continue with the search as long as the value does not degrade or, stop when the search space is exhausted.

Feature selection methods are broadly categorised into three groups: Filters, Wrappers and embedded methods. Each of these methods is discussed here.

#### **2.2.10.1 Filters**

The filter methods use heuristics to evaluate the importance of each feature. An attribute evaluator is used to evaluate the importance of the features by assigning a weight to each feature and a ranker method is used to rank the features in the entire dataset based on the weight assigned. One advantage with filter methods is that they select features independently of the machine learning algorithm model, this makes the filter methods very fast and, the features selected can be used as an input to any machine learning models. Filter methods are more practical to use especially on large data since they don't

require any learning algorithms to filter hence perform much faster than wrapper methods. Some of the commonly used filter techniques include the relief algorithm (RA), information gain (IG), the threshold number of misclassification (TNoM) score, the signal-to-noise ratio, correlation based feature selection and fast correlation based filter [60]. We shall discuss some of the commonly used techniques here.

#### **2.2.10.1.1 Correlation-based Feature Selection (CFS)**

This is a filter algorithm for ranking features using a correlation based heuristic evaluation function [50]. CFS works on the premise that feature selection can be achieved on the basis of how the features are correlated with one another. Based on the heuristic evaluation function, features are accepted to be relevant if they are highly correlated with the target attribute and uncorrelated with each other. Likewise features that have low correlation with the class attribute are considered as irrelevant features and features that are highly correlated with either of the remaining features are considered as redundant features and should be ignored. Examples of the heuristic search strategies used in CFS are forward selection, backward elimination, and best first strategy.

The forward selection search strategy starts search with an empty set of features and successively adds the features as the search proceeds forward through the entire feature vector. Backward elimination begins search on the full set of features, then successively removes irrelevant features as the search proceeds backward through the search space. For the best first search strategy, the search begins with an empty feature subset or the entire feature subset. Correlation based Feature Selection is widely recommended for its ability to quickly identify relevant features and its ability to screen irrelevant, redundant, and noisy features from a dataset of features in machine learning experiments [50]. The

drawback to CFS is that it fails to take into consideration the interaction between features [61].

#### **2.2.10.1.2 Relief**

Relief algorithm is a type of feature weighting algorithm that assigns different weights to each feature according to the relevance (weight) of feature [62]. Relief algorithm works by first sampling instances randomly from the dataset and then update their relevance values. The relevance value is based on the difference between the selected instance and the two nearest instances of the same and opposite class [63]. The first nearest neighbour (also called nearest hit) to the selected feature is from the same class and the second nearest neighbour (also called nearest miss) to the selected feature is from a different class. The algorithm estimates the quality of features (relevance or weight) according to how well their values distinguish between instances that are near to each other [61]. Relief can search on discrete and continuous features and can capture local dependencies that other methods miss. The drawback is that it is limited to two-class problems. Relief also does not identify redundant features [63].

#### **2.2.10.1.3 Variance Thresholds**

The variance thresholds method evaluates the variance of each feature in the feature dataset then ranks the features based on the computed variance. Features with higher variance value are selected or a certain number of the top features with the largest variance are selected. The method assumes that the higher the variance a feature has the more useful information it contains.

#### **2.2.10.1.4 Information Gain**

Information gain is used to tell how important a chosen attribute of the feature vectors is to the class attribute. Information gain is the amount of information gained by knowing

the value of a feature, also called the entropy of the distribution. The attribute which has the highest Information gain will always split first. Information gain is biased towards selecting attributes with large number of values which could result to overfitting.

#### **2.2.10.1.5 Gain Ratio**

This method is a modification of the information gain to reduce the bias by taking into account the number and size of the branches when choosing the relevant attributes. However, gain ratio has a problem of overcompensating attributes by choosing those attributes that have low intrinsic information. The fix to this is to choose attributes with greater than average information gain.

#### **2.2.10.2 Wrappers**

The wrapper methods are applicable when the researcher wants to use a particular machine learning algorithm to train a model. In such a scenario, the features selected using filter methods may not be the most optimal set of features for the target algorithm. Wrapper methods use a subset evaluator that creates a set of all possible subsets from the feature space provided. The evaluator uses a search technique such as random search, breadth first search, depth first search or hybrid search such as best first search etc., to search for the subset. Then the evaluator applies a classification algorithm like Naïve Bayes to induce classifiers from the features in each subset and finally select the subset of features with which the classification algorithm has the best performance.

Wrapper is one of the approaches used for feature selection which uses a target learning algorithm to evaluate different features sets [50]. The wrapper methods apply a learning algorithm to the data in order to evaluate the worth of features. According to Isabelle and Elisseeff [64], the wrapper methodology makes use of the prediction performance of a given learning machine to evaluate the relative usefulness of subsets of features. Hall

[50] notes that an induction algorithm is used to estimate the significance of features in the wrapper approach.

Wrappers methods give better predictive accuracy compared to the filters methods due to its ability to optimize feature selection for each learning algorithm used [64]. However, they are not advisable to run for big databases containing many features since they use of learning algorithm to evaluate every combination of features thus making it very expensive especially in terms of time it takes to execute. Wrappers are less generalizable since the process of feature selection is based on a specific learning algorithm [50]. The wrapper methods for feature selection fall under three categories: Step forward feature selection, Step backwards feature selection and Exhaustive feature selection.

#### **2.2.10.2.1 Step forward feature selection**

The step forward feature selection method is an iterative method. In the first iteration, the step forward feature selection is used to select the feature that performs the best out of all the features after evaluating the performance of the classifier with respect to each feature. In the next iteration, the first feature is combined with each of the other features at a time and their performance evaluated. The combination of the two features that give the best algorithm performance is selected. In the subsequent iterations, this process is repeated continuously until all the specified number of features are selected. The method is implemented using a Sequential Feature Selector (SFS) which belongs to the family of greedy search algorithms.

#### **2.2.10.2.2 Step Backwards Feature Selection**

The step backwards feature selection works the reverse of the step forward feature selection.



In the first step, we start with all features and then one feature (the least significant feature) is removed from the feature set in a round-robin fashion and the performance of the classifier is evaluated, the set of features that gives the best performance is retained. In the next iteration, one feature is removed in a round-robin fashion and except the second features, the performance of all the combination of features is evaluated. In the other subsequent steps, this process is repeated continuously until all the specified number of features remain in the dataset. The method is implemented using a Sequential Backward Selector (SBS) which also belongs to the family of greedy search algorithms.

#### **2.2.10.2.3 Exhaustive Feature Selection**

Although the wrapper methods are based on greedy search algorithms, exhaustive feature selection is the greediest algorithm of all. The algorithm evaluates the performance of a machine learning algorithm against all possible combinations of the features in the dataset. The feature subset that produces the best performance is then selected. However, this method performs slower compared to step forward and step backward methods since it evaluates all feature combinations hence not preferred for large datasets.

#### **2.2.10.3 Embedded Methods**

Embedded methods (also called hybrid methods) use a mixture of the filter and wrapper methods by implementing algorithms with in-built feature selection methods [65]. The hybrid approach is the latest approach in feature selection approaches which combines both filter and wrapper methods. The approach was as a result of the challenge with wrapper methods that they require greater computational resources and perform slower compared to filter methods yet they give better predictive accuracy. Some of the widely known hybrid approaches using the genetic algorithm (GA) and a classifier that have been successfully used include a hybrid of GA and a neural network classifier (GANN),

incorporating GA and the support vector machine (SVM) classifier (GASVM), and combining GA via the weight voting classifier.

### **2.3 Literature review of students' performance related work**

Prediction of students' academic performance has been one of the most popularly researched area in the recent past [66]. It provides an opportunity for academic institutions to help students improve and attain their academic goals. There are two major components in predicting student's academic performances; the attributes, also called features, and the prediction methods. Machine learning techniques have been widely used to explore student data attributes from educational settings with a view of understanding the student better and the environments in which they learn from. According to Paulo & Silva [10], the education sector offers a fertile ground for researchers in academic fields due to multiple sources of data such as the traditional databases and online content. Substantial amount of work has been done in the area of prediction of academic performance in education. The literature borders on university admission, student performance, and academic related problem [11]. These studies differ in terms of their target classes, factors, prediction techniques applied and the target levels of education. This section will focus on the two main issues: the factors used in predicting students' academic performance and the prediction methods used in predicting students' academic performance.

#### **2.3.1 Factors used in predicting students' academic performance**

A systematic review of previous studies on predicting student academic performance prediction models has been used to identify the factors used in prediction of academic performance. The factors include student's demographic factors such as gender, family background, age, disability, high school background, social factors and psychometric

factors such as student interest and family support [66]. Bhardwaj and Pal [17] noted that academic performances is not always reliant on students' own effort but other factors that have significant influence over their academic performance. Table 2.3.2 (column three) shows systematic review of the factors used in predicting students' academic performance and the prediction methods used.

### **2.3.2 Prediction Methods used for Predicting Students' Academic Performance**

Prediction is a positivist theory which is aligned with the systematic reduction of a positivist approach [3]. This approach gives the researcher the ability to control and predict. Predictive modelling has been used for predicting student performance in the educational sector [66]. There are several tasks that are used in order to build predictive models, the most popular are classification and regression tasks (see section 2.2.6). Classification task is the most popularly used task in predicting students' performance [66]. Classification uses machine learning algorithms to predict students' performance. The algorithms used include Naive Bayes, Decision tree, Artificial Neural Networks, K-Nearest Neighbour and Support Vector Machine [20] [66]. A description of these algorithms is given in section 2.2.7. A summary of the machine learning algorithm used for predicting students' performance is shown in Table 2.3.1. Oladokun et al [11] applied Artificial Neural Network (ANN) to build a model for predicting academic performance of secondary school students before being considered for university admission using multilayer perceptron topology. The input variables included parental background, gender, ordinary level subjects' scores, subject's combination, matriculation examination, scores, type of school, location of school and age on admission. The data consisted of 112 records that spanned five generations of graduates from University of

Ibadan-engineering department. The results showed that the model predicted more than 70% of prospective students' academic performance correctly.

Livieris et al [9] conducted a study to predict secondary school students' academic performance in final examinations in the course of Mathematics. The study compared the effectiveness of two wrapper-based semi-supervised learning approach: self-training and Yet Another Two Stage Idea (YATSI) methods with neural network classifier in prediction of performance. The author utilized data containing 2 attributes on the performance of 3,716 students collected by the Microsoft showcase school Avgoulea-Linardatou during the years 2007 to 2016. The findings revealed that use of semi-supervised algorithms which utilize fewer labeled and many unlabelled data helps improve prediction accuracy and develop reliable prediction models.

Paulo & Silva [10] applied Business Intelligence and Data Mining techniques to predict performance in mathematics and Portuguese language courses for secondary school student's. The input variables consisted of 33 attributes that included mark reports, students' demographic, social and school related attributes such as student's age, alcohol consumption and mother's education. Four data mining techniques were applied to construct the model: Decision Trees, Random Forest, Neural Networks and Support Vector Machine. The results showed that the prediction accuracy could be improved by including the first and second year grades. The study also revealed that other than student achievement in past evaluations, other attributes such as student absences, parent's job, parent's education, and alcohol consumption are very relevant in predicting student performances.

In a related study by Osmanbegović and Suljić [40] on predicting student performance, three data mining techniques for classification were applied; Bayesian classifier, neural networks and decision trees. Input data included data on student's gender, distance, GPA, scholarship, learning materials and grade importance collected from 257 students of the Faculty of Economics in Bosnia and Herzegovina in a survey conducted between 2010- 2011. The study found out that Naïve Bayes classifier outperformed decision tree and neural network methods in prediction.

Khasanah and Harwati [56] conducted a comparative study to predict students' academic performance using Bayesian network and decision tree classification algorithms. Using feature selection, the most influential student attributes were used which included gender, origin, father education, father occupation, mother education, mother occupation, senior high school type, senior high school department, senior high school final grade, attendance, GPA and drop out. The data consisted of 178 student data collected from student data base from Universitas Islam Indonesia's information system. The performance parameter used to compare both algorithm was accuracy rate. The best prediction was obtained from Bayesian network classification algorithm.

Khan et al [47] applied J48 decision tree algorithm on student data containing previous performance to build a model to predict the student final grade based on Secondary School Certificate (SSC) – part one marks from Islamabad Capital Territory in Pakistan. The required data was extracted from Federal Board of Intermediate and Secondary Education student database for the years 2005, 2006, 2009, 2010, and 2012. The predictive model obtained was able to correctly classify 1268 student out of 1500 with a prediction accuracy of 84.53%.

Shahiri, et al [20] conducted a comparative study on how the different prediction algorithms can be used to identify the most significant attributes in a student's data while predicting students' academic performance. The study noted that Neural Network and Decision Tree are the two methods highly used methods under the classification techniques for predicting students' performance. Agrawal, et al [12] applied Neural Networks and Bayesian classification algorithms on a dataset containing marks of 80 Bachelor of Engineering, Information Technology (B.E. I.T) students from semester 3 to semester 6 to predict the performance of students. The study used feature selection technique to select the highly influential features. They included student living location, grade in secondary education and medium of teaching. The results showed that neural networks outperformed Bayesian classification. Sharma and Vishwakarma [67] developed a model based on previous student performances to predict final student performance by applying the ID3 decision tree algorithm on student data from Gyan Ganga Institute of Technology and Sciences in India. The dataset consists of 70 record and five attributes namely roll number, name, assignm1, assignm2, midsem1, midsem2 and final examination. The model achieved accuracy of 90%.

Guo, et al. [19] used deep learning neural network technique to develop a model for predicting students' academic performance which was trained on a 120,000 student dataset collected from 100 junior high schools in Hubei province. The model was found to work effectively with diverse student factors and variables that correlate in complicated nonlinear ways. Kabakchieva [30] used Decision Tree, Neural Network and the k-Nearest Neighbor algorithms in a comparative study to develop a performance prediction model for Bulgarian universities students. The input dataset contained 10067

instances and 14 attributes grouped into student personal data, pre-university data and university-performance data. The attributes included gender, birth year, birth place, living place and country, type of previous education, profile and place of previous education, total score from previous education, university admittance exam and achieved score, total university score at the end of the first year and number of failures. The is Neural Network model achieved the highest accuracy of 73.59%, Decision Tree model achieved 72.74% and the k-nearest neighbor model achieved 70.49%. Asif, et al [8] conducted a case study that used student data of four academic cohorts that consisted of 347 undergraduate students to predict the graduation performance in 4th year at university using pre-university marks and marks of first and second year courses only. The study used decision tree algorithm.

Kaur and Singh [68] applied Naïve Bayes and J48 decision tree classification techniques using WEKA software in prediction of student performance. The study used dummy data set that consisted of 52 instances and 9 attributes that included gender, hometown, family income, previous semester grade, attendance, medium (language) and senior secondary grade, seminar performance and sports. The results showed that Naïve Bayes provide better accuracy at 63.59 % than j48 which provide 61.53% accuracy

Lin [42] conducted a study to compare the quality and accuracy of chosen machine learning algorithms for predicting student dropout in institutions of higher education. The input data included gender, state, citizenship, academic major, ethnic group, age, student aid, family contribution, financial need, loan received, awarded scholarship and cal grant receiver. The study applied Decision Trees, Decision Rule Learning algorithms, Lazy Instance-based Nearest Neighbor algorithms, Function-based algorithms, Naive

Bayes method and Bayesian Networks. It was observed that machine learning algorithms performed better in predictive models.

Nichat and Raut [21] applied decision tree algorithm on student data from course evaluation questionnaires to develop a predictive model to predict the performance of student and recommend to the teacher the topics the student is weak or need to study again. Sundar [24] applied Bayesian network classifiers to build a students' academic prediction model. The input dataset contained 48 records and attributes: student id, name, quota in which student joins, previous semester performance, performance in internal exam, performance in seminars, assignment, attendance, co-curriculum activities and end of semester marks. The results showed that AODEsr algorithm achieved the highest overall accuracy of 64.6%.

Bhardwaj and Pal [17] build a model for prediction of students' academic performance using Bayes classification algorithm. A dataset of 300 records of student data from colleges and institutions colleges affiliated with Dr. R. M. L. Awadh University in Faizabad in India was used to learn the model. The attributes consisted of sex, student category, medium of teaching, student food habit, student other habit, living location, hostel, family size, family status, family income, senior secondary education grade, type of college, father's and mother's qualifications, father's and mother's occupation and BCA grade. The study found out that student performance is highly reliant on the student grade obtained in Senior Secondary Examination and living location. Goker and Bulbul [69] developed a prediction model for predicting student performance. They applied Naive Bayes method on records of 220 students and achieved 86.66% accuracy.



Gadhavi and Patel [70] used univariate linear regression to build a model that predict grade of final examination in particular subject. The model was trained on 181 records of students in one subject and tested on same data set. The attributes included average of unit test and sessional examination marks. Xing, et al [48] observed that learning curve for some classical models such as statistical models, artificial neural networks and Bayes Networks may be more difficult due to their complexity compared to others like rule-based models and decision trees. However, models that are easily understood by users often comes at the price of decreased performance. Therefore, trade-offs between model understandability and model performance need to be taken into account [48]. In terms of implementation, decision trees algorithms are fast to learn and to make prediction, they do not require any special treatments of the data and their predictions are often accurate for a broad range of problems.

**Table 2.3.2 Factors used in predicting students' academic performance and prediction methods**

No	Source	Factors	Machine Learning Technique	Instances
1	Oladokun et al [11]	UME score, O'level results, further math, age at entry, time before admission, parents education, zone of secondary school attended, type of secondary school, location of school and gender.	Neural Network - Multilayer Perceptron	112
2	Livieris et al [9]	Secondary stage type, Oral grade of the first test, second test and final examination of the first and second semester, Final grade of the first and second semester and Grade in the final examinations	Neural Networks	3716
3	Paulo & Silva [10]	Sex, age, school, address, parents cohabitation status, mothers education, mothers job, fathers education, fathers job, family size, guardian, family relationship, reason for choice of school, travel time, study time, failures, school support, family support, activities, extra	Decision Trees Random Forest Neural Networks Support Vector Machine	395 and 649 (for mathematics and Portuguese language course)

		paid classes, internet, nursery, higher education interest, romantic, free time, going out with friends, alcohol consumption, health status, absences, first, second and third period grades.		
4	Osmanbegović and Suljić [40]	Gender, family, distance, high school, GPA, entrance exam, scholarships, time, materials, internet, grade importance and earnings	Naïve Bayes Neural Networks Decision Trees	257
5	Khasanah and Harwati [56]	gender, origin, father education, father occupation, mother education, mother occupation, senior high school type, senior high school department, senior high school final grade, attendance and GPA	Bayesian Network, Algorithm Decision Trees	178
6	Khan et al [47]	Student marks in SSC-I, final grade in SSC-II and number of students	J48-Decision Tree	1500
7	Shahiri, et al [20]	Internal assessments, psychometric factors, external assessment, CGPA, student demographic, high school background, scholarship, social network interaction, extra-curricular activities	Decision Tree Neural Networks Naïve Bayes K-nearest Neighbour Support Machine Vector	Not Indicated

		and soft skills		
8	Agrawal, et al [12]	Student's grade in secondary education, living location and medium of teaching.	Neural Networks Bayesian Network Algorithm	80
9	Guo, et al. [19]	Not Assigned	Deep Learning Neural Network	Not Assigned
10	Sharma and Vishwakarma [67]	roll number, name, assignm1, assignm2, midsem1, midsem2 and final performance of the students in that semester	ID3 Decision Tree	70
11	Kabakchieva [30]	Gender, birth year, birth place, living place and country, type of previous education, profile and place of previous education, total score from previous education, university admittance exam and achieved score, total university score at the end of the first year and number of failures	Decision Trees Neural Network K-Nearest Neighbor	10067
12	Asif, et al [8]	4th year grade, HSC examination total marks, HSC examination mathematics marks, marks for units: MPC, CT-153, CT-157, CT-158, HS-205/206, MS-121, CS-251, CS-252, CT-251, CT-254, CT-255, CT-257, EL-238	Decision Trees	347

		and HS-207		
13	Kaur and Singh [68]	gender, hometown, family income, previous semester grade, attendance, medium(language) and senior secondary grade, seminar performance and sports.	Naïve Bayes J48 Decision Tree	52
14	Lin [42]	gender, state, citizenship, academic major, ethnic group, age, student aid, family contribution, financial need, loan received, awarded scholarship and cal grant receiver	Decision Trees Decision Rule Learning algorithms Lazy Instance-based Nearest Neighbor algorithms Function-based algorithms Naive Bayes method Bayesian Networks	5943
15	Nichat and Raut [21]	Not Assigned	Decision Tree	Not Assigned
16	Sundar [24]	student id, name, quota in which student joins, previous semester performance, performance in internal exam, performance in seminars, assignment, attendance, co-curriculum activities and end of semester marks	Bayesian Network Classifiers	48

17	Bhardwaj and Pal [17]	sex, student category, medium of teaching, student food habit, student other habit, living location, hostel, family size, family status, family income, students grade in senior secondary education, student's college type, father's qualification, mother's qualification, father's occupation, mother's occupation and grade obtained in BCA	Bayesian Network Classifiers	300
18	Goker and Bulbul [69]	Not Assigned	Naïve Bayes	220
19	Gadhavi and Patel [70]	average of unit test and sessional examination marks	Univariate Linear Regression	181

### 2.3.3 Gap Analysis

In effect, many studies have been carried out on the topic of prediction of students' academic performance in secondary schools using machine learning and data mining techniques [9] [10]. However, majority of these studies have been carried out in the developed countries mostly in the European countries such as Portugal [9] [10]. To the best of our knowledge none of these related studies has ever been carried out in Kenya which was the focus of the current study. The differences in terms the academic environments between the developed and the developing world make it difficult to domesticate the research findings from such studies to the current study. The two differ in terms of: (i) Education systems of study – whereas in Kenya, secondary school

education consists of four years of schooling preceding eight years of primary education, in Portugal and some other European countries, secondary education consists of three years of schooling preceding nine years of basic education. In Nigeria, secondary school education consists of six years preceding six years of primary education [9] [10] [11]. (ii) Grading system – Kenya uses an expanded letter grade ranging from A to E as follows: A is expanded to A, A-; B is expanded to B+, B, B-; C is expanded to C+, C, C-; D is expanded to D+, D, D- and; E. These grades are based on a numeric 12-point scale where A is equivalent to 12 points representing excellent and, E is equivalent to 1 point representing poorest, contrary to this, in European countries such as Portugal, France or Venezuela they use a 20-point grading scale where 0 is the lowest score and 20 is the highest score. The grading system in Nigeria consists of nine grades for each subject, these include distinction grades - A1, A2, A3; credit grades - C4, C5, C6; pass grades - P7, P8 and Failure grade - F9 [9] [10] [11]. (iii) Mode of evaluation in Kenya is an exit examination called Kenya Certificate of Secondary Education (KCSE) administered at the end of the four years secondary school schooling whereas in Portugal, France, Venezuela and other European countries, students are evaluated in three periods during the years of schooling and the last evaluation corresponds to the final grade, in Nigeria, students sit for the General Certificate of Education Examination (GCE) also called Senior Secondary Certificate Examination (SSCE) or the Ordinary Level Examination at the end of the six years secondary school period [9] [10] [11]. (iv) Although there is a vast publications on students' academic performance, however, in most of the literature reviewed do not constitute a conclusive list of students' attributes that may potentially influence their performance and the quality of the prediction model [9] such as students' social and cultural characteristics. (v) This study also postulates that due to differences in terms of environmental, political, social and cultural characteristics, students' attributes

may differ considerably from one place to another and thus the need to conduct fresh empirical findings [9] [10]. (vi) In other studies, on prediction of students' academic performances in secondary schools, the predictions were based on performance in a single subject such as mathematics which could introduces biasness in the study if generalized since some students may be good in certain subjects and poor in other subjects and hence the subjects don't all carry equal weightings [9].

## **2.4 Theoretical Framework**

A theoretical framework is described as a way of seeing and understanding the phenomenon being studied [71]. There are several strategies and methods that can be used to build theory. These strategies and methods are informed by theoretical assumptions about what makes for knowledge [71]. According to Jamal et al [72] there are two cycles in theory building. One cycle involves identification of the units, also called variables or concepts, of theory whose interactions constitute the theory and, the other cycle involves determining the boundaries in which the theory operate within. Lynham [73], identified four steps as necessary for the development of theory for performance. The steps include: description of the units of the theory, specification of the laws of interaction of the theory, determination of the boundaries of the theory and identification of the system states of the theory [73]. This study applied the two cycles in developing a conceptual framework. The first cycle involved identifying the units or factors that affect students' academic performance, this is discussed in depth in the subsequent subsections. In the second cycle on determining the boundaries under which the theory should operate, the study was limited to the secondary school student's environment in Kenya which is a developing country where this study was conducted.



## **2.4.1 Identifying the Factors Affecting Students' Academic Performance**

Previous studies into the factors affecting academic performance of students can be grouped into two main approaches. The first approach include those studies that have conceptualized the factors affecting student academic performance in the form of theories [72]. On the other hand, there are those studies that research on student academic performance in relation to the factors that influence academic performance. In this section, we will discuss the factors based on these two approaches.

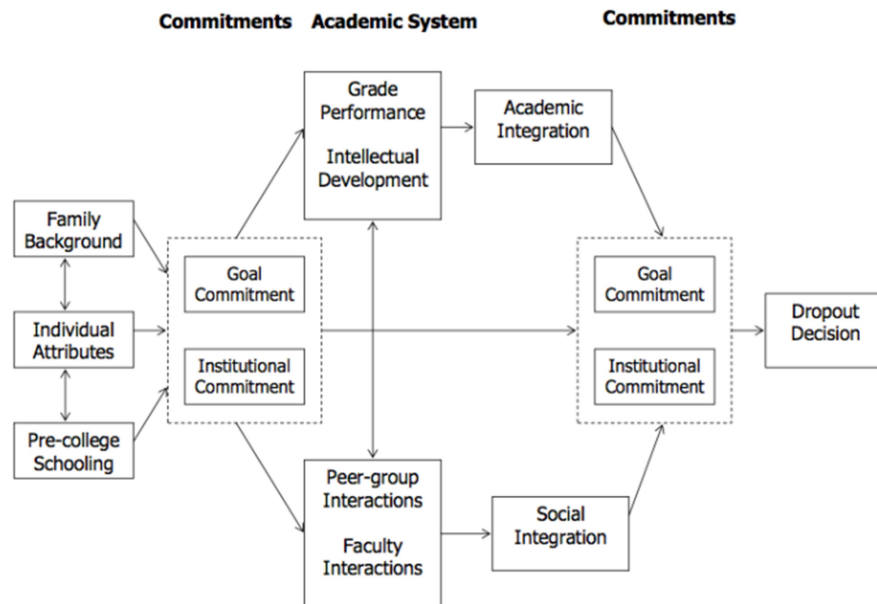
### **2.4.1.1 Theoretical Perspective on Factors Affecting Students' Academic Performance**

Various ground breaking theories in the area of student performance and attrition have been proposed by various scholars, among them Tinto's integration theory [74] which is the most cited model on student departure and performance. Three common theoretical frameworks regarding students' academic performance are discussed: Tinto's Longitudinal Theory of Institutional Departure, Bean's Longitudinal Student Attrition Model and Ogude, Kilfoil and Du Plessis student academic development and excellence model (SADEM). Each of these theories provides conceptual underpinnings for the literature on prediction of students' academic performance.

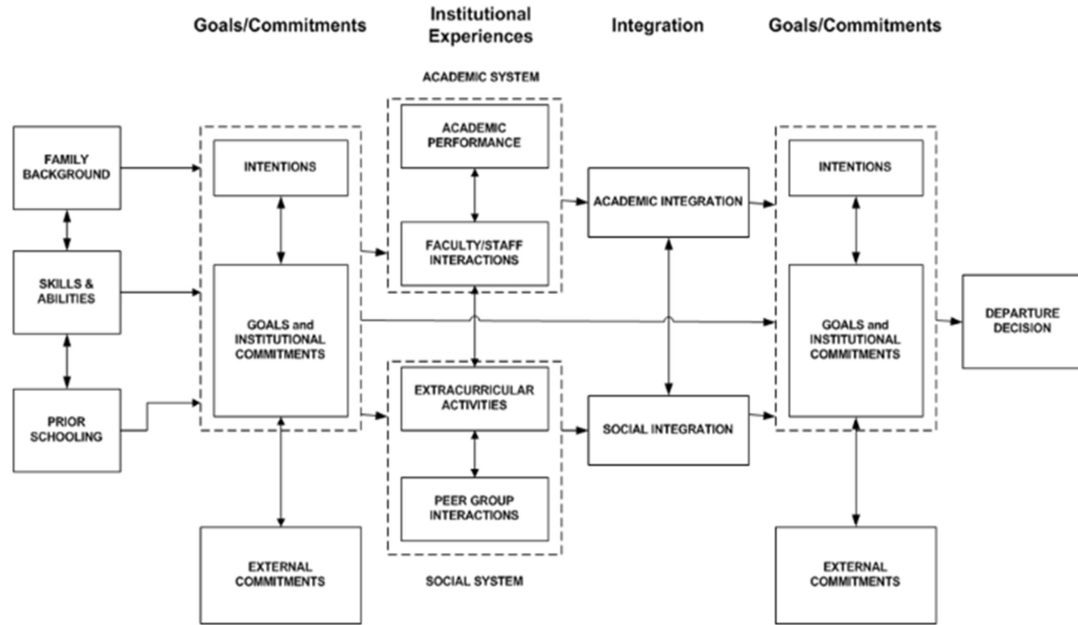
#### **2.4.1.1.1 Tinto's Integration Theory**

Tinto's model [74] provides a theoretical framework for understanding students' academic performance and departure from higher education prior to completion, it is among the most widely cited theories in education [74]. The theory drew from Spady's sociological theory which is based on a basic assumption that student failure or dropout from institution is best explained by an interaction process between individual student and the institutions environment [72]. This process is linked to the factors that promote

academic integration such as academic potential, grade performance and factors that promote social integration of the students such as family background and peer support. Tinto's theory makes an assumption that student failure to complete study is as a result of the failure by the student to sufficiently integrate into the different aspects of the institution such as academics or social systems. According to Tinto's theory, the factors that affect student performance can be summarized into three broad categories: factors related to academic integration which include student academic progress, intellectual development and lecturers' committed to teaching and helping students; factors related to social integration which include student's self-esteem and the quality of relationship with other students and lecturers and; factors related to students' pre-entry attributes which include family background, academic ability, sex, race and prior.



**Figure 2.4.1.1 Tinto Conceptual Schema for Dropout in College [39]**



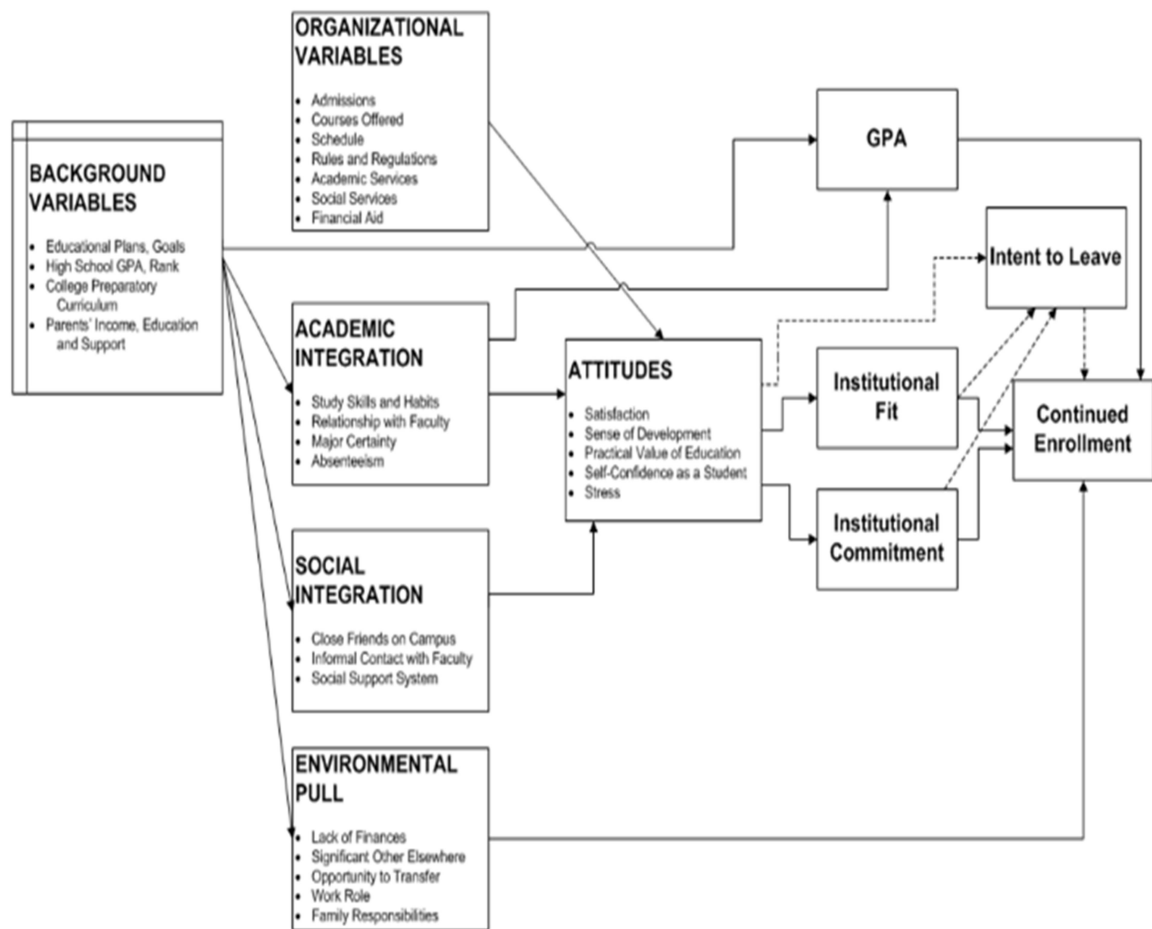
**Figure 2.4.1.2: Tinto's Longitudinal Model of Institutional Departure [75]**

However, Tinto's theoretical model requires validation when being applied at various types of institutions because it only describes departure process from an institution and not departure from higher education system [74] [75].

#### **2.4.1.1.2 Bean's Longitudinal Student Attrition Model**

Bean's model [76] presents an improved Tinto's model [75] of student failure to successfully complete studies. The theory advocates for integration of the student background characteristics into Tinto's model [75] in order for the students to understand their integration into a new institutional environment. Bean's model [76] asserts that the most influential factors that affect student progression or student attrition are the external environment factors such as family responsibilities, finances and encouragements and not the social integration factors such as university memberships and friends [74] [75] [76]. Bean's model [76] further includes the students sociological aspects such as

background characteristics, academic integration and social integration of the student with the institution, work responsibilities and family responsibilities, economic aspects such as student finances, organizational aspects such as admissions criteria, rules and regulations, academic advising, course scheduling and offering, and financial assistance, and psychological aspects such as students' attitudes, self-beliefs and academic intent.



**Figure 2.4.1.1.2: Bean Longitudinal Student Attrition Model [76]**

### 2.4.1.1.3 Ogude, Kilfoil and Du Plessis student academic development and excellence model (SADEM)

In an attempt to improve student retention, performance, and throughput rates at the University of Pretoria, Ogude, et al. [77] developed the student academic development and excellence model (SADEM). The model targets all years of undergraduate study while prioritizing the first year. To improve student retention, this model starts with identification of three organizational sub-levels and associated projects. These include institutional readiness projects which includes a teaching charter, an early warning system, student finance and academic promotions; faculty readiness projects which includes the educational model, faculty academic culture and student success, and resources for large classes; student readiness projects which includes collaboration with feeder schools and the design of survey instruments to determine academic readiness, effective mentoring and tutorial support.

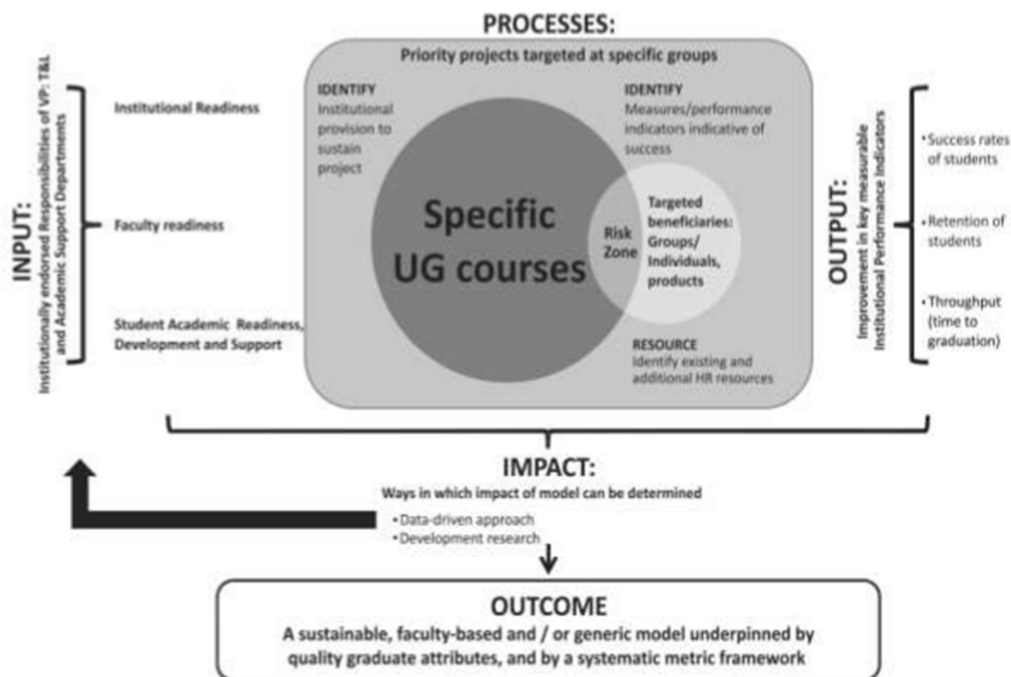


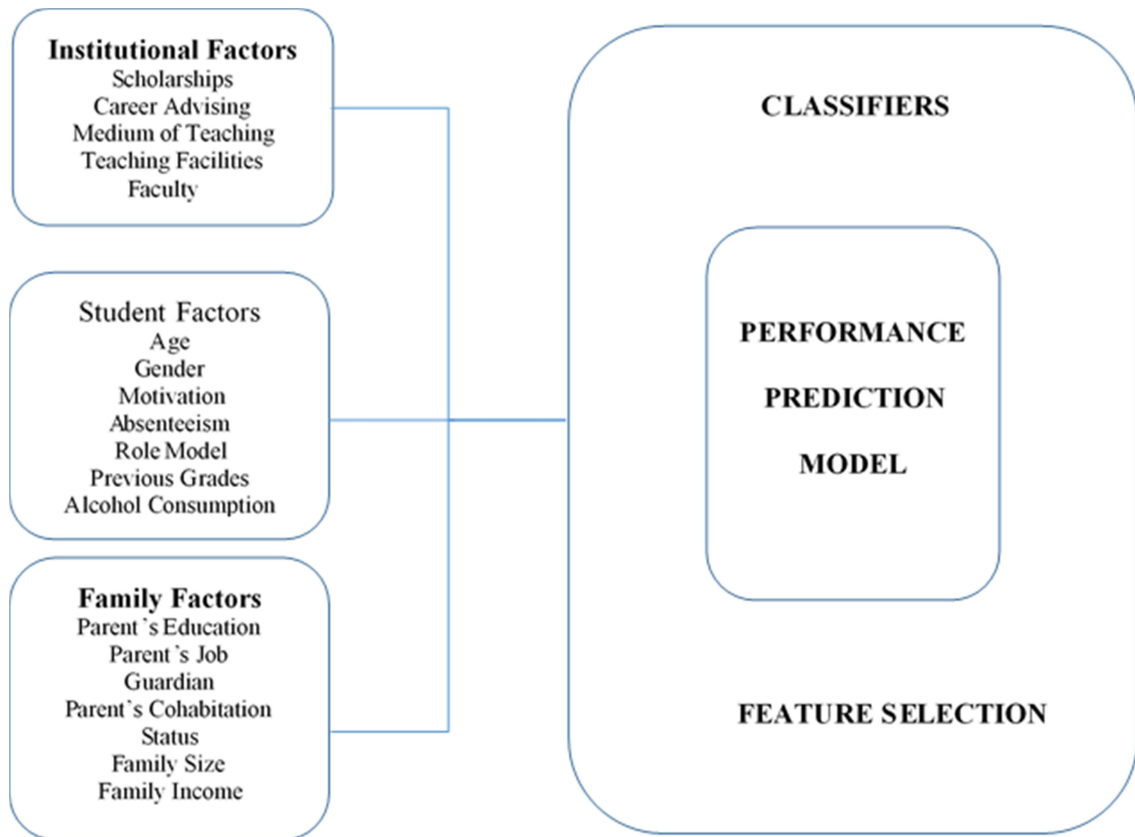
Figure 2.4.1.1.3 Student academic development and excellence model [77]

#### **2.4.1.2 Summary on Theoretical Perspective**

It is evident from the three theorist discussed above that factors affecting student academic performance are quite diverse and often differ from one theory to another. Tinto's theory insists on academic and social integration factors, and pre-entry characteristics as the most influential on student academic performance. Bean's model identified student background characteristics and other external environment factors such as family responsibilities, finances and encouragements as the most influential factors affecting student performance. The student academic development and excellence model (SADEM) by Ogude, et al. [77] identified institutional-related factors, faculty-related factors and student-related factors as the most influential factors affecting student performance. Tinto and Bean's theories share common belief that pre-admission factors such as family background and prior schooling have significant influence on student academic performance.

#### **2.4.2 Conceptual Framework**

In developing a conceptual model, the study drew much from the three categories of factors that affect student academic performance as identified by Ogude, Kilfoil and Du Plessis student academic development and excellence model (SADEM) [77] ,they include institutional-related factors, faculty-related factors and student-related factors, and also it was guided by literature review in section 2.3.1. Figure 2.4.2 shows the conceptual framework for the study.



**Figure 2.4.2 Conceptual Framework**

As shown in figure 2.4.2, the factors affecting student academic performance have been categorized into three namely; institutional factors, student factors and family related factors. Some selected factors from each category are presented next.

### **Institutional Related Factors**

Several institutional factors were identified as being influential on student academic performance. Improving these institutional factors can help students in the learning and make a more accommodative living environments that boost academic success. These include student scholarships or financial aid, student academic support through career advisors, learning facilities such as teaching laboratories, faculty-to-student interactions, medium of teaching and co-curriculum activities.

### **Student Related Factors**

Students play a critical role in determining their academic performance. Several factors that relate to students include age, gender, student interest and motivation, role models, absenteeism, previous examination grades, alcohol consumption, social network interactions.

### **Family Related Factors**

These are the external factors that the student joined the institution with. They include father's education, mother's education, father's job, mother's job, guardian, parent's cohabitation status, family size, family income and marital status.

## **2.5 Summary of Literature Review**

This chapter started by discussing the process of building a machine learning model for prediction of students' academic performance. This was followed by a discussion on the different types of machine learning and machine learning techniques used for academic performance prediction, data representation, feature selection and evaluation of machine learning models. Finally, a review of previous work on prediction of students' academic performance was presented followed by theoretical frameworks and the conceptual framework used to develop the prediction model for students' academic performance.



## **CHAPTER THREE**

### **RESEARCH METHODOLOGY**

#### **3.1 Introduction**

Research methodology is the systematic approach to carrying out a research, it consists of the theory and basis of philosophical assumption that form the foundation of how to conduct research [78] [79]. The purpose of research methodology is to guide the researcher on how to proceed from the findings of empirical research to make inference about the truth [80]. This chapter therefore gives a description of the research methodology used. We present a description of the research philosophy adopted, research design, sampling procedure and sample size, research instruments used, data collection procedures and data analysis.

#### **3.2 Research Philosophy**

According to Ponterotto [81], the selected research philosophy guides the researcher in making philosophical assumptions about the research and the selection of tools, instruments, participants, and methods used in the study. Research philosophy gives the direction on how to carry out research. There are four main research philosophies: positivism, pragmatism, realism and interpretivism. These philosophies are characterised through their ontology, epistemology and methodology [82]. As shown in Figure 3.2, there is need for the researcher to position research philosophy at the initial stages of the research.





**Figure 3.2: The Research Pyramid, 2010**

**Source: Jonker and Pennink [83]**

The research philosophy for this study was positivism. Positivism philosophy is based on the notion that research can be objective and that the researcher is independent. Its primary goal of inquiry is an explanation that ultimately leads to prediction and control of phenomena [81]. Positivism research paradigm utilizes mainly quantitative techniques and mostly the research design is experimental [84]. According to Jonker and Pennink [83], the philosophical approach has effect on the choice of research design and research methods to be used to find solutions to the research questions. Therefore, in choosing the research philosophy, the researcher reflected through the methodological paradigms available, their relevance to the research problem and compatibility with the research design.


### **3.3 Research design**

The research design for this study was experimental research design. Experimental research design has been extensively used in the natural sciences [85] [86]. It provides a solid foundation for advancement in the hard sciences [85] and are often touted as the

most rigorous research design setting the gold standard against which other designs are judged with respect to internal validity [86]. In experimental research design, the researcher is allowed to control different situations, and are preferred in research where there is time priority in a causal relationship. There are five main types of research designs, they include experimental design, cross-sectional or social survey design, longitudinal design, case study design and comparative design. Experimental research design was used due to its suitability to the research problem.

According to Saunders, Lewis and Thornhill [79], a research design is used to represent a framework for data collection and data analysis. A study conducted by Levy and Ellis [85] on experimental and quasi-experimental studies grouped experimental design into four research categories: the lab experiment also known as true-experiment, the quasi-experiment also known as the field-experiment, the factorial design and the ex-post facto design. They also noted that the common types of experimental designs for experiments are the pretest-posttest with control group design and the Solomon four-group design [85]. According to the study, the researcher in the pretest-posttest with control group design randomly assigns the participants to two groups; the experimental group and the control group. The experimental group then undergoes the prescribed treatment while the control group serves as the benchmarking point of comparison and receives no treatment at all as shown in Figure 3.3. The study noted that the pretest-posttest with control group design was better in controlling threats to internal validity.

**Figure 3.3: Pretest-posttest with control group design**

		Time (t) →		
		t <sub>1</sub>	t <sub>2</sub>	t <sub>3</sub>
		Measure	Treatment	Measure
Randomly Assigned	Group A (The Experimental Group)	M <sub>A<sub>t1</sub></sub>	T <sub>x</sub>	M <sub>A<sub>t3</sub></sub>
	Group B (The Control Group)	M <sub>B<sub>t1</sub></sub>	-No-	M <sub>B<sub>t3</sub></sub>
In an ideal case – desired observed differences		No Diff	-	Sig. Diff
In an ideal case – graphical representation				

**Source: Levy and Ellis [85]**

### 3.4 Location of Study

The study was carried out in five public tertiary institutions located within the republic of Kenya. The institutions included: Masinde Muliro University of Science and Technology, Nairobi Technical Training Institute, St. John’s Kilimambogo Teachers Training College, Kenya Medical Training College - Machakos Campus and Sigalagala National Polytechnic.

### 3.5 Target Population

Target population refers to the entire group to be studied within a definite area. The target population was secondary school form four leavers who are currently in tertiary institutions. For the purposes of the study, the respondents comprised of recent students from secondary school who studied and completed their secondary school studies in Kenya.

### 3.6 Sampling Techniques

The study used stratified sampling technique to select the five categories of public tertiary training institutions namely: University, Technical Training Institute (TTI), Kenya Medical Training College (KMTC), Teacher Training College (TTC) and National Polytechnic. Each category was equated to a stratum. Purposive sampling

technique was used to select one institute from each category to participate in the study (see Table 3.7). In the category of universities, Masinde Muliro University of Science and Technology was selected; in the category of technical training institutes, Nairobi Technical Training Institute was selected; in the category of Kenya medical training colleges, Kenya Medical Training College - Machakos Campus was selected; in the category of teacher training colleges, St. John's Kilimambogo Teachers Training College was selected and in the category of national polytechnics, Sigalagala National Polytechnic was selected. Finally, random sampling was used to select the respondents to take part in the study from each sampled institute. Random sampling was used so as to ensure the study establish useful target population and achieve external validity [87].

In selecting the choice of the sampling techniques, the study borrowed from other previous related studies. According to Khalid et al [88], it may be impossible for the researcher to study the whole population of interest, hence researchers use sample which is a subset of the population then generalize the research findings. The study noted that although there are several sampling techniques that can be used to get a sample, the choice of the right technique is dictated by the nature of the study and the specific research questions to be addressed [88]. The commonly used sampling techniques include random sampling, systematic sampling, stratified sampling, and cluster sampling. In random sampling, the researcher selects a sample at random from the entire population.

Stratified random sampling divides the data into different strata on the bases of factors available such as income levels etc. and a random sample is then drawn from each stratum. Purposive sampling technique is a type of non-probability sampling used in

qualitative and quantitative research techniques and is considered one of the most effective technique when studying a certain cultural domain with knowledgeable experts within [89]. The technique is also referred as judgment sampling in some literature. In purposive sampling, the researcher makes an independent decision on what needs to be investigated by virtue of acquired knowledge or experience and does not require any underlying theories [89]. This ensures the quality of data gathered is maintained as well as reliability and competence of the informant [89].

### 3.7 Sampling Size

Cochran's Sample Size Formula was used to get the required sample size for this study. The Cochran formula is used to calculate an ideal sample size given a desired level of precision and confidence. The Cochran's  $n_0 = \frac{Z^2 pq}{e^2}$  Sample Size Formula is given as: Where:  $n_0$  is the desired sample size,  $p$  is the (estimated) proportion of the students whose secondary school performance will be predicted,  $q$  is  $1 - p$  denoting students whose performance will not be predicted and  $e$  is the desired level of precision or degree of accuracy.

The study used a 95% confidence level for student's sample which corresponds to a standard normal deviation of 1.96 from the Z-table,  $p = 0.5$  (50%) and  $q = 0.5$  (50%) and  $e = 5\%$  (0.05). substituting these values on the Cochran formula we get 384.16 as the sample size computed as follows:

$$n_0 = \frac{(1.96)^2(0.5)(0.5)}{(0.05)^2} = 384.16$$

Since the sample size involves students, the researcher rounded-off 384.16 to the next whole number to get 385 respondents. Therefore, the sample population using the Cochran formula should have at least 385 students. Table 3.7 shows the target population for the study.

**Table 3.7 Target Population**

<b>Category of Institution</b>	<b>Institution</b>	<b>Target Number</b>
University	Masinde Muliro University of Science and Technology	385
Polytechnic	Sigalagala National Polytechnic	385
Medical Training College	Kenya Medical Training College - Machakos Campus	385
Teacher Training College	St. John's Kilimambogo Teachers Training College	385
Technical Training Institute	Nairobi Technical Training Institute	385
<b>Total</b>		<b>1,925</b>

### **3.8 Instruments of Data Collection**

According to Birmingham et al [90], research instruments are devices for obtaining relevant information to a research project. This study used questionnaire as the data collection instrument. A questionnaire is an effective and efficient research instrument of eliciting information from individuals regarding their views and opinion on particular research issues. It is widely used in research because of its ability to provide cheap and effective way of collecting data from respondents in a structured and manageable way [90]. The suitability of the questionnaire as the research instrument for this study was informed by a number of factors with the key consideration being that the study required information on a range of subjects which necessitated the need to use questionnaires as the most convenient method. According to Birmingham et al [90], questionnaire are found to be more suitable for study that require information on a range of subjects that require to ask respondents questions. Again, compared to other forms of collecting data such as interviews, content analysis, focus groups and observation, questionnaires are



usually inexpensive to administer; easy to develop, easy to analyse especially closed-ended questionnaires and can be sent simultaneously to a great number of respondents hence time-efficient [91].

The questionnaire consisted of a total of 35 questions printed in 4 standard A4 size sheets (see Appendix I). The questions were further divided into 6 sections where section 1 consisted of general questions (such as type of institution, date etc.), section 2 was on students' demographic attributes, section 3 collected data on students' family information, section 4 collected data on co-curriculum information, section 5 collected data on secondary school students' academic performance and section 6 collected data on secondary school demographic features. All the instruments that were used to collect data are attached in Appendix I. This was a structured questionnaire that consisted of both closed and open questions.

### **3.9 Validity and Reliability of Research Instruments**

#### **3.9.1 Validity of Research Instrument**

According to Kimberlin et al [92], validity is the truthfulness of the research findings or the degree to which the research instrument measures what it purports to measure. This study used face validity and content validity to measure validity of the research instruments. According to Mohajan and Haradhan [87], content validity is used to assess the degree to which the research questions on the research instrument and the scores obtained from these questions are a representative of all possible questions that could have been asked about the content. Since there is no statistical test to determine whether a measure adequately covers a content area or represents a construct adequately, content validity usually depends on the judgment of experts in the field. Face validity refers to

the extent to which a test appears to measure what it expected to measure [87]. Face validity depends entirely on the assessor’s level of expertise and familiarity concerning the subject matter, the expert can describe the appearance of validity without empirical testing.

The research tools in this study were subjected to three domain experts to assess the content validity and also three domain experts to measure face validity of the tools used. The assessments were on a scale of 1-10 with 1 being poor and 10 being excellent. Average of the ratings from the three experts were computed and a verdict taken. The instruments are deemed to be valid if the rating is 0.6 (60%) and above. The average rating for face validity was 73% and content validity was rated at 70% as tabulated in Table 3.9.1.

**Table 3.9.1: Validity of Research Instruments**

<b>Expert</b>	<b>Face Validity Score</b>	<b>Content Validity Score</b>	<b>Decision Taken</b>
Valid Expert I	7	6	Accepted
Expert II	7	7	Accepted
Expert III	8	8	Accepted
<b>Average</b>	<b>7.3 (73%)</b>	<b>7.0 (70%)</b>	

Validation of research instrument in quantitative research focuses on how to reduce errors in the measurement process and ensures that the research tool measures what it was designed to measure. Validity of a research instrument in qualitative research measures how a researcher uses certain procedures to check for the accuracy of the

research findings. In general, validity of a research instrument is used to check credibility and transferability (replicability) of the study. According to [87], external validity (transferability validity) can be improved through strategies such as using random selection and heterogeneous groups to select the sample representation of population. Other types of validity including; construct validity, criterion-related validity, convergent validity, concurrent validity, predictive validity and discriminant validity.

### **3.9.2 Reliability of Research Instrument**

Mohajan and Haradhan [87], defined reliability as the ability of a measure to remain the same. Although there are many different ways of estimating the reliability of research measures, this study used pilot testing of research instrument and Cronbach Alpha to gauge the reliability of the questionnaire. Cronbach Alpha is used to measure the internal consistency (or reliability) of the measuring instrument such as a questionnaire. The purpose of the pilot study was to identify any weaknesses, if any, in the questionnaire used in this study. Use of a pilot testing or pretesting of research instrument is widely used as a means of identifying sources of errors in an instrument and refining the measure to minimize the effects of the error [92]. The questionnaire was first reviewed by supervisors and professionals then piloted on a small set of 15 university students from Technical University of Mombasa to test the ease of use, clarity and readability. Then reliability test was conducted within SPSS in order to measure the Cronbach Alpha internal consistency of the questionnaire. The Cronbach Alpha value was found to be 0.743 and Cronbach's Alpha Based on Standardized Items was 0.729 as shown in Table 3.10. For a questionnaire to be considered reliable, the Cronbach Alpha value must be more than 0.6, hence for this study, the questionnaire was rated above 0.6 hence considered reliable.

**Table 3.10: Reliability Statistics**

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items
0.743	0.729

In research, reliability estimates are used to evaluate the stability of measures, internal consistency of measurement instruments, and inter-rater reliability of instrument scores [92] [87]. According to the classical test theory on reliability [92], any score obtained by a measuring instrument comprises of the true score which would have been received if the measurement were perfectly accurate of which is unknown and, the error involved in the measurement process. It is therefore upon the researcher to correctly identify the sources of measurement errors that are deemed most detrimental to the usefulness of result interpretation

### **3.10 Data Collection for Prediction Model Development**

Data collection is the first step in prediction model development which involves collecting raw data from the respondents. This process involved identifying the data sources, collecting the data and digitising the raw data.

#### **3.10.1 Data Sources**

The data was collected from recent secondary school form four leavers who were enrolled in tertiary institutions namely: university, polytechnic, teacher training college, medical training college and technical training college. The data represented information on student demographic data, family data, socio-economic details, previous academic performance and other environmental factors at secondary school level of study. Appendix VII gives a sample of the data collected in CSV format.

### **3.10.2 Data Collection Procedure**

The data collection exercise was carried out between the months of January 2019 and April 2019. This exercise involved two phases: phase one started with the researcher first obtaining authorization from relevant government bodies and offices before commencing on data collection. After obtaining the letter of approval (see appendix II) from the university, the researcher personally travelled to the respective offices to get the required letters and permit. First was a research authorization letter (see appendix III) and a permit (see appendix IV) to carry out research from the National Commission for Science, Technology and Innovation (NACOSTI), then letters of authorization from the County Commissioner and County Director of Education in each of the counties sampled. Finally, an introductory letter was presented to the head of each institutions selected to obtain authorization to collect data.

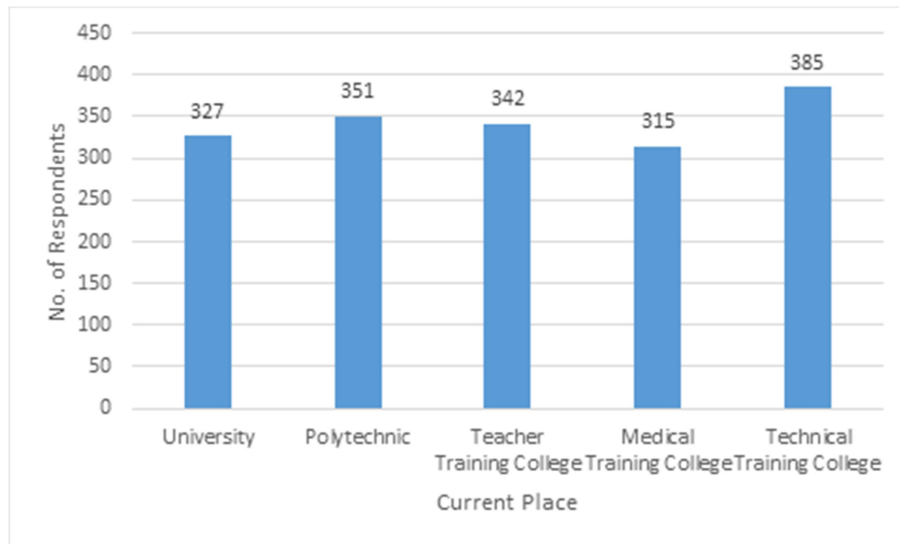
Phase two involved delivery of questionnaires to the institutions, administering of the questionnaires to respondents and collection of the same. The researcher used research assistants to administer and collect the questionnaires from the respondents. The exercise took a period of one month.

### **3.11 Data Analysis**

Prediction modelling development is interested with selecting the best combination of predictors to be included in a model in such a way that makes the predictions as accurate as possible [93]. Prediction model development is therefore not interested with unravelling casual associations between the predictors and the outcomes. Whereas inferential statistics is used when the researcher want to infer the behaviour of the entire population from a subset of sample data [94], the goal of using descriptive knowledge (or statistics) in this study was to give a simple description that summarizes the data [95].

Statistical Package for Social Science (SPSS) software was used to give a simple description that summarizes the data. SPSS comes with a powerful suite of tools for statistical and data analysis.

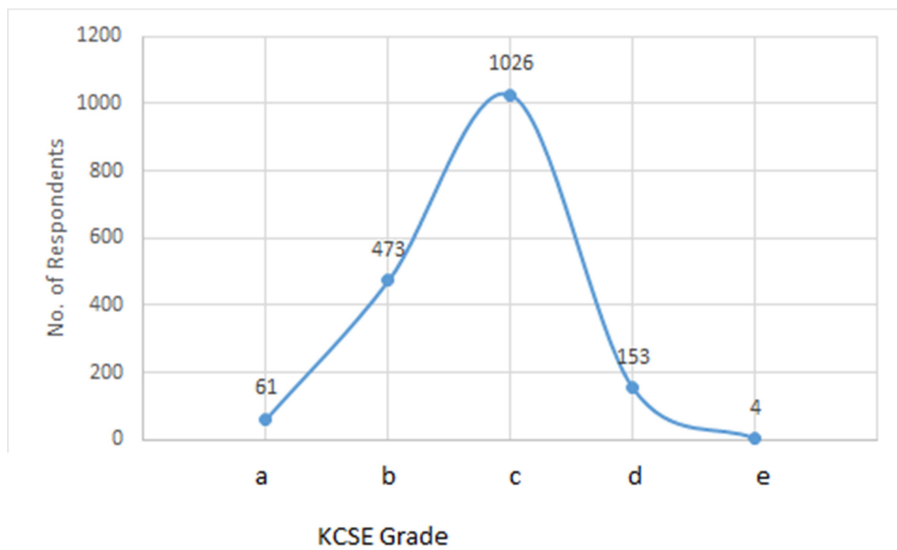
The total number of valid instances collected was 1720 out of 1925 as shown in figure 3.11.1. The training data was later extended to 5,199 instances using data augmentation techniques.



**Figure 3.11.1: Number of respondents per Institutions**

The distribution of the KCSE grades scored by students that were sampled is displayed in figure 3.11.2.

**Figure 3.11.2: Grade Distribution**



Mji & Glencross [96] described data analysis as a process that involves pre-processing of data collected to manageable proportions either through some techniques such as feature selection and, identification of patterns and themes in the data. In this study, machine learning process was used to infer about the predictive capabilities using WEKA (Waikato Environment for Knowledge Analysis) tool. WEKA is the most popular suite of machine learning software [97]. It is an open-source software that was developed in New Zealand by Waikato University. It provides a collection of data mining and machine learning algorithms and stores data in a flat file format called ARFF (Attribute Relation File Format). WEKA is used under the GNU license for knowledge analysis and implements almost all machine learning algorithms. The machine learning algorithms used to build the models in this study were naïve bayes classifier, decision tree J48 classifier and neural network - multilayer perceptron classifier.

### **3.12 Machine Learning Methodology for Developing Prediction Model**

Machine learning methodology follows the study the machine learning process to develop prediction models. The methodology provides a structured approach to developing prediction models. Machine learning process is made up of six steps: creating the data set, data pre-processing, feature selection, training models, model selection and final model. Figure 3.11 shows the processes that are followed to develop machine learning models.

**Stage One - Student data set:** This phase deals with raw data collection from various data sources. In this phase, data received from the respondents is digitized and proof read to ensure completeness.

**Stage Two - Data pre-processing:** Digitized data is converted to the desired study data set for machine learning. The data is then input into the machine learning tools for further analysis. Incomplete or missing data are handled at this stage.

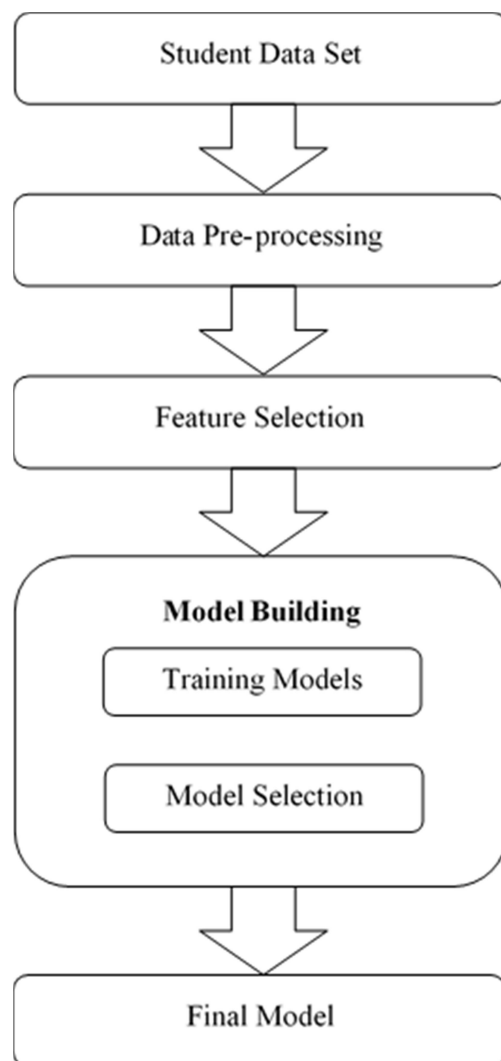
**Stage Three - Feature selection:** also known as dimensionality reduction, is used find the optimal feature subset by removing irrelevant features. This ensures that the resulting model is not overly complicated due to too many features. The objective of this process is to increase prediction accuracy.

**Stage Four - Training models:** also known as learning algorithms, involves training machine learning algorithms using optimal feature subset to build candidate prediction models. The models are trained by going through the training data iteratively.



**Stage Five - Model selection:** This is the last phase in model development and involves selecting the final model. Successive modeling is used to train the models iteratively using the optimal feature subset to get the best performing model.

**Stage Six - Final Model:** After the process of training and selecting models, the final model is then presented. The Predictive Toxicology Mark-up Language (PTML) was used to present the final model.



**Figure 3.11 Machine Learning Processes [98]**

### **3.13 Model Validation**

The study performed two types of model validation: internal validation and external validation. To test for internal validity of the model, the study used 10-fold cross-validation technique. Cross-validation has over the years been used as a standard way of evaluating the performance of machine learning algorithms due to its ability to reduce the variance. 10-fold cross validation works by generating many models and then taking an average of all the models other than relying on a single model,

External validity helps in making the final model generalizable by ensuring the model not overfitted on the training data. In this study, external validation and reliability of the modelling algorithm was evaluated using the following validation metrics: accuracy, precision, recall, F-measure, specificity, sensitivity and AUC - ROC curve. Confusion matrix, also called error metrics, was used. This matrix is presented in a table layout and is used to visualize the performance of a machine learning algorithm. The confusion matrix is used to derive the other measures such as accuracy, precision, recall, F-measure, specificity and sensitivity.

According to [99], model validation is substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model. The validation often requires several sets of experimental conditions to define the domain of model's intended applicability. For a model to be considered valid, the prediction accuracy must be within the acceptable range. Determining whether a model is absolutely valid is touted to be costly and time consuming, however, the best way to determine validity is by conducting tests and evaluations until sufficient confidence is obtained [99].

### **3.14 Ethical Considerations**

The ethical issues that were addressed in this study were; privacy and confidentiality, access and acceptability, and informed consent. In addressing the issue of access and acceptance, the researcher obtained approval from the Directorate of Postgraduate Studies of Masinde Muliro University of Science and Technology (see Appendix II), then secured a research permit from the National Commission for Science, Technology and Innovation (NACOSTI) (see Appendix IV) and a letter for research authorization (see Appendix III). Then permission was sought from the County Commissioners and County Directors of Education in each of the counties that were sampled and finally permission was sought from all the Principals and Directors of the institutions that were visited. In terms of informed consent, all the participants were informed in the beginning about the nature and purpose of the study and also that their participation was on voluntary basis. On privacy and confidentiality, the data collected from the research was purely for research purposes and it was made anonymous by the researcher by not including unique identifiable traits about the respondents.

## **CHAPTER FOUR**

### **RESULTS AND DISCUSSION**

#### **4.1 Introduction**

The main purpose of this study was to develop a machine learning model for prediction of students' academic performance for secondary schools in the developing countries. The chapter presents the research findings, results, and discussions of the study data collection. The presentation in this chapter follows the methodology for developing prediction model as discussed in chapter three. The chapter starts with a discussion on the student data set and a summary of the respondents that took part in the study. This is followed by the data pre-processing stage. The process of selecting the optimal feature subset for the student data set and the feature selection techniques used to rank the features is presented. This is followed by development of the prediction models, and the process of selecting the final prediction model for prediction of students' academic performance. The results of the three machine learning algorithms used to create the models and the performance evaluation metrics applied on the prediction models are also presented. Finally, the chapter concludes with a discussion and summary of the results.

#### **4.2 Student Data Set**

The total number of respondents that participated in filling the questionnaires were 1925 out of which 1720 were found to be complete and valid, these were used to create the student data set. The data set consisted of 1720 records and 62 attributes. This formed that training data set. Each record, or instance, represents attributes from a single respondent. Using data augmentation techniques, the training data set was scientifically extended to 5,199 records.

### 4.3 Pre-Processing the Student Data Set

This is the first step after collecting data. It involves digitization of the raw data, checking for inconsistencies and missing values in the raw data, and data conversion to the desired format for machine learning using data pre-processing tools.

#### 4.3.1 Digitization

Initially the data from the questionnaires was captured and stored in excel worksheets. Excel tool stores data in a tabular format. Tabular data is represented in a column-row format. Each column represent a student attributes and each row represent a single student record. The data consisted of attribute name (in short form), description and domain (associated values of each attribute) as shown in Table 4.3.1.

**Table 4.3.1 Attribute Description**

No	Attribute	Description	Domain
1	Institution	Category of Institute	{kmtc,polytechnic,ttc,tti,university}
2	County	County	{machakos,kakamega,kiambu,nairobi}
3	Gender	Gender	{female,male}
4	Age	Age	{ Below 14 yrs (1), 14-18 yrs(2), above 18yrs(3)}
5	Disability	Disability	{yes,no}
6	Religion	Religion	{muslim,christian,others}
7	LP	Lived with Parents	{yes,no}
8	WPC	Witnessed Parent Conflicts	{yes,no}
9	FS	Family Structure	{singleparent,nuclear,extended,step}
10	DF	Difficulties Paying Fees	{yes,no}
11	Sponsor	Sponsor	{parents,guardian,others}
12	PE	Parents Employment	{ both(1),one(2),none(3)}
13	FE	Father's Education	{none(1),primary education(2),secondary education(3),postsecondary(4),degree and

			above(5)}
14	ME	Mothers Education	{none(1),primary education(2),secondary education(3),postsecondary(4),degree and above(5)}
15	CA	Participated in Curriculum Activities	{yes,no}
16	CF	Frequency of Participation	{1,2,3,4,5}
17	NSF1	Number of subjects in Form 1	{7,8,9,10,11,12,13,14,15,16}
18	NSF2	Number of subjects in Form 2	{6,7,8,9,10,11,12,13,14,15}
19	NSF3	Number of subjects in Form 3	{4,5,6,7,8,9,10,11,12}
20	NSF4	Number of subjects in Form 4	{4,5,6,7,8,9,10,11,12}
21	Specialization	Specialization	{yes,no}
22	YS	Year of Specialization	{1,2,3,4}
23	CS	Changed School	{yes,no}
24	RC	Repeated Class	{yes,no}
25	F1G	Form 1 Grade	{a,b,c,e,d}
26	F2G	Form 2 Grade	{a,b,c,e,d}
27	F3G	Form 3 Grade	{a,b,c,e,d}
28	MG	Mock Grade	{a,b,c,e,d}
29	LS	Learning Styles Used	{ one(1),two(2),three(3)}
30	AS	Assessment Style	{ formal(1),informal(2),all(3)}
31	ELS	Effect of Learning Style	{very low(1),low(2),moderate(3),high(4),very high(5)}
32	EAS	Effect of Assessment Style	{very low(1),low(2),moderate(3),high(4),very

			high(5)}
33	Absences	Absences in months	{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,18,24}
34	EC	Examination Challenges	{yes,no}
35	AD	Access to Drugs	{yes,no}
36	RM	Role Model	{yes,no}
37	EA	Effect of Absences	{very low(1),low(2),moderate(3),high(4),very high(5)}
38	ETA	Effect of Teacher Absenteeism	{very low(1),low(2),moderate(3),high(4),very high(5)}
39	EFS	Effect of Failure to Cover Syllabus	{very low(1),low(2),moderate(3),high(4),very high(5)}
40	ECA	Effect of Co-Curriculum Activities	{very low(1),low(2),moderate(3),high(4),very high(5)}
41	ED	Effect on Access to Drug	{very low(1),low(2),moderate(3),high(4),very high(5)}
42	EEC	Effect of Examination Challenges	{very low(1),low(2),moderate(3),high(4),very high(5)}
43	ERM	Effect of Role Model	{very low(1),low(2),moderate(3),high(4),very high(5)}
44	TS	Type of School	{national,extracounty,subcounty,county}
45	Residence	Residence	{boarding(1),day(2),both(3)}
46	SC	School Composition	{girls,mixed,boys}
47	TL	Teaching Laboratory	{yes,no}
48	Library	Library	{yes,no}
49	CL	Computer Laboratory	{yes,no}

50	Electricity	Availability of Power/Electricity in school	{yes,no}
51	Internet	Access to Internet	{yes,no}
52	STL	Status of Teaching Laboratory	{worst(1),worse(2),bad(3),good(4),better(5),best(6)}
53	SL	Status of Library	{worst(1),worse(2),bad(3),good(4),better(5),best(6)}
54	SCL	Status of Computer Laboratory	{worst(1),worse(2),bad(3),good(4),better(5),best(6)}
55	SE	Status of Power/Electricity	{worst(1),worse(2),bad(3),good(4),better(5),best(6)}
56	SI	Status of Internet	{worst(1),worse(2),bad(3),good(4),better(5),best(6)}
57	ETL	Effect of Teaching Laboratory	{very low(1),low(2),moderate(3),high(4),very high(5)}
58	EL	Effect of Library	{very low(1),low(2),moderate(3),high(4),very high(5)}
59	ECL	Effect of Computer Laboratory	{very low(1),low(2),moderate(3),high(4),very high(5)}
60	EE	Effect of Power/Electricity	{very low(1),low(2),moderate(3),high(4),very high(5)}
61	EI	Effect of Internet	{very low(1),low(2),moderate(3),high(4),very high(5)}
62	Class (KCSE)	KCSE Grade	{a, b, c, d, e}

Figure 4.3.1 shows a sample of the data set in excel format.



The image shows a spreadsheet titled 'StudentDataSet' with 62 columns and 28 rows of data. The columns are labeled with letters and abbreviations, including Q, R, S, T, U, V, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF, BG, BH, BI, BJ, BK, and KCSE. The data consists of numerical values (0-5) representing different attributes and grades. The 'KCSE' column shows the final examination grades for each student.

**Figure 4.3.1 Student Data Set**

The class attribute or predicted attribute was KCSE. It consists of five classes (also called grades) which represent the possible final examination grades a student can score. Table 4.3.2 shows the grouping of the grades into five categories. The rest of the attribute were predictor variables.

**Table 4.3.2 Class Attribute Grouping**

Values/Grades	Category/Group
A, A-	A
B+, B, B-	B
C+, C, C-	C
D+, D, D-	D
E	E

### 4.3.2 Missing Data

Each complete student record consisted of 62 attributes. However, some of the student's records were missing some attributes. However, majority of the records were complete. Out of the 1925 records, 1720 of the records were found to be complete with valid data

which makes 89.3% of the total responses. 205 records were found to be incomplete. Since the missing data represented potential predictors (attributes) for the class attribute, this rendered them incomplete and therefore unsuitable for analysis and prediction of the class attribute hence they were excluded from the training dataset.

#### **4.3.3 Data Conversion**

The student data set was then converted to the desired format for machine learning. First, the data in excel format was imported to SPSS, it was then pre-processed to CSV (comma delimited) format, then to attribute-relation file format (arff) which was the desired format for machine learning. It was finally imported to WEKA tool which provides the machine learning environment for learning the algorithms (See Appendix VII).

#### **4.4. Feature Selection**

This section presents the results of the findings of the feature selection process. The process employed to find the optimal feature subset is described. This is followed by feature ranking process using three feature selection techniques in WEKA: information-gain based feature selection technique, correlation-based feature selection technique, and one rule feature selection technique. Finally, the section ends with a discussion on the findings from the feature selection process.

##### **4.4.1 Feature Selection Process**

Feature selection is the process of selecting relevant features (also called attributes or variables) in a data set to improve machine learning results. Kira and Rendell [100] described feature selection as the problem of choosing a smaller subset of features from the feature vector that is ideally necessary and sufficient to describe a target concept.

In this study, feature selection is used to achieve two objectives: first to remove unnecessary or irrelevant attributes by assessing the relevance of the variable and secondly to address the problem of class imbalance. Firstly, if the input data contains too many variables, this may make the predictive model overly complicated instead of improving prediction performance. Redundant features degrade performance of the classifiers both in speed and prediction accuracy [100]. Secondly, class imbalance problem in machine learning may arise when the number of instances of one class far exceeds the other. Most machine learning algorithms works best when the number of instances of each classes are roughly equal. Although this problem has been a challenge in machine learning and has attracted significant research in the recent past [101], a number of techniques have been crafted in order to overcome the class imbalance problem. They include resampling, new algorithms, data augmentation and feature selection techniques [97]. In this study, feature selection and data augmentation techniques are applied to overcome the class imbalance problem.

There are two methods that can be applied to remove unnecessary and redundant features from the feature vector. The first method involves removing the features manually using domain knowledge [101]. The second method is to use a technique (feature selection technique) that ranks the usefulness of each feature based on its relevance to the class attribute. This study used both methods. Using the manual process (domain knowledge), the researcher removed two features: institution and county. The two features refer to the present location of the respondent as at the time the data was being collected thus not related to the data concerning secondary school information. The second method which involves the ranking of the features is discussed next.

#### **4.4.2 Using Feature Selection Techniques to Rank Features**

The feature selection techniques are categorized into three main categories: the wrapper methods, filter methods and embedded methods as discussed in section 2.2.10. In this study, feature selection was done using three filter methods namely: Information-gain feature selection, Correlation-based feature selection and One Rule (OneR) technique. These methods evaluate the usefulness of each feature in relation to the class attribute in order to get a better understanding of the significance of each feature.

All the experimentations were carried out in the WEKA machine learning environment. Feature selection in WEKA involves two major steps, first step involves selection of the attribute evaluator and the second step involves selection of a search method. Attribute evaluator is a technique which evaluates each attribute in the dataset in the context of the class attribute. The search method is the technique used to navigate different combinations of features in the dataset before settling on the relevant features. Some attribute evaluator techniques require the use of specific search methods. The results obtained from the experiments using the three techniques are presented next.

##### **4.4.2.1 Feature Selection Using Information-Gain Based Technique**

The first experiment involved using Information Gain Based Feature Selection technique to evaluate the significance of each attribute to the class attribute. Information Gain Based Feature Selection technique works by calculating the information gain (or entropy) for each attribute for the output variable. Attributes that contribute a higher information gain value are selected (ranked as most significant) and those with lower information gain value are considered not to add much information and are ranked as least significant. The values vary from 0 (no information) to 1 (maximum information). WEKA supports information gain based feature selection technique using the InfoGainAttributeEval attribute evaluator which uses the ranker search method.

The InfoGainAttributeEval technique was applied on the student data set, the results are presented in Figure 4.4.2.1.

**Figure 4.4.2.1 WEKA Information Gain Based Feature Selection Technique Results**

==== Run information ====			
Evaluator: weka.attributeSelection.InfoGainAttributeEval			
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1			
Relation: StudentPerformanceDataSet			
Instances: 1720			
Attributes: 60			
Evaluation mode: evaluate on all training data			
==== Attribute Selection on all input data ====			
Search Method:			
Attribute ranking.			
Attribute Evaluator (supervised, Class (nominal): 60 KCSE):			
Information Gain Ranking Filter			
Ranked attributes:	0.03417	53 SE	0.01379 45 TL
0.30719 26 MG	0.03217	30 EAS	0.01346 39 ED
0.23445 25 F3G	0.03136	32 EC	0.01344 41 ERM
0.11178 23 F1G	0.02933	44 SC	0.01321 2 Age
0.10406 24 F2G	0.0291	43 Residence	0.01303 56 EL
0.0988 12 ME	0.02899	14 CF	0.01243 36 ETA
0.09609 11 FE	0.02841	27 LS	0.01188 33 AD
0.07421 16 NSF2	0.0284	46 Library	0.01187 37 EFS
0.06065 10 PE	0.02659	47 CL	0.01133 58 EE
0.05947 50 STL	0.02482	29 ELS	0.01076 13 CA
0.059 51 SL	0.0247	17 NSF3	0.01056 7 FS
0.05705 15 NSF1	0.0243	4 Religion	0.01029 19 Specialization
0.0545 54 SI	0.0227	57 ECL	0.01027 38 ECA
0.05252 31 Absences	0.02199	59 EI	0.00858 9 Sponsor
0.05238 42 TS	0.02181	18 NSF4	0.00621 48 Electricity
0.05033 52 SCL	0.02125	1 Gender	0.00392 21 CS

0.04833	28 AS	0.02072	34 RM	0.00307	3 Disability
0.04651	20 YS	0.01963	55 ETL	0.00163	22 RC
0.04557	8 DF	0.01446	35 EA	0.00156	5 LP
0.04024	49 Internet	0.01412	40 EEC	0.00106	6 WPC
Selected attributes:					
26,25,23,24,12,11,16,10,50,51,15,54,31,42,52,28,20,8,49,53,30,32,44,43,14,27,46,47,29, 17,4,57,59,18,1,34,55,35,40,45,39,41,2,56,36,33,37,58,13,7,19,38,9,48,21,3,22,5,6: 59					

From the results of the experiment in figure 4.4.2.1, 60 attributes and 1720 instances were used in this experiment to rank the attributes using the computed information gain value. The predicted (class) attribute was KCSE grade. According to the information gain feature selection technique, the features that were ranked as top ten most influential factors in predicting student academic performance were: mock examination grade, form three grade, form one grade, form two grade, mother's education, father's education, number of subjects in form 2 (before specialization), parent's employment, status of school teaching laboratories and library. WPC (parental conflicts in the family) was ranked the least influential factors in predicting student academic performance according to information gain feature selection technique.

#### 4.4.2.2 Feature Selection Using Correlation-Based Technique

The second experiment involved using correlation based feature selection technique to evaluate the significance of each attribute to the class attribute. Correlation based feature selection calculates the correlation (Pearson's correlation coefficient) between each attribute and ranks their output. According to Doshi and Chaturvedi [102], a feature is said to be good if it is highly correlated to the class attribute and not much correlated with other features of the class. WEKA supports correlation based feature selection with the CorrelationAttributeEval technique which uses a ranker search method. The

technique selects the attributes that have a moderate-to-high positive or negative correlation (close to -1 or 1) and removes the attributes with a low correlation (value close to zero).

Using the CorrelationAttributeEval technique, an experiment was performed on WEKA environment where the technique was applied on the training data set, the results from the experiment are shown in Figure 4.4.2.2.

**Figure 4.4.2.2 WEKA Correlation-Based Feature Selection Method Results**

==== Run information ====					
Evaluator: weka.attributeSelection.CorrelationAttributeEval					
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1					
Relation: StudentPerformanceDataSet					
Instances: 1720					
Attributes: 60					
Evaluation mode: evaluate on all training data					
==== Attribute Selection on all input data ====					
Search Method:					
Attribute ranking.					
Attribute Evaluator (supervised, Class (nominal): 60 KCSE):					
Correlation Ranking Filter					
Ranked attributes:	0.0989	43 Residence	0.0483	14 CF	
0.342	26 MG	0.0789	20 YS	0.0473	39 ED
0.2841	25 F3G	0.0774	16 NSF2	0.0439	57 ECL
0.1904	8 DF	0.0769	44 SC	0.0411	53 SE
0.1763	23 F1G	0.0731	19 Specialization	0.0405	48 Electricity
0.1612	32 EC	0.0727	50 STL	0.0394	17 NSF3
0.1594	49 Internet	0.0705	51 SL	0.0377	58 EE
0.1341	28 AS	0.0703	1 Gender	0.0318	35 EA
0.1278	15 NSF1	0.0698	9 Sponsor	0.0301	37 EFS
0.1222	10 PE	0.0665	2 Age	0.0296	22 RC
0.1208	31 Absences	0.0636	54 SI	0.0281	55 ETL

0.1188	4 Religion	0.0619	42 TS	0.0273	56 EL
0.1151	34 RM	0.0599	30 EAS	0.0261	36 ETA
0.1142	47 CL	0.0553	29 ELS	0.0254	41 ERM
0.1104	12 ME	0.0529	7 FS	0.0237	3 Disability
0.1074	11 FE	0.0527	18 NSF4	0.0234	6 WPC
0.1032	24 F2G	0.0498	21 CS	0.0215	38 ECA
0.1009	27 LS	0.0496	45 TL	0.0144	40 EEC
0.1001	46 Library	0.049	59 EI	0.0115	5 LP
0.1001	13 CA	0.0486	52 SCL	0.0105	33 AD
Selected attributes:					
26,25,8,23,32,49,28,15,10,31,4,34,47,12,11,24,27,46,13,43,20,16,44,19,50,51,1,9,2,54,4 2,30,29,7,18,21,45,59,52,14,39,57,53,48,17,58,35,37,22,55,56,36,41,3,6,38,40,5,33: 59					

From the results of the experiment in figure 4.4.2.2, again 60 attributes and 1720 instances were used in the experiment to rank the attributes based on the correlation based feature selection value. The predicted (class) attribute was KCSE grade. According to the correlation based feature selection technique, the features that were ranked as top ten most influential factors in predicting student academic performance were: mock examination grade, form three grade, challenges in paying school fees, form one grade, challenges during examination period, internet, assessment style used, number of subjects in form one, parent's employment, and absenteeism from school. Access to drugs and other related substances was ranked the least influential factors in predicting student academic performance according to the correlation based feature selection technique.

#### 4.4.2.3 Feature Selection Using One Rule Technique

The third experiment involved using One Rule (OneR) feature selection technique to evaluate the significance of each attribute to the class attribute. OneR uses the accuracy of a single-attribute classifier. WEKA supports the learner-based feature selection



technique by the OneRAttributeEval technique. Applying the OneRAttributeEval technique on the student data set, the attributes were ranked as shown in Figure 4.4.2.3.

**Figure 4.4.2.3 WEKA OneR Feature Selection Technique Results**

==== Run information ====			
Evaluator: weka.attributeSelection.OneRAttributeEval -S 1 -F 10 -B 6			
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1			
Relation: studentData-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last			
Instances: 1720			
Attributes: 60			
Evaluation mode: evaluate on all training data			
==== Attribute Selection on all input data ====			
Search Method:			
Attribute ranking.			
Attribute Evaluator (supervised, Class (nominal): 60 KCSE):			
OneR feature evaluator			
Ranked attributes:	59.826	2 Age	59.826 55 ETL
68.256 26 MG	59.826	13 CA	59.826 53 SE
63.14 25 F3G	59.826	5 LP	59.826 54 SI
62.791 12 ME	59.826	6 WPC	59.826 45 TL
62.267 11 FE	59.826	7 FS	59.826 44 SC
61.628 16 NSF2	59.826	8 DF	59.826 43 Residence
61.163 28 AS	59.826	9 Sponsor	59.826 35 EA
60.174 24 F2G	59.826	58 EE	59.826 38 ECA
60.116 4 Religion	59.826	30 EAS	59.826 36 ETA
60 15 NSF1	59.826	1 Gender	59.826 39 ED
59.884 17 NSF3	59.826	48 Electricity	59.826 34 RM
59.826 20 YS	59.826	49 Internet	59.826 40 EEC
59.826 19 Specialization	59.826	47 CL	59.826 41 ERM
59.826 59 EI	59.826	51 SL	59.826 42 TS
59.826 21 CS	59.826	46 Library	59.826 37 EFS
59.826 27 LS	59.826	50 STL	59.767 23 F1G

59.826	22 RC	59.826	52 SCL	59.651	29 ELS
59.826	14 CF	59.826	33 AD	59.593	31 Absences
59.826	32 EC	59.826	56 EL	59.186	18 NSF4
59.826	3 Disability	59.826	57 ECL	58.372	10 PE
Selected attributes:					
26,25,12,11,16,28,24,4,15,17,20,19,59,21,27,22,14,32,3,2,13,5,6,7,8,9,58,30,1,48,49,47, 51,46,50,52,33,56,57,55,53,54,45,44,43,35,38,36,39,34,40,41,42,37,23,29,31,18,10: 59					

From the results of the experiment in figure 4.4.2.3, the same number of attributes and instances, i.e., 60 attributes and 1720 instances were applied in the experiment to rank the attributes based on the OneR feature selection value. The predicted (class) attribute was KCSE grade. According to the OneR feature selection technique, the features that were ranked as top ten most influential factors in predicting student academic performance were: mock examination grade, form three grade, mother's education, father's education, number of subjects in form two, assessment style used, form two grade, religion, number of subjects in form one and number of subjects in form three. Parent's employment was ranked the least influential factors in predicting student academic performance according to the OneR feature selection technique.

#### 4.4.3 Discussion of Feature Selection

As observed from the results of the three experiments, each method evaluated for the usefulness of each attribute differently. This is also the case in most of the previous studies on feature selection [25] [57]. Table 4.4.3 shows the results from the three experiments. Feature MG or the mock examination grade has the highest impact on the class attribute since it was ranked the most influential feature in predicting student academic performance. Feature F3G or form three examination grade was also rated as the second most important feature in predicting the KCSE grade according to the

findings from the three experiments conducted. According to information gain feature selection, the top five influential features are: MG, F3G, F1G, F2G, ME, FE, NSF2, PE, STL and SL. The least influential features are: WPC, LP, RC, Disability and CS. The top ten features using the correlation-based attribute evaluator feature selection are: MG, F3G, DF, F1G, EC, internet, AS, NSF1, PE and absences. The least influential features according to correlation based feature selection are: AD, LP, EEC, ECA and WPC. According to oneR, the results show: MG, F3G, ME, FE, NSF2, AS, F2G, Religion, NSF1 and NSF3 as the most influential features while: PE, NSF4, Absences, ELS and F1G are ranked the least influential.

In order to select the optimal features subset for predicting the class attribute, the results from the three techniques were tabulated as shown in Table 4.4.3. Then the average of each attribute was computed from the set of three values instead of selecting one technique or method over others. The values represent the measure of goodness (merit) which was used as the evaluation metrics. A similar approach was used by Osmanbegović and Suljić [40] where they noted that each method accounted for the relevance of attributes in a different way. The results with average values are presented in Table 4.4.3.

**Table 4.4.3 Summary of Feature selection**

<b>Ranker</b>	<b>Attribute</b>	<b>CBFS</b>	<b>IGBFS</b>	<b>OneR</b>	<b>Average</b>
1	MG	0.342	0.30719	68.256	22.96839667
2	F3G	0.2841	0.23445	63.14	21.21951667
3	ME	0.1104	0.0988	62.791	21.00006667
4	FE	0.1074	0.09609	62.267	20.82349667
5	NSF2	0.0774	0.07421	61.628	20.59320333
6	AS	0.1341	0.04833	61.163	20.44847667
7	F2G	0.1032	0.10406	60.174	20.12708667
8	Religion	0.1188	0.0243	60.116	20.08636667
9	NSF1	0.1278	0.05705	60	20.06161667
10	DF	0.1904	0.04557	59.826	20.02065667
11	F1G	0.1763	0.11178	59.767	20.01836
12	Internet	0.1594	0.04024	59.826	20.00854667
13	EC	0.1612	0.03136	59.826	20.00618667
14	CL	0.1142	0.02659	59.826	19.98893
15	RM	0.1151	0.02072	59.826	19.98727333
16	STL	0.0727	0.05947	59.826	19.98605667
17	SL	0.0705	0.059	59.826	19.98516667
18	LS	0.1009	0.02841	59.826	19.98510333
19	Library	0.1001	0.0284	59.826	19.98483333
20	Residence	0.0989	0.0291	59.826	19.98466667
21	YS	0.0789	0.04651	59.826	19.98380333
22	NSF3	0.0394	0.0247	59.884	19.9827
23	SI	0.0636	0.0545	59.826	19.98136667
24	TS	0.0619	0.05238	59.826	19.98009333
25	CA	0.1001	0.01076	59.826	19.97895333
26	SC	0.0769	0.02933	59.826	19.97741
27	SCL	0.0486	0.05033	59.826	19.97497667
28	EAS	0.0599	0.03217	59.826	19.97269
29	Gender	0.0703	0.02125	59.826	19.97251667

30	Specialization	0.0731	0.01029	59.826	19.96979667
31	Age	0.0665	0.01321	59.826	19.96857
32	Sponsor	0.0698	0.00858	59.826	19.96812667
33	CF	0.0483	0.02899	59.826	19.96776333
34	SE	0.0411	0.03417	59.826	19.96709
35	EI	0.049	0.02199	59.826	19.96566333
36	ECL	0.0439	0.0227	59.826	19.9642
37	FS	0.0529	0.01056	59.826	19.96315333
38	TL	0.0496	0.01379	59.826	19.96313
39	ED	0.0473	0.01346	59.826	19.96225333
40	CS	0.0498	0.00392	59.826	19.95990667
41	EE	0.0377	0.01133	59.826	19.95834333
42	ETL	0.0281	0.01963	59.826	19.95791
43	Electricity	0.0405	0.00621	59.826	19.95757
44	EA	0.0318	0.01446	59.826	19.95742
45	EFS	0.0301	0.01187	59.826	19.95599
46	EL	0.0273	0.01303	59.826	19.95544333
47	ERM	0.0254	0.01344	59.826	19.95494667
48	ETA	0.0261	0.01243	59.826	19.95484333
49	ECA	0.0215	0.01027	59.826	19.95259
50	RC	0.0296	0.00163	59.826	19.95241
51	EEC	0.0144	0.01412	59.826	19.95150667
52	Disability	0.0237	0.00307	59.826	19.95092333
53	WPC	0.0234	0.00106	59.826	19.95015333
54	AD	0.0105	0.01188	59.826	19.94946
55	LP	0.0115	0.00156	59.826	19.94635333
56	Absences	0.1208	0.05252	59.593	19.92210667
57	ELS	0.0553	0.02482	59.651	19.91037333
58	NSF4	0.0527	0.02181	59.186	19.75350333
59	PE	0.1222	0.06065	58.372	19.51828333

The biggest advantage of taking the average of the three sets is that the researcher is able to determine the overall impact of each attribute since the individual methods use mutually incompatible metrics. As observed from the summaries above, the target class attribute KCSE grade has a strong correlation with attributes MG (Mock Examination Grade). MG was shown as the most predictive in all the tests. Mock grade is the last test grade a student scores before sitting for the final KCSE examination in secondary school. Mock examinations are often used as an indicator of how well a student is prepared for the final examination. Thus this is a significant factor for predicting KCSE examination grade.

The second most influential feature according to the feature selection techniques used was F3G (Form Three performance). F3G is the performance of the student in the previous class before the final year in a four-year course cycle. ME (mother's education) and FE (father's education) again have a lot of influence on student performance especially through parental motivation and encouragement. Most students would like to do at least better than what their parents scored in their final examination. The results have also shown that NSF2 (number of subjects in form two) is very important since this is the final year before students select their subject majors(specialization) which is normally done in form three.

The attributes that had the least impact based on the average value were: PE (Parents Employment), NSF4 (number of subjects in form four), Absences, ELS (Effect of learning style) and F1G (form one performance). To further understand the attributes that play an important role in predicting the target attribute KCSE, the tree structure generated by J48 on selected features was used as shown in Figure 4.4.3. The structure revealed that MG as the most significant attribute and was selected as the root node. The principle rule when constructing a decision tree is that the attribute that returns the

highest information gain becomes the root node. Information gain is calculated using the formulae:

$$\text{Gain}(p, T) = \text{Entropy}(p) - \sum_{x=0}^n (p(x) - \text{Entropy } p(x))$$

Where  $p(x)$  is the probability of feature  $x$ .





Successive modelling technique was used to find the optimal feature subset. In successive modelling, the process of selecting the optimal subset is done iteratively until the algorithms reach the optimal performance. The procedure started with an initial feature subset of the top three most predictive features, i.e., MG, F3G and ME. These were used to train the classifiers. The performance of each model was recorded based on the three features. The process was performed repeatedly adding the next most significant feature in every subsequent iteration until each classifier attained its optimal performance level. The results of each model are presented in the next section.

#### **4.5 Model Development**

The primary objective of this study was to develop a students' academic performance prediction model that classifies student academic performance into one of the five classes: a, b, c, d and e where each class represents the possible target academic grade a student may score in their secondary school exit examination - KCSE. This section presents the results of the prediction models. The section begins with training of the three models using the features discussed in the previous section 4.4 as per the procedure outlined in section 4.4.4. The performances of each model building iterations are recorded. This is followed by a discussion on the optimal feature subset. Finally, there is a section on data augmentation and a discussion on the best prediction model.

##### **4.5.1 Training of the Predictive Models**

Machine learning involves training the prediction models using training data. Training is the process of building a prediction model from previously known data. There are many classification learning algorithms (also called classifiers) in machine learning used to implement prediction of student performance, this study mainly focused on three most

commonly used algorithms: decision tree classifiers, Naive Bayes classifier and neural networks classifiers to predict student academic performance [40] [66] [103]. These three classifiers were identified from literature review (see section 2.3) as the most widely-used among the machine learning community [10] [42] and are based on different computational methods [101].

The researcher conducted several experiments in order to develop successive models that were later used to find out the optimal set of features that would provide the optimal model performance. The initial set of features for training the model comprised of the top three ranked features from the feature ranking list in the previous section 4.4. A summary of the models performance in terms of prediction accuracy (classification accuracy) was recorded as shown in table 4.4.3. In each successive experiment, a new feature was successively added to the feature subset until the classifiers achieved optimal performance. All the models were built using WEKA tool which is the most popular suite of machine learning software [97]. The candidate classifiers used in WEKA for training the models were J48 classifier for C4.5 decision tree, multilayer perceptron classifier representing neural network and Naïve Bayes classifier. 10-fold cross-validation was used to calculate classification accuracy of each model developed. Results of each classifier are presented next.

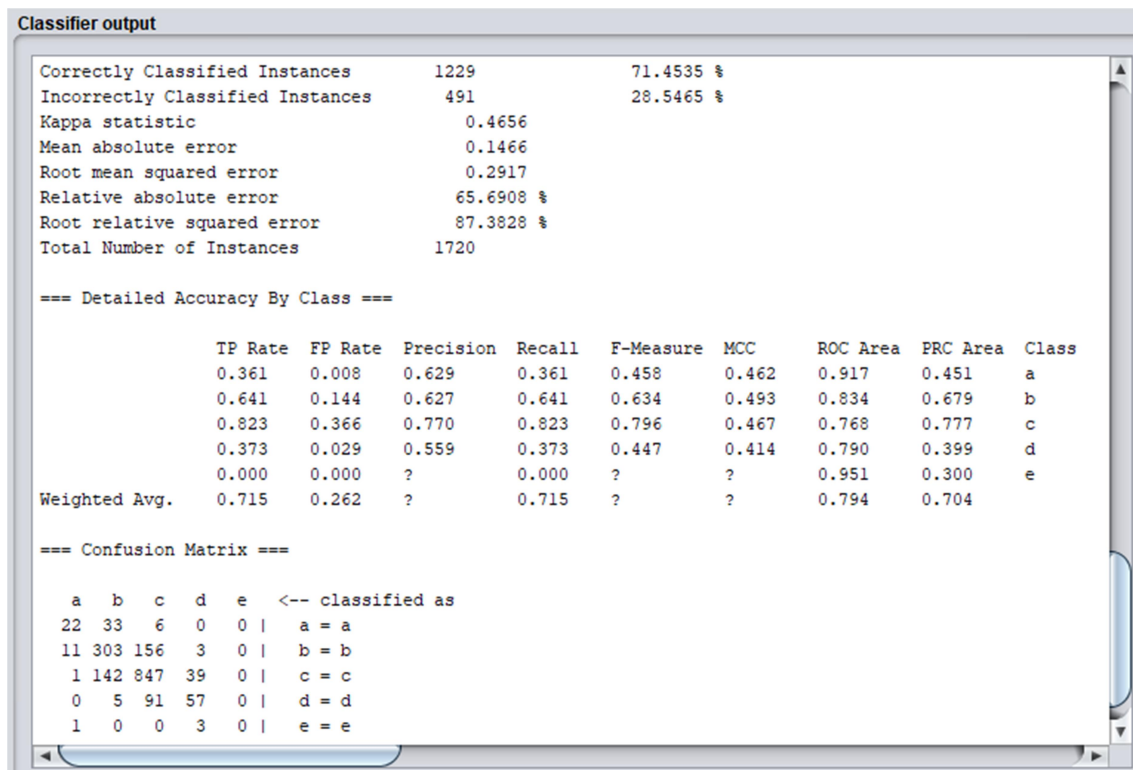
#### **4.5.1.1 Naïve Bayes Model**

This is a machine learning technique that rely on the Bayes' Theorem given by:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where  $P(A)$  indicates the probability of event A occurring,  $P(B)$  indicates the probability of event B occurring,  $P(A|B)$  indicates the probability of A occurring given that B has already occurred.

Naïve bayes classifier was applied to the training data set containing the initial feature subset and the results of the first iteration were recorded. In the second iteration, the next most significant feature was added to the data subset. The results of the second iteration were recorded. This process was performed repeatedly adding an additional feature in every new iteration and recording the results (classification accuracy) of each iteration (see table 4.5.1.1) until the model (or classifier) attained its optimal performance. The output of the classifier's optimal performance is shown in figure 4.5.1.



**Figure 4.5.1.1: Naïve Bayes Model showing the Optimal Prediction Performance**

The results of classification accuracy for each successive iteration are presented in Table 4.5.1.1.

**Table 4.5.1.1 Performance of the Naïve Bayes Classifier on the Top Ranked Twenty Features**

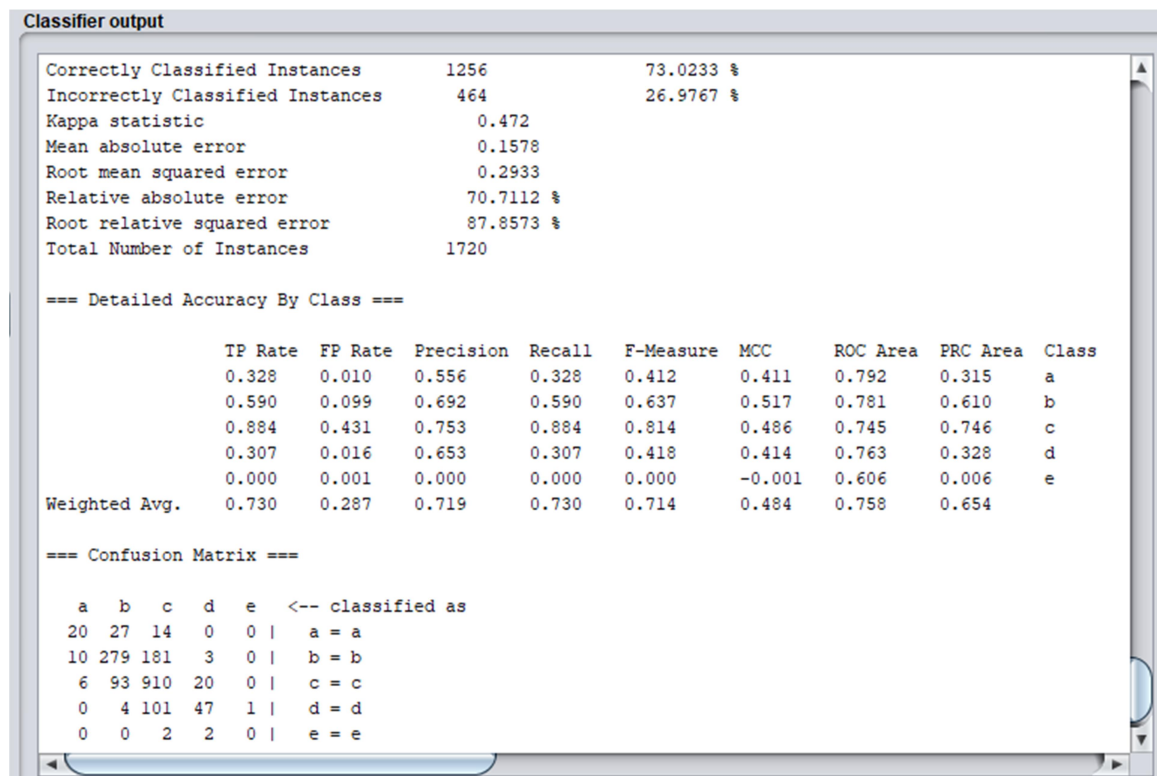
<b>Ranker</b>	<b>Attribute</b>	<b>Naïve Bayes (%)</b>
1	MG	67.97
2	F3G	70.29
3	ME	70.93
4	FE	71.34
5	NSF2	71.28
6	AS	<b>71.45</b>
7	F2G	70.00
8	Religion	70.23
9	NSF1	70.47
10	DF	70.30
11	F1G	69.77
12	Internet	70.06
13	EC	70.12
14	CL	70.12
15	RM	70.06
16	STL	69.77
17	SL	69.94
18	LS	70.17
19	Library	70.61
20	Residence	70.41

#### **4.5.1.2 J48 Decision Tree Model**

In our second experiment, J48 decision tree classifier was applied to the initial feature subset. J48 decision tree uses a tree structure to build classification models [47]. In building the tree structure, the training data set is broken down into distinct groups through a recursive process. The objective of breaking the data is to maximize the distance among groups. The tree that is constructed consists of nodes and branches

where the nodes represents attributes and each branch represents a value that the node can take [5] [45].

WEKA implements decision tree algorithm using J48 classifier. J48 is an improvement of C4.5 decision tree classifier. J48 comes with inbuilt capabilities that can handle missing values as well as pruning of the tree. J48 classifier was trained using successive modeling technique using the same procedure used in the previous section 4.5.1.1. The output of the classifier's optimal performance is shown in figure 4.5.1.2.



**Figure 4.5.1.2 J48 Decision Tree Model showing the Optimal Prediction Performance**

The results of classification accuracy for each successive iteration are shown in table 4.5.1.2.

**Table 4.5.1.2 Performance of the J48 Classifier on the Top Ranked Twenty Features**

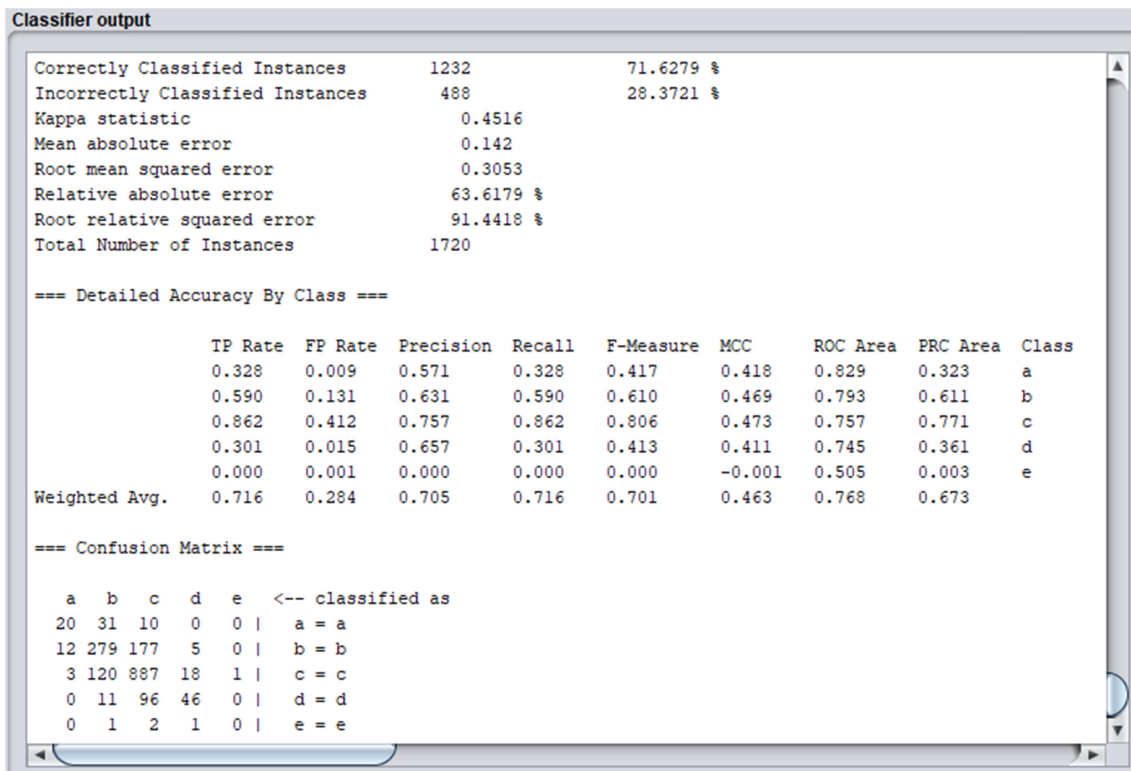
<b>Ranker</b>	<b>Attribute</b>	<b>J48 Decision Tree (%)</b>
1	MG	68.26
2	F3G	70.40
3	ME	71.57
4	FE	71.51
5	NSF2	70.52
6	AS	69.83
7	F2G	70.52
8	Religion	70.87
9	NSF1	70.87
10	DF	72.50
11	F1G	71.74
12	Internet	71.45
13	EC	72.27
14	CL	<b>73.02</b>
15	RM	72.21
16	STL	72.21
17	SL	71.74
18	LS	71.40
19	Library	71.05
20	Residence	71.45

#### **4.5.1.3 Multilayer Perceptron – Neural Networks Model**

In our third experiment, we applied neural network algorithm to the initial feature subset. WEKA tool implements the Neural Networks algorithm using Multilayer Perceptron classifier. According to Doshi & Chaturvedi [102], multilayer perceptron is one of the most widely used and popular neural networks. The classifier consists of several layers of nodes divided into input, hidden and output layers. Multilayer perceptron uses feed

forward propagation to match input to suitable output. It also uses back propagation to implement supervised learning.

In implementing the Multilayer Perceptron classifier, the same procedure used in the previous two experiments was followed (see section 4.5.1.1 and 4.5.1.2). The output of the classifier's optimal performance is shown in figure 4.5.1.3.



**Figure 4.5.1.3: Multilayer Perceptron Model showing the Optimal Prediction Performance**

The results of classification accuracy for each successive iteration using multilayer perceptron classifier are shown in table 4.5.1.3.



**Table 4.5.1.3 Performance of the Multilayer Perceptron Classifier on the Top Ranked Twenty Features**

<b>Ranker</b>	<b>Attribute</b>	<b>Multilayer Perceptron (%)</b>
1	MG	68.26
2	F3G	70.81
3	ME	70.76
4	FE	71.10
5	NSF2	<b>71.63</b>
6	AS	69.94
7	F2G	70.06
8	Religion	67.97
9	NSF1	67.85
10	DF	68.66
11	F1G	67.67
12	Internet	65.81
13	EC	65.47
14	CL	66.74
15	RM	65.87
16	STL	65.00
17	SL	66.92
18	LS	66.51
19	Library	66.22
20	Residence	66.04

#### **4.5.2 Using Successive Modelling to Select the Optimal Feature Subset**

This section presents the most significant features in predicting academic performance for secondary school students. The process of selecting the optimal feature subset was implemented in two phases. Phase I involved running several experiments using feature selection techniques to determine the relevance of each feature. Three experiments were conducted to rank the features in three sets using three feature selection techniques:

information-gain feature selection (see section 4.4.2.1), correlation-based feature selection (see section 4.4.2.2) and OneR technique (see section 4.4.2.3). Since each technique accounted for the relevance of each feature using different methods, an average value of the results from the three sets was used as the overall value representing the relevance of each feature to the target class. The same approach was used by Osmanbegović and Suljić [40]. The average value was then used to rank the features from the most relevant to the least relevant feature (see table 4.4.3).

Phase II involved selecting the optimal feature subset by successive modeling. Using successive modeling, training of the models started with an initial smaller feature set of the top ranked features (see table 4.4.5.1, 4.5.1.2 and 4.5.1.3) and proceeded successively adding a new feature in each subsequent iteration until the model reaches the optimal performance. The combination of features that gave the best classifier performance was then selected as the optimal feature subset. Table 4.5.2 presents the successive results of performance of the three models. The highest performance for each model is shown in bold.

**Table 4.5.2 Comparing Performance of the Models**

<b>Ranker</b>	<b>Attribute</b>	<b>Naïve Bayes (%)</b>	<b>J48 Decision Tree (%)</b>	<b>Multilayer Perceptron (%)</b>
1	MG	67.97	68.26	68.26
2	F3G	70.29	70.40	70.81
3	ME	70.93	71.57	70.76
4	FE	71.34	71.51	71.10
5	NSF2	71.28	70.52	<b>71.63</b>
6	AS	<b>71.45</b>	69.83	69.94
7	F2G	70.00	70.52	70.06
8	Religion	70.23	70.87	67.97
9	NSF1	70.47	70.87	67.85
10	DF	70.30	72.50	68.66
11	F1G	69.77	71.74	67.67
12	Internet	70.06	71.45	65.81
13	EC	70.12	72.27	65.47
14	CL	70.12	<b>73.02</b>	66.74
15	RM	70.06	72.21	65.87
16	STL	69.77	72.21	65.00
17	SL	69.94	71.74	66.92
18	LS	70.17	71.40	66.51
19	Library	70.61	71.05	66.22
20	Residence	70.41	71.45	66.04

From the results, it can be observed that addition of more features was halted after the first 20 features. This was because neither of the remaining new alternatives improved upon the current performance after the first set of 14 features. For naïve bayes classifier, the optimal performance was achieved with the first set of six features since further addition of features did not improve current performance, J48 performance does not

improve classification performance after the first set of 14 features while for multilayer perceptron performance does not improve performance after the first set of five features.

## **4.6 Discussion of Research Findings**

### **4.6.1 Selection of the Optimal Feature Subset**

As observed from the results of experiments carried out in the previous section 4.5.2, it can be observed that the three classifiers achieved their highest performance within the range of first 5 -14 features. Naïve Bayes model achieved optimal classification performance of 71.45% with the first top six features namely: MG, F3G, ME, FE, NSF2 and AS. These are the optimal feature subset based on the Naïve Bayes classifier. J48 model achieved optimal classification performance of 73.02% with the first top 14 features namely: MG, F3G, ME, FE, NSF2, AS, F2G, Religion, NSF1, DF, F1G, Internet, EC and CL. This are the optimal feature subset based on the J48 classifier. Multilayer perceptron model achieved optimal classification performance of 71.63% with the first five features namely: MG, F3G, ME, FE and NSF2. This are the optimal feature subset based on the Multilayer perceptron classifier.

However, since the classifiers attained the highest performance within the range of 5 – 14 features, and that the best performance was achieved by J48 classifier using 14 features, it is only reasonable therefore to consider the 14 top features as the most predictive features of the class attribute as given by the best performing classifier. Thus the optimal feature subset is given by the features: MG (mock examination grade), F3G (form three grade), ME (mother’s education), FE (father’s education), NSF2 (number of subjects in form two (before specialization)), AS (Assessment Style), F2G (form two grade), Religion, NSF1 (number of subjects in form one), DF (difficulties in paying

school fees), F1G (form one grade), Internet, EC (challenges during examination period) and CL (laboratory).

#### **4.6.2 Finding the Best Prediction Model**

This section presents the criteria used to select the final model for predicting students' academic performance. According to Ruta & Gabrys [104], the quality of the selected model relies on the strength and diversity of selection criterion used. In order to identify the best prediction model, the study made use of three selection approaches. In the first approach, the study analysed and compared individual performance metrics for each classifier. In the second approach, the study employed the use of data augmentation techniques to improve the classifier performance, remove any class imbalance and overfitting problem, and ensure the final model is generalizable. In the third approach the study used majority voting methodology (Voting Technique) to select the best classifier. Each criterion is described here.

##### **4.6.2.1 Using Performance Metrics**

The results and findings from the experiments conducted to create the predictive models are presented and discussed next. Table 4.6.2.1 shows the classification results of the three classification models using two metrics: correctly classified instances and incorrectly classified instances.

**Table 4.6.2.1 Performance of Models Based on the optimal Feature Subset**

<b>Evaluation Metric</b>	<b>Naïve Bayes</b>	<b>J48</b>	<b>MLP</b>
Total of correctly classified student grades	1229	1256	1232
Total of Incorrectly classified student grades	491	464	488
Correctly classified grades as grade a (class a)	22	20	20
Correctly classified grade as grade b (class b)	303	279	279
Correctly classified grade as grade c (class c)	847	910	887
Correctly classified grade as grade d (class d)	57	47	46
Correctly classified grade as grade e (class e)	0	1	0

From the results above, naïve bayes classifier correctly classified 1229 instances and incorrectly classified 491 records. Out of this, 22 instances were correctly classified as belonging to class a, 303 instances were correctly classified as belonging to class b, 847 instances were correctly classified as belonging to class c and 57 instances were correctly classified as belonging to class d. J48 classifier correctly classified 1256 instances and incorrectly classified 464 instances. 20 instances were correctly classified as belonging to class a, 279 instances were correctly classified as belonging to class b, 910 instances were correctly classified as belonging to class c, 47 instances were correctly classified as belonging to class d and 1 instance was correctly classified as belonging to class e. The multilayer perceptron classifier classified 1232 instances correctly and 488 instances incorrectly. Out of this, 20 instances were correctly classified as belonging to class a, 279 instances were correctly classified as belonging to class b, 887 instances were correctly classified as belonging to class c and 46 instances were correctly classified as belonging to class d.

As shown in table 4.6.2.1, J48 classifier had the highest total number of correctly classified instances and the least total number of incorrectly classified instances.

Multilayer perceptron model was second with 1232 correctly classified instances while naïve bayes was ranked third with 1229 correctly classified instances. Based on the two metrics, J48 classifier performed better than naïve bayes classifier and multilayer perceptron classifier.

The performance of the three models was again compared using the following performance evaluation metrics: precision, recall, f-measure, ROC curve, TP rate and FP rate. Table 4.6.2.2 presents the results of various performance metrics for each prediction model.

**Table 4.6.2.2 Performance Metrics of the Models**

<b>Evaluation Metric</b>	<b>J48</b>	<b>MLP</b>	<b>Naïve Bayes</b>
Precision	0.719	0.705	-
Recall	0.730	0.716	0.715
F-Measure	0.714	0.701	-
ROC Area	0.484	0.463	-
TP Rate	0.730	0.716	0.715
FP Rate	0.287	0.284	0.262

In order to understand the impact of each measure on the models, an analysis of each metric is presented next.

#### **4.6.2.1.1 Precision (Confidence)**

Precision (or confidence) represents the proportion of predicted positive cases that are actually positives [105]. It measures the exactness of a classifier. The formulae for calculating precision is shown below:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

J48 model achieved the best weighted precision rate of 73%. Class c of the J48 model achieved the highest precision rate of 75.3% (see figure 4.5.1.2). Using the precision value, it means that the model incorrectly classified fewer instances as correctly classified instances. The good performance is attributed to the high number of instances for class c. However, the overall performance of the classifier is also affected by the under-represented classes. Under representation may lead to class imbalance. Class imbalance occurs when some classes have fewer instances than others. For example, class e has four instances hence greatly affected by class imbalance issues. However, the use of feature selection technique as earlier discussed helps in handling the class imbalance issues [97]. The other technique that was used to handle the class imbalance problem was use of data augmentation techniques. This is discussed in the next section.

Multilayer perceptron model was the second best classifier. It achieved a weighted precision rate of 70.5%. Class c of the model achieved the highest precision rate of 75.7%. Class c of the naïve bayes model achieved the highest precision rate of 77% but for class e, it could not compute the precision rate due to low number of instances in the students that scored grade e since the value for true positive and false positive was zero. This affected the weighted precision rate for the entire classifier.

#### **4.6.2.1.2 Recall (Sensitivity)**

Recall or sensitivity represents the proportion of real positive cases that are correctly predicted positive [105]. Recall measures the completeness of a classifier. One of the desirable features of recall is that it is a reflection of how many of the relevant cases are



predicted positive [105]. Recall is given by the number of correctly classified instances divided by the number of all relevant instances which should have been classified as positive.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

As shown in table 4.6.2.2, J48 achieved the best weighted recall value of 73.0%, class c attained the highest recall of 88.4%. This confirms the ability of the model to correctly predict students' grades. Multilayer perceptron model achieved a weighted recall probability of 71.6% where class c achieved the best recall value of 86.2%. Naïve bayes model attained a recall value of 71.5%, the best class was c which achieved recall value of 82.3%

#### **4.6.2.1.3 F-Measure**

F-measure or F-score is used to measure Recall and Precision at the same time. This makes it easier to compare precision and recall of two models at the same time. It is calculated as follows:

$$\text{F - Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

F-measure is preferred for imbalanced datasets because it combines precision metric and recall metric to get balanced average value. In our experiment, J48 model achieved the best F-measure value of 71.4% while multilayer perceptron model achieved F-measure value of 70.1%. This values are considered reasonable enough to determine the performance of the five classes independently.

#### 4.6.2.1.4 Specificity

The specificity metric measures the percentage of actual negatives that are classified correctly as being negatives [106]. Specificity is calculated as follows:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

#### 4.6.2.1.5 ROC Area

The ROC curve graph displays the performance of a classification model at all classification thresholds. ROC is a plot of sensitivity against specificity (inverse recall). The curve shows the capability of the model to classify given instances into target classes. ROC is one of the known reliable measure because it's not affected by class imbalance. This makes ROC more suitable for this study since some classes were affected by the class imbalance problem. As observed from the three experiments, the values for ROC remained relatively high with the highest being 79.4% and the lowest being 76.8%. This indicates that the performance of the models is reliable and generalizable. The ROC curves for J48 model are presented next.

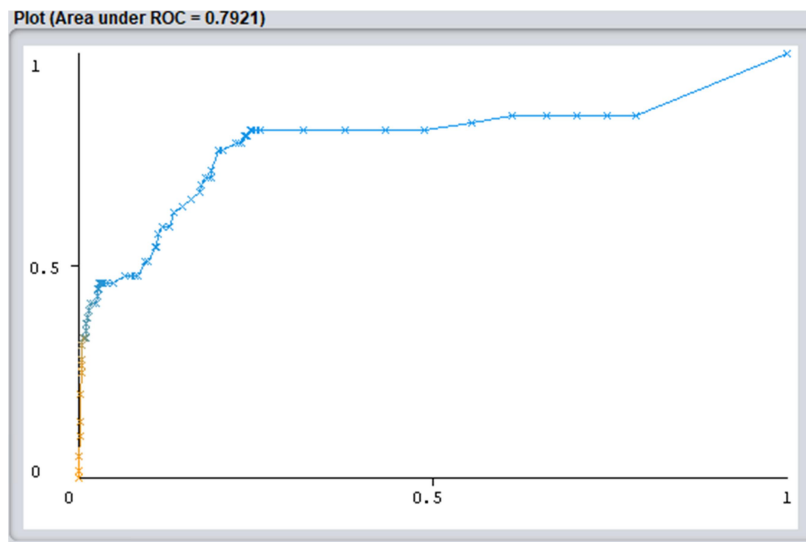
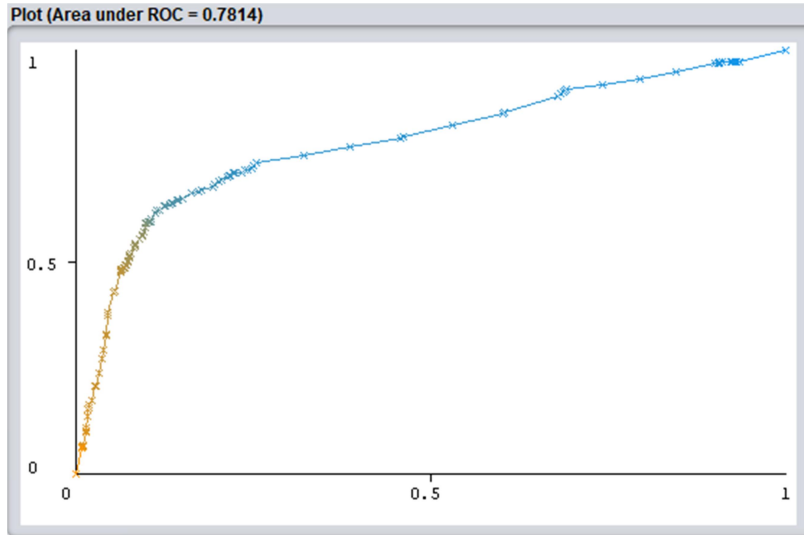
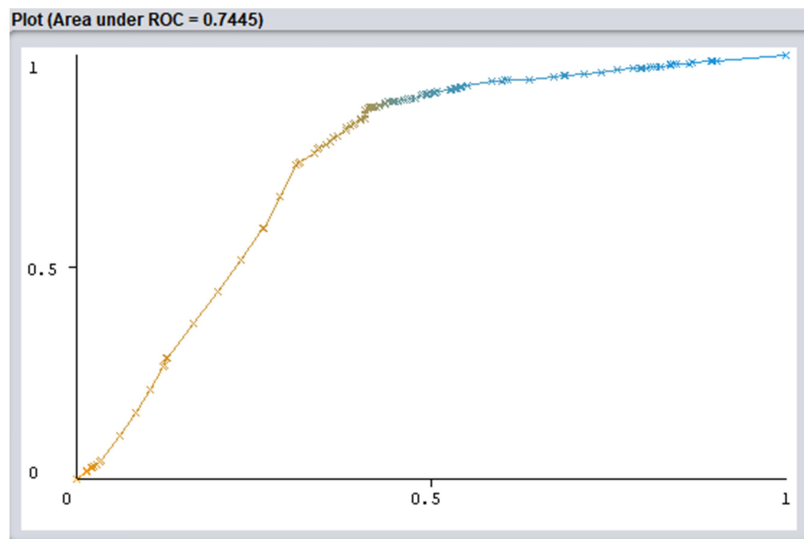


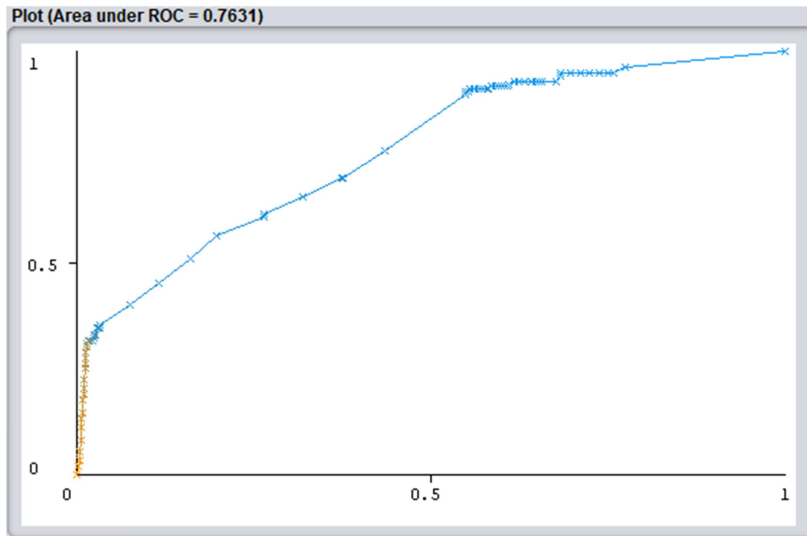
Figure 4.6.2.1.5.1 Class a ROC Curve for J48 Classifier



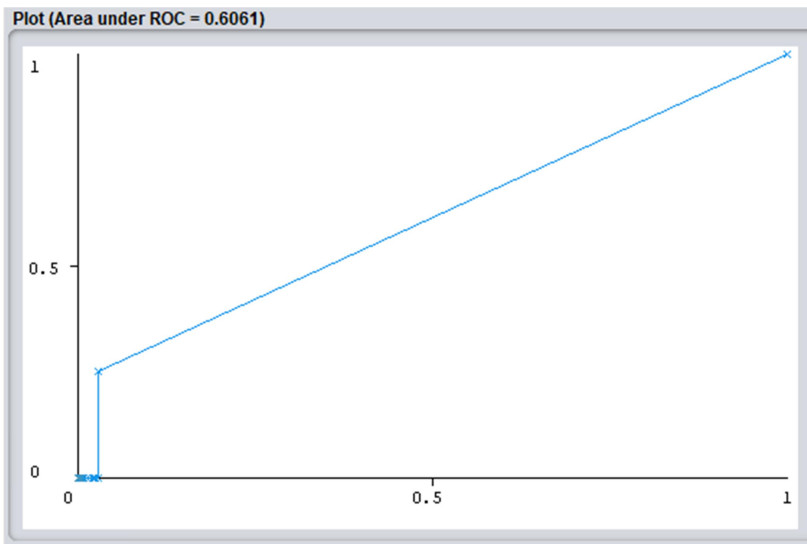
**Figure 4.6.2.1.5.2 Class b ROC Curve for J48 Classifier**



**Figure 4.6.2.1.5.3 Class c ROC Curve for J48 Classifier**



**Figure 4.6.2.1.5.4 Class d ROC Curve for J48 Classifier**



**Figure 4.6.2.1.5.5 Class e ROC Curve for J48 Classifier**

From the ROC curves, it can be observed that the AUC ranges from 0.7921 to 0.6061 which is fairly reasonable enough to conclude that J48 is a good classifier.

#### 4.6.2.1.6 Accuracy

The prediction accuracy of each model is computed as the ratio of number of correct predictions to the total number of predictions made (or total number of input samples) as shown below:

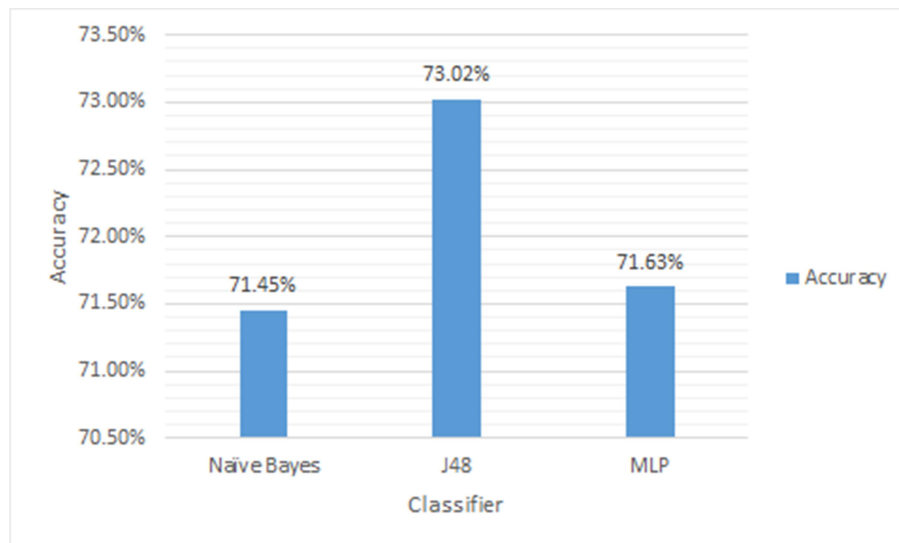
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The accuracy of the three classifiers is presented in table 4.6.2.1.6

**Table 4.6.2.1.6 Prediction Accuracy of Classifiers**

Evaluation Metric	Naïve Bayes%	J48%	MLP%
Accuracy	71.45	73.02	71.63

The J48 model achieved the highest accuracy of 73.02% followed by multilayer perceptron with 71.63% and naïve bayes had the least 71.45%. Figure 4.5.2.6 shows a comparison of the model performances.



**Figure 4.6.2.1.6 Prediction Accuracy of Classifiers**

Among the three classifiers, J48 decision Tree classifiers outperforms Naïve Bayes classifier and Multi Perceptron classifiers in terms of prediction accuracy.

#### **4.6.2.2 Using Data Augmentation to Compare Models' Performance**

Data augmentation is a technique used by researchers to extend the current size of the training data by artificially creating more new training data from training data. According to Iosifidis & Ntoutsi [107] , data augmentation is a process of generating more training data based on the information gathered from the training data corpus. This strategy allows researchers to significantly increase the size of the training data without collecting new data. Data augmentation improves the performance of machine learning models such as deep neural networks that often require large size of training data to fit the model.

Data augmentation is useful when the size of training data is small hence the need to increase the data in order to improve the ability to fit the model and for generalization of models. It is applied to training data as a technique to overcome the overfitting problem in machine learning. In machine learning, data augmentation has been used to address the under-representation of some groups or classes in the training data [107]. Under-representation occurs due to class imbalance where some classes have few number of instances compared to others. In this case, the focus is to improve the number of correctly classified instances of the under-represented class, reduce the classification error for the under-represented classes and avoid degrading the overall model performance.

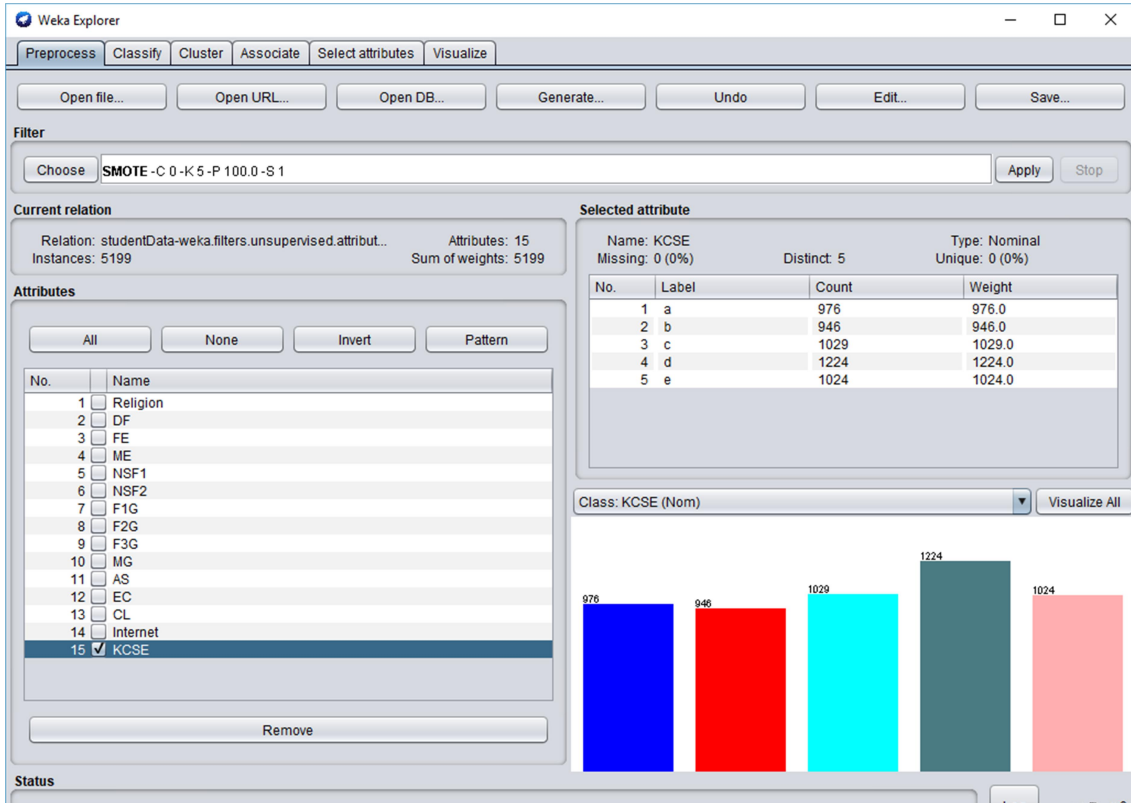
There are two techniques of data augmentation that are commonly used, the oversampling technique and synthetic minority oversampling technique (SMOTE) [107]. Oversampling technique uses a simple strategy that works by randomly duplicating the number of the under-represented instances in order to achieve class balance. SMOTE applies the k-nearest neighbour algorithm on the training data set of the under-represented class to find the k-nearest under-represented neighbours.

This study used SMOTE technique to implement data augmentation. All the experiments were conducted within WEKA machine learning environment. The process involved running several iterations using SMOTE algorithm in order to achieve class balance (see figure 4.6.2.2.1). Since SMOTE places the synthetic minority data at the bottom of the file (data base), a second experiment was performed in WEKA using unsupervised randomize technique to randomize the instances. Randomize algorithm ensures that the training instances are randomly represented across all the folds when using 10-fold cross validation (see figure 4.6.2.2.2). If the instances are not randomized across the folds, the resultant model could suffer the overfitting problem. Overfitting problem occur if each fold is holding data that predominantly belongs to a single class. Table 4.6.2.2.1 shows the class instances before and after data augmentation.

**Table 4.6.2.2.1 Class instances before and after data augmentation**

Class	Data Augmentation	
	Before	After
A	61	976
B	473	946
C	1029	1029
D	153	1224
E	4	1024
Total Instances	<b>1720</b>	<b>5199</b>

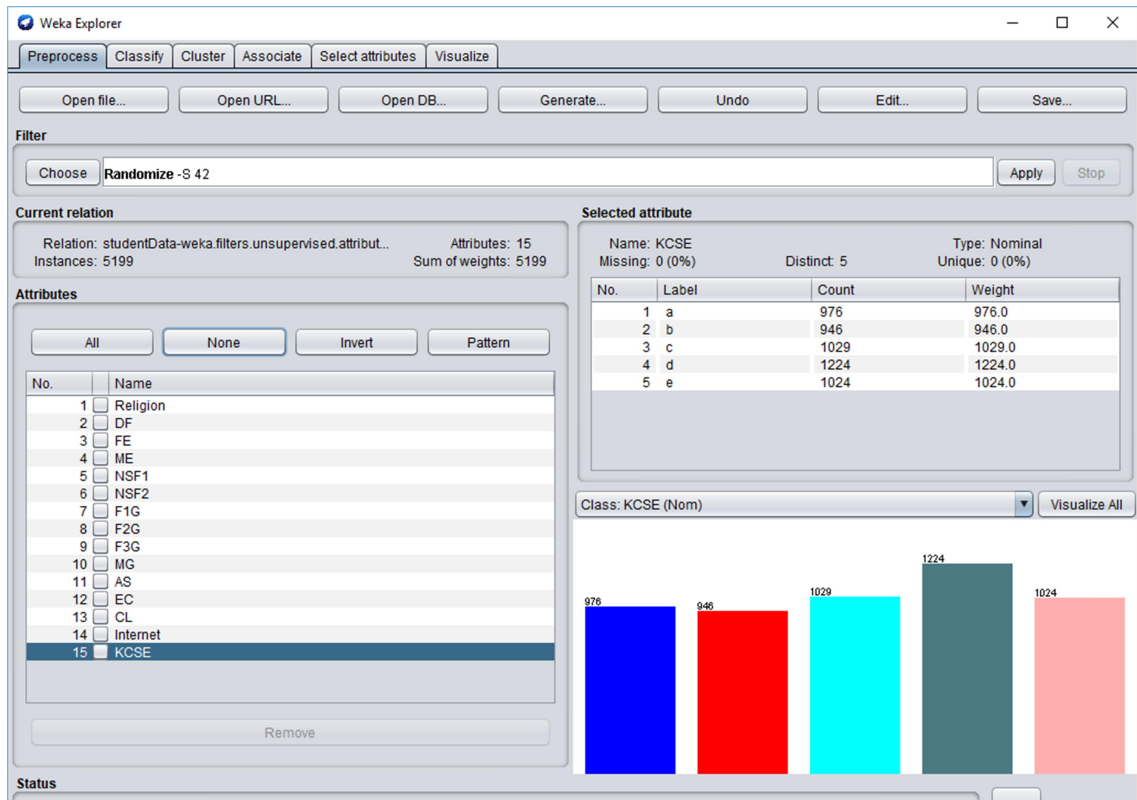
Figure 4.6.2.2.1 shows the output from WEKA after applying SMOTE technique on the class attribute.



**Figure 4.6.2.2.1 Data Augmentation using SMOTE Technique**

Figure 4.6.2.2.2 shows the output from WEKA after applying Randomize technique on the entire data set after data augmentation process.





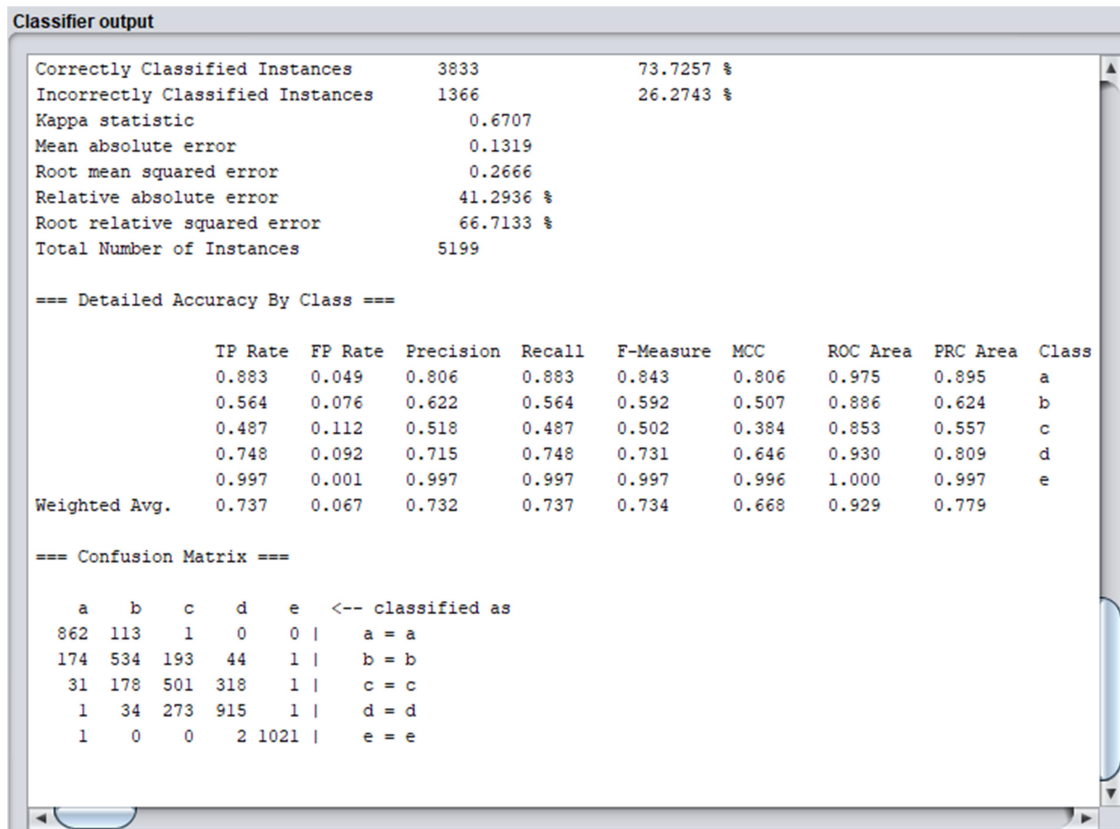
**Figure 4.6.2.2.2 Randomizing File data using Randomize Technique**

The classifiers were then trained again using the data obtained after data augmentation. The total number of instances used in the experiments was 5,199. The optimal feature subset for each classifier (see section 4.6.2) was applied. Table 4.6.2.2 shows a comparison of the performance of the three classifiers before and after data augmentation.

**Table 4.6.2.2 Performance of the three classifiers before and after data augmentation**

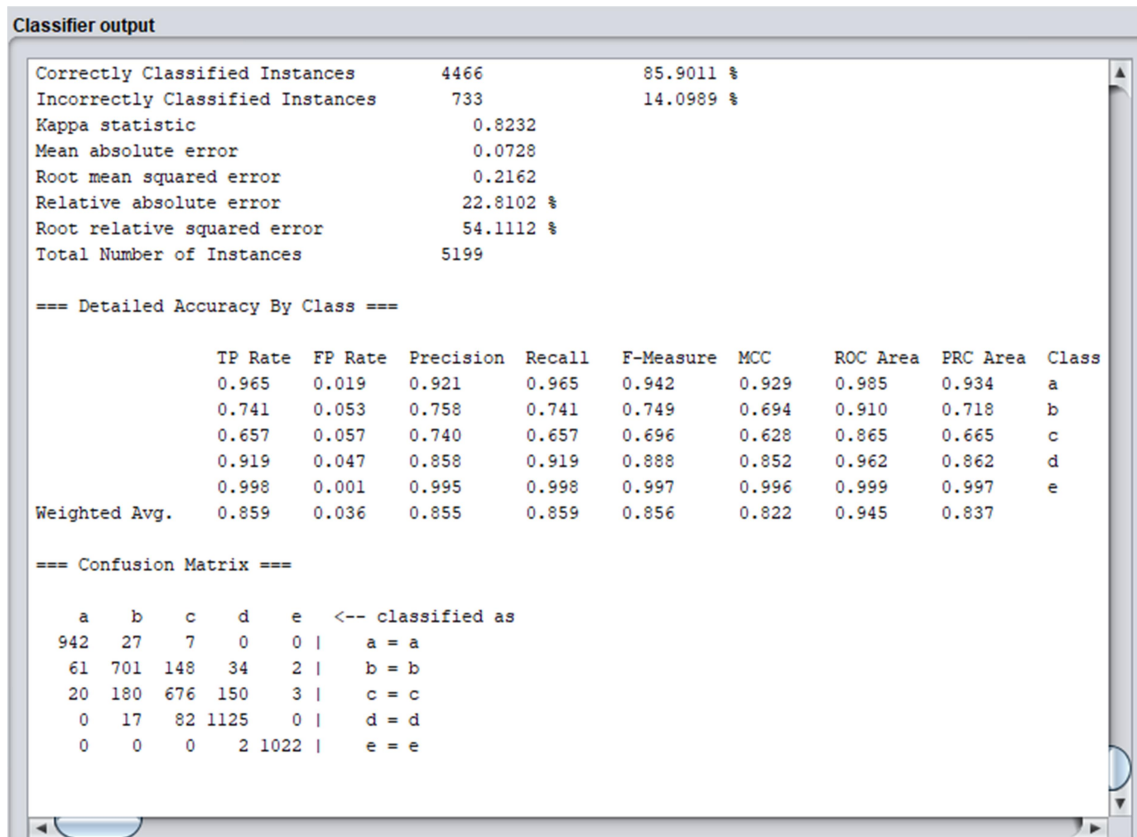
Evaluation Metric	Performance Before Data Augmentation			Performance After Data Augmentation		
	Naïve Bayes	J48	MLP	Naïve Bayes	J48	MLP
Correctly Classified Instances	1229	1256	1232	3833	4466	4105
Incorrectly Classified Instances	491	464	488	1366	733	1094
Accuracy	71.45	73.02	71.63	73.73	<b>85.90</b>	78.96
Precision	-	0.719	0.705	0.732	0.855	0.784
Recall	0.715	0.730	0.716	0.737	0.859	0.790
F-Measure	-	0.714	0.701	0.734	0.856	0.783
ROC Area	-	0.484	0.463	0.929	0.945	0.938
TP Rate	0.715	0.730	0.716	0.737	0.859	0.790
FP Rate	0.262	0.287	0.284	0.067	0.036	0.055

Figure 4.6.2.2.3 shows the performance of Naïve Bayes classifier after data augmentation. The classification accuracy of the classifier improved from 71.45% to 73.73%.



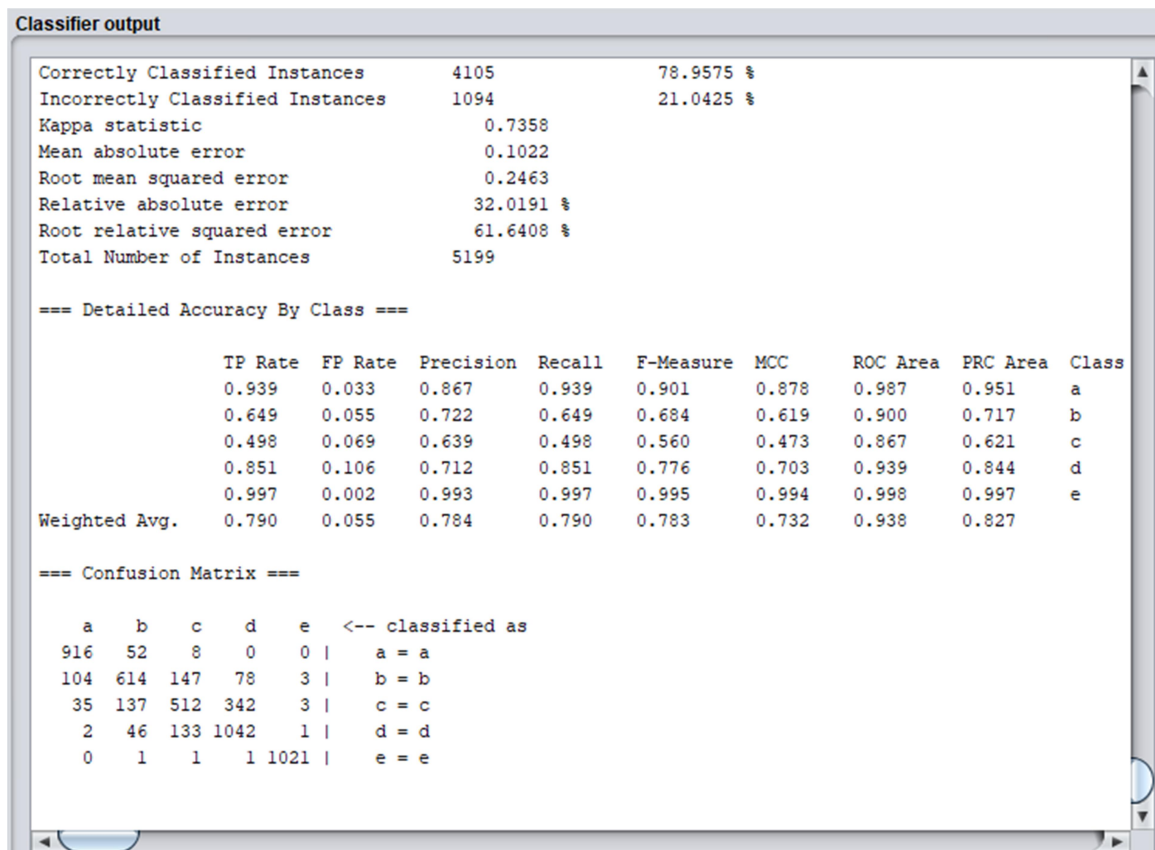
**Figure 4.6.2.2.3 Naïve Bayes classifier output after data augmentation**

Figure 4.6.2.2.4 shows the performance of J48 classifier after data augmentation. The performance in terms of classification accuracy improved from 73.02% to 85.9%.



**Figure 4.6.2.2.4 J48 classifier output after data augmentation**

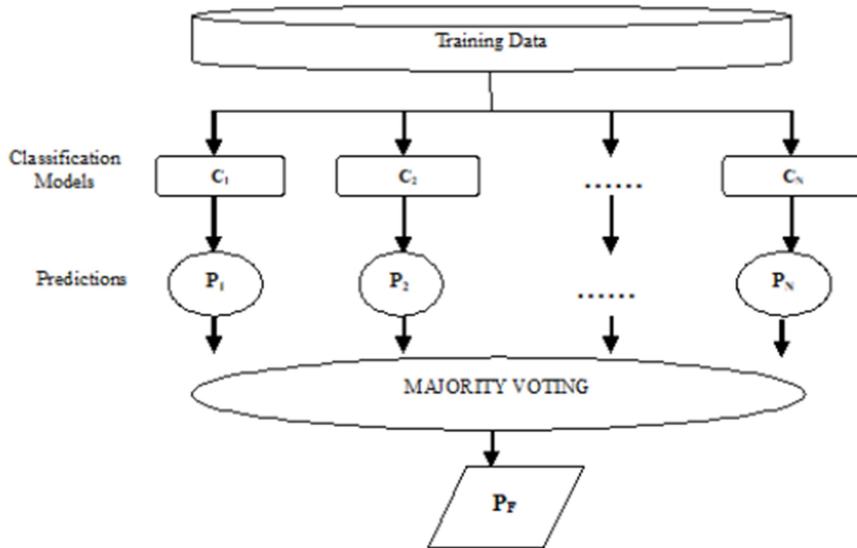
Figure 4.6.2.2.5 shows the performance of Multilayer Perceptron classifier after data augmentation. The performance in terms of classification accuracy improved from 71.63% to 78.96%.



**Figure 4.6.2.2.5 Multilayer Perceptron classifier output after data augmentation**

#### 4.6.2.3 Using Voting Technique

Classifier voting is a method used to combine predictions of different classifiers. The final prediction is decided by majority votes where the classifier that obtains highest votes is selected [108]. Majority voting is a classifier selection methodology [104] that is used to choose the classifier that gets the majority votes. It is used to choose the best classification model out of multiple classification models [104]. According to Salini & Jeyapriya [109], majority voting technique improves the performance of machine learning classifiers by combining predictions of different classification models (classifiers). The majority voting model [109] is described in figure 4.6.2.3 below.



**Figure 4.6.2.3 Majority Voting Process [109]**

According to [108], majority voting can improve prediction of student's class grade by comparing the accuracy of different machine learning classifiers using simple voting technique to choose the most accurate classifier. In this study, three experiments were performed using majority voting process. The first experiment was to predict student academic grade using Naïve Bayes classifier (see figure 4.6.2.2.3). The second experiment was to predict student academic grade J48 classifier (see figure4.6.2.2.4). The third experiment was to predict student academic grade using Multilayer Perceptron classifier (see figure 4.6.2.2. 5). The number of instances classified for each class were recorded in tables 4.6.2.3.1, 4.6.2.3.2, 4.6.2.3.3, 4.6.2.3.4 and 4.6.3.2.5. Majority voting selection criteria was applied to the results of the three classifiers in order to compare the similarities between the classifiers and select the best performing model. Other than the number of the classified instances, several other performance metrics were used including validation accuracy, recall, precision and F-measure as shown in table 4.6.2.3.6.

#### 4.6.2.3.1 Classification for Grade A

Table 4.6.2.3.1 presents the results of classification for grade a.

**Table 4.6.2.3.1 Classification of Grade A using Majority Voting**

Grade A classified as:	Classifier			Comments
	Naïve Bayes	J48	Multilayer Perceptron	
A	862	942	916	Correctly classified as grade a
B	113	52	27	Incorrectly classified as grade b
C	1	8	7	Incorrectly classified as grade c
D	0	0	0	Incorrectly classified as grade d
E	0	0	0	Incorrectly classified as grade e

From table 4.6.2.3.1, grade a was correctly classified as a by Naïve Bayes classifier with 862 instances (votes), incorrectly classified as b, c, d and e by 113,1,0 and 0 instances respectively. Using J48 classifier, grade a was correctly classified as a with 942 instances (votes), incorrectly classified as b, c, d and e by 52,8,0 and 0 instances respectively. Finally, using multilayer perceptron classifier, grade a was correctly classified as a with 916 instances (votes), incorrectly classified as b, c, d and e by 27,7,0 and 0 instances respectively. From the results, it is evident that majority of the instances were correctly classified to truly belong to grade a by J48 decision trees classifier.

#### 4.6.2.3.2 Classification for Grade B

Table 4.6.3.2 presents the results of classification for grade b.

**Table 4.6.2.3.2 Classification of Grade B using Majority Voting**

Grade B classified as:	Classifier			Comments
	Naïve Bayes	J48	Multilayer Perceptron	
A	174	61	104	Incorrectly classified as grade a
B	534	<b>701</b>	614	Correctly classified as grade b
C	193	148	147	Incorrectly classified as grade c
D	44	34	78	Incorrectly classified as grade d
E	1	2	3	Incorrectly classified as grade e

From table 4.6.2.3.2, grade b was correctly classified as b by Naïve Bayes classifier by 534 instances (votes), incorrectly classified as a, c, d and e by 174,193,44 and 1 instances respectively. Using J48 classifier, grade b was correctly classified as b by 701 instances (votes), incorrectly classified as a, c, d and e by 61,148,34 and 2 instances respectively. Using multilayer perceptron classifier, grade b was correctly classified as b with 614 instances (votes), incorrectly classified as a, c, d and e by 104,147,78 and 3 instances respectively. From the results, majority of the instances were correctly classified to truly belong to grade b by J48 decision trees classifier.

#### **4.6.2.3.3 Classification for Grade C**

Table 4.6.3.3 presents the results of classification for grade c.



**Table 4.6.2.3.3 Classification of Grade C using Majority Voting**

Grade C classified as:	Classifier			Comments
	Naïve Bayes	J48	Multilayer Perceptron	
A	31	20	35	Incorrectly classified as grade a
B	178	180	137	Incorrectly classified as grade b
C	501	<b>676</b>	512	Correctly classified as grade c
D	318	150	342	Incorrectly classified as grade d
E	1	3	3	Incorrectly classified as grade e

From table 4.6.2.3.3, grade c was correctly classified as c by Naïve Bayes classifier by 501 instances (votes), incorrectly classified as a, b, d and e by 31,178,318 and 1 instances respectively. Using J48 classifier, grade c was correctly classified as c by 676 instances (votes), incorrectly classified as a, b, d and e by 20,180,150 and 3 instances respectively. Using multilayer perceptron classifier, grade c was correctly classified as c with 512 instances (votes), incorrectly classified as a, b, d and e by 35,137,342 and 3 instances respectively. The results show that majority of the instances were correctly classified to truly belong to grade c by J48 decision trees classifier.

#### **4.6.2.3.4 Classification for Grade D**

Table 4.6.3.4 presents the results of classification for grade d.

**Table 4.6.2.3.4 Classification of Grade D using Majority Voting**

Grade D classified as:	Classifier			Comments
	Naïve Bayes	J48	Multilayer Perceptron	
A	1	0	2	Incorrectly classified as grade a
B	34	17	46	Incorrectly classified as grade b
C	273	82	133	Incorrectly classified as grade c
D	915	<b>1125</b>	1042	Correctly classified as grade d
E	2	0	1	Incorrectly classified as grade e

From table 4.6.2.3.4, grade d was correctly classified as d by Naïve Bayes classifier by 915 instances (votes), incorrectly classified as a, b, c and e by 1,34,273 and 2 instances respectively. Using J48 classifier, grade d was correctly classified as d by 1125 instances (votes), incorrectly classified as a, b, c and e by 0,17,82 and 0 instances respectively. Using multilayer perceptron classifier, grade d was correctly classified as d with 1042 instances (votes), incorrectly classified as a, b, c and e by 2,46,133 and 1 instances respectively. The results show that majority of the instances were correctly classified to truly belong to grade d by J48 decision trees classifier.

#### 4.6.2.3.5 Classification for Grade E

Table 4.6.3.5 presents the results of classification for grade e.

**Table 4.6.2.3.5 Classification of Grade E using Majority Voting**

Grade E classified as:	Classifier			Comments
	Naïve Bayes	J48	Multilayer Perceptron	
A	1	0	0	Incorrectly classified as grade a
B	0	0	1	Incorrectly classified as grade b
C	0	0	1	Incorrectly classified as grade c
D	2	2	1	Incorrectly classified as grade d
E	1021	<b>1022</b>	1021	Correctly classified as grade e

From table 4.6.2.3.5, grade e was correctly classified as e by Naïve Bayes classifier by 1021 instances (votes), incorrectly classified as a, b, c and d by 1,0,0 and 2 instances respectively. Using J48 classifier, grade e was correctly classified as e by 1022 instances (votes), incorrectly classified as a, b, c and d by 0,0,0 and 2 instances respectively. Using multilayer perceptron classifier, grade e was correctly classified as e with 1021 instances (votes), incorrectly classified as a, b, c and e by 0,1,1 and 1 instances respectively. The results show that majority of the instances were correctly classified to truly belong to grade e by J48 decision trees classifier.

#### 4.6.2.3.6 Comparing Precision for Different Classes

Based on the results of various performance metrics including precision, recall and f-measure as shown in tables 4.6.2.3.6.1, 4.6.2.3.6.2, 4.6.2.3.6.3, J48 classifier has been

voted the best in all classes. This is also the case in terms of overall accuracy as earlier shown in Table 4.6.2.2.2.

**Table 4.6.2.3.6.1 Comparing Precision for Different Classes**

Class	Classifier		
	Naïve Bayes	J48	Multilayer Perceptron
A	0.806	<b>0.921</b>	0.867
B	0.622	<b>0.758</b>	0.722
C	0.518	<b>0.740</b>	0.639
D	0.715	<b>0.858</b>	0.712
E	0.997	<b>0.995</b>	0.993
<b>Weighted Average</b>	<b>0.732</b>	<b>0.855</b>	<b>0.790</b>

**Table 4.6.2.3.6.2 Comparing Recall for Different Classes**

Class	Classifier		
	Naïve Bayes	J48	Multilayer Perceptron
A	0.883	<b>0.965</b>	0.939
B	0.564	<b>0.741</b>	0.649
C	0.487	<b>0.657</b>	0.498
D	0.748	<b>0.919</b>	0.851
E	0.997	<b>0.998</b>	0.997
<b>Weighted Average</b>	<b>0.737</b>	<b>0.859</b>	<b>0.790</b>

**Table 4.6.2.3.6.3 Comparing F-measure for Different Classes**

Class	Classifier		
	Naïve Bayes	J48	Multilayer Perceptron
A	0.843	<b>0.942</b>	0.901
B	0.592	<b>0.749</b>	0.684
C	0.502	<b>0.696</b>	0.560
D	0.731	<b>0.888</b>	0.776
E	0.997	<b>0.997</b>	0.995
<b>Weighted Average</b>	<b>0.734</b>	<b>0.856</b>	<b>0.783</b>

#### 4.6.3 Selected Prediction Model

In the previous sections of this chapter, the study has have presented experimental results of three classification models: Naïve Bayes, Decision Trees, and Neural Networks in order to evaluate the classification capability of each training algorithms. These results were used as a basis for selecting the most efficient prediction model. 10-fold cross-validation was used to evaluate classification accuracy. In all experiments, WEKA was used to evaluate the classifiers and for comparisons.

Using feature selection technique, more experiments were performed to find the performance of the three classifiers using the features that were ranked as the most predictive. After data augmentation process, J48 Decision Tree classifier outperformed all other models with an overall 85.9% classification accuracy. These results are shown in Table 4.6.2.2.2. The voting technique was then used to choose the best classifier using majority voting methodology (see section 4.6.2.3). The results revealed that J48 classifier was the best. Though there was no particular model that stands above all the other

models by a very high margin, however, it is evident that J48 model predicts better than the other two models in this study.

The other evaluation metrics that were used to confirm the success of machine learning prediction model were accuracy, precision, recall, specificity, F-measure and ROC curve. The performance of the prediction models was validated using 10-fold cross-validation method in all experiments conducted in this study. Cross validation is a systematic way of doing a repeated hold-out validation. Although there are several ways of evaluating a model such as; evaluating on an independent test dataset or, using hold-out method, cross-validation has over the years been used as a standard way of evaluating performance of machine learning algorithms due to its ability to reduce the variance. 10-fold cross-validation divides the data set into 10 parts (folds), then 9 folds are used to train the model and the last fold used to test the model. This is done iteratively 10 times with each time it is repeated, a different fold from the previous one (hold-out fold) is used as the test fold. Then an average results of the 10 folds is then taken and used to build the final model.

Even with various experiments on different attributes, it was not enough to outperform the performance by J48 Decision Tree model. The study therefore concludes that J48 Decision Tree model is a satisfying choice for a classifier for prediction of students' academic performance in secondary school based on the student data set used in this study.

#### **4.7 Summary of Results and Discussion**

This section presents the summarized results on data analysis, feature selection, prediction model development and a discussion on the same. The data set for the study was collected from recent secondary school graduates between January 2019 and April 2019 using questionnaires. It consisted of 1720 instances and 60 features. After data augmentation, the number of instances was increased to 5,199 (see table 4.6.2.2.1). The attributes represent information on individual student characteristics, family characteristics and institutional characteristics.

Experiments were carried out using three feature selection techniques in order to identify and remove unnecessary or irrelevant attributes. The techniques used were: Information-gain feature selection, Correlation-based feature selection and One Rule feature selection technique. Each experiment was conducted in WEKA machine learning environment. The optimal feature subset was selected through successive modeling and consisted of features: MG (mock examination grade), F3G (form three grade), ME (mother's education), FE (father's education), NSF2 (number of subjects in form two before specialization), AS (Assessment Style), F2G (form two grade), Religion, NSF1 (number of subjects in form one), DF (difficulties in paying school fees), F1G (form one grade), Internet, EC (challenges during examination period) and CL (laboratory).

Finally, three machine learning algorithms: naïve bayes, decision trees, and neural network classifier were used to train the models. Evaluation metrics used to evaluate the models were accuracy, precision, recall, F-measure, specificity and ROC curve. 10-fold cross-validation was used in all experiments. From the results of all the experiments conducted, J48 model achieved the highest classification accuracy of 85.9% followed by multilayer perceptron with 78.96% classification accuracy and naïve bayes had the least classification accuracy of 73.73%. The chapter concludes by selecting the J48 Decision

Tree model as a best classifier for prediction of students' academic performance for secondary school.



## **CHAPTER FIVE**

### **FINAL MODEL FOR PREDICTION STUDENTS' ACADEMIC PERFORMANCE**

#### **5.1 Introduction**

In this study, a model for predicting student academic performance was created using machine learning methodology. This marks the final phase in the development of a prediction model using machine learning process. This chapter therefore presents the final model for prediction of students' academic performance in secondary schools. From the results, it was noted that the prediction model developed using decision tree algorithm performed better than the models developed using naïve bayes algorithm and neural network algorithm. It was therefore selected as the best model for prediction of students' academic performance in secondary schools. The chapter begins with a discussion on different ways used to represent prediction models, the structure of the predictive model using Predictive Toxicology Mark-up Language (PTML) and presentation of the six elements of the selected predictive model. The elements include model description, model parameter, model attributes, model performance, class attribute and confusion matrix [110]. The last section gives a summary of the chapter.

#### **5.2 Predictive Model Presentations**

There are several machine learning tools that have been developed to provide the functionality to generate predictive models. They include freeware tools such as WEKA and commercial tools such as SPSS. The models generated from this development tools come with different types of representation. The main objective of predictive model representation is therefore to provide a standard way of representing predictive model and makes them easier to process and understand [110]. Several approaches have been

proposed for managing predictive data and predictive models. They include use of object oriented database, Extensible Mark-up Language (XML) and Predictive Toxicology Mark-up Language (PTML) [110]. The object oriented database uses objects and classes to represent the records. However, results from machine learning and data mining can have different types of classes, patterns and different formats of model presentation. XML format provides a basic method to present the model and describe the information. PTML provides a standard representation of predictive models that allows for sharing or reusability of models. This study used PTML to represent the predictive model for students' academic performance.

### **5.3 Predictive Model Structure**

Predictive Toxicology Mark-up Language (PTML) has been used to represent the predictive model for students' academic performance in this study. The PTML structure presents a simpler representation that is able to hold predictive models information [110]. The model was initially generated using WEKA tool then converted to Extensible Mark-up Language (XML) structure. XML structure is used to represented predictive model with minimal tags that are necessary to describe the model and for further analysis. The XML was then converted to the PTML structure. The model structure for PTML consists of six elements that include model description, model parameter, model attributes, model performance, class attribute and confusion matrix. Each of these elements is discussed next.

#### **5.3.1 Model Description**

The PTML model description section give the general information of the model such as date when the model was developed, name of author, software version and the file name for WEKA model. The description of the model is shown in figure 5.3.1.

```
<modelDescription>
  <Name>AcademicPerformancePredictionModel</Name>
  <Date>31-5-2019</Date>
  <Version>1.0</Version>
  <Author>Musau</Author>
  <Description> Model Generated with Weka 3.8.3</Description>
  <wekaModel>J48DT.model</wekaModel>
</modelDescription>
```

**Figure 5.3.1 Performance Predictive Model Description**

From the figure, the model was generated using Weka 3.8.3 version on 31-05-2019. The J48 decision tree classifier was used to generate the model.

### 5.3.2 Model Parameters

Model parameters include information such as the name of classifier used, validation method used and the number of folds. This section gives a summary of the parameters used to generating the predictive model. Figure 5.3.2 shows the parameters of the model.

```
<modelParameter>
  <option Classifier="weka. classifiers. trees. J48"></option>
  <option TestMode>10fold-cross-validation</option>
  <option Fold="10"></option>
  <option NumberofLeaves ="635"></option>
  <option SizeofTree ="819"></option>
</modelParameter>
```

**Figure 5.3.2 Predictive Model Parameters**

### 5.3.3 Model Attributes

The model attributes are presented in figure 5.3.3. They include the data set, attributes and attribute selection techniques that were used during the development of the students’ academic performance prediction model.

```

<modelAttributes>
  <DataSet> StudentPerformanceDataSet.arff</DataSet>
  <FeatureSelectionAlgorithm> InfoGainAttributeEval
  CorrelationAttributeEval OneRAttributeEval
</FeatureSelectionAlgorithm>
  <FeatureSearchMethod> weka.attributeSelection.Ranker -T
</FeatureSearchMethod>
  <TotalNumberOfInstances>5199</TotalNumberOfInstances>
  <NumberOfAttributes>60</NumberOfAttributes>
  <NumberOfAttributesSelected>14</NumberOfAttributesSelected>
  <Features>
    <Name>MG</Name>
    <Type>Nominal</Type>
  </Features>
  <Features>
    <Name>F3G</Name>
    <Type>Nominal</Type>
  </Features>
  <Features>
    <Name>ME</Name>
    <Type>Nominal</Type>
  </Features>
  <Features>
    <Name>FE</Name>
    <Type>Nominal</Type>
  </Features>
  <Features>
    <Name>NSF2</Name>
    <Type>Nominal</Type>
  </Features>
  <Features>
    <Name>AS</Name>
  </Features>

```

```
<Type>Nominal</Type>
</Features>
<Features>
  <Name>F2G</Name>
  <Type>Nominal</Type>
</Features>
<Features>
  <Name>Religion</Name>
  <Type>Nominal</Type>
</Features>
<Features>
  <Name>NSF1</Name>
  <Type>Nominal</Type>
</Features>
<Features>
  <Name>DF</Name>
  <Type>Nominal</Type>
</Features>
<Features>
  <Name>FIG</Name>
  <Type>Nominal</Type>
</Features>
<Features>
  <Name>Internet</Name>
  <Type>Nominal</Type>
</Features>
<Features>
  <Name>EC</Name>
  <Type>Nominal</Type>
</Features>
<Features>
  <Name>CL</Name>
  <Type>Nominal</Type>
```

```
</Features>
</modelAttributes>
```

**Figure 5.3.3 Predictive Model Attributes**

### 5.3.4 Model Performance

The model performance is a presentation of the results generated by the model. It is used to demonstrate the overall quality of the predictive model. To illustrate the model performance, other performance evaluation metrics were used which included correctly classified instances and incorrectly classified instances as shown in Figure 5.3.4.

```
<modelPerformance>
  <modelType>Classification</modelType>
  <CorrectlyClassifiedInstances>4466
</CorrectlyClassifiedInstances>
  <IncorrectlyClassifiedInstances>733
</IncorrectlyClassifiedInstances>
  <Accuracy>85.90</Accuracy>
</modelPerformance>
```

**Figure 5.3.4 Predictive Model Performance**

### 5.3.5 Class Attribute

The class attribute element gives further model performance information. Additional evaluation metrics are presented that include true positive rate, false positive rate, precision, recall and receiver operating characteristic (ROC) area (see Figure 5.3.5).

```
<classAttribute>
  <Details>
    <TPRate>0.859</TPRate>
```

```
<FPRate>0.036</FPRate>
<Precision>0.855</Precision>
<Recall>0.859</Recall>
<F-Measure>0.856</F-Measure>
<ROC Area>0.945</ROC Area>
</Details>
</classAttribute>
```

**Figure 5.3.5 Class Attributes**

### 5.3.6 Confusion Matrix

The confusion matrix is used to give an overview of correctly and incorrectly classified instances to the class attribute. As shown in see Figure 5.3.6, 942 instances were correctly classified as class a, 701 instances were correctly classified as class b, 676 instances were correctly classified as class c, 1125 instances were correctly classified as class d and 1022 instances were correctly classified as class e.

```
<ConfusionMatrix>
  <Array> Class a Class b Class c Class d Class e </Array>
  <Array> 942      27      7      0      0 Class a </Array>
  <Array> 61  701  148  34   2 Class b </Array>
  <Array> 20  180  676  150  3 Class c </Array>
  <Array> 0   17   82  1125  1 Class d </Array>
  <Array> 0   0   0   2  1022 Class e </Array>
</ConfusionMatrix>
```

**Figure 5.3.6 Confusion Matrix**

#### **5.4 Chapter Summary**

The objective of this chapter was to represent the predictive model. The model consists of six elements: model description, model parameters, model attributes, model performance, class attributes and confusion matrix as described above. Predictive Toxicology Mark-up Language (PTML) has been used to represent the predictive model for students' academic performance.



## CHAPTER SIX

### SUMMARY, CONCLUSION AND RECOMMENDATIONS

#### 6.1 Introduction

This chapter presents the objectives and achievements of the study. The chapter starts by systematically reviewing the research objectives and how the study has addressed each objective by answering the research questions associated with the objective. Finally, the chapter gives a conclusion, the contributions of the study, recommendations for future work and limitations of the study.

The objectives of the study were:

- i. To analyse existing studies on students' academic performance prediction
- ii. To find out significant factors that affect students' academic performance
- iii. To develop a model for students' academic performance prediction in Kenya
- iv. To validate the students' academic performance prediction model

#### 6.1.1 Objective 1: To analyse existing studies on students' academic performance prediction

This research objective lead to the formulation of the research question:

*What are the algorithms used in the prediction of students' academic performance?*

Table 2.3.2 presents a summary of previous work on prediction of students' academic performance that was reviewed. The study undertook a systematic literature review of studies on students' academic performance prediction done both locally and internationally in countries such as Kenya, Nigeria, and European countries such as Portugal. Based on the review of previous studies as presented in chapter two, this study

was able to compile a conclusive list of all the factors that affect students' academic performance in secondary school. These factors are listed in table 2.3.2 (column three).

A systematic literature review of the machine learning techniques was conducted. The study found out that there are four types of machine learning techniques; supervised learning, unsupervised learning, semi-supervised learning and reinforced learning. The study identified several machine learning algorithms that were used in previous studies to predict performance. They include: Naïve Bayes, Neural Network, K-Nearest Neighbor, Decision Trees, Support Vector Machine, Lazy Instance-based Nearest Neighbor algorithms, Random Forest, Function-based algorithms, Deep Learning Neural Network and Decision Rule Learning algorithms (see Table 2.3.2). The study further identified the most commonly used algorithms for classification problem in machine learning and data mining as Naïve Bayes, decision trees and neural networks [40]. The same conclusion was given by [10] [42] that these classifiers are widely-used among the machine learning community and they all differ in terms of the computational methods used in each of them [101].

### **6.1.2 Objective 2: To find out significant factors that affect students' academic performance**

In order to address objective two, the study sought to answer the research question: *How to find out the most significant factors for predicting students' academic performance?*

In addressing this objective, the study used feature selection techniques in machine learning to find the optimal feature subset. As described in chapter two, feature selection is a process of selecting relevant feature in a data set in order to improve machine learning results [100]. Several experiments were conducted using three feature selection

techniques to identify and remove unnecessary or irrelevant attributes by assessing the relevance of each attribute. The study used information-gain feature selection, correlation-based feature selection and One Rule feature selection techniques. WEKA machine learning environment was used in all the experiments conducted. The results of the three experiments were compared as shown in Table 4.4.3. The average of the three sets of results produced by the three feature selection techniques was then used to rank all the attributes in terms of the relevance (average value) instead of selecting one technique or method over others. A similar approach was used by Osmanbegović and Suljić [40] after they noted that each method accounted for the relevance of attributes in a different way.

Finally, in order to get the optimal feature subset, the researcher applied successive modeling. The results are presented in Table 4.5.2. All the classifiers attained optimal performance within the range of the first set of 5 – 14 features as ranked in table 4.4.3, and that the best performance in terms of classification accuracy was produced by J48 classifier using a set of 14 features. The researcher therefore considered it reasonable to conclude that the top 14 features are the most predictive features of the class attribute. Therefore, going by the aforementioned, the study identified the optimal feature subset as consisting of features: MG (mock examination grade), F3G (form three grade), ME (mother's education), FE (father's education), NSF2 (number of subjects in form two (before specialization)), AS (Assessment Style), F2G (form two grade), Religion, NSF1 (number of subjects in form one), DF (difficulties in paying school fees), F1G (form one grade), Internet, EC (challenges during examination period) and CL (laboratory).

### **6.1.3 Objective 3: To develop a model for students' academic performance prediction in Kenya**

The research question for this objective was: *How to model student academic performance based on significant factors?* The purpose of this objective was to build a prediction model for prediction of students' academic performance into one of the five possible classes or grades: a, b, c, d and e. To achieve this objective, the study started by collecting primary data set (see appendix VII). Using questionnaires (see appendix I), the data collected consisted of 1720 instances and 62 features (see Table 4.3.1). Further analysis and experiments on the data were carried out. Data augmentation was applied on the data to extend the existing training data set from 1720 records to 5,199 records (see table 4.6.2.2.1). The attributes represent information on individual student characteristics, family characteristics and institutional characteristics.

Three feature selection techniques namely info-gain based evaluator, correlation-based attribute evaluator and One Rule feature selection techniques were applied on the data set to identify and remove unnecessary or irrelevant attributes (see Table 4.4.3). Selection of the optimal feature subset by successive modeling resulted to 14 features that included: MG, F3G, ME, FE, NSF2, AS, F2G, Religion, NSF1, DF, FIG, Internet, EC and CL. All the experiments were conducted in the WEKA machine learning environment.

To train the prediction models, the study then used the top three commonly used machine learning algorithms for prediction of student performance. They included naïve bayes, decision trees and neural network classifier. Review of related studies revealed these classifiers as the most widely-used among the machine learning community [10] [42] and are based on different computational methods [101]. The results from the

experiments carried out using these classifiers are presented in figure 4.5.1.1, figure 4.5.1.2, figure 4.5.1.3, table 4.5.1.1, table 4.5.1.2, table 4.5.1.3, table 4.5.2, table 4.6.2.1 and table 4.6.2.2.2. From the results of all the experiments conducted, J48 prediction model achieved the highest classification accuracy of 84.90% followed by multilayer perceptron with 78.96% classification accuracy and naïve bayes had the least classification accuracy of 73.73%. This study therefore concludes that J48 Decision Tree model is the best classifier and a satisfying choice for a classifier for prediction of students' academic performance in secondary school.

#### **6.1.4 Objective 4: To validate the students' academic performance prediction model**

The following research question was formulated to address this objective: *How to validate a students' academic performance prediction model?* Several performance evaluation metrics were used to validate the models as shown in table 4.6.2.1, table 4.6.2.2 and table 4.6.2.2.2. The metrics included accuracy, precision, recall, F-measure, specificity and ROC curve (see section 4.6.2). 10-fold cross-validation was used in all experiments. Based on the results from the experiments, J48 model was selected as the best model that had the highest performance based on all the evaluation metrics used.

## **6.2 Summary of the Conclusion**

### **6.2.1 Most Significant Factors Affecting Students' Academic Performance**

The study found out that the most significant factors for predicting KCSE grade for secondary school students are MG (mock examination grade), F3G (form three grade), ME (mother's education), FE (father's education), NSF2 (number of subjects in form two (before specialization)), AS (Assessment Style), F2G (form two grade), Religion, NSF1 (number of subjects in form one), DF (difficulties in paying school fees), F1G

(form one grade), Internet, EC (challenges during examination period) and CL (laboratory).

### **6.2.2 Best Machine Learning Algorithms for Modelling Academic Performance Prediction Model**

The study revealed that the most widely-used machine learning algorithms among the machine learning community for prediction of student academic performance in secondary school are Naïve Bayes, Decision Trees and Neural Network.

### **6.2.3 Academic Performance Prediction Model**

J48 Decision tree prediction model was identified as the best and most suitable for prediction of students' academic performance in secondary school. It achieved the highest classification accuracy of 84.90%. This study therefore recommends use of J48 Decision Tree model for prediction of students' academic performance in secondary school for developing countries like Kenya.

## **6.3 Contribution of the Thesis**

This study contributes to the body of knowledge in several ways. The contributions of this study are also in tandem with the main themes of information and communication technologies for development (ICT4D) on promoting digital inclusion [111]. ICT4D is a research field that has gained prominence in the recent past due to its contribution towards socio-economic development of developing communities worldwide through: deployment and use ICT technologies in resource constrained areas particularly in underdeveloped/ developing regions of the world and; through experimental interventions which yield a range of benefits both tangible and intangible [112].

In terms of the technology, this study contributes to the body of knowledge in ICTs inform of a machine learning prediction model. The model which was build using

machine learning techniques has the ability to classify student academic performance and predict future grades for students in secondary schools. In terms of promoting socio-economic development, this study contributes towards providing quality education for all [1] through the use of this model for early identification of risky and non-performing students. The long-term effect is the increased number of secondary school students transiting to tertiary education, this will ensure developing countries have skilled manpower thus promoting economic growth. Finally, the study provides a comprehensive and analytical review of student' academic performance prediction, factors for predicting students' academic performance, algorithms used for prediction of student academic performance, feature selection techniques for student data and evaluation techniques for student performance prediction models.

#### **6.4 Limitations of the Study**

The study data concentrated on attributes that represent information about secondary school students hence the data may differ from that of students from other levels of education such as primary schools or universities. This means the model may not be directly applicable to other levels of study due to differences in the prevailing conditions. Again, the study was carried out in Kenya and focussed on factors affecting student academic performance in Kenya. Since other countries may differ in terms of education systems and other environmental settings, therefore the results of this study may not be easily generalizable to other countries.

#### **6.5 Recommendations for Future Work**

This study recommend further research on this topic that looks into the possibility of increasing the amount of data (instances) especially for the students that scored lower grades such as grade E so as to address the problem of class imbalance.

The study was conducted within secondary school students' environment in Kenya, however, it would be worthy trying to see if similar results are achievable in other developing countries other than Kenya. Again, though this study may not guarantee similar results if the model is applied to other academic levels such as primary schools and tertiary environments, the study further recommend future studies to be conducted on other levels of education including primary schools and tertiary institutions since prediction of student performance is a continuous process at all levels of study and has the potential to improve student performance and the quality of education.



## REFERENCES

- [1] W. B. Unicef, "Abolishing School Fees in Africa: Lessons from Ethiopia, Ghana, Kenya, Malawi and Mozambique," *Development Practice in Education*, 2009.
- [2] A. Brudevold-Newman, "The Impacts of Free Secondary Education: Evidence from Kenya.," Working Paper, 2016.
- [3] K. Braa and R. Vidgen, "Interpretation, intervention, and reduction in the organizational laboratory: a framework for in-context information system research," *Accounting, Management and Information Technologies*, vol. 9, no. 1, pp. 25-47, 1999.
- [4] F. Ahmad, I. NurHafieza and A. Azwa, "The prediction of students' academic performance using classification data mining techniques," *Applied Mathematical Sciences*, vol. 9, no. 129, pp. 6415-6426, 2015.
- [5] B. K. Baradwaj and S. Pal, "Mining Educational Data to Analyze Students' Performance," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, pp. 63-69, 2011.
- [6] P. Kayur, "Lowering the barrier to applying machine learning," in *Adjunct proceedings of the 23rd annual ACM symposium on User interface software and technology*, October, 2010.
- [7] K. Karthikeyan and P. Kavipriya, "Karthikeyan, Dr K., and Dr K. Karthikeyan. "On Improving Student Performance Prediction in Education Systems using Enhanced Data Mining Techniques," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 5, pp. 935-941, May 2017.
- [8] R. Asif, A. Merceron and M. K. Pathan, "Predicting Student Academic Performance at Degree Level: A Case Study," *International Journal of Intelligent Systems Technologies and Applications*, vol. 01, pp. 49-61, December 2014.
- [9] I. E. Livieris, K. Drakopoulou, V. T. Tampakas, T. A. Mikropoulos and P. Pintelas, "Predicting secondary school students' performance utilizing a semi-supervised learning approach," *Journal of Educational Computing Research*, p. 0735633117752614, 2018.
- [10] C. Paulo and A. Silva, "Using data mining to predict secondary school student

- performance,” pp. 5-12, 2008.
- [11] V. O. Oladokun, A. T. Adebajo and O. E. Charles-Owaba, “Predicting students academic performance using artificial neural network: A case study of an engineering course,” *The Pacific Journal of Science and Technology*, vol. 8, no. 1, pp. 72-79, 2008.
- [12] H. Agrawal and H. Mavani, “Student Performance Prediction using Machine Learning,” *International Journal of Engineering Research & Technology (IJERT)*, vol. 4, no. 03, pp. 111-113, March 2015.
- [13] O. Usman and A. Adenubi, “Artificial Neural Network (ANN) model for predicting students’ academic performance,” *Journal of Science and Information Technology*, vol. 1, pp. 23-37, October 2013.
- [14] S. S. A. Naser, “Predicting learners performance using artificial neural networks in linear programming intelligent tutoring system,” *International Journal of Artificial Intelligence & Applications*, pp. 3(2), 65, 2012.
- [15] J. Xu, Y. Han, D. Marcu and . M. a. Schaar, “Progressive Prediction of Student Performance in College Programs,” *AAAI*, pp. 1604-1610, 2017.
- [16] R. R. Kabra and R. S. Bichkar, “Performance prediction of engineering students using decision trees,” *International Journal of Computer Applications*, vol. 36(11), pp. 8-12, Dec 2011.
- [17] B. K. Bhardwaj and S. Pal, “Data Mining: A prediction for performance improvement using classification,” *International Journal of Computer Science and Information Security*, vol. 9, no. 4, April 2011.
- [18] J. Xu, H. M. Kyeong and V. D. S. Mihaela, “A machine learning approach for tracking and predicting student performance in degree programs,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11(5), pp. 742-753, 2017.
- [19] Guo, Bo, R. Zhang, G. Xu, C. Shi and L. Yang, “Predicting students performance in educational data mining,” in *Educational Technology (ISET), 2015 International Symposium*, 2015.
- [20] Shahiri, A. Mohamed and W. Husain, “A review on predicting student's performance using data mining techniques,” in *Procedia Computer Science*, 2015.
- [21] A. A. Nichat and A. B. Raut , “Predicting and Analysis of Student Performance

- Using Decision Tree Technique,” *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 4, pp. 7319-7327, 2017.
- [22] L. E. Livieris, K. Drakopoulou and P. Panagiotis, “Predicting students' performance using artificial neural networks,” in *8th PanHellenic Conference with International Participation Information and Communication Technologies in Education*, 2012.
- [23] L. K. Lau, “Institutional factors affecting student retention,” *Education-Indianapolis then Chula Vista*, pp. 124(1), 126-136, 2003.
- [24] P. P. Sundar , “A comparative study for predicting student’s academic performance using Bayesian Network Classifiers,” *IOSR Journal of Engineering (IOSRJEN)* , vol. 3, no. 2, pp. 37-42 , Feb 2013.
- [25] X. Ma, Y. Yang and Z. Zhou, “Using Machine Learning Algorithm to Predict Student Pass Rates In Online Education,” in *Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing*, April 2018.
- [26] T. M. Mitchell, *Machine learning*, vol. 45(37), Burr Ridge: IL:McGraw Hill, 1997, pp. 870-877.
- [27] M. Kuhn and J. Kjell, “Applied predictive modeling,” in *Springer*, New York, 2013.
- [28] M. Plagge, “Using artificial neural networks to predict first-year traditional students second year retention rates,” in *Proceedings of the 51st ACM Southeast Conference*, April, 2013.
- [29] S. Dawson, J. Jovanovic, G. Gašević and A. Pardo, “From prediction to impact: evaluation of a learning analytics retention program,” in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, March 2017.
- [30] D. Kabakchieva, “Student performance prediction by using data mining classification algorithms,” *International Journal of Computer Science and Management Research*, pp. 1(4), 686-690, 2012.
- [31] L. G. Moseley and D. M. Mead, “Predicting who will drop out of nursing courses: a machine learning exercise,” *Nurse education today*, pp. 28(4), 469-475, 2008.
- [32] Z. Iqbal, J. Qadir, A. Noor and F. Kamiran, “Machine Learning Based Student Grade Prediction: A Case Study,” *arXiv preprint arXiv:1708.08744*, 2017.

- [33] M. Imran, R. K. Muhammad and A. Ajith, “An ensemble of neural networks for weather forecasting,” *Neural Computing & Applications*, vol. 13.2, pp. 112-122, 2004.
- [34] M. Lernverfahren and zur Koreferenz Resolution, “A Machine Learning Approach for Coreference Resolution,” in *Decision Tree Machine Learning Approach for Coreference Resolution*.
- [35] T. M. Mitchell, “The Discipline of Machine Learning,” July 2006.
- [36] S. O. Danso, “An exploration of classification prediction techniques in data mining: the insurance domain,” *Master Degree thesis*, 2006.
- [37] A. Blum, *Machine learning theory*, Carnegie Melon Universit, School of Computer Science, 2007.
- [38] B. Hssina, A. Merbouha and H. Ezzikouri, “A comparative study of decision tree ID3 and C4. 5,” *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 13-19, 2014.
- [39] Aftab, Cheung, Kim, Thakkar and Yeddanapudi, “Information Theory, Information Theory and the Digital Age, 6.933 – Final Paper,” *Information Theory and The Digital revolution*, 2001.
- [40] E. Osmanbegović and M. Suljić , “Data mining approach for predicting student performance,” *Economic Review: Journal of Economics and Business*, vol. 10, no. 1, pp. 3-12, 2012.
- [41] A. D. Kumar and V. Radhika, “A Survey on Predicting Student Performance,” *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 5 (5), pp. 6147-6149, 2014.
- [42] S. H. Lin, “Data mining for student retention management,” *Journal of Computing Sciences in Colleges*, pp. 27(4), 92-99, 2012.
- [43] Y. Chen and M. Zhang, “MOOC student dropout: pattern and prevention,” in *Proceedings of the ACM Turing 50th Celebration Conference-China*, May 2017.
- [44] G. Kostopoulos, S. Kotsiantis and P. Pintelas, “Estimating student dropout in distance higher education using semi-supervised techniques,” in *Proceedings of the 19th Panhellenic Conference on Informatics*, October 2015.
- [45] A. Dey, “Machine Learning Algorithms: A Review,” *International Journal of*

- Computer Science and Information Technologies*, vol. 7, no. 3, pp. 1174-1179., 2016.
- [46] P. K. Dushyant, “Lowering the barrier to applying machine learning (Doctoral dissertation),” Washington, 2013.
- [47] B. Khan, M. S. Hayat and M. Daud, “Final Grade Prediction of Secondary School Student using Decision Tree,” *International Journal of Computer Applications*, vol. 115, no. 21, pp. 32-36, 2015.
- [48] W. Xing, G. Rui, P. Eva and G. Sean, “Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory,” *Computers in Human Behavior*, vol. 47, pp. 168-181, 2015.
- [49] T. Guenther, I. Mueller, M. Preuss, R. Kruse and B. A. Sabel, “A treatment outcome prediction model of visual field recovery using self-organizing maps,” *IEEE transactions on biomedical engineering*, pp. 56(3), 572-581, 2009.
- [50] M. A. Hall, “Correlation-based feature selection for machine learning,” 1999.
- [51] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 14, no. 2, pp. 1137-1145, 1995.
- [52] S. Narkhede, “Understanding Confusion Matrix,” 9 May 2018. [Online]. Available: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>. [Accessed 10 september 2018].
- [53] S. Narkhede, “Understanding AUC - ROC Curve,” 26 June 2018. [Online]. Available: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>. [Accessed 10 September 2018].
- [54] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” *Statistics surveys*, vol. 4, pp. 40-79, 2010.
- [55] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.
- [56] A. U. Khasanah and Harwati, “A Comparative Study to Predict Student’s Performance Using Educational Data Mining Techniques,” *IOP Conference*

*Series: Materials Science and Engineering*, vol. 215, no. 1, 2017.

- [57] M. Ramaswami and R. Bhaskaran, "A Study on Feature Selection Techniques in Educational Data Mining," *Journal of Computing*, vol. 1, no. 1, pp. 7-11, 2009.
- [58] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence 97*, Vols. 1-2, pp. 245-271, 1997.
- [59] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *The 20th international conference on machine learning (ICML-03)*, 2003.
- [60] C. E. Zwillig and M. Y. Wang, "Covariance based outlier detection with feature selection," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016.
- [61] N. Sánchez-Marño, A. A.-B. Alonso-Betanzos and M. Tombilla-Sanromán, "Filter methods for feature selection—a comparative study," in *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, Berlin, Heidelberg, 2007.
- [62] M. S. Mohamad, S. Omatu, S. Deris and S. Z. M. Hashim, "A model for gene selection and classification of gene expression data," *Artificial Life and Robotics*, vol. 11, no. 2, pp. 219-222, 2007.
- [63] R. Kohavi and H. J. George, "Wrappers for feature subset selection," *Artificial intelligence 97*, Vols. 1-2, pp. 273-324, 1997.
- [64] G. Isabelle and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, pp. 1157-1182, 3 March 2003.
- [65] S. Kaushik, 1 December 2016. [Online]. Available: <https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>. [Accessed 11 April 2019].
- [66] J. Feng, "Predicting Students' Academic Performance with Decision Tree and Neural Network," 2019.
- [67] G. Sharma and V. K. Santosh, "Analysis and Prediction of Student's Academic Performance in University Courses," *International Journal of Computer Applications*, vol. 160, no. 4, 2017.

- [68] G. Kaur and W. Singh, "Prediction of student performance using weka tool," *An International Journal of Engineering Sciences*, vol. 17, pp. 8-16, 2016.
- [69] H. Goker and I. B. Halil, "Improving an early warning system to prediction of student examination achievement," in *2014 13th International Conference on Machine Learning and Applications*, USA, 2014.
- [70] M. Gadhavi and P. Chirag, "Student final grade prediction based on linear regression," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 3, pp. 274-279, 2017.
- [71] S. A. Lynham, "Quantitative research and theory building: Dubin's method," *Advances in developing human resources*, vol. 4, no. 3, pp. 242-276, 2002.
- [72] M. P. Jama, M. L. E. Mapesela and A. A. Beylefeld, "Beylefeld, A. A., M. P. Jama, and M. L. E. Mapesela. "Theoretical perspectives on factors affecting the academic performance of students," *South African Journal of Higher Education*, vol. 22, no. 1, pp. 992-1005, 2008.
- [73] S. A. Lynham, "Advances in Developing Human," *Advances in Developing Human Resources*, vol. 4, no. 3, pp. 221-241, 2002.
- [74] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of educational research*, pp. 45(1), 89-125, 1975.
- [75] H. T. Khuong, *Evaluation of a conceptual model of student retention at a public urban commuter university*, Doctoral dissertation, Loyola University Chicago, 2014.
- [76] J. P. Bean, *The application of a model of turnover in work organizations to the student attrition process*, *The review of higher education*, 1983, pp. 6(2), 129-148.
- [77] N. A. Ogude, W. Kilfoil and G. Du Plessis, *An institutional model for improving student retention and success at the University of Pretoria*, *The International Journal of the First Year in Higher Education*, 2012, pp. 3(1), 21-34.
- [78] C. R. Kothari, "Research methodology: Methods and techniques," *New Age International*, 2004.
- [79] M. Saunders, P. Lewis and A. Thornhill, "Research methods for business students," 2009.
- [80] G. Perri and C. Bellamy, "Principles of Methodology: Research Design in Social

- Science,” *Journal of Multidisciplinary Evaluation*, vol. 8, no. 18, 2012.
- [81] J. G. Ponterotto, “Qualitative research in counseling psychology: A primer on research paradigms and philosophy of science,” *Journal of counseling psychology*, vol. 52, no. 2, pp. 126-136, 2005.
- [82] E. G. Guba, *The paradigm dialog*, Sage publications, 1990.
- [83] J. Jonker and B. Pennink, “The essence of research methodology: A concise guide for master and PhD students in management science,” *Springer Science & Business Media*, 2010.
- [84] S. Pather and D. Remenyi, “Some of the philosophical issues underpinning research in information systems: from positivism to critical realism,” in *Proceedings of the 2004 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, October, 2004.
- [85] Y. Levy and T. J. Ellis, “A guide for novice researchers on experimental and quasi-experimental studies in information systems research,” *Interdisciplinary Journal of Information, Knowledge and Management*, vol. 6, pp. 151-162, 2011.
- [86] S. P. Bates , “Types of Research Designs,” October 2006. [Online]. Available: <http://www.socialresearchmethods.net/kb/destypes.php>. [Accessed 28 May 2018].
- [87] H. Mohajan, “Two Criteria for Good Measurements in Research: Validity and Reliability,” 2018.
- [88] K. Khalid, H. Hilman and D. Kumar, “Get along with quantitative research process,” *International Journal of Research in Management*, vol. 2, no. 2, pp. 15-29, 2012.
- [89] M. D. Tongco, *Purposive Sampling as a Tool for Informant Selection, Ethnobotany Research and applications*, 2007.
- [90] P. Birmingham and D. Wilkinson, *Using research instruments: A guide for researchers*, Routledge, 2003.
- [91] M. Zohrabi, “Mixed Method Research: Instruments, Validity, Reliability and Reporting Findings,” *Theory & practice in language studies*, vol. 3, no. 2, pp. 254-262, 2013.
- [92] C. L. Kimberlin and G. W. Almut, “Validity and reliability of measurement



- instruments used in research,” *American Journal of Health-System Pharmacy*, vol. 65(23), pp. 2276-2284, Dec, 2008.
- [93] K. van, M. J. Sander, F. J. Dankers, A. Traverso and L. Wee, “Preparing data for predictive modelling,” in *Fundamentals of Clinical Data Science*, Cham, Springer, 2019, pp. 75-84.
- [94] S. Vergura, G. Acciani, V. Amoruso, G. E. Patrono and F. Vacca, “Descriptive and inferential statistics for supervising and monitoring the operation of PV plants,” *IEEE Transactions on Industrial Electronics* 56, vol. 56, no. 11, pp. 4456-4464, 2009.
- [95] M. J. Albers, “Quantitative Data Analysis - In the Graduate Curriculum,” *Journal of Technical Writing and Communication* 47, no. 2, pp. 215-233, 2017.
- [96] A. Mji and M. J. Glencross, “The role of a research resource centre in the training of social science researchers,” *South African Journal of Higher Education* 15, vol. 2, pp. 179-185, 2001.
- [97] M. Koutina and K. L. Kermanidis, “Predicting postgraduate students’ performance using machine learning techniques,” *Artificial intelligence applications and innovations. Springer, Berlin, Heidelberg*, pp. 159-168, 2011.
- [98] F. Y. Osisanwo, J. E. T. Akinsola and O. Awodele, “Supervised machine learning algorithms: classification and comparison,” *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128-138, 2017.
- [99] R. G. Sargent, “Validation and Verification of simulated models,” in *Simulation Conference (WSC), Proceedings of the 2009 Winter*, Dec, 2009.
- [100] K. Kira and L. A. Rendell, “The feature selection problem: Traditional methods and a new algorithm,” *AAAI*, vol. 2, pp. 129-134, 1992.
- [101] M. Pojon, “Using machine learning to predict student performance,” MS thesis, 2017.
- [102] M. Doshi and S. K. Chaturvedi, “Correlation based feature selection (CFS) technique to predict student Performance,” *International Journal of Computer Networks & Communications*, vol. 6, no. 3, pp. 197-206, 2014.
- [103] A. A. Saa, A.-E. Mostafa and S. Khaled, “Factors Affecting Students’ Performance in Higher Education: A Systematic Review of Predictive Data Mining

- Techniques,” *Technology, Knowledge and Learning*, pp. 1-32, 2019.
- [104] D. Ruta and B. Gabrys, “Classifier selection for majority voting,” *Information fusion*, vol. 6, no. 1, pp. 63-81, 2005.
- [105] D. M. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.
- [106] M. Mvurya, “Investigating prediction modelling of academic performance for students in rural schools in Kenya,” *Diss. University of Cape Town*, 2016.
- [107] V. Iosifidis and E. Ntoutsi, “Dealing with Bias via Data Augmentation in Supervised Learning Scenarios,” *Jo Bates Paul D. Clough Robert Jäschke*, p. 24, 2018.
- [108] I. H. M. Paris, L. S. Affendey and N. Mustapha, “Improving Academic Performance Prediction using Voting Technique in Data Mining,” *International Journal of Computer and Information Engineering*, vol. 4, no. 2, pp. 306-309, 2010.
- [109] A. Salini and U. Jeyapriya, “A Majority Vote Based EnsembleClassifier for Predicting Students Academic Performance,” *International Journal of Pure and Applied Mathematics*, vol. 118, no. 24, 2018.
- [110] M. Makhtar, “Contributions to Ensembles of Models for Predictive Toxicology Applications. On the Representation, Comparison and Combination of Models in Ensembles,” *Diss. University of Bradford*, 2012.
- [111] V. Bhadauria, R. Mahapatra and S. Nerur, “ICT4D: Exploring Emergent Themes,” *Emergent Research Forum (ERF)*, 2018.
- [112] J. Donner and K. Toyama, “Persistent themes in ICT4D Research: priorities for inter-methodological exchange,” in *57th Session of the International Statistics Institute*, Durban, South Africa, 2009.
- [113] A. Nurhuda and D. Rosita , “Prediction Student Graduation on Time Using Artificial Neural Networks on Data Mining Students STMIK Widya Cipta Dharma Samarinda,” in *Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government*, June 2017.
- [114] T. Beaubouef and J. Mason , “Why the High Attrition Rate for Computer Science

Students: Some Thoughts and Observations,” June 2005.

- [115] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “From data mining to knowledge discovery in databases,” in *AI magazine*, 1996, pp. 17(3), 37.
- [116] L. C. Borges, J. Bernardino and V. M. Marques, “Comparison of Data Mining techniques and tools for data classification,” pp. 113-116, July 2013.
- [117] A. Dhond, A. Gupta and V. Sanjeev, “Data mining techniques for optimizing inventories for electronic commerce,” in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, August 2000.
- [118] Y. L. Liu, Y. Ren and R. Dew, “Monetising user generated content using data mining techniques,” in *Proceedings of the Eighth Australasian Data Mining Conference*, December 2009.
- [119] N. Rehman, “Data Mining Techniques Methods Algorithms and Tools,” *International Journal of Computer Science and Mobile Computing*, vol. 6, no. 7, pp. 227-231, July 2017.
- [120] T. Power, R. McCormick and E. Asbeek-Brusse, “English in Action (EIA) (2017) A Quasi-Experimental Study of the Classroom Practices of English Language Teachers and the English Language Proficiency of Students, in Primary and Secondary Schools in Bangladesh.,” EIA QE Report , Dhaka, Bangladesh, 2017.
- [121] S. Ameri, M. J. Fard, R. B. Chinnam and C. Reddy, “Survival analysis based framework for early prediction of student dropouts,” in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, October 2016.
- [122] R. Barber and M. Sharkey, “Course Correction: Course correction using analytics to Predict Course Success,” in *Proceedings of the 2nd international conference on learning analytics and knowledge*, April 2012.
- [123] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “The KDD process for extracting useful knowledge from volumes of data,” *Communications of the ACM*, pp. 39(11), 27-34, 1996.
- [124] U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework,” in *KDD (Vol. 96)*, August 1996, pp. 82-88.

- [125] P. Branco, L. Torgo and R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains,” *ACM Computing Surveys (CSUR)*, pp. 49(2), 31, 2016.
- [126] R. W. Bybee, “The case for STEM education: Challenges and opportunities,” NSTA press, 2013.
- [127] J. Cheng, “Data-Mining Research in Education.,” *arXiv preprint arXiv:1703.10117*, 2017.
- [128] S. Domínguez-Almendros, N. Benítez-Parejo and A. R. Gonzalez-Ramirez, “Logistic regression models,” *Allergologia et immunopathologia*, pp. 39(5), 295-305, 2011.
- [129] F. Garcia and A. E. Retamar, “Towards building a bus travel time prediction model for Metro Manila,” in *Region 10 Conference (TENCON)*, November 2016.
- [130] R. A. Huebner, “A Survey of Educational Data-Mining Research,” *Research in higher education journal*, p. 19, 2013.
- [131] E. Mukhwana, S. Oure, S. Kiptoo, A. Kande, R. Njue, J. Too and D. K. Some, *State of University Education in Kenya*, Commission for University Education. Discussion Paper,4,3 , 2016.
- [132] C. I. Muntean, F. M. Nardini, F. Silvestri and R. Baraglia, *On learning prediction models for tourists paths*, *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2015, pp. 7(1), 8.
- [133] J. S. Plasman and M. A. Gottfried, *Applied STEM coursework, high school dropout rates, and students with learning disabilities*, Educational policy,0895904816673738, 2016.
- [134] J. Quinn, *Drop-out and completion in Higher Education in Europe*, European Union, 2013.
- [135] C. Romero and S. Ventura, *Data mining in education*, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2013, pp. 3(1), 12-27.
- [136] F. Ruiz-Ugalde, G. Cheng and M. Beetz, “Prediction of action outcomes using an object model. In Intelligent Robots and Systems (IROS),” *2010 IEEE/RSJ International Conference on*, pp. 1708-1713, October 2010.
- [137] J. Sarraipa, F. Ferreira, E. Marcelino-Jesus, A. Artifice, C. Lima and M. Kaddar, “Technological Innovations tackling Students dropout,” in *Proceedings of the 7th*

*International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, December 2016.

- [138] P. P. Sundar, “A Comparative Study for Predicting Student’s Academic Performance Using Bayesian Network Classifiers,” *IOSR Journal of Engineering (IOSRJEN) e-ISSN*, pp. 2250-3021, 2013.
- [139] M. B. Kerby, “Toward a new predictive model of student retention in higher education: An application of classical sociological theory,” *Journal of College Student Retention: Research, Theory & Practice*, pp. 17(2), 138-161, 2015.
- [140] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, *From data mining to knowledge discovery in databases*, AI magazine, 1996, pp. 17(3), 37.
- [141] U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework,” in *KDD*, August 1996.
- [142] H. J. Miller and J. Han, “Geographic data mining and knowledge discovery,” *Taylor & Francis*, 2001.
- [143] D. T. Larose, “Introduction to data mining,” *John Wiley & Sons, Inc*, pp. 1-26, 2005.
- [144] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “The KDD process for extracting useful knowledge from volumes of data,” *Communications of the ACM*, pp. 39(11), 27-34., 1996.
- [145] M. Goebel and L. Gruenwald, “A survey of data mining and knowledge discovery software tools,” *ACM SIGKDD explorations newsletter*, pp. 1(1), 20-33, 1999.
- [146] L. B. Klinkenberg, “A quantitative analysis of a mandatory student success course on first-time full-time student college academic progress and persistence,” 2013.
- [147] D. A. De Vaus, “Research design in Social Research,” London, SAGE, pp. 1-16.
- [148] N. L. Leech and A. J. Onwuegbuzie, “A typology of mixed methods research designs,” vol. 43, no. 2, pp. 265-275, March, 2009.
- [149] C. V. N. Index, “The zettabyte era—trends and analysis,” 2013. [Online]. Available: <https://www.itproportal.com/features/the-importance-of-big-data-and-analytics-in-the-era-of-digital-transformation/>. [Accessed 24 July 2018].
- [150] T. M. Mitchell, “Machine learning and data mining,” *Communications of the ACM*, vol. 42(11), pp. 30-36, 1999.

- [151] J. Cavazo, C. Dubach, F. Agakov, E. Bonilla, M. F. O'Boyle, G. Fursin and O. Temam, "Automatic performance model construction for the fast software exploration of new hardware designs," in *Proceedings of the 2006 international conference on Compilers, architecture and synthesis for embedded systems*, October 2006.
- [152] M. Vahdat, G. Alessandro, O. Luca, A. Davide, F. Mathias and R. Matthias, "Advances in learning analytics and educational data mining," in *ESANN2015*, 2015.
- [153] P. Thakar , A. Mehta and Manisha, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue," *International Journal of Computer Applications*, vol. 110, no. 15, p. 60, January 2015 .
- [154] O. Otach, "Abolishing school fees in Africa: Lessons from Ethiopia, Ghana, Kenya, Malawi and Mozambique," *Kenya Literature Bureau.*, 2008.
- [155] A. Elbadrawy, R. S. Studham and G. Karypis, "Collaborative multi-regression models for predicting students' performance in course activities," in *Fifth International Conference on Learning Analytics And Knowledge*, 2015.
- [156] W. Xindong, K. Vipin, R. Q. J, G. Joydeep, Y. Qiang, M. Hiroshi, J. M. Geoffrey, N. Angus, L. Bing, S. Y. Philip, Z. Zhi-Hua, S. Michael, J. H. David and S. Dan, "Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, GeTop 10 algorithms in data mining," in *Knowledge and information systems*, 2008.
- [157] C. Machinery, "Computing machinery and intelligence-AM Turing," *Mind*, vol. 59.236, p. 433, 1950.
- [158] S. B. Kotsiantis, C. J. Pierrakeas and I. D. Zaharaki, "Kotsiantis, S., C. Pierrakeas, and P. Pintelas. "Efficiency of machine learning techniques in predicting students' performance in distance learning systems," *Educational Software Development Laboratory Department of Mathematics, University of Patras, Greece*, 2002.
- [159] L. Pat, "Selection of relevant features in machine learning," in *The AAAI Fall symposium on relevance*, 1994.
- [160] H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," *IEEE Transactions on Knowledge & Data Engineering*, vol. 4, pp. 491-502, 2005.

- [161] V. Iosifidis and E. Ntoutsi, "Dealing with Bias via Data Augmentation in Supervised Learning Scenarios," *Jo Bates Paul D. Clough Robert Jäschke*, p. 24, 2018.
- [162] D. Silva, I. Nunes, D. Spatti, R. A. Flauzino, L. H. B. Liboni and S. F. dos Reis Alves, "Artificial neural network architectures and training processes," in *Artificial Neural Networks*, Springer, Cham, 2017.
- [163] J. M. Braxton, A. V. Shaw Sullivan and R. M. Johnson, *Appraising Tinto's theory of college student departure*, Higher Education-New York-Agathon Press Incorporated, 1997, pp. 12, 107-164.

## APPENDICES

### Appendix I: Student Academic Performance Prediction Questionnaire

The purpose of this questionnaire is to facilitate collection of data that will be used to develop a machine learning model for prediction of secondary school students' academic performance in Kenya. The respondents should be students in tertiary institutions who have studied and completed their secondary school studies in Kenya.

#### Section I: General Information

Date:                    \_\_\_ / \_\_\_ / 2019

Institution:

.....

County: .....

#### Section II: Student Demographic Information

1. Gender:            Male [ ]                    Female [ ]

2. What was your age while in Form IV?

Below 14 Years [ ]      14 – 18 Years [ ]                    Above 18 Years [ ]

3. Did you have any form of disability while in high school?    Yes [ ]    No [ ]

4. What is your religion?    Christian [ ]    Muslim [ ]    Others [ ]

If            your            answer            is            **others**,            please            specify:

.....

#### Section III: Family Information

5. Did you live with your parents while in high school?    Yes [ ]    No [ ]

If **No**, where were your parent(s)?

Deceased [ ]    Divorced [ ]    Separated [ ]    Am Adopted [ ]    Others [ ]

If            your            answer            is            **others**,            please            specify:

.....

6. Have you ever witnessed conflicts between your parents?    Yes [ ]    No [ ]

If **Yes**, how often?    Very often [ ]    Often [ ]    Rarely [ ]

7. What kind of family structure do you come from?

Nuclear Family [ ]    Single parent [ ]    Extended Family [ ]    Step Family [ ]

8. Which of the following best describes your family structure? Tick [✓] appropriately

Description	Yes	No
I live with my parents, brothers and sisters only		



I live with my brothers, sisters and one parent		
I live with my parents, brothers, sisters, cousins, grandparents, uncles and aunts		
I live with a step parent		
I live with people who are not my real parents		

9. Did you have any difficulties in fees payment? Yes [ ] No [ ]
10. Who used to pay your school fees? Parents [ ] Guardian [ ] Others [ ]

If your answer is **others**, specify: .....

11. Where your parents employed while you were in high school?

Yes (Both) [ ] Yes (One) [ ] No [ ]

If your answer is **No**, what did he/she/they do to earn a living:

.....

12. Select by checking [√] appropriately your parents level of education

Parent	Degree and above	Diploma	Certificate	KCSE	KCPE	School Dropout	No formal education
Father							
Mother							

#### Section IV: Co-Curricular Activities

13. Did you participate in any co-curricular activities like games, drama etc? Yes [ ] No [ ]

If **Yes**, how often did you participate in co-curricular activities?

Daily [ ] Once in a week [ ] Once monthly [ ] Not Often [ ] Never [ ]

14. Did you ever become a member of any school team? Yes [ ] No [ ]

15. Did you ever represent your school in any co-curricular activity? Yes [ ]

No [ ]

If **Yes**, Up to which level? Zonal [ ] District [ ] Provincial [ ] National [ ]

#### Section V: Academic

16. Indicate the number of subject taken in each of the following levels of study

Level	No. of Subjects Taken
Form I	
Form II	
Form II	
Form IV	

17. Were all the subjects indicated above examined in the final KCSE examination?

Yes [ ] No [ ]

If No, specify how many were not examined in KCSE: .....

18. Did your school allow students to specialize in specific subjects?

Yes [ ] No [ ]

If Yes, at what level did you start to specialize?

Form I [ ] Form II [ ] Form III [ ] Form IV [ ]

19. Did you ever change schools while in high school? Yes [ ] No [ ]

If Yes, how many times? .....

20. Did you ever repeat a class while in high school? Yes [ ] No [ ]

If Yes, which Form: .....

21. Fill the table below by putting a check [√] against the average grade you scored at the end of the said level of study in secondary school

LEVEL	A (A,A-)	B (B+,B,B-)	C (C+,C,C-)	D (D+,D,D-)	E
Form I					
Form II					
Form III					
Mock Exam					
KCSE					

22. Check [√] appropriately against the learning styles that best characterised your learning in high school

Learning Style	Yes	No
Use of graphics such as charts, graphs, diagrams, demonstration, PowerPoint presentations, blackboard, whiteboard or flipcharts		
Discussion with study groups or teacher, use of tapes or recordings or lectures		



29. To which extend do you think the following affected your academic performance?

Resource	Very Low	Low	Moderate	High	Very High
Absenteeism from class					
Teacher Absenteeism from class					
Failure to complete the syllabus					
Co-curricular activities					
Access drugs and alcohol					
Challenges during examination period					
Presence or Absence of a role model					

#### Section VI: Secondary School Demographics

30. Type of School: National [ ] Extra-County [ ] County [ ] Sub-County [ ]

31. Was the school boarding or a day school?

Boarding [ ] Day [ ] Both [ ]

32. What was the social composition of your school?

Mixed School [ ] Boys Only [ ] Girls Only [ ]

33. Did the school have the following facilities:

Facility	Yes	No
Teaching Laboratories		
Library		
Computer Laboratory		
Access to Power/Electricity		
Access to Internet		

34. If your answer is **Yes** in any of the items listed above, in a scale of 1-6 (1-worst, 2-worse, 3-Bad, 4-Good, 5-Better, 6-Best), how would you rate the state of following resources in the school by then? Tick [√] appropriately

Resource	1	2	3	4	5	6
Teaching Laboratories						
Library						

Computer Laboratory						
Access Power/Electricity						
Access to Internet						

35. In your own opinion, to which extent do you think the presence or absence of the following affected your academic performance? Tick [√] appropriately

<b>Resource</b>	<b>Very High</b>	<b>High</b>	<b>Moderate</b>	<b>Low</b>	<b>Very Low</b>
Teaching Laboratories					
Library					
Computer Laboratory					
Access Power/Electricity					
Access to Internet					

## Appendix II: Letter of Approval



MASINDE MULIRO UNIVERSITY OF SCIENCE AND TECHNOLOGY (MMUST)

Tel: 056-30870  
Fax: 056-30153  
E-mail: [directordps@mmust.ac.ke](mailto:directordps@mmust.ac.ke)  
Website: [www.mmust.ac.ke](http://www.mmust.ac.ke)

P.O Box 190  
Kakamega – 50100  
Kenya

### Directorate of Postgraduate Studies

Ref: MMU/COR: 509099

Date: 20<sup>th</sup> December, 2018

Obadiah Matolo Musau  
SIT/H/09/11  
P.O. Box 190-50100,  
KAKAMEGA.

Dear Mr. Musau,

#### RE: APPROVAL OF PROPOSAL

I am pleased to inform you that the Directorate of Postgraduate Studies has considered and approved your Ph.D proposal entitled: *'Machine Learning Model for Prediction of Students' Academic Performance, Kenya* and appointed the following as supervisors:

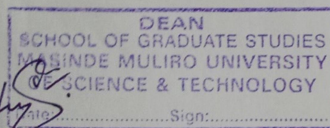
1. Dr. Kelvin Omieno - School of Computing and Informatics, MMUST
2. Dr. Raphael Angulu - School of Computing and Informatics, MMUST

You are required to submit through your supervisor(s) progress reports every three months to the Director Postgraduate Studies. Such reports should be copied to the following: Chairman, School of Computing and Informatics Graduate Studies Committee and Chairman, Computer Science Department. Kindly adhere to research ethics consideration in conducting research.

It is the policy and regulations of the University that you observe a deadline of three years from the date of registration to complete your Ph.D thesis. Do not hesitate to consult this office in case of any problem encountered in the course of your work.

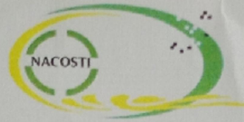
We wish you the best in your research and hope the study will make original contribution to knowledge.

Yours Sincerely,



Prof. John Obiri  
DIRECTOR, DIRECTORATE OF POSTGRADUATE STUDIES

### Appendix III: Letter of Research Authorization



#### NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY AND INNOVATION

Telephone: +254-20-2213471,  
2241349, 3310571, 2219420  
Fax: +254-20-318245, 318249  
Email: dg@nacosti.go.ke  
Website : www.nacosti.go.ke  
When replying please quote

NACOSTI, Upper Kabete  
Off Waiyaki Way  
P.O. Box 30623-00100  
NAIROBI-KENYA

Ref. No. **NACOSTI/P/19/01375/27626**

Date: **17<sup>th</sup> January, 2019**

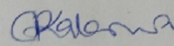
Obadiah Matolo Musau  
Masinde Muliro University of Science and Technology  
P. O Box 190-50100  
**KAKAMEGA**

#### RE: RESEARCH AUTHORIZATION

Following your application for authority to carry out research on “*Machine learning model for prediction of students’ academic performance, Kenya*” I am pleased to inform you that you have been authorized to undertake research in **all Counties** for the period ending **17<sup>th</sup> January, 2020**.

You are advised to report to **the County Commissioners and the County Directors of Education, all Counties** before embarking on the research project.

Kindly note that, as an applicant who has been licensed under the Science, Technology and Innovation Act, 2013 to conduct research in Kenya, you shall deposit **a copy** of the final research report to the Commission within **one year** of completion. The soft copy of the same should be submitted through the Online Research Information System.

  
**GODFREY P. KALERWA MSc., MBA, MKIM**  
**FOR: DIRECTOR-GENERAL/CEO**

Copy to:

The County Commissioners  
All Counties.

The County Directors of Education  
All Counties.

**Appendix IV: Research Permit**


**THE SCIENCE, TECHNOLOGY AND INNOVATION ACT, 2013**

The Grant of Research Licenses is guided by the Science, Technology and Innovation (Research Licensing) Regulations, 2014.


**CONDITIONS**

1. The License is valid for the proposed research, location and specified period.
2. The License and any rights thereunder are non-transferable.
3. The Licensee shall inform the County Governor before commencement of the research.
4. Excavation, filming and collection of specimens are subject to further necessary clearance from relevant Government Agencies.
5. The License does not give authority to transfer research materials.
6. NACOSTI may monitor and evaluate the licensed research project.
7. The Licensee shall submit one hard copy and upload a soft copy of their final report within one year of completion of the research.
8. NACOSTI reserves the right to modify the conditions of the License including cancellation without prior notice.

National Commission for Science, Technology and innovation  
P.O. Box 30623 - 00100, Nairobi, Kenya  
TEL: 020 400 7000, 0713 788787, 0735 404245  
Email: dg@nacosti.go.ke, registry@nacosti.go.ke  
Website: www.nacosti.go.ke



**REPUBLIC OF KENYA**



**National Commission for Science, Technology and Innovation**

**RESEARCH LICENSE**

Serial No.A **22754**

**CONDITIONS: see back page**


**THIS IS TO CERTIFY THAT:**  
**MR. OBADIAH MATOLO MUSAU**  
of MASINDE MULIRO UNIVERSITY OF SCIENCE AND TECHNOLOGY, 56428-200 NAIROBI, has been permitted to conduct research in All Counties

**on the topic: MACHINE LEARNING MODEL FOR PREDICTION OF STUDENTS' ACADEMIC PERFORMANCE, KENYA**

**for the period ending:**  
**17th January, 2020**

*[Handwritten Signature]*  
Applicant's Signature

**Permit No. : NACOSTI/P/19/01375/27626**  
**Date Of Issue : 17th January, 2019**  
**Fee Received :Ksh 2000**



*[Handwritten Signature]*  
**Director General**  
**National Commission for Science, Technology & Innovation**



## Appendix V: Sample Source Code for Student Performance Prediction Model

```
// Generated with Weka 3.6.9
```

```
package weka.classifiers;
import weka.core.Attribute;
import weka.core.Capabilities;
import weka.core.Capabilities.Capability;
import weka.core.Instance;
import weka.core.Instances;
import weka.core.RevisionUtils;
import weka.classifiers.Classifier;

public class WekaWrapper
    extends Classifier {

    /**
     * Returns only the toString() method.
     *
     * @return a string describing the classifier
     */
    public String globalInfo() {
        return toString();
    }

    /**
     * Returns the capabilities of this classifier.
     *
     * @return the capabilities
     */
    public Capabilities getCapabilities() {
        weka.core.Capabilities result = new weka.core.Capabilities(this);

        result.enable(weka.core.Capabilities.Capability.NOMINAL_ATTRIBUTES);
    }
}
```

```

result.enable(weka.core.Capabilities.Capability.NUMERIC_ATTRIBUTES);
result.enable(weka.core.Capabilities.Capability.DATE_ATTRIBUTES);
result.enable(weka.core.Capabilities.Capability.MISSING_VALUES);
result.enable(weka.core.Capabilities.Capability.NOMINAL_CLASS);
result.enable(weka.core.Capabilities.Capability.MISSING_CLASS_VALUES);

result.setMinimumNumberInstances(0);

return result;
}

/**
 * only checks the data against its capabilities.
 *
 * @param i the training data
 */
public void buildClassifier(Instances i) throws Exception {
    // can classifier handle the data?
    getCapabilities().testWithFail(i);
}

/**
 * Classifies the given instance.
 *
 * @param i the instance to classify
 * @return the classification result
 */
public double classifyInstance(Instance i) throws Exception {
    Object[] s = new Object[i.numAttributes()];

    for (int j = 0; j < s.length; j++) {
        if (!i.isMissing(j)) {
            if (i.attribute(j).isNominal())
                s[j] = new String(i.stringValue(j));

```

```

        else if (i.attribute(j).isNumeric())
            s[j] = new Double(i.value(j));
        }
    }

    // set class value to missing
    s[i.classIndex()] = null;

    return WekaClassifier.classify(s);
}

/**
 * Returns the revision string.
 *
 * @return the revision
 */
public String getRevision() {
    return RevisionUtils.extract("1.0");
}

/**
 * Returns only the classnames and what classifier it is based on.
 *
 * @return a short description
 */
public String toString() {
    return "Auto-generated classifier wrapper, based on weka.classifiers.trees.J48
(generated with Weka 3.6.9).\n" + this.getClass().getName() + "/WekaClassifier";
}

/**
 * Runs the classifier from commandline.
 *
 * @param args the commandline arguments

```

```

*/
public static void main(String args[]) {
    runClassifier(new WekaWrapper(), args);
}
}

class WekaClassifier {

    public static double classify(Object[] i)
        throws Exception {

        double p = Double.NaN;
        p = WekaClassifier.N7f58c3a0(i);
        return p;
    }

    static double N7f58c3a0(Object []i) {
        double p = Double.NaN;
        if (i[9] == null) {
            p = 1;
        } else if (i[9].equals("b")) {
            p = WekaClassifier.N2275ebda1(i);
        } else if (i[9].equals("a")) {
            p = WekaClassifier.N4b92aaaa23(i);
        } else if (i[9].equals("c")) {
            p = WekaClassifier.N39a20bdb27(i);
        } else if (i[9].equals("d")) {
            p = WekaClassifier.N35bc8dd238(i);
        } else if (i[9].equals("e")) {
            p = 4;
        }
        return p;
    }

    static double N2275ebda1(Object []i) {
        double p = Double.NaN;

```

```

if (i[1] == null) {
    p = 1;
} else if (i[1].equals("no")) {
p = WekaClassifier.N476f13672(i);
} else if (i[1].equals("yes")) {
p = WekaClassifier.N652db7c410(i);
}
return p;
}

static double N476f13672(Object []i) {
    double p = Double.NaN;
    if (i[8] == null) {
        p = 1;
    } else if (i[8].equals("a")) {
p = WekaClassifier.N48a0058e3(i);
} else if (i[8].equals("b")) {
    p = 1;
} else if (i[8].equals("c")) {
p = WekaClassifier.N51721af56(i);
} else if (i[8].equals("d")) {
    p = 1;
} else if (i[8].equals("e")) {
    p = 1;
}
return p;
}

static double N48a0058e3(Object []i) {
    double p = Double.NaN;
    if (i[18] == null) {
        p = 1;
    } else if (i[18].equals("1")) {
        p = 1;
    } else if (i[18].equals("2")) {
        p = 2;
    }
}

```

```

    } else if (i[18].equals("3")) {
        p = 1;
    } else if (i[18].equals("4")) {
        p = 1;
    } else if (i[18].equals("5")) {
        p = 1;
    } else if (i[18].equals("6")) {
        p = WekaClassifier.N631fd4fc4(i);
    }
    return p;
}
static double N631fd4fc4(Object []i) {
    double p = Double.NaN;
    if (i[11] == null) {
        p = 2;
    } else if (i[11].equals("1")) {
        p = 2;
    } else if (i[11].equals("2")) {
        p = 0;
    } else if (i[11].equals("3")) {
        p = WekaClassifier.N6e39d87e5(i);
    }
    return p;
}
static double N6e39d87e5(Object []i) {
    double p = Double.NaN;
    if (i[4] == null) {
        p = 0;
    } else if (i[4].equals("7")) {
        p = 0;
    } else if (i[4].equals("8")) {
        p = 0;
    } else if (i[4].equals("9")) {
        p = 1;
    }
}

```

```

    } else if (i[4].equals("10")) {
        p = 0;
    } else if (i[4].equals("11")) {
        p = 0;
    } else if (i[4].equals("12")) {
        p = 0;
    } else if (i[4].equals("13")) {
        p = 0;
    } else if (i[4].equals("14")) {
        p = 0;
    } else if (i[4].equals("15")) {
        p = 0;
    } else if (i[4].equals("16")) {
        p = 0;
    }
    return p;
}

static double N51721af56(Object []i) {
    double p = Double.NaN;
    if (i[12] == null) {
        p = 2;
    } else if (i[12].equals("yes")) {
        p = 2;
    } else if (i[12].equals("no")) {
        p = WekaClassifier.N1936e9207(i);
    }
    return p;
}

static double N1936e9207(Object []i) {
    double p = Double.NaN;
    if (i[4] == null) {
        p = 1;
    } else if (i[4].equals("7")) {
        p = 1;
    }
}

```

```

    } else if (i[4].equals("8")) {
        p = 2;
    } else if (i[4].equals("9")) {
        p = 0;
    } else if (i[4].equals("10")) {
        p = 0;
    } else if (i[4].equals("11")) {
        p = WekaClassifier.N49376ce8(i);
    } else if (i[4].equals("12")) {
        p = 1;
    } else if (i[4].equals("13")) {
        p = 2;
    } else if (i[4].equals("14")) {
        p = 1;
    } else if (i[4].equals("15")) {
        p = 1;
    } else if (i[4].equals("16")) {
        p = 1;
    }
    return p;
}

```

.  
.  
.

```

static double N1f7ba0e741(Object []i) {
    double p = Double.NaN;
    if (i[3] == null) {
        p = 2;
    } else if (i[3].equals("1")) {
        p = 2;
    } else if (i[3].equals("2")) {

```



```
    p = 2;
  } else if (i[3].equals("3")) {
    p = 3;
  } else if (i[3].equals("4")) {
    p = 1;
  } else if (i[3].equals("5")) {
    p = 2;
  }
  return p;
}
}
```

## Appendix VI: J48 Pruned Decision Tree

MG = b

| AS = 1

| | Internet = no

| | | NSF1 = 7: c (1.0)

| | | NSF1 = 8: b (1.0)

| | | NSF1 = 9: b (0.0)

| | | NSF1 = 10: c (7.0)

| | | NSF1 = 11

| | | | NSF2 = 6: b (0.0)

| | | | NSF2 = 7: b (0.0)

| | | | NSF2 = 8: c (7.0/2.0)

| | | | NSF2 = 9: b (0.0)

| | | | NSF2 = 10: d (12.0)

| | | | NSF2 = 11

| | | | | F3G = a: b (3.0/1.0)

| | | | | F3G = b

| | | | | | ME = 1: c (16.0/4.0)

| | | | | | ME = 2

| | | | | | | FE = 1: b (2.0)

| | | | | | | FE = 2: b (2.0)

| | | | | | | FE = 3: b (5.0/1.0)

| | | | | | | FE = 4: c (3.0)

| | | | | | | FE = 5: c (2.0)

| | | | | | | ME = 3

| | | | | | | EC = yes: b (9.0)

| | | | | | | EC = no

| | | | | | | | FE = 1: a (0.0)

| | | | | | | | FE = 2: a (0.0)

| | | | | | | | FE = 3: b (4.0/1.0)

| | | | | | | | FE = 4: a (8.0)

| | | | | | | | FE = 5: a (0.0)

| | | | | | | ME = 4: b (12.0/2.0)

| | | | | ME = 5: c (3.0/1.0)  
| | | | | F3G = c  
| | | | | F2G = b  
| | | | | FE = 1: c (7.0/2.0)  
| | | | | FE = 2: c (2.0)  
| | | | | FE = 3: b (6.0/1.0)  
| | | | | FE = 4: c (3.0/1.0)  
| | | | | FE = 5: c (2.0)  
| | | | | F2G = a: b (1.0)  
| | | | | F2G = c  
| | | | | DF = no  
| | | | | EC = yes  
| | | | | CL = yes: c (2.0)  
| | | | | CL = no: d (2.0)  
| | | | | EC = no: b (3.0)  
| | | | | DF = yes: d (13.0/1.0)  
| | | | | F2G = d: b (2.0/1.0)  
| | | | | F2G = e: c (0.0)  
| | | | | F3G = d: b (0.0)  
| | | | | F3G = e: b (0.0)  
| | | | NSF2 = 12: b (0.0)  
| | | | NSF2 = 13: d (8.0)  
| | | | NSF2 = 14: b (0.0)  
| | | | NSF2 = 15: b (0.0)  
| | | NSF1 = 12  
| | | | F3G = a: b (1.0)  
| | | | F3G = b  
| | | | DF = no: b (38.0/7.0)  
| | | | DF = yes  
| | | | | F1G = a  
| | | | | NSF2 = 6: b (0.0)  
| | | | | NSF2 = 7: b (0.0)  
| | | | | NSF2 = 8: b (1.0)  
| | | | | NSF2 = 9: b (0.0)

| | | | | | NSF2 = 10: b (2.0)  
| | | | | | NSF2 = 11: b (0.0)  
| | | | | | NSF2 = 12  
| | | | | | CL = yes: b (13.0/4.0)  
| | | | | | CL = no: c (3.0)  
| | | | | | NSF2 = 13: b (0.0)  
| | | | | | NSF2 = 14: b (0.0)  
| | | | | | NSF2 = 15: b (0.0)  
| | | | | | F1G = b: c (11.0/3.0)  
| | | | | | F1G = d: c (0.0)  
| | | | | | F1G = c: c (1.0)  
| | | | | | F1G = e: c (0.0)  
| | | | F3G = c  
| | | | DF = no  
| | | | Religion = muslim: b (2.0)  
| | | | Religion = christian: d (10.0/1.0)  
| | | | Religion = others: d (0.0)  
| | | | DF = yes  
| | | | CL = yes: c (9.0/1.0)  
| | | | CL = no: b (5.0/1.0)  
| | | F3G = d: b (0.0)  
| | | F3G = e: b (0.0)  
| | | NSF1 = 13  
| | | F3G = a: c (1.0)  
| | | F3G = b: d (8.0/1.0)  
| | | F3G = c: c (5.0/1.0)  
| | | F3G = d: d (0.0)  
| | | F3G = e: d (0.0)  
| | | NSF1 = 14: c (1.0)  
| | | NSF1 = 15: b (0.0)  
| | | NSF1 = 16: b (1.0)  
| | Internet = yes  
| | | DF = no  
| | | FE = 1

| | | | | F1G = a  
 | | | | | ME = 1: c (3.0/1.0)  
 | | | | | ME = 2: a (0.0)  
 | | | | | ME = 3: a (18.0/2.0)  
 | | | | | ME = 4: a (2.0)  
 | | | | | ME = 5: a (0.0)  
 | | | | | F1G = b: b (6.0/1.0)  
 | | | | | F1G = d: a (0.0)  
 | | | | | F1G = c: a (0.0)  
 | | | | | F1G = e: a (0.0)  
 | | | | FE = 2: b (5.0/1.0)  
 | | | | FE = 3: b (26.0/3.0)  
 | | | | FE = 4  
 | | | | | F3G = a: c (1.0)  
 | | | | | F3G = b  
 | | | | | F1G = a  
 | | | | | ME = 1: a (0.0)  
 | | | | | ME = 2: a (0.0)  
 | | | | | ME = 3: a (0.0)  
 | | | | | ME = 4: a (16.0/3.0)  
 | | | | | ME = 5: b (2.0)  
 | | | | | F1G = b: b (4.0)  
 | | | | | F1G = d: a (0.0)  
 | | | | | F1G = c: a (0.0)  
 | | | | | F1G = e: a (0.0)  
 | | | | | F3G = c  
 | | | | | NSF1 = 7: c (0.0)  
 | | | | | NSF1 = 8: c (0.0)  
 | | | | | NSF1 = 9: c (0.0)  
 | | | | | NSF1 = 10: c (0.0)  
 | | | | | NSF1 = 11: c (3.0)  
 | | | | | NSF1 = 12: b (6.0/2.0)  
 | | | | | NSF1 = 13: c (0.0)  
 | | | | | NSF1 = 14: c (0.0)

| | | | | NSF1 = 15: c (0.0)  
| | | | | NSF1 = 16: c (0.0)  
| | | | | F3G = d: a (0.0)  
| | | | | F3G = e: a (0.0)  
| | | | FE = 5: b (73.0/4.0)  
| | | DF = yes  
| | | | F3G = a: b (3.0/1.0)  
| | | | F3G = b  
| | | | | F2G = b  
| | | | | EC = yes: c (13.0/4.0)  
| | | | | EC = no: b (36.0/11.0)  
| | | | | F2G = a: b (6.0)  
| | | | | F2G = c: c (4.0)  
| | | | | F2G = d: b (0.0)  
| | | | | F2G = e: b (0.0)  
| | | | F3G = c  
| | | | | ME = 1  
| | | | | FE = 1: c (3.0)  
| | | | | FE = 2: d (7.0)  
| | | | | FE = 3: d (0.0)  
| | | | | FE = 4: d (0.0)  
| | | | | FE = 5: d (0.0)  
| | | | | ME = 2: c (5.0/1.0)  
| | | | | ME = 3: b (7.0/1.0)  
| | | | | ME = 4: c (3.0/1.0)  
| | | | | ME = 5: c (1.0)  
| | | | F3G = d: b (0.0)  
| | | | F3G = e: b (0.0)  
| AS = 2  
| | NSF2 = 6: b (0.0)  
| | NSF2 = 7: b (1.0)  
| | NSF2 = 8  
| | | F1G = a: a (2.0/1.0)  
| | | F1G = b: b (2.0)

| | | F1G = d: b (0.0)  
| | | F1G = c: b (0.0)  
| | | F1G = e: b (0.0)  
| | NSF2 = 9: b (3.0/1.0)  
| | NSF2 = 10: b (52.0/3.0)  
| | NSF2 = 11: b (12.0)  
| | NSF2 = 12  
| | | FE = 1: b (1.0)  
| | | FE = 2: b (1.0)  
| | | FE = 3  
| | | | Internet = no: c (2.0)  
| | | | Internet = yes: b (9.0/1.0)  
| | | FE = 4: b (5.0)  
| | | FE = 5  
| | | | CL = yes: a (59.0/4.0)  
| | | | CL = no: b (2.0)  
| | NSF2 = 13: b (2.0)  
| | NSF2 = 14: b (0.0)  
| | NSF2 = 15: b (0.0)  
| AS = 3  
| | F1G = a  
| | | F3G = a  
| | | | DF = no  
| | | | | F2G = b: a (128.0/8.0)  
| | | | | F2G = a: b (3.0)  
| | | | | F2G = c: b (1.0)  
| | | | | F2G = d: a (0.0)  
| | | | | F2G = e: a (0.0)  
| | | | DF = yes: b (7.0/1.0)  
| | | F3G = b  
| | | | FE = 1  
| | | | | Internet = no  
| | | | | ME = 1  
| | | | | | CL = yes: b (9.0/3.0)

| | | | | | CL = no: c (2.0)  
| | | | | | ME = 2: a (0.0)  
| | | | | | ME = 3: a (11.0)  
| | | | | | ME = 4: b (2.0)  
| | | | | | ME = 5: a (0.0)  
| | | | | Internet = yes: a (73.0/3.0)  
| | | | FE = 2  
| | | | | NSF1 = 7: b (0.0)  
| | | | | NSF1 = 8: b (0.0)  
| | | | | NSF1 = 9: c (1.0)  
| | | | | NSF1 = 10: b (0.0)  
| | | | | NSF1 = 11  
| | | | | DF = no: b (2.0)  
| | | | | DF = yes: c (2.0)  
| | | | | NSF1 = 12: b (8.0)  
| | | | | NSF1 = 13: b (0.0)  
| | | | | NSF1 = 14: b (0.0)  
| | | | | NSF1 = 15: b (0.0)  
| | | | | NSF1 = 16: b (0.0)  
| | | | FE = 3: b (16.0/3.0)  
| | | | FE = 4  
| | | | | NSF2 = 6: a (0.0)  
| | | | | NSF2 = 7: a (0.0)  
| | | | | NSF2 = 8  
| | | | | DF = no: b (3.0)  
| | | | | DF = yes: a (4.0)  
| | | | | NSF2 = 9: a (16.0/1.0)  
| | | | | NSF2 = 10: b (23.0/1.0)  
| | | | | NSF2 = 11  
| | | | | NSF1 = 7: b (0.0)  
| | | | | NSF1 = 8: b (0.0)  
| | | | | NSF1 = 9: b (0.0)  
| | | | | NSF1 = 10: b (0.0)  
| | | | | NSF1 = 11



| | | | | | EC = yes: c (2.0)  
| | | | | | EC = no: b (7.0/1.0)  
| | | | | | NSF1 = 12: b (2.0)  
| | | | | | NSF1 = 13: b (0.0)  
| | | | | | NSF1 = 14: b (0.0)  
| | | | | | NSF1 = 15: b (0.0)  
| | | | | | NSF1 = 16: b (0.0)  
| | | | | NSF2 = 12  
| | | | | | EC = yes: b (3.0)  
| | | | | | EC = no: a (73.0/2.0)  
| | | | | NSF2 = 13: a (0.0)  
| | | | | NSF2 = 14: b (3.0/1.0)  
| | | | | NSF2 = 15: a (0.0)  
| | | | FE = 5  
| | | | | NSF1 = 7: b (0.0)  
| | | | | NSF1 = 8: b (1.0)  
| | | | | NSF1 = 9: a (17.0/1.0)  
| | | | | NSF1 = 10: b (3.0)  
| | | | | NSF1 = 11  
| | | | | | DF = no: c (2.0)  
| | | | | | DF = yes: b (6.0)  
| | | | | NSF1 = 12  
| | | | | | Religion = muslim  
| | | | | | ME = 1: b (0.0)  
| | | | | | ME = 2: b (0.0)  
| | | | | | ME = 3: b (0.0)  
| | | | | | ME = 4: b (11.0)  
| | | | | | ME = 5: a (13.0/4.0)  
| | | | | | Religion = christian: b (63.0/2.0)  
| | | | | | Religion = others: b (1.0)  
| | | | | NSF1 = 13: b (2.0)  
| | | | | NSF1 = 14: b (8.0)  
| | | | | NSF1 = 15: b (1.0)  
| | | | | NSF1 = 16: b (0.0)

| | | F3G = c  
| | | | NSF1 = 7: a (0.0)  
| | | | NSF1 = 8: c (1.0)  
| | | | NSF1 = 9: a (33.0)  
| | | | NSF1 = 10: a (67.0/1.0)  
| | | | NSF1 = 11: a (87.0/1.0)  
| | | | NSF1 = 12  
| | | | | EC = yes: c (4.0/1.0)  
| | | | | EC = no  
| | | | | | NSF2 = 6: b (0.0)  
| | | | | | NSF2 = 7: b (0.0)  
| | | | | | NSF2 = 8: b (2.0)  
| | | | | | NSF2 = 9: a (4.0/1.0)  
| | | | | | NSF2 = 10: b (21.0/4.0)  
| | | | | | NSF2 = 11: b (1.0)  
| | | | | | NSF2 = 12  
| | | | | | | ME = 1: a (0.0)  
| | | | | | | ME = 2: a (0.0)  
| | | | | | | ME = 3: b (2.0)  
| | | | | | | ME = 4: a (4.0)  
| | | | | | | ME = 5: b (1.0)  
| | | | | | | NSF2 = 13: b (0.0)  
| | | | | | | NSF2 = 14: b (0.0)  
| | | | | | | NSF2 = 15: b (0.0)  
| | | | NSF1 = 13: c (4.0)  
| | | | NSF1 = 14: c (2.0)  
| | | | NSF1 = 15: a (0.0)  
| | | | NSF1 = 16: a (0.0)  
| | | F3G = d: a (0.0)  
| | | F3G = e: a (0.0)  
| | F1G = b  
| | | NSF1 = 7: b (0.0)  
| | | NSF1 = 8: c (1.0)  
| | | NSF1 = 9: b (0.0)

| | | NSF1 = 10  
| | | | F3G = a: b (0.0)  
| | | | F3G = b: b (4.0/1.0)  
| | | | F3G = c: d (2.0)  
| | | | F3G = d: b (0.0)  
| | | | F3G = e: b (0.0)  
| | | NSF1 = 11  
| | | | F2G = b  
| | | | | EC = yes: c (9.0/3.0)  
| | | | | EC = no: b (18.0/1.0)  
| | | | F2G = a: a (1.0)  
| | | | F2G = c: c (3.0/1.0)  
| | | | F2G = d: b (0.0)  
| | | | F2G = e: b (0.0)  
| | | NSF1 = 12: b (58.0/9.0)  
| | | NSF1 = 13  
| | | | EC = yes: a (2.0/1.0)  
| | | | EC = no: c (5.0/1.0)  
| | | NSF1 = 14: b (25.0/2.0)  
| | | NSF1 = 15: b (1.0)  
| | | NSF1 = 16: b (0.0)  
| | F1G = d: c (1.0)  
| | F1G = c  
| | | ME = 1: c (0.0)  
| | | ME = 2: c (1.0)  
| | | ME = 3: c (3.0/1.0)  
| | | ME = 4: c (0.0)  
| | | ME = 5: b (2.0)  
| | F1G = e: a (0.0)  
MG = a  
| Internet = no  
| | F1G = a  
| | | CL = yes: b (34.0/9.0)  
| | | CL = no: a (3.0/1.0)

| | F1G = b: c (3.0)  
 | | F1G = d: b (0.0)  
 | | F1G = c: b (0.0)  
 | | F1G = e: b (0.0)  
 | Internet = yes: a (353.0/5.0)  
 MG = c  
 | F3G = a: c (0.0)  
 | F3G = b  
 | | FE = 1  
 | | | CL = yes  
 | | | | Internet = no  
 | | | | | ME = 1  
 | | | | | NSF2 = 6: d (0.0)  
 | | | | | NSF2 = 7: d (0.0)  
 | | | | | NSF2 = 8: d (24.0)  
 | | | | | NSF2 = 9: d (0.0)  
 | | | | | NSF2 = 10: d (0.0)  
 | | | | | NSF2 = 11  
 | | | | | F2G = b  
 | | | | | | DF = no: b (3.0/1.0)  
 | | | | | | DF = yes: d (22.0/4.0)  
 | | | | | | F2G = a  
 | | | | | | DF = no: d (2.0)  
 | | | | | | DF = yes: c (3.0)  
 | | | | | | F2G = c: d (0.0)  
 | | | | | | F2G = d: d (0.0)  
 | | | | | | F2G = e: d (0.0)  
 | | | | | NSF2 = 12: c (4.0/1.0)  
 | | | | | NSF2 = 13: c (1.0)  
 | | | | | NSF2 = 14: d (0.0)  
 | | | | | NSF2 = 15: d (0.0)  
 | | | | | ME = 2  
 | | | | | EC = yes: b (6.0)  
 | | | | | EC = no: c (3.0/1.0)

| | | | | ME = 3: c (2.0)  
| | | | | ME = 4  
| | | | | F1G = a: d (9.0/1.0)  
| | | | | F1G = b: b (3.0/1.0)  
| | | | | F1G = d: d (0.0)  
| | | | | F1G = c: d (0.0)  
| | | | | F1G = e: d (0.0)  
| | | | | ME = 5: c (3.0)  
| | | | | Internet = yes  
| | | | | F2G = b: c (12.0/1.0)  
| | | | | F2G = a: b (3.0/1.0)  
| | | | | F2G = c: b (4.0/1.0)  
| | | | | F2G = d: c (0.0)  
| | | | | F2G = e: c (0.0)  
| | | | | CL = no: c (19.0/4.0)  
| | | | | FE = 2  
| | | | | CL = yes  
| | | | | F1G = a: c (20.0/4.0)  
| | | | | F1G = b  
| | | | | Internet = no: b (7.0)  
| | | | | Internet = yes: c (3.0)  
| | | | | F1G = d: c (0.0)  
| | | | | F1G = c: c (1.0)  
| | | | | F1G = e: c (0.0)  
| | | | | CL = no: c (16.0)  
| | | | | FE = 3: c (29.0/4.0)  
| | | | | FE = 4  
| | | | | F1G = a  
| | | | | NSF2 = 6: b (0.0)  
| | | | | NSF2 = 7: c (1.0)  
| | | | | NSF2 = 8: b (2.0/1.0)  
| | | | | NSF2 = 9: b (5.0/1.0)  
| | | | | NSF2 = 10: b (8.0)  
| | | | | NSF2 = 11

| | | | | DF = no: c (2.0)  
| | | | | DF = yes  
| | | | | Internet = no  
| | | | | EC = yes  
| | | | | F2G = b: b (4.0/1.0)  
| | | | | F2G = a: c (4.0/1.0)  
| | | | | F2G = c: b (0.0)  
| | | | | F2G = d: b (0.0)  
| | | | | F2G = e: b (0.0)  
| | | | | EC = no: c (3.0)  
| | | | | Internet = yes: b (6.0/1.0)  
| | | | NSF2 = 12: c (8.0/2.0)  
| | | | NSF2 = 13: b (0.0)  
| | | | NSF2 = 14: b (2.0)  
| | | | NSF2 = 15: b (0.0)  
| | | F1G = b: c (20.0/4.0)  
| | | F1G = d: c (0.0)  
| | | F1G = c: c (3.0)  
| | | F1G = e: c (0.0)  
| | FE = 5  
| | | NSF2 = 6: b (0.0)  
| | | NSF2 = 7: b (2.0/1.0)  
| | | NSF2 = 8  
| | | | AS = 1: c (3.0)  
| | | | AS = 2: b (0.0)  
| | | | AS = 3: b (4.0)  
| | | NSF2 = 9: b (11.0)  
| | | NSF2 = 10: b (23.0/1.0)  
| | | NSF2 = 11: c (13.0/1.0)  
| | | NSF2 = 12  
| | | | Internet = no: b (6.0/2.0)  
| | | | Internet = yes: c (8.0/2.0)  
| | | NSF2 = 13: d (1.0)  
| | | NSF2 = 14: b (2.0)

| | | NSF2 = 15: b (0.0)  
| F3G = c  
| | F1G = a  
| | | DF = no  
| | | | NSF2 = 6: c (0.0)  
| | | | NSF2 = 7: c (1.0)  
| | | | NSF2 = 8  
| | | | | NSF1 = 7: c (0.0)  
| | | | | NSF1 = 8: c (0.0)  
| | | | | NSF1 = 9: c (0.0)  
| | | | | NSF1 = 10: c (0.0)  
| | | | | NSF1 = 11: c (5.0)  
| | | | | NSF1 = 12: b (8.0/2.0)  
| | | | | NSF1 = 13: c (0.0)  
| | | | | NSF1 = 14: c (0.0)  
| | | | | NSF1 = 15: c (0.0)  
| | | | | NSF1 = 16: c (0.0)  
| | | | NSF2 = 9: c (2.0)  
| | | | NSF2 = 10: c (18.0/2.0)  
| | | | NSF2 = 11  
| | | | | ME = 1: b (4.0/1.0)  
| | | | | ME = 2: c (3.0/1.0)  
| | | | | ME = 3: b (1.0)  
| | | | | ME = 4: c (7.0)  
| | | | | ME = 5  
| | | | | FE = 1: d (0.0)  
| | | | | FE = 2: d (0.0)  
| | | | | FE = 3: b (2.0)  
| | | | | FE = 4: b (5.0/2.0)  
| | | | | FE = 5: d (4.0)  
| | | | NSF2 = 12  
| | | | | FE = 1: c (2.0)  
| | | | | FE = 2: c (1.0)  
| | | | | FE = 3

| | | | | AS = 1: c (2.0)  
| | | | | AS = 2: c (0.0)  
| | | | | AS = 3: b (4.0/2.0)  
| | | | | FE = 4  
| | | | | Internet = no: c (2.0)  
| | | | | Internet = yes: b (10.0/1.0)  
| | | | | FE = 5: c (6.0)  
| | | | NSF2 = 13: b (1.0)  
| | | | NSF2 = 14: c (1.0)  
| | | | NSF2 = 15: c (0.0)  
| | | DF = yes  
| | | | CL = yes  
| | | | | NSF2 = 6: d (0.0)  
| | | | | NSF2 = 7: c (1.0)  
| | | | | NSF2 = 8: c (10.0)  
| | | | | NSF2 = 9: d (0.0)  
| | | | | NSF2 = 10: c (7.0/2.0)  
| | | | | NSF2 = 11  
| | | | | FE = 1  
| | | | | | EC = yes: c (5.0)  
| | | | | | EC = no: d (4.0/1.0)  
| | | | | | FE = 2: d (36.0/2.0)  
| | | | | | FE = 3: c (7.0/1.0)  
| | | | | | FE = 4  
| | | | | | EC = yes: c (3.0)  
| | | | | | EC = no: d (26.0/2.0)  
| | | | | | FE = 5: b (2.0/1.0)  
| | | | | NSF2 = 12  
| | | | | ME = 1  
| | | | | | AS = 1: b (2.0)  
| | | | | | AS = 2: b (0.0)  
| | | | | | AS = 3: c (2.0)  
| | | | | ME = 2  
| | | | | | AS = 1: c (2.0)



| | | | | | AS = 2: b (0.0)  
| | | | | | AS = 3: b (4.0)  
| | | | | | ME = 3  
| | | | | | FE = 1: c (0.0)  
| | | | | | FE = 2: b (2.0)  
| | | | | | FE = 3: c (2.0)  
| | | | | | FE = 4: c (0.0)  
| | | | | | FE = 5: c (2.0)  
| | | | | | ME = 4  
| | | | | | EC = yes: c (3.0/1.0)  
| | | | | | EC = no: d (7.0)  
| | | | | | ME = 5: c (0.0)  
| | | | | NSF2 = 13: c (6.0/2.0)  
| | | | | NSF2 = 14: c (3.0/1.0)  
| | | | | NSF2 = 15: d (0.0)  
| | | | CL = no: c (29.0/6.0)  
| | F1G = b  
| | | F2G = b  
| | | | NSF2 = 6: c (0.0)  
| | | | NSF2 = 7: c (5.0/1.0)  
| | | | NSF2 = 8  
| | | | FE = 1  
| | | | | CL = yes: c (4.0)  
| | | | | CL = no: d (5.0)  
| | | | | FE = 2: c (2.0)  
| | | | | FE = 3: c (3.0/1.0)  
| | | | | FE = 4  
| | | | | | Internet = no: c (4.0)  
| | | | | | Internet = yes: d (12.0/1.0)  
| | | | | FE = 5: c (5.0)  
| | | | NSF2 = 9: c (4.0)  
| | | | NSF2 = 10  
| | | | | ME = 1: b (1.0)  
| | | | | ME = 2: c (3.0)

| | | | | ME = 3: c (5.0)  
| | | | | ME = 4: b (3.0/1.0)  
| | | | | ME = 5  
| | | | | EC = yes: d (5.0)  
| | | | | EC = no: b (2.0/1.0)  
| | | | NSF2 = 11  
| | | | | ME = 1  
| | | | | AS = 1  
| | | | | | EC = yes  
| | | | | | | CL = yes  
| | | | | | | | Internet = no: c (4.0)  
| | | | | | | | Internet = yes: d (10.0/1.0)  
| | | | | | | | CL = no: d (33.0/3.0)  
| | | | | | | EC = no  
| | | | | | | NSF1 = 7: c (0.0)  
| | | | | | | NSF1 = 8: c (0.0)  
| | | | | | | NSF1 = 9: c (0.0)  
| | | | | | | NSF1 = 10: c (0.0)  
| | | | | | | NSF1 = 11: c (13.0/2.0)  
| | | | | | | NSF1 = 12: d (2.0)  
| | | | | | | NSF1 = 13: c (0.0)  
| | | | | | | NSF1 = 14: c (0.0)  
| | | | | | | NSF1 = 15: c (0.0)  
| | | | | | | NSF1 = 16: c (0.0)  
| | | | | | AS = 2: d (0.0)  
| | | | | | AS = 3: c (14.0/1.0)  
| | | | | ME = 2  
| | | | | NSF1 = 7: c (0.0)  
| | | | | NSF1 = 8: c (0.0)  
| | | | | NSF1 = 9: c (0.0)  
| | | | | NSF1 = 10: c (0.0)  
| | | | | NSF1 = 11  
| | | | | | FE = 1  
| | | | | | | EC = yes: c (4.0/1.0)

| | | | | | | | EC = no: b (4.0)  
 | | | | | | | | FE = 2: c (11.0/1.0)  
 | | | | | | | | FE = 3: c (3.0)  
 | | | | | | | | FE = 4: c (1.0)  
 | | | | | | | | FE = 5: c (0.0)  
 | | | | | | | | NSF1 = 12: b (2.0)  
 | | | | | | | | NSF1 = 13: c (0.0)  
 | | | | | | | | NSF1 = 14: c (0.0)  
 | | | | | | | | NSF1 = 15: c (0.0)  
 | | | | | | | | NSF1 = 16: c (0.0)  
 | | | | | | ME = 3  
 | | | | | | | | DF = no: d (88.0/12.0)  
 | | | | | | | | DF = yes: c (19.0/1.0)  
 | | | | | | ME = 4: c (18.0/2.0)  
 | | | | | | ME = 5  
 | | | | | | | | DF = no: b (6.0/2.0)  
 | | | | | | | | DF = yes: c (2.0)  
 | | | | | NSF2 = 12  
 | | | | | | FE = 1  
 | | | | | | | | AS = 1: d (45.0/6.0)  
 | | | | | | | | AS = 2: d (0.0)  
 | | | | | | | | AS = 3: b (2.0/1.0)  
 | | | | | | FE = 2: c (11.0/2.0)  
 | | | | | | FE = 3  
 | | | | | | | | CL = yes: c (17.0/4.0)  
 | | | | | | | | CL = no: d (4.0/1.0)  
 | | | | | | FE = 4  
 | | | | | | | | Religion = muslim: b (6.0/1.0)  
 | | | | | | | | Religion = christian  
 | | | | | | | | ME = 1  
 | | | | | | | | DF = no: d (7.0)  
 | | | | | | | | DF = yes: c (2.0)  
 | | | | | | | | ME = 2  
 | | | | | | | | EC = yes: d (4.0)

| | | | | | | | EC = no: b (5.0)  
| | | | | | | | ME = 3  
| | | | | | | | Internet = no: d (3.0)  
| | | | | | | | Internet = yes: c (3.0)  
| | | | | | | | ME = 4  
| | | | | | | | AS = 1  
| | | | | | | | Internet = no: c (3.0/1.0)  
| | | | | | | | Internet = yes: d (2.0)  
| | | | | | | | AS = 2: c (0.0)  
| | | | | | | | AS = 3: c (7.0/2.0)  
| | | | | | | | ME = 5: c (1.0)  
| | | | | | Religion = others: d (0.0)  
| | | | | FE = 5  
| | | | | AS = 1: c (9.0/2.0)  
| | | | | AS = 2: b (3.0/1.0)  
| | | | | AS = 3  
| | | | | EC = yes  
| | | | | | | | Internet = no: b (2.0)  
| | | | | | | | Internet = yes: c (2.0)  
| | | | | | | | EC = no: b (11.0/1.0)  
| | | | NSF2 = 13  
| | | | | EC = yes: c (6.0)  
| | | | | EC = no: b (10.0/2.0)  
| | | | NSF2 = 14: c (15.0/1.0)  
| | | | NSF2 = 15: c (0.0)  
| | | F2G = a: b (1.0)  
| | | F2G = c  
| | | | AS = 1  
| | | | FE = 1  
| | | | | NSF2 = 6: d (0.0)  
| | | | | NSF2 = 7: c (3.0)  
| | | | | NSF2 = 8: d (8.0/1.0)  
| | | | | NSF2 = 9: d (0.0)  
| | | | | NSF2 = 10: d (19.0/1.0)

| | | | | NSF2 = 11  
 | | | | | ME = 1  
 | | | | | CL = yes  
 | | | | | EC = yes: c (3.0)  
 | | | | | EC = no  
 | | | | | DF = no: c (2.0/1.0)  
 | | | | | DF = yes: b (4.0/1.0)  
 | | | | | CL = no: d (7.0)  
 | | | | | ME = 2  
 | | | | | DF = no: d (12.0)  
 | | | | | DF = yes: c (2.0)  
 | | | | | ME = 3: c (2.0)  
 | | | | | ME = 4: b (2.0/1.0)  
 | | | | | ME = 5: d (0.0)  
 | | | | | NSF2 = 12: d (69.0/9.0)  
 | | | | | NSF2 = 13: c (1.0)  
 | | | | | NSF2 = 14: d (0.0)  
 | | | | | NSF2 = 15: d (0.0)  
 | | | | | FE = 2  
 | | | | | ME = 1: c (6.0/1.0)  
 | | | | | ME = 2  
 | | | | | NSF1 = 7: d (0.0)  
 | | | | | NSF1 = 8: d (0.0)  
 | | | | | NSF1 = 9: d (0.0)  
 | | | | | NSF1 = 10: d (0.0)  
 | | | | | NSF1 = 11  
 | | | | | DF = no: d (13.0/1.0)  
 | | | | | DF = yes  
 | | | | | Internet = no: c (7.0/3.0)  
 | | | | | Internet = yes: d (8.0/1.0)  
 | | | | | NSF1 = 12: c (6.0/1.0)  
 | | | | | NSF1 = 13: d (0.0)  
 | | | | | NSF1 = 14: d (0.0)  
 | | | | | NSF1 = 15: c (1.0)

| | | | | | NSF1 = 16: d (0.0)  
 | | | | | | ME = 3: c (4.0)  
 | | | | | | ME = 4: c (0.0)  
 | | | | | | ME = 5: c (0.0)  
 | | | | | FE = 3  
 | | | | | | Internet = no  
 | | | | | | NSF1 = 7: d (0.0)  
 | | | | | | NSF1 = 8: d (0.0)  
 | | | | | | NSF1 = 9: d (0.0)  
 | | | | | | NSF1 = 10: d (0.0)  
 | | | | | | NSF1 = 11  
 | | | | | | EC = yes  
 | | | | | | | ME = 1: c (0.0)  
 | | | | | | | ME = 2: d (6.0/2.0)  
 | | | | | | | ME = 3: c (3.0)  
 | | | | | | | ME = 4: c (0.0)  
 | | | | | | | ME = 5: c (0.0)  
 | | | | | | | EC = no: d (31.0/5.0)  
 | | | | | | | NSF1 = 12: d (53.0/3.0)  
 | | | | | | | NSF1 = 13: c (1.0)  
 | | | | | | | NSF1 = 14: c (1.0)  
 | | | | | | | NSF1 = 15: d (0.0)  
 | | | | | | | NSF1 = 16: d (0.0)  
 | | | | | | Internet = yes: c (12.0/3.0)  
 | | | | | FE = 4  
 | | | | | | NSF2 = 6: d (0.0)  
 | | | | | | NSF2 = 7: c (1.0)  
 | | | | | | NSF2 = 8  
 | | | | | | ME = 1: c (0.0)  
 | | | | | | ME = 2: c (0.0)  
 | | | | | | ME = 3: c (4.0/1.0)  
 | | | | | | ME = 4: d (2.0)  
 | | | | | | ME = 5: c (0.0)  
 | | | | | | NSF2 = 9: c (2.0/1.0)

| | | | | NSF2 = 10: d (1.0)  
| | | | | NSF2 = 11  
| | | | | ME = 1: d (1.0)  
| | | | | ME = 2  
| | | | | Internet = no: c (3.0)  
| | | | | Internet = yes: d (17.0)  
| | | | | ME = 3: c (4.0/1.0)  
| | | | | ME = 4: d (42.0/3.0)  
| | | | | ME = 5: c (1.0)  
| | | | | NSF2 = 12: d (56.0/5.0)  
| | | | | NSF2 = 13: c (1.0)  
| | | | | NSF2 = 14: d (0.0)  
| | | | | NSF2 = 15: d (0.0)  
| | | | FE = 5: c (21.0/3.0)  
| | | AS = 2: c (10.0/1.0)  
| | | AS = 3  
| | | | DF = no: c (18.0/5.0)  
| | | | DF = yes  
| | | | | NSF2 = 6: d (0.0)  
| | | | | NSF2 = 7: d (0.0)  
| | | | | NSF2 = 8: c (4.0/1.0)  
| | | | | NSF2 = 9: c (1.0)  
| | | | | NSF2 = 10: d (0.0)  
| | | | | NSF2 = 11  
| | | | | EC = yes  
| | | | | NSF1 = 7: c (0.0)  
| | | | | NSF1 = 8: c (0.0)  
| | | | | NSF1 = 9: c (0.0)  
| | | | | NSF1 = 10: c (0.0)  
| | | | | NSF1 = 11: c (9.0/1.0)  
| | | | | NSF1 = 12: d (7.0)  
| | | | | NSF1 = 13: c (0.0)  
| | | | | NSF1 = 14: c (0.0)  
| | | | | NSF1 = 15: c (0.0)

| | | | | | | NSF1 = 16: c (0.0)  
 | | | | | | | EC = no: d (49.0/5.0)  
 | | | | | | | NSF2 = 12: c (16.0/4.0)  
 | | | | | | | NSF2 = 13: c (1.0)  
 | | | | | | | NSF2 = 14: c (3.0)  
 | | | | | | | NSF2 = 15: d (0.0)  
 | | | F2G = d: c (2.0)  
 | | | F2G = e: d (0.0)  
 | | F1G = d: c (2.0/1.0)  
 | | F1G = c: c (109.0/19.0)  
 | | F1G = e: d (0.0)  
 | F3G = d  
 | | FE = 1: d (8.0/1.0)  
 | | FE = 2: c (5.0)  
 | | FE = 3: c (4.0)  
 | | FE = 4  
 | | | DF = no: c (2.0)  
 | | | DF = yes: d (2.0)  
 | | FE = 5: c (0.0)  
 | F3G = e: c (0.0)  
**MG = d**  
 | F2G = b  
 | | CL = yes  
 | | | AS = 1  
 | | | | NSF1 = 7: d (0.0)  
 | | | | NSF1 = 8: d (0.0)  
 | | | | NSF1 = 9: d (0.0)  
 | | | | NSF1 = 10: d (1.0)  
 | | | | NSF1 = 11: d (57.0/2.0)  
 | | | | NSF1 = 12  
 | | | | | FE = 1: d (9.0)  
 | | | | | FE = 2: c (2.0)  
 | | | | | FE = 3: c (1.0)  
 | | | | | FE = 4: c (2.0)



| | | | FE = 5: c (1.0)  
| | | | NSF1 = 13: d (0.0)  
| | | | NSF1 = 14: d (0.0)  
| | | | NSF1 = 15: d (0.0)  
| | | | NSF1 = 16: d (0.0)  
| | | AS = 2: b (1.0)  
| | | AS = 3  
| | | | ME = 1: c (2.0/1.0)  
| | | | ME = 2: c (1.0)  
| | | | ME = 3: d (2.0)  
| | | | ME = 4: b (3.0)  
| | | | ME = 5: b (0.0)  
| | CL = no  
| | | NSF1 = 7: e (0.0)  
| | | NSF1 = 8: e (0.0)  
| | | NSF1 = 9: e (0.0)  
| | | NSF1 = 10: e (0.0)  
| | | NSF1 = 11  
| | | | FE = 1: c (2.0)  
| | | | FE = 2: d (2.0)  
| | | | FE = 3: c (0.0)  
| | | | FE = 4: c (0.0)  
| | | | FE = 5: c (0.0)  
| | | NSF1 = 12: e (1024.0/2.0)  
| | | NSF1 = 13: e (0.0)  
| | | NSF1 = 14: e (0.0)  
| | | NSF1 = 15: e (0.0)  
| | | NSF1 = 16: e (0.0)  
| F2G = a: b (1.0)  
| F2G = c: d (354.0/15.0)  
| F2G = d  
| | F3G = a: d (0.0)  
| | F3G = b: d (0.0)  
| | F3G = c: c (3.0)

| | F3G = d: d (6.0/1.0)

| | F3G = e: d (0.0)

| F2G = e: c (1.0)

MG = e: e (1.0)

Number of Leaves : 635  
Size of the tree : 819  
Time taken to build model: 0.19 seconds  
Instances: 5199  
Attributes: 15  
    Religion  
    DF  
    FE  
    ME  
    NSF1  
    NSF2  
    F1G  
    F2G  
    F3G  
    MG  
    AS  
    EC  
    CL  
    Internet  
    KCSE  
Test mode: 10-fold cross-validation

## Appendix VII: Student Performance Data Set

@relation 'studentData-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.attribute.Remove-R1-3,5-7,9-10,13-14,17-22,29-31,33,35-42,44-45,48,52-59-weka.filters.unsupervised.attribute.Remove-R11,14-16,19-20'

@attribute Religion {muslim,christian,others}

@attribute DF {no,yes}

@attribute FE {1,2,3,4,5}

@attribute ME {1,2,3,4,5}

@attribute NSF1 {7,8,9,10,11,12,13,14,15,16}

@attribute NSF2 {6,7,8,9,10,11,12,13,14,15}

@attribute F1G {a,b,d,c,e}

@attribute F2G {b,a,c,d,e}

@attribute F3G {a,b,c,d,e}

@attribute MG {b,a,c,d,e}

@attribute AS {1,2,3}

@attribute EC {yes,no}

@attribute CL {yes,no}

@attribute Internet {no,yes}

@attribute KCSE {a,b,c,d,e}

@data

muslim,no,4,3,12,9,a,b,a,b,3,yes,yes,no,a  
muslim,no,5,3,11,11,b,a,b,b,3,no,yes,yes,a  
christian,no,4,5,12,12,a,a,a,a,3,no,yes,yes,a  
christian,yes,1,2,13,13,a,a,b,b,3,yes,yes,yes,a  
christian,yes,5,5,11,11,b,b,c,b,3,yes,yes,no,a  
muslim,yes,4,3,11,11,b,b,a,a,3,yes,yes,yes,a  
muslim,yes,5,5,11,11,a,a,a,a,1,no,yes,yes,a  
others,yes,5,4,14,12,a,b,c,b,1,yes,yes,yes,a  
muslim,no,2,2,12,12,a,a,a,a,1,no,yes,no,a  
christian,no,4,4,10,10,a,b,c,a,1,yes,yes,no,a  
christian,yes,2,2,11,11,a,a,b,a,1,yes,yes,yes,a  
muslim,no,5,5,11,11,a,a,a,a,1,no,no,no,a

christian,no,5,5,12,12,b,b,c,c,1,no,yes,yes,a  
christian,no,4,5,12,9,a,b,c,b,3,no,yes,yes,a  
christian,no,5,5,11,11,a,a,a,a,1,yes,no,yes,a  
christian,yes,1,1,12,12,a,a,a,a,3,no,yes,yes,a  
muslim,yes,5,5,12,12,b,c,b,c,2,yes,yes,no,a  
christian,no,1,1,12,12,a,a,a,a,1,no,no,no,a  
christian,yes,1,1,12,12,a,a,a,a,3,no,yes,yes,a  
christian,no,5,5,13,10,a,b,b,a,3,no,yes,yes,a  
. . .  
christian,yes,5,4,7,7,c,c,d,d,1,no,no,no,e