A MODEL FOR PREDICTING STUDENTS ACADEMIC PERFORMANCE IN PUBLIC SECONDARY SCHOOLS IN KITUI WEST CONSTITUENCY.

SUBMITTED BY: PETER MUTUA NDAMBUKI

REG NO: 19/05539

A FINAL DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTERS OF SCIENCE IN DATA ANALYTICS IN THE FACULTY OF COMPUTING AND INFORMATION MANAGEMENT AT KCA UNIVERSITY

PRESENTED TO: Dr. SIMON MWENDIA

SUBMISSION DATE: 17/10/2021

DECLARATION

I declare that, this dissertation is my original work and has not been previously published or submitted elsewhere for award of a degree. I also declare that this contains no materials written or published by other people except where due references are made and the author duly acknowledged.

Student Name: PETER MUTUA NDAMBUKI Reg. No. 19/05539

Sign:  Date: October, 2021

I do hereby confirm that I have examined the master's Proposal of

PETER MUTUA NDAMBUKI

And have approved it for examination.

Sign: _____ Date _____

DR. Simon MWENDIA

ABSTRACT

In the present era of data deluge, institutions have accumulated huge amounts of data in their databases. Educational institutions all over the world are not an exception, having as well accumulated large amounts of data in their various educational management information systems databases of various forms and formats. The accumulation of such data in various educational institutions has led to the rise of two research fields namely; Educational data mining and learning analytics in an effort to discover hidden knowledge (insights) that can greatly improve operations in educational institutions. Among the hidden knowledge include but not limited to; predicting students' performance, students' drop out, discovering students interest which could avert popular student's unrest in various institutions etc. This study seeks to take advantage of such an opportunity and develop a model using dataset obtained from public secondary schools in Kitui west constituency that can be used to predict students' academic performance. There has been attempts from various researchers all over the globe to address this problem. Although such studies achieved some level of success, various limitation discussed in details in the empirical review militated against the performance of the earlier models. Desk research methodology was used to extract relevant secondary data from various schools' departments within Kitui west constituency. Then preprocessing which includes feature selection after which the cleaned dataset was loaded to staging Data Lake in Hadoop. Data was queried from the Data Lake to python using Pyspark where data analysis procedures took place. Dataset consisting of optimal subset of features was used to train four machine-learning algorithms: Gradient boost classifier, Random forest classifier, Decision tree classifier and Deep Neural Network classifier. Generally, Decision tree and Random forest classifiers registered the best performance overall, with an accuracy of 97%, but after stratified Kfold cross validation, Decision tree classifier's performance proved more stable with an average of 97% compared to Random forest classifier with 93%. Thus, Decision tree classifier was recommended for deployment in predicting students 'academic performance for its reliable accuracy and relatively good precision on predicting the study's target group. The developed Model will place students in to two groups: PASS and FAIL. The aim being to arouse an initiation of intervention from various stakeholders to reduce dismal performance among public secondary schools in Kitui west constituency.

Key words: Educational datamining, learning analytics, Machine learning, feature selection, desk research, diagnostic research, experimental research, data deluge, minimum redundancy maximum relevance.

ACKNOWLEDGMENT

I acknowledge the support I received from my supervisor Dr. Simon MWENDIA, who guided me in writing my proposal and dissertation. My interest for carrying out this study is to develop a model for predicting students' academic performance, and hence address the issue of student dismal performance in public secondary schools in Kitui west constituency by designing appropriate intervention to assist students improve their performance. Special acknowledgment goes to my mentor Samuel KATHINUKU who has supported me and encouraged me to enroll for my Master's program. I cannot forget my family, my wife and children who always provided a conducive environment during the coursework period, writing of my proposal and finally my dissertation, and above all, God who guided and took care of me.

ACRONYMS AND ABBREVIATION

UK – United Kingdom

EMIS – Educational Management Information System

EDM – Educational Datamining

LA – Learning Analytics

M.o.E – Ministry of Education

KCPE – Kenya Certificate of Primary Education

KCSE – Kenya Certificate of Secondary Education

UN – United Nations

MDG – Millennium Development Goals

NARC – National Rainbow Coalition

ECDE – Early Childhood Development Education

DM – Datamining

WEKA – Wakito Environment for Knowledge Analysis

TIVET – Technical Institutes of Vocational Education Training

KCAU - Kenya College of Accountancy University

DNN – deep neural network

NN – neural network

TABLE OF CONTENTS

LIST OF TABLES

.

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Education is actually the means through which trusted true believes (knowledge), skills and attitudes are passed from generation to the other. It enables students to discover and or leverage on their potentials for a brighter future in various sectors of the economy. It is considered as the key to shaping human lives based on its capacity to transform and enrich their lives irrespective of social and economic status. (Sekeroglu B. et al. 2019) acknowledges education as necessary for productive good life as it equips students with values and excellence, motivates their self-assurance and keeps them updated as far as needs of the present world are concerned. Education is one of the fundamental human rights all over the world, enshrined in various state's constitutions due to its importance in shaping human life. For example, in United Kingdom (UK), human rights act 1998 part 2; education is a fundamental human right protocol 1, article 2 – *Right to education* (Human rights act 1998). In the constitution of Kenya, 2010 chapter four (The bill of rights), education is a human right under section 43 (economic and social rights), (f) – *right to education for all* (Constitution of Kenya, 2010).

Education institutions take students as their raw material to be processed to a refined finished product ready for the market. The extent to which refinement has been achieved is measured by academic and /or co-curricular performance as well as character formation. Student performance refers to the measure of achievements in learning assessments and co-curricular activities (Usamah et al. 2013). Academic performance is the ability to remember learnt facts and be in a position to communicate them in writing exams (Mbithi j. 2017). It is assumed that the better a student can remember and communicate the knowledge, the better they are in a position to independently apply in their own unique way, the knowledge in solving problems in the real world *(i.e. the more creative and innovative they become).*

From previous literatures in education and psychology (Hellas A. et al. 2018), research work aimed at determining factors that contribute to academic performance date back to late 1910s. The focus being to obtain the most predictive attributes/features on the target variable and identify the best prediction method. Various test like; verbal memory tests, reading backward test, hard directions test, test of suggestibility (recall of details of a picture or

tendency to make egocentric reactions in free association experiment) etc. were administered on Vassar freshmen consecutively in the years, 1918 – 1920. (Montague M. et al. 1918). Later on, (Gerald E. and Mark G. 1989) carried out a study to identify variables that predict computer aptitude to help employers and educators in selecting potential employees and students respectively. Several explanatory variables were included in their study that included; prior computer training, high school achievement, problem solving skills and cognitive styles. Major objective of their study was to identify computer information system majors and nurture interest of students in the fields.

During the present era of data deluge, institutions have accumulated huge amounts of data in their databases. Educational institutions all over the world are not an exception, having as well accumulated large amounts of data in their various educational management information systems (EMIS) databases of various forms and formats. The accumulation of such data in various educational institutions has led to the rise of two research fields namely; Educational data mining (EDM) and learning analytics (LA) in an effort to discover hidden knowledge (insights) that can greatly improve operations in the educations institutions. Among the hidden knowledge include but not limited to; predicting students' performance, students' drop out, discovering students interest which could avert popular student's unrest in various institutions etc. (Amrieh E. et. al, 2016), (Mgala M. 2016) and (Ha, D.T. et al, 2020).

There has been attempts from various researchers all over the globe to develop models for predicting students' academic performance at various levels using educational dataset accumulated over time, based on either of the two earlier mentioned research fields. Although such studies achieved some level of success, various limitation discussed in details in the empirical review militated against the performance of these earlier models. This study seeks to address in particular limitation concerning efficiency and accuracy of such models. On efficiency, these models used unnecessarily large dataset leading to high computational cost; an appropriate feature selection method, which removes redundant features, will be used to address this issue. Secondly, the accuracy of these models has been less than 100%; this study seeks to develop a model, hopefully with the highest accuracy at the lowest computational cost.

1.2 Statement of the problem

The efficiency of any education systems all over the world is the ability to produce graduates who can fit well and compete fairly for limited job opportunities in various sectors (Frans R. 2018). This has not been the case in various countries as various factors militate against efficiency of education systems such as poverty, lack of enough human resource etc. This has led to students' dismal performance in academics occasioned in most institutions of learning all over the globe. Such a challenge is experienced more in the developing countries, Kenya being one of them, because of high prevalence of factors mentioned above. Dismal performance or underperformance refers to a case where a particular student fails to score more that an agreed threshold within a particular academic period. (Mgala M. 2016).

Education consumes a large chunk of funds from various countries budget all over the world, e.g. In Kenya, from the 2020/2021 budget; education as allocated Ksh.497B out of 2.91T. Despite such huge expenditure coupled with efforts from various stakeholders in various countries, students' performance is still wanting (below average) all over the globe and worse in developing countries thus a need to address this issue. Going by the statistics of the recently released KCPE results by the cabinet secretary, Ministry of Education (M.o.E), Republic of Kenya, out of 1,179,192 pupils who sat for the 2021 KCPE, 889,011 scored below 300 marks out of 500 implying that about 76% of the pupils failed in their KCPE. The trend is even worse in secondary schools, where out of 699,745 students who sat for KCSE in the year 2019, 129,746 students scored C+ and above (attained university entry grade). These results imply that 81.5% of student who sat for their KCSE in the year 2019 failed to secure university admission. Driven by these statistics, this study seeks to collect data from public secondary schools in Kitui west constituency and develop a model for predicting students' academic performance with an aim of identifying in advance, students likely to fail for intervention measures to be put in place to address the problem.

Previously, studies aimed at solving this problem have been carried out as discussed in details under empirical review in the next chapter. These studies, although achieved some level of success had the following limitations. First accuracy achieved by these studies was less than 100% and was not the same. For example, studies carried out by (Saa A. et al, 2010) achieved an accuracy of 75.52% for a decision tree classifier, (Amra I and Maghari A. 2017) achieved an accuracy of 93.6% for a Naïve Bayesian classifier and (Youfsafzi B et al, 2020) achieved an accuracy of 94.39 for a decision tree classifier. Secondly, studies were carried out in

developed countries using public datasets as opposed to dedicated students' datasets. Thirdly, the datasets used were obtained from higher levels of education (University level) and very few if not none from the lower levels of study (i.e. primary and secondary). Last but not the least, most of these studies used the same data analysis tool WEKA, as opposed to a different tool like python. The above stated statistics on student's performance in Kenyan main exams in both primary and secondary levels and the limitations of the previously conducted studies on the topic guided this study.

Educational institutions as stated earlier have accumulated valuable data in their repositories during their daily operations in this era of data deluge. Within this data, that seems of less use to most of educational institutions especially in the lower levels and mostly in the developing countries, lies valuable insights that can inform decision that may lead to sound administration and management of such institutions. Thus, this study seeks to demonstrate how data analytics methods and techniques can be used to address this issue in a developing country. Using educational dataset, data analytics methods and techniques, this study sought to develop a model that can be used by educational institution in taking their monitory and evaluation to the next level. The model will aid in revealing insights concerning students' academic performance. Which in turn may lead to initiation of appropriate interventions to address dismal performance e.g. improving academics quality, correct planning that leads to successful service delivery within the institutions etc.

In addition to addressing the problem of students' dismal performance by arousing initiation of appropriate intervention from various concerned education stakeholders, this study contributes to the field of data analytics through addressing limitations mentioned above, of earlier models developed in an attempt to address this problem, though at different levels. For example, developing a model, hopefully with the highest accuracy at the lowest computational cost among other improvements.

1.3 Main objective

The study's main objective is to develop a model for predicting students' academic performance in public secondary schools in Kitui west constituency.

1.4 Specific objectives

i. To identify the main features that affect students' academic performance in public secondary schools.
ii. To develop a predictive model for students' academic performance using the optimal subset of features identified from the main features with appropriate data mining methods.
iii. To validate the developed model.

1.5 Research questions

- What are the main features that affect students' academic performance in public secondary schools?
- How will the predictive model be developed using appropriate data mining methods?
- How will the model for predicting students' academic performance be validated?

1.6 Significance of the study

The research findings would go a long way in improving students' performance in secondary schools in Kitui west constituency. This is through providing actionable insights to various key education stakeholders that will enable them optimally contribute to students' academic performance.

To the students, the study results would assist them in generating vital knowledge about their performance that will guide them in adjusting accordingly to appropriate study skills and learning approaches. This knowledge could include but not limited to; various weak points of every learner, factors that contribute to their academic performance etc. These results would serve as early warning indicators of any form of dismal performance to various groups of learners.

Leveraging on the study results, teachers/tutors would seek to; identify the most appropriate learning behaviors that suites various learners' groups to facilitate corresponding interventions that effectively addresses the learning needs of each of the groups. e.g. learners who need extra assistance so that appropriate intervention procedures be initiated to save them from dropping further especially towards the main exam, adjust the curriculum and class sessions to facilitate students' learning plans, identify various students' weak points and factors that affect their performance and assist them in interpreting the early warning signals guiding them accordingly.

Education administrators using the study results would adjust the curriculum and class sessions to facilitate students' learning plans, arrange for extra tutoring resources and remedial classes to various groups of students etc.

Once schools implement the findings of this study, i.e. use the developed model to predict student academic performance, intervention will be initiated on time to address dismal performance among students. This will facilitate the school in excelling on its core business, which ultimately improves its reputation.

To the researchers, the finding of this study would serve as a guide to their research in either the same sector or other related fields. For example, if the mentioned limitations of the previous models are fully addressed i.e. significantly improving the model's accuracy, the models set up can be adopted with dataset from a different sector like medical where high accuracy is paramount to assist in diagnosis. The reviewed work would also make it easier for researchers to learn and identify some of the variables that affect students' academic performance, which they can use in developing advanced models in future.

Accurately predicted results can assist the Governments through their ministries of education in improving academic quality and better services delivery. This is possible if the Government provides centralized educational management information systems (EMIS) for various levels of study that are well synchronized for effective tracking of learner's progress records.

1.7 Motivation of the Study

As mentioned earlier, Education is a basic need enshrined in constitutions in various countries all over the world, this is because of the impact it has in human life (Sekeroglu B. et al. 2019). In the United Nations (UN) millennium development goals (MDGs) *education for all* was declared the goal number two immediately after *poverty and hunger eradication* (Brundrett, 2011), (Bruns and Rakotomalala, 2003). Thus, developing countries' Governments allocate huge amount of money from their budget in order to attain the goal. However, despite all these efforts, it has not been possible to meet the goal fully. Though a commendable growth from 83% in 2000 to 91% in 2015 (Motala S. et al. 2015). Kenya's spends a large amount of funds to facilitate and boost education in its attempt to achieve the goal. In an effort to achieve this goal, the Kenyan government (NARC) in 2003 introduced free primary education that led to 1.2 million admissions in primary schools in the country. The enrolment is set to improve more by the policy that seeks to in cooperate early childhood development education (ECDE) to the

basic education. To ensure 100% transition from primary to secondary education, the government in the year 2008 introduced free day secondary school and subsidized boarding secondary education (Ministry of Devolution and Planning. 2013). Despite all these efforts, statistics presented earlier on this study shows that student academic performance is wanting (majority of the students perform below the pass mark). These findings motivated the researcher to develop a model for predicting students' academic performance, in public secondary schools in Kitui County, earlier before the main national exams to alert, arouse and direct appropriate intervention from the relevant stakeholders to address the problem of mass failure of students.

1.8 Scope of the Study

The scope of this research is limited to the sampled public secondary schools in Kitui west constituency. Data collected by the researcher will only be used in this study in an effort to obtain viable research findings. Although the scope is limited to the sampled public secondary schools, research findings can be replicated and applied in any secondary school in Kenya to predict students' academic performance prior to the main exam.

This study aims at achieving the objectives stated above by developing a binary classification model using supervised machine learning algorithms. That is, after the model is developed and trained with training data, the model it put learners in to two categories, those that have attained PASS mark and those that didn't – FAILED.

1.9 Outline of the Study

The study is organized into several sections, with *chapter one* covering background information of the study, statement of the problem being addressed by the study, the objectives of the study and research question of the study, the significance of the study and motivation of the study. The second section, which is *chapter two*, will cover the introduction of the literature review, theoretical review, and conceptual framework, operationalization of the variables, and the summary of the chapter. The third section, which is *chapter three* of the study, will cover the introduction of the methodology, research design, target population, procedure for data collection and data processing and analysis.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of literature related to the research problem identified in the previous chapter. The main aim of the review is to identify factors/features that influence student academic performance and later on, highlight research gaps to be addressed by the study. The section will also cover the theories on which this study was based, identify various machine-learning models used by different researchers to predict student academic performance with their respective accuracies and lastly outline the relationship between various factors/features identified during the literature review in the form of a conceptual framework.

2.2 Theoretical Review

Research in to factors that influence students' academic performance of students who dropout from various institutions of learning concentrates strongly on; survey studies enquiring directly from students the factors that lead to their drop out and studies that investigate students' academic performance with respect to features that affect their academic performance like gender, personal characteristics, finances etc. Other studies conceptualize these features into theories e.g. Spady (1970), Tinto (1975) and Bean (1980).

2.2.1 Spady's sociological theory (1970)

This theory sought to justify student's retention within a given institution of learning. Its major assumption is that; students' dropout can be explained through the process of student individual interaction with the institution's environment. During the interaction, students' skills, interest and attitudes are subjected to institutions demands, expectations and influences which in turn determines whether a student will dropout or get assimilated to the social and academic system of a given educational institution. Factors associated with the process of assimilation are similar to those that influence academic and social integration. Although this theory does not fit well with the current study, as its main objective is to justify students' retention, it sheds some light into factors that influence student academic performance. The context of this theory is the South Africa higher education environment (Spady, 1970).

In 1975, Tinto within the framework of Durkheim's theory which emphasizes that, individual likelihood to commit suicide is determined by their level of integration into the

society, echoed the sentiments of spady's theory by asserting that student's dropout can be attributed to insufficient integration to various aspects of a given institution, the key ones being academic and social systems. This led to Tinto's integration theory 1975 (Tinto, 1975). Later on in 1980, another theory was developed by Bean known as Bean's psychological theory, which emphasizes on the background characteristic of a student as the key to understanding students' retention. This include student's behavior and attitudes, which are key factors that influence students' academic performance (Bean, 1980). From the three theories that seek to explain students' retention, it is clear that factors that influence students' academic performance are complex and multidirectional. In addition, two features come out clearly as the key features that influence students' dropout i.e. academic integration and social integration. These theories do not directly fit to the current study but from their investigation into academic integration with an effort to justify student retention, features that influence student academic performance must be included.

2.2.2 Constructivism Theory

Also referred to as the social constructivism theory of engagement seeks to implement cooperative learning through interactions in which students are expected to share skills obtained through experience or learnt in a classroom setting from a resource person. The theory emphasizes collaboration learning, students' engagement, students' interaction, educational quality, students' satisfaction, social media use and their impact on students' academic performance. The main aim of the theory is to determine how learning processes can be efficiently implemented in a classroom setting and various ways to construct knowledge. Under constructivism classroom, teacher's role is to guide students in building their knowledge and regulate the classroom environment to ensure maximum learning takes place. This is achieved by allowing students discussion during class sessions, stimuli variation during content delivery, encouraging peer interactions and varying content being delivered etc. This theory focuses more on factors that affect student performance within the classroom setting and encourages student-centered mode of learning. Thus, other factors that influence students' performance outside the classroom are not considered (Al-Rahmi, 2020)

2.2.3 Walberg's Theory of Academic Achievement

This study is premised on Walberg's theory of academic achievement and the concept of Educational Data Mining (EDM). Which serve as a guide to the processes involved in student

academic performance prediction, clearly highlight, and lead the whole process toward a well-defined goal (prediction of students' academic achievements)

This theory advances the idea of educational outcomes being influenced by psychological characteristics of the individual student and the immediate psychological environment, i.e. behavioral, attitudinal and cognitive (Rugutt, J. K., & Chemosit, C. C., 2005). This theory is anchored on the nine main variables believed to influence educational outcome which includes; ability or prior achievement of the student, student motivation, student's developmental level (age), quality of instruction to which the student is subjected, classroom climate, home climate, peer groups in which the student belongs and access to mass media while outside the school. The nine variables proposed by the Walberg theory of educational achievement rhyme very well with the factors that influence academic performance by (Mgala M. 2016; Obadiah M. et al., 2019). Thus, the theory provides the basis upon which this study is going to be established. This is by providing the avenues through which relevant educational data is obtained to advance the processes involved in answering the research questions stated in the previous chapter.

2.2.4 Educational Data Mining (EDM)

In the previous chapter, two fields of research concerned with carrying out analysis on educational data with an aim of understanding learners and the environment in which learning takes place have briefly been mentioned. These include; Learning Analytics (LA) and Educational Data mining (EDM). According to the Society for Learning Analytics Research, Learning Analytics (LA) is defined as; "the measurement, collection, analysis and reporting of data about learners and their context, for the purpose of understanding and optimizing learning and the environment in which it occurs" (Siemens, G., & Baker, R. S. D., 2012). While on the other hand, according to Educational Data Mining Society, Educational Data Mining (EDM) is defined as; "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students, and the settings they learn in" (Siemens, G., & Baker, R. S. D., 2012).

Although the two fields share the same goals of improving planning and selection of interventions, improving assessments and understanding educational problems. As well, as improving the quality of analysis of humongous educational data to support research and practice. EDM has been chosen as the one that best suites this study for reasons clearly outlined by (Siemens, G., & Baker, R. S. D., 2012) which include-

EDM advocates for automated discovery as human judgement is leveraged as a support tool for discovery. This is because of knowledge discovery being purely data-driven after the stage of feature selection by human. EDM places greater emphasis on reducing to components then analyzing such components individually. This is in agreement with splitting data into training and testing samples cross validation during the models evaluation and feature selection to identify the optimal subset of features that are more predictive to the target class. Its popularity with communities that have conducted student academic prediction in the past makes it preferred as guidance is guaranteed from the previous literature. Finally, in line with the goal of this study; predicting student academic performance, its focus on automation guarantees support to education stakeholders who will be using the model developed.

EDM's origin can be traced from human learning which has been in existence close to a century. In human learning students are studied in a lab setting as they perform various tasks while EDM uses data mining techniques (DM) to discover hidden patterns, relationships, associations or insights/knowledge (Alnoukari, M., & El Sheikh, A. 2012), in historical data obtained from educational institutions, that can contribute in understanding the learners and their learning environment for appropriate interventions. EDM has five varied approaches as proposed by (Baker, 2010) which include; distillation of data for human judgement, discovery within models, clustering, relationship mining and prediction. This study adopted the prediction approach as it seeks to use data collected from public secondary schools in Kitui west constituency, to build a model for predicting student academic performance for initiation of intervention measures on time to assist affected students. Thus for this study EDM will be the best choice as historical data collected from targeted public secondary schools in Kitui west constituency will be used to build a students' academic performance prediction model based on its well laid down procedures. These laid down procedures can be traced back to Knowledge Discovery Process Models (Alnoukari, M., & El Sheikh, A., 2012).

2.2.5 Machine learning Techniques in EDM

As mentioned earlier in this chapter, EDM as a recent growing field of research is in simple terms an application of Data Mining techniques in education. EDM differs with application of data mining techniques in other sectors of the economy e.g. banks, security, transport, communication, business, genetics, medicine etc. in the following three aspects i.e. *Data* - Various types of data in education are unique in structure, formats and relationships. These variations can be attributed to the education system of delivery e.g. (Intelligent Tutoring

System, e-learning or face-to-face etc.). The education system itself, that vary from one country to another e.g. in Kenya we have the 8:4:4 system of education, level of education under focus e.g. in Kenya under 8:4:4 system there are three distinct levels of education e.g. Primary, Secondary and Higher levels of education etc. This makes the application of DM in education a special case that requires its own approach and thus cannot be generalized with other domains. *Objective* – The main aim of applying DM in other domain areas mentioned earlier is to increase profits, a measurable quantity that is determined by the increase in the sales. EDM on the other hand has applied unquantifiable objectives like improving the student learning, predicting students' performance etc. as well as pure research objectives, such as understanding the learning process and the environment it occurs, searching for a deeper understanding of an educational issue e.g. student discipline, student drop-out rate etc. *Techniques*- Due to the special characteristics of educational data as mentioned earlier, different data mining approaches are required. As a result, some DM techniques may be adopted directly, while others have to be adapted to the unique problem being studied (Livieris et al, 2019, Mgala M. 2016, Gajwani & Chakraborty, 2021, Yousafzai et al, 2020).

Data mining process follows any of the nine knowledge discovery process models (KD) also referred to as knowledge discovery in data process models (KDD) advanced by different proponents as shown in (Alnoukari & Sheikh 2012). Among the proposed models include; Knowledge Discovery in Databases (KDD) by Fayyad et al, 1996, CRoss-Industry-Standard Process for Data Mining (CRISP-DM) by CRISP-DM, 2000), Information Flow in a Data Mining Life Cycle by Ganesh et al. 1996 etc.

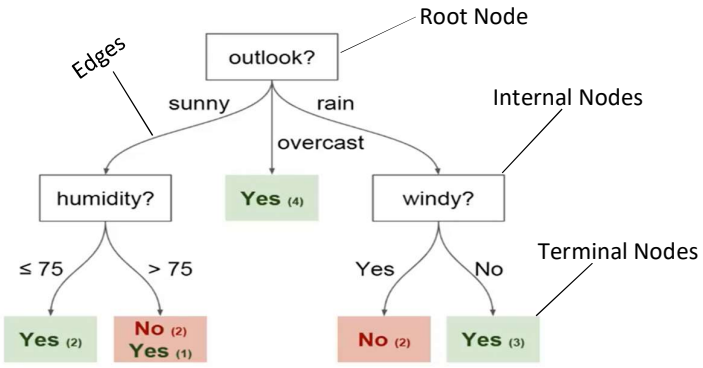
In each of the above mentioned knowledge discovery process models, machine-learning techniques are used to facilitate the process of knowledge discovery from data. Mitchell, 1997 provided a definition for Machine learning as: "A process in which a computer program learns from experience E with respect to some class of tasks T and performance measure P and its performance at tasks in T, as measured by P, improves with experience E." i.e. the ability of a computer to learn from experience (data). From such experiences computer program can find dependencies that are difficult for human beings to discover such as patterns, associations etc. that helps in prediction in terms of structured data or reveal-hidden structures in unstructured data that guides classification or clustering procedures. Machine learning techniques fall under four broad categories namely; supervised machine learning, semi-supervised machine learning, unsupervised machine learning and reinforcement machine learning techniques. Among the just mentioned machine learning categories, supervised and

unsupervised machine learning techniques are the commonly used in various applications to facilitated decision making in various sectors within different economies all over the world. Supervised machine learning techniques are applied where data records have known labels and correct targets while unsupervised are applied where data records have neither labels nor target class. Problems addressed by supervised machine learning techniques fall under two categories, which include Classification and regression with examples such as prediction, recommendation etc. while equivalent categories in unsupervised machine learning techniques include clustering and association with examples such as customer segmentations, products placement in shelfs etc. as mentioned earlier. Therefore, this study is a binary supervised machine learning approach where student will be placed in to two groups; those that will *pass* and those that will *fail*, using a labeled dataset obtained from sampled public secondary schools in Kitui west Constituency. Thus an appropriate knowledge discovery model was selected from the ones outlined in (Alnoukari & Sheikh 2012) with supervised machine learning algorithm with an history of good performance from the empirical review carried earlier on in this chapter. The following supervised machine learning techniques identified from literature as described above were used to facilitate the achievement of the study's main objective.

(a) *Decision tree algorithm* - belongs to supervised learning category of machine learning. There are two types of decision trees algorithms; *regression trees* – used to solve regression problems and *classification trees* – used to solve classification problems. Classification tree consist of nodes, edges and leafs. There are three types of nodes; *root node* – without incoming link and zero or more outgoing link, *internal nodes* – each with exactly an incoming link and at least two outgoing links and *terminal node* – each with exactly one incoming link and without outgoing link. The structure of the tree is as follows: - *Nodes* – splits for the value of a certain attribute, *Root node* – is a node that performs the first split, *Leaves* – Terminal nodes that predicts the outcome and *Edges* – outcome of a split to the next node.

FIGURE 1

Structure of a Decision tree



Source: Müller, & Guido, 2016, James et al. 2013

Figure 3 is an example of a decision tree applied in making a decision on which day is good for playing tennis using a weather pattern dataset given in the table below;

TABLE 1

Dataset Description

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Source: Müller, & Guido, 2016, James et al. 2013

Two statistical models, Entropy and Information gain guide the selection of best attribute for the root node and subsequent growing of the tree (maximizing information gain over the split). Entropy H(s) is the measure of impurity or uncertainty in a given dataset. Its computation is as guided by the formula 1 below (Müller, & Guido, 2016, James et al. 2013, Madhavan, S. 2015).

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i \text{ ----------(1)}$$

Where: - S – is the set of all instances in the dataset

N – number of distinct class values

P_i – event probability

Information Gain (IG) – is the measure of the information given out by a particular feature or variable about the outcome. It is given by the formula 2 below.

$$Gain\ (A,S) = H(S) - \sum_{j-1}^{v} \frac{|S_j|}{|S|}.H(S) = H(S) - H(A,S) \text{ ---------- (2)}$$

Where: -

$|S_j|$ - number of instances with j value of attribute A

S – total number of instances in dataset S

V – set of distinct values of an attribute A

H (S_j) – entropy of subset of instances of attribute A

H (A, S) – entropy of an attribute A

Decision trees learn in a hierarchy layout facilitated by if/else questions (tests) that lead to a decision making concerning a given problem at hand.

Ensemble methods are used to combine multiple single machine learning algorithms (known as weak learners) to obtain a powerful model. These methods include, boosting and bagging or bootstrapping. Random forest and gradient boosted trees are a good example of ensemble models that have proved effective for both classification and regression over a wide range of datasets. They are built by combining multiple decision trees using any of the earlier mentioned methods.

(b) *Random forest* is a more powerful prediction model constructed by bagging or bootstrapping using a number of decision trees as the building blocks. Decision trees suffer from high variance. That is, if training data is randomly split into two parts then fitted to a decision tree, for each of the parts quite different results are obtained. Bagging or bootstrap aggregation is a generally accepted procedure for minimizing variance in statistical learning methods. Random forest algorithm obtained its name from the manner in which randomness is introduced in the trees building. This is achieved in two ways; selecting a number of features (*max_features*) in each split test or selecting the data points to be used in the tree building. For the later, *bootstrap sampling* is used in which a data point is repeatedly picked from the initial dataset in a random manner with replacement. This leads to (*n_samples*) called *bootstrap*

samples of the same size as the initial dataset but with some data point missing, approximately a third of the original dataset due to drawing data points with replacement. Trees are built from these *bootstrap samples,* which makes them very independent from each other in the 'forest'. The idea being each tree will produce good results on generalization and over fit but each at a different potion of the dataset, as single decision trees are prone to overfitting. In case of regression these results are averaged while for classification, a 'soft voting' associated with prediction of each single tree in the entire 'forest' in terms of probability for the possible target outcome, are averaged and the class with the highest probability is chosen as the predicted. This significantly reduces the over fitting of the trees in the 'forest' while retaining the predictive power of the trees. Therefore, random forest corrects the high variance in decision trees thus producing better results. Random forest has a higher accuracy and more robust to errors and outliers. Error in random forest converges as long as the number of trees in the forest increases (Müller, & Guido, 2016, James et al. 2013).

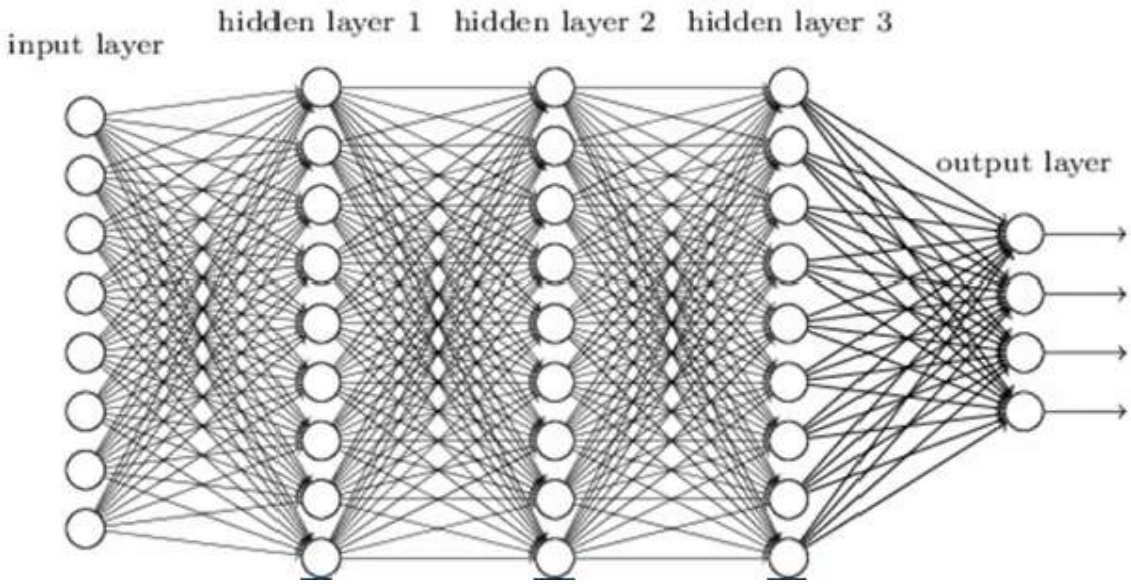
(c) *Gradient boosted regression trees* - Boosting is another approach for improving predictions of a decision tree, it works in a similar manner as bagging except the fact that in boosting, trees are grown sequentially i.e. each tree is grown using information from previous grown trees and tries to correct the mistakes of the previous trees. Unlike random forest algorithm, there is no randomization in gradient boosted trees instead, shallow trees of depth ranging from one to five are used which makes it economical in terms of memory and faster in prediction. It is more sensitive to parameter setting unlike random forest. If the parameters a well-tuned, gradient boosted trees tend to produce better accuracy than random forest. These parameters include; first the number of trees to include (*n_estimators*), this increases the complexity of the ensemble as well as the prediction accuracy as the higher the number of trees included, the higher the chances of correcting mistakes on the training sample. Second is the extent to which the trees are allowed to grow i.e. (*max_depth*). As mentioned earlier, for this ensemble algorithm the depth is only allowed within a minimum of one and a maximum of five. Last but not the least is the (*learning_rate*) which controls the extent to which a tree controls the mistakes of its predecessor. The higher the learning rate the stronger the correction, which increases the complexity of the ensemble algorithm (Müller, & Guido, 2016, Madhavan, S. 2015)

(d) *Artificial Neural Networks* (ANN). It falls under the reinforcement learning type of machine learning and a vital stepping-stone to understanding deep learning. It works with both labeled and unlabeled datasets. ANN is modeled after the biological neural network and allows computers to learn in a similar manner as humans. Reinforcement learning is applied in areas

like; pattern recognition, time series predictions, signals processing etc. Human brain consist of billions of connected neurons with dendrites through which it receives inputs and based on those inputs it produces an electrical signal which is outputted through one of its parts known as *axiom* to other cells. Neural Networks (NN) helps computers solve those problems that are easier for human but very hard for computers i.e. image recognition, speech recognition etc. The simplest form of a neural network is the perceptron, consisting of at least one input, a processor and one output. A perceptron follows the '*feed forward model*' meaning inputs sent into the neuron is processed and an output generated (Géron, A. 2019). Perceptron's processor performs four tasks on the inputs; receiving the inputs, putting some weights on them, summing up the inputs then generating an output. This simplest form of neural network is limited to linear functions, thus a need for more robust forms of NN. The Multilayer perceptron consisting of an input layer, one or more hidden layer(s) and an output layer in which activation function such as rectified *linear unit* (relu) and *tangens hyperbolicus* (tanh) are used to make the model more powerful than a linear model. These nonlinear activation functions allow the model to learn from more sophisticated functions than just a linear model. A Neural Network with many hidden layers is referred to as Deep Neural Network (DNN). There is no threshold in terms of the number of hidden layers a neural network should have to be considered as a deep neural network. However, deep neural networks are applied in solving complex problems like vision recognition. This concept is applied in this study to develop a model for predicting students' academic performance after sufficiently training it with an educational dataset obtained from Kitui west constituency. Figure below shows the architecture of a deep neural network (Géron, A. 2019, Müller, & Guido).

FIGURE 2.

Deep Neural Network



Source: Wang et al. 2019

2.2.6 Feature Selection

Also referred as feature reduction, is one of the most important data pre-processing step in machine learning. Its main aim is to determine from the dataset, variables which are more predictive to the target class. At this step, irrelevant and redundant features are removed leading to an optimal subset of features. Through theory and practice as discussed in the empirical study later on in this chapter, feature selection stage of data pre-processing has been proven significant in the whole process of machine learning. This is because it improves the *accuracy* of the machine-learning model by facilitating the dropping of irrelevant and redundant features, improves the *efficiency* of the model by reducing the training/learning time and other processing resources required because of reduction in the dataset dimension and enhances development of *simple models* that are easier to understand. Which ensures better models diagnosis and interpretation. Thus as the complexity of the model development process decreases, *workload* on monitoring and maintenance of features within the data pipeline also decreases. (Punlumjeak & Rachburee, 2015, Ramaswami & Bhaskaran, 2009, Billah & Waheed, 2020 etc)

Feature selection methods fall under three categories namely; filters, wrappers and embedded. Filter methods are independent of the learning phase where the classification algorithm is trained. Actually, these methods select the optimal subset of features, which will be used in the training phase. They rely on various monotonic metrics that are statistically oriented to access predictive strength and relevance of a given set of variables. They are monotone metrics as their value vary depending on the number of attribute being considered. They search for simple, individual relations between the independent variables and the target class (the independent variable), the simplest being correlation, and then rank them based on such relation. In the simplest case mentioned earlier when applied in supervised machine learning, involves assessment of every individual attribute based on its correlation with the target class. One major weakness of this approach is that it produces a subset of features that are highly correlated to the target class without considering the correlation among the selected predictive variables themselves, which could result to redundancy. Some filter selection methods include; minimum Redundancy Maximum Relevance (mRMR), Correlation-based attribute evaluation, Chi-Square attribute evaluation, Gain-Ratio attribute evaluation, Information-Gain attribute evaluation, Relief attribute evaluation, Symmetrical Uncertainty Attribute evaluation etc. (Wah et. al, 2018, Billah & Waheed 2020, Moreira et al. 2019, Vercellis 2009 etc.)

Wrappers on the other hand relies mostly on the classifier in the attribute selection process. Regardless of the attribute ranking, wrappers using a search engine, work through all possible combinations of variables within the entire set and selects the particular set composed of only predictive attribute that provide the highest predictive performance for a given classifier. Wrappers are associated with high computational cost as they are supposed to carry out an assessment of each possible combination of variables identified by the search engine each time through the entire training phase. They are also associated with increasing overfitting by relying on the classifier and the classification algorithm used affects their performance. Also, attribute selected using a given classifier may not work the same way with another classifier. Classifiers such as Genetic Algorithm (GA), Support Vector Machine (SVM) etc. fall under this category.

In embedded methods, attribute selection process lies within the algorithm and thus selection of optimal subset of features take place during the model generation phase. A good example under this category are the classification trees (decision tree induction algorithm).

Once applied, these algorithms produce a model as well as subset of predictive attributes deemed relevant for the classification model induction.

For both filters and wrappers, a search strategy is very crucial, the simplest search strategy is the *exhaustive search* in which all possible subsets of attributes are evaluated and the best subset is selected. The other simple and more efficient search strategy based on *greedy sequential technique* are the *forward* and *backward* selection. In the forward, also referred to as bottom-up search strategy, the process begins with an empty set of attributes. Then attributes are subsequently introduced one at a time based on a ranking obtained from an appropriate relevance indicator. The stoppage point is determined by either the accuracy remaining unchanged for any addition of an attribute or after reaching a prefixed threshold relevance index. The backward strategy is the reverse of the forward. It begins with a full attribute set, and then attributes are removed one at a time again based on a preferred relevance indicator. The algorithm stops when the relevance index of all the attributes still included in the model is higher than a prefixed threshold. The last but not the least search strategy under this category is forward-backward search strategy. It is a trade of between the earlier discussed strategies. The best attribute among those excluded is introduced and the worst attribute among those included is eliminated. Again, threshold values for the included and excluded attributes determine the stopping criterion. (Moreira et. al. 2019, Vercellis 2009, Zaffar et. al. 2019, Jalota & Agrawal 2021, Zaffar et. al, 2017 and Echegaray-Calderon & Barrios-Aranibar, 2015). For this study, minimum Redundancy Maximum Relevance (mRMR), a filter based feature selection technique was used for feature selection on educational dataset obtained from public secondary schools in Kitui west constituency. Filter method as mentioned earlier have the advantage of high computational efficiency and can be generalized to various machine-learning models.

Minimum Redundancy Maximum Relevance (mRMR) as mentioned earlier is a filter feature selection method, developed by (Peng et al. 2005). It controls for redundancy among selected features while preserving the relevant ones for the model development. This is achieved by selecting features that are more predictive to the target class and have minimal redundancy (are independent of each other). Under the mRMR framework, the idea is first to get features that have high correlation (maximum relevance) with the target class, then among the selected features obtain the ones that have low correlation within them (minimum redundancy). Both relevance and redundancy are measured based on mutual information.

Mutual information's mean value of all features $\{Z_i\}$ with the target class $\{C\}$ is used as a measure of maximum relevance. This provides a strategy for selecting a feature set K composed of m features a subset of $\{Z_i\}$ that are highly correlated to the target class $\{C\}$ (features with maximum relevance). The equation 3 below can be used to obtain the mutual information between two random variables within a given set.

$$I(Y,X) = \int_{\Omega Y}\int_{\Omega X}(P(X,Y)\log\frac{P(X,Y)}{P(X)P(Y)})dxdy \quad (3)$$

In case the variables Y and X are categorical (discrete), the mutual information formula above takes the form shown in equation 4 below.

$$I(Y,X) = \sum_{y\epsilon\Omega Y}\sum_{x\epsilon\Omega X}P(X,Y)\log\frac{P(X,Y)}{P(X)P(Y)} \quad (4)$$

Where: $\boldsymbol{\Omega}$Y is the sample space for Y, $\boldsymbol{\Omega}$X is the sample space for X, P(x) is the probability density for X, P(y) is the probability density for Y and P (x, y) is the joint probability density.

Due to the high computational cost involved in estimating the probability density for continuous variables as shown in equation 3 above, the mutual information for such case is replaced with F – statistic F (Y, X). Hence, the corresponding mutual information mean value takes form shown in equations 5 and 6 below.

$$V_D(s) = \frac{1}{|s|}\sum_{z_i\epsilon S}MI(C,Z_i) \quad (5)$$

$$V_C(s) = \frac{1}{|s|}\sum_{z_i\epsilon S}F(C,Z_i) \quad (6)$$

Where: V_D (S) is the mutual information mean value for discrete variable, members of set S and V_C (S) the corresponding value for continuous variables, members of set S. This approach only selects features that are relevant to the target class and does not test for dependency among the selected features. Thus, if used alone could lead to a subset of dependent feature relevant to the target class, which enhances redundancy. The second approach implemented using equations 7 and 8 below addresses the redundancy issue among the selected features as mentioned earlier.

$$W_D(s) = \frac{1}{|s|^2}\sum_{z_i,z_j,\epsilon S}MI(Z_i,Z_j) \quad (7)$$

For continuous variables (features), this part is implemented using Pearson correlation as shown in equation 6 below.

$$W_C(s) = \frac{1}{|s|^2}\sum_{z_i,\epsilon s}\rho(Z_i, Z_j) - (8)$$

Based on the above information, Peng et al, 2005 proposed four variants of mRMR that can be organized in to two categories. These categories include;

i) Discrete variables (features)
   a) Mutual information difference (MID) = $V_D(S) - W_D(S)$
   b) Mutual information quotient (MIQ) = $\frac{V_D(S)}{W_D(S)}$

ii) Continuous variables (features)
   c) Mutual information difference (FCD) = $V_C(S) - W_C(S)$
   d) Mutual information quotient (FCQ) = $\frac{V_C(S)}{W_C(S)}$

A weakness that emanates from the different measures used for relevance and redundancy due to difference in the scales of the selected measures affects the trade-off between the relevance part and the redundancy. The quotient variants are not affected much as compared to the difference variance. Hence this study adopted MIQ for the same reason among others mention in this study (Peng et al. 2005, Zhao et al. 2019, Billah et al. 2020).

2.3. Empirical Review.

As stated earlier at the beginning of this chapter, various empirical literature relevant to the study's problem were reviewed with a keen interest on their objectives to determine their usefulness to the current study. From this review also, research gaps were identified which the study sought to address. Lastly, factors that influence students' academic performance were identified from the review as well as various machine-learning models used by different researchers to predict student academic performance with their respective accuracy.

Various types of datasets classified according to their sources have been used by researchers to address various aspects of learners and the learning process which include; predicting students' academic performance, predicting students likely to drop out, understanding the learning process, learners and the environment in which learning occurs, advancing scientific knowledge about learners and the learning process etc. These types of data include; Massive Open Online Courses (MOOCs) datasets, e-learning datasets, traditional

classroom datasets and intelligent tutoring system datasets. These datasets were obtained from any of the three levels; primary level, secondary level and higher level with most of research being concentrated at the higher level.

A study carried out by (Amra, I. & Maghari A. 2017) used educational dataset from secondary schools obtained from the ministry of education in Gaza strip in the year 2015. The dataset consisted of 500 students' records with 10 attributes, which were reduced to 8 attributes after a process of feature selection by observation. Thus their optimal subset of features consisted of; student gender, student date of birth, student field of specialization, student city of residence, name of the secondary school attended, student marital status, Father's job and student performance status (a pass or incomplete). Then two supervised machine learning classification algorithms (KNN and Naïve Bayesian) were trained with 70% of the dataset and evaluated on the remaining 30% on metrics such as Recall, Precision and Accuracy. According to the results of their study, Naïve Bayesian was the best with a recall, precision and accuracy of 93.6%, 94.65% and 93.17% respectively. From this study, the following research gaps were identified; (i) the dataset used was small in terms of both records (500) and attributes (8). (ii) Key attributes in students' academic performance prediction were either under-represented or missing i.e. student behavioral factors. (iii). the study used only two among many supervised machine learning algorithms. (iv) the dataset used was obtained from public data source as opposed to a dedicated student's dataset.

Another study by (Saa, A. et. al, 2019) used a dataset related to student's information consisting of 34 attributes and 56,000 records obtained from a private university in United Arab of Emirates (UAE). They identified four categories of the most important features that are more predictive on students' academic performance which include; students' demographics, course and instructor information, student general information and student previous performance information. Feature selection was carried out using Information Gain that ranked attributes in the order of predictive ability on the target class. Out of the initial 34 attributes, they were able to obtain an optimal subset of features that consisted of only 15 features, which include; high school name, university requirement average, course name, attendance warning, number of absences, prerequisite average, student program, high school merit, high school percentage, mathematics average, has discount, student nationality, offering college, Gender and physical average. This study compared five commonly used data mining technique in EDM namely; Support Vector Machine (SVM), Decision Tree (DT), Regression, Naïve Bayes (NB) and Artificial Neural Networks (ANN). These algorithms were evaluated on Accuracy, Recall and

Precision metrics, with Recall and Precision being broken into the categories of the target class i.e. High performer, Low performer and Failure. The study concluded that Random Forest, a DT based ensemble algorithm was the best overall with an accuracy of 75.52% hence it was the proposed algorithm best for student performance prediction. The following research gaps were identified from the study: 1) Data source limited to student information system only. 2) More historical records to be included to enhance the prediction levels of student academic performance. 3) Key attributes in students' academic performance prediction were either under-represented or missing i.e. student behavioral factors. 4) Model performance way too low.

A study by (Sharma, D., & Aggarwal, D. 2021) investigated the influence parental factors have on students' academic performance. They used python to examine a dataset that consisted of about 400 college students randomly selected from Jagan Institute of Management Studies, Delhi, India. The results of the study showed that, parental factors such as; job of the parent, cohabitation status of the parents, internet connection and paid classes at home, family size, education of mother and father etc., have a significant impact on students' academic performance. Twelve features related to parental influence and students' previous academic performance were used in this study. Which include; family size, cohabitation status of parents, Mothers education, Father's education, Mother's job, Father's job, guardian of the student, family educational support, extra paid classes, internet access at home, quality of the family relationships and grades (the target class), given by the average of semester 1, 2 and 3 marks obtained by the student. Feature selection (dimensionality reduction) by correlation coefficient whose value ranges between -1 to +1 was used in this study to obtain the optimal subset of features that are more predictive to the student grades. Only variables with the greatest correlation with final student grade were considered useful for the study. The study used factors that have positive correlation coefficients on the grades and linear regression as the basic machine learning technique upon which their model was build. Then compared it with other techniques such as random forest, support vector machine, gradient boost and Naïve Bayes (as their baseline). Models were evaluated on Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as they are easier to interpret. The linear regression model was the best overall, although the performance of both support vector machine (SVM) and gradient boost was good. Research gaps similar to the previously reviewed studies were identified. The following research gaps were identified from the study: 1) More historical records to be included to enhance the prediction levels of student academic performance. 2) Key attributes in students'

academic performance prediction were either under-represented or missing i.e. student behavioral factors. 3) Low sample size (400 subjects) for efficient training of machine learning algorithm.

In their study (Yousafzai, B. et. al, 2020) proposed a prediction system using machine learning techniques for student grade and marks. The system was developed using students' performance historic dataset obtained from Federal Board of Intermediate and Secondary Education Islamabad Pakistan. The dataset consisted of 106 features and 80,000 records. Feature selection by genetic algorithm (GA), a metaheuristic feature selection method was used to obtain an optimal feature subset that consisted of 29 features. GA method puts attributes into high and low ranks, with the low ranked features being dropped. The old/new datasets consisting of all features and only features that are more predictive to target class (highly ranked) respectively were used to train the decision tree classifier and the K-nearest neighbor regression model. The decision tree classifier using 10-fold cross-validation attained an accuracy of 94.39% while K-nearest neighbor regression attained an accuracy of 85.75 % on the original dataset i.e. without feature selection. While on the new dataset consisting of optimal subset of features obtained by GA method, decision tree classifier improved to an accuracy of 96.64% while K-nearest neighbor regression attained an accuracy of 89.92%. The following research gaps were identified from the study: 1) Model developed was unnecessarily complex. 2) High computational cost feature selection method used. 3) Key attributes in students' academic performance prediction were either under-represented or missing i.e. student behavioral factors.

It is worthy to note that research on educational data mining in the African continent has not been carried out as in other developed countries. This can be attributed to the fact that the continent is composed of developing countries, which lack technology, expertise and resources to effectively implement such studies. Their educational setting also does not allow for easier access to educational data for such research as most of the institutions store their data in various educational management information systems databases and others in hardcopy files. Therefore, for a researcher to venture in these fields of research and in African countries, must be willing, ready and prepared to extract the relevant data from such hard copy files in various institutions of learning especially at the lower level of primary and secondary education. This is very tedious, expensive (in terms of time and funds) and involving hence the reasons for low research in these fields in the developing countries.

Despite the above-mentioned challenges, (Qazdar A. et. al 2019) successfully conducted a study aimed at developing a model for predicting students' academic performance based on machine learning techniques. The dataset used consisted of 478 physics students obtained from the School Management System "MASSAR" (SMS – MASSAR) of H.E.K high school in Morocco within 2016 – 2018 academic periods. They used the correlation coefficient method to obtain the optimal subset of features, which consisted of 15 features. The new dataset was then used to develop a multiple regression model I & II whose results where then used to predict the students Bac score. Bac is a certificate issued by the Government of Morocco through its ministry of education for various majors (Physics, Computer, Statistics etc.). The two models were evaluated in two stages; stage one involved metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) and Relative Squared Error (RSE). The two models combined performance achieved; MAE, RMSE, RAE and RSE of 0, 1.25, 0.51 and 0.75 respectively on the calculated GB (Grade of Bac). Stage two involved discussion and interpretation of the results by a pedagogical committee established by the administrators.

Kenya has not been left behind; a study carried out by (Obadiah M. et al., 2019) sought to develop a model for predicting students' academic performance in secondary schools using machine-learning algorithms Dataset used was collected through a questionnaire from fourth form leavers of five public institutions in Kenya and consisted of 1720 records and 60 attributes (features). These attributes were reduced to 15 that consisted the optimal subset of features obtained through appropriate feature selection techniques e.g. Gain ratio, Info gain and One R-test. The new dataset was then used to train Naïve Bayes, J48 decision tree and Neural Network multilayer perceptron supervised machine learning classification algorithms in WEKA (Waikato Environment for Knowledge Analysis). Models were evaluated on; accuracy, Precision, Recall, Error rate and ROC area. J48 Decision tree performed better than the other models in predicting students KCSE performance in secondary schools.

Another study by (Ogwoka, et al. 2015), then students at Jomo Kenyatta University of Agriculture and Technology, used a dataset consisting of 173 students obtained from students' management information system of the Technical University of Mombasa-Kenya. Feature extraction was carried out using WEKA and the resultant optimal subset of features used to train K-Nearest Neighbor and decision tree. They used the second semester's results as the testing dataset and evaluated the two models obtained on; Accuracy, Recall, Precision F-measure and Kappa. Their model realized a much higher accuracy of 98.8439%.

Last but not the least is the study carried out by (Mgala M. 2016) that proposed a model for predicting students' academic performance that included a mobile interface whose design was based on user needs. The mobile interface was to enhance the model usability in developing countries like Kenya where electricity and computers are not common among the local primary schools. The study was based in Kwale County in Kenya. Two datasets were used one consisting of 2426 records and 22 attributes obtained from 54 primary schools in the rural areas using a questionnaire. The second dataset consisted of 1105 records with 19 attributes obtained from eleven primary schools located adjacent to a town in Kwale County. Feature selection by filter algorithms (Information Gain, ReliefF and Gain ratio) as opposed to wrapper algorithms were used on WEKA. The optimal subset of features was used in WEKA to build perdition models with logistic regression, multilayer perceptron, J48, Sequential Minimal Optimization (SMO), Naïve Bayes and Random forest algorithms. These models were evaluated on Recall, Specificity value, ROC area, F-Measure, Cohen's kappa value and RMSE value metrics. Logistic regression outperformed the other models, thus recommended by this study for student academic performance prediction. The table below presents a summary of empirical review of earlier models on the study topic and the associated research gaps.

TABLE 2

Summary of Empirical Review with Research Gaps Identified.

Author(s)	Level Country	Sample size	ML Model	Performance	Research Gap(s)
Amra, I. & Maghari A.2017	Secondary Gaza	500	KNN & Naïve Bayesian	NB; R=93.6%, P=94.65% A=93.17%	- Key attributes in model development were left out. - Used only two machine-learning algorithms. - Accuracy of the developed model not equal to 100%
Saa, A. et. al, 2019	University UAE	56,000	SVM, DT, Regression, NB & ANN	Accuracy; DT 75.52% the best	- Accuracy of the developed model is way below 100%

Sharma, D., & Aggarwal, D. 2021	College Delhi, India	400	LR (linear regression),RF, GB & NB	LR was the best on MAE & RMSE metrics	- Low sample size (400 subjects) for efficient training of machine learning algorithm - Feature selection method prone to redundancy as it only selects features highly correlated to the target class.
Yousafzai, B. et. al, 2020	Secondary Pakistan	80,000	KNN & DT	DT 96.64% KNN 89.92%.	- Key attributes in model development left out – students behavioral factors.
Qazdar A. et. al 2019	High school Morocco	478	Multiple regression	MAE = 0 RMSE =1.25	- Low sample size (478 subjects) for efficient training of machine learning algorithm - Feature selection method prone to redundancy (Correlation coefficient feature selection)
Ogwoka, et al. 2015	University Kenya	173	DT &KNN	DT was the best	- Too low sample size (173 subjects) for efficient training of ML algorithms.
Mgala M. 2016	Primary Kenya	2426 & 1105	LR, MLP, RF SMO, NB & J48	LR- 90% sensitivity	- Data collection method (survey) was subjective as opposed to objective and was for a particular instance as opposed to historical.

2.3.1 Features that Affect Students' Academic Performance.

Factors that influence student academic performance fall under two broad categories namely; Individual factors and institutional factors. (Mgala M. 2016, Obadiah M. et al., 2019, Mgala, & Mbogho, 2014 etc.). Individual factors include; student performance, behavior of the student, student background and student attitudes while institutional factors include; family of the student, type of school attended by the student and the community from which the student belongs. With some additional features based on domain Knowledge within the categories mentioned, this study adopts these features as they represent almost every aspect of student academic performance.

(a) *Institutional factors* – these are factors that concern the community and family where the student belongs and the kind of school in which the student attends. These factors indirectly affect the student academic performance by contributing to the impact that students' individual factors have on their academic performance. From the empirical review, several studies (Mgala M. 2016, Obadiah M. et al., 2019, Amra, I. & Maghari A. 2017, Ramaswami & Bhaskaran 2009 etc.) have identified family background as one of the major factors that influence students' academic performance. These factors can be split into family resources, family practices and family structure. *Family practices* involves all activities parents/caregivers should do to support good performance for their children. Parental involvement, expectation and support of their children and the institution in which they are schooling has been identified by research (Mgala M. 2016) as a key factor that influence academic performance. On the other hand, *family structure* includes; number of members (especially dependents) in the family, whether the students come from a single parent or a total orphan and the nature of cohabitation of the parents in the family. Studies (Sharma, D., & Aggarwal, D. 2021, Mgala M. 2016) have established that these factors significantly affect student academic performance. Last but not the least under family background is the *family resources*, this can further be split into *social, human and financial* factors. Which from the empirical review, studies have represented the financial and social resources in terms of mother's job and father's job (Sharma, D., & Aggarwal, D. 2021). Which will determine the student social status that have been identified by various studies e.g. (Bornstein and Bradley, 2014) as a factor that contribute to students' academic performance. The parent's level of education has been considered a resource as educated parents assist their children in various aspects of learning as opposed to illiterate parents.

Factors associated with the school where the student attends include; the school structure, school policies and practices, school resources and students' characteristics. Indicators associated with student characteristics that affect student academic performance according to (Mgala M. 2016) are; students in need of urgent intervention to save them from registering dismal performance, students from problematics homes, transfer cases, and the proportion of marginalized students. School policies and practices spell out the environment that exist within the school. They determine whether it is conducive for learning or not. These involves all policies enacted by the school administration to support teaching and learning and may include; exam policies that spells out exam administration guidelines, language policy encourages student to enhance their command to the recommended languages in a given country, e.g. in Kenya, students are encouraged to practice to perfect, communication in English and Kiswahili. Discipline policy that spell out the expected code of conduct of students and teachers within the school compound and beyond, with respective punitive measures to be taken on anyone who violates any of the laid down regulations. Last but not the least is the teaching methods used in a given school that determine the content delivery procedures that is usually guided by the type of students admitted within a given academic period.

Factors concerned with school structure include: school size in terms of student's population, school type this refers to whether the school is private or public, boarding or day, boys or girls school and in Kenya public schools have categories depending on how they have been equipped by the government (school resources) i.e. National schools which are fully equipped with almost all required resources for learning. These schools are for students who pass well in their KCPE i.e. with 400+ out of 500 marks. The next category is extra county and county schools, partially equipped by the government. They have the necessary resources for a student to pass and admit those who attained the pass mark and above i.e. 250+ marks. Then the last category, the sub-county schools that are for anyone who yearns for secondary education. The idea behind their establishment in the year 2008 was to ensure 100% transition from primary to secondary level. They are all mixed day school thus tend to be very cheap. Most of the students found in these schools have issues concerning either their performance, discipline, family, or fees payment. Most bright students whose parents cannot afford fees charged at national schools, extra and county schools are admitted to these kind of schools and still make it to the universities.

Community factors that influence student performance can be classified in to three categories namely; community support, security within the community and community

practices. Factors related to community support involves measures put in place by the provincial administration to guard students, ensuring they are disciplined even when out of school, support given by the community to teachers and the school administration etc. Security is very crucial for peaceful learning. Communities faced with insecurities emanating from ethnic clashes and other causes, interfere with the learning and hence performance in the national exams of students in such communities. In Kenya, Northeastern region is a good example due to frequent terrorist attacks. As far as community practices are concerned, those which affect student performance include; female genital mutilation (FGM) and early marriages which has been a concern from the government, non-governmental organizations (NGOs) and members of the civil society to sensitize communities to leave such practices. In some occasions, communities allow politics to spill over into schools interfering with the conducive learning environment etc.

(b) *Student Individual factors* – These refers to student characteristics which have been identified by various studies to influence students' academic performance. It is worthy to note that as mentioned earlier, these factors draw their input from the institutional factors discussed earlier on in this chapter. They have been branded with different variables names by various studies but in their actual sense they represent; student behaviour, student background, student attributes and student performance (Hirokawa, S. 2018, Mgala M. 2016, Hellas, A. et al. 2018 & Ashraf A. et al., 2018).

Students' behavioral factors have been identified by various studies (Mgala & Mbogho 2015, Hirokawa, 2018, Amrieh, et. al, 2016, Gajwani & Chakraborty 2021 etc.) as factors that are highly predictive on students' academic performance. Hirokawa, 2018 has shown that behavioral factors a lone account for about 75% on students' academic performance prediction. Thus, behavioral factors are key to students' academic prediction and should not be left behind while developing models for predicting students' academic performance. These factors include; student discipline, student commitments to learning activities, student's attitude towards themselves, life and the learning process etc. Appropriate intervention driven by insights obtained from studies such as this one, should guide the development of strategic student's engagements geared towards improving students' academic performance.

Student's previous performance has also been very instrumental in the development of students' academic performance prediction models. It is linked to indicators like truancy, transfers, student's attainment and achievements.

Student's background, which include; experiences in the past, demographic and health have also been identified as factors, which influence student academic performance. Student gender has been a key factor to consider when making any decision concerning students' academic performance. Like in Kenya, girls' university entry is a grade lower than that of boys; this implies that in Kenya boys perform better than girls do. The health status of a student affects their performance by increasing their absenteeism from class as they seek medical attention. Experiences faced by the student in the past like denial by parent(s), mistreatment, bullying, poverty, child labour etc. might change the student's attitude towards learning that negatively affect their commitment to learning process thus registering dismal performance. Child labour is very common among developing countries, which has called for regulation to be enacted to make it illegal. Like in Kenya, labour laws prohibit employment of any under 16 years old child. They are referred to as *minors*, who should be actively engaged in various institutions of learning.

Student attributes which include; self-perception, values and goals, affect a lot on student academic performance (Conley, 2012). The way a student perceives the entire education process is determined by their values in life, which are anchored on the student's goals. Students who set higher targets that are realistic and achievable, and then focus all their efforts towards attaining them always perform well in their academics. Table below presents, factors that generally influence students' academic performance.

TABLE 3

Features that Generally Influence Students' Academic Performance

Variable Category	Sub-Category	Variable	Symbol	Value
Institutional Variables	School Variables	- School type	SST	Categorical
		- School level	SLV✓	"
		- Teacher's attitude towards students	TAS	"
		- Teacher's Commitment	TCM	"
		- Teacher's Absenteeism	TAB	"
		- School facilities	SFC✓	"
		- Trained teachers	TTN	"
	Family Variables	- Family type	FTP	"
		- Mother's education	MED	"
		- Father's education	FED	"
		- Parent/Guardian encouragement	PGE	"
			PLP	"

		- Parental level of participation - Parents state of harmony - Family income - Family size - Difficulties in fees payment	PHM FIN FSZ DFP	“ “ “ “
	Community Variables	- Community security - Community support	CSR CSP	“ “
Student Individual Variables	Student behavior	- Student discipline - Student English Command - Days Absent - Consulting teachers - Completing Assignment	SDP✓ SEC ABS TCS ASC✓	“ “ “ “ “
	Student background	- Student age - Student religion - Student gender - Student status - Study time at Home	AGS✓ RGS GDS STS STH	“ “ “ “ “
	Student Attitude	- Attitude towards education	ATE✓	“
	Student Performance	- Challenges with exams - KCPE points (Entry status) - Subjects in Form One - Subjects in Form Two - Subjects in Form Three - Form 1 average points - Form 2 average points - Form 3 average points	CWE✓ ENT SBF1 SBF2 SBF3✓ AGF1✓ AGF2✓ AGF3	“ Numeric “ “ “ “ “ “
Student Academic Performance	KCSE Results	- (≥ 7 Points) PASS (1) - (≤ 6 Points) FAIL (0)	SAP	Categorical

Next section highlights the gaps identified from the review, which the study will seek to address.

2.3.2 Research Gaps.

This study, using educational dataset obtained from public secondary schools in Kitui west constituency will seek to address the following gaps identified during the empirical review: -

1. Models developed, though at different levels of study have not been able to achieve 100% accuracy. This is due to the limitations mentioned above which this study seeks

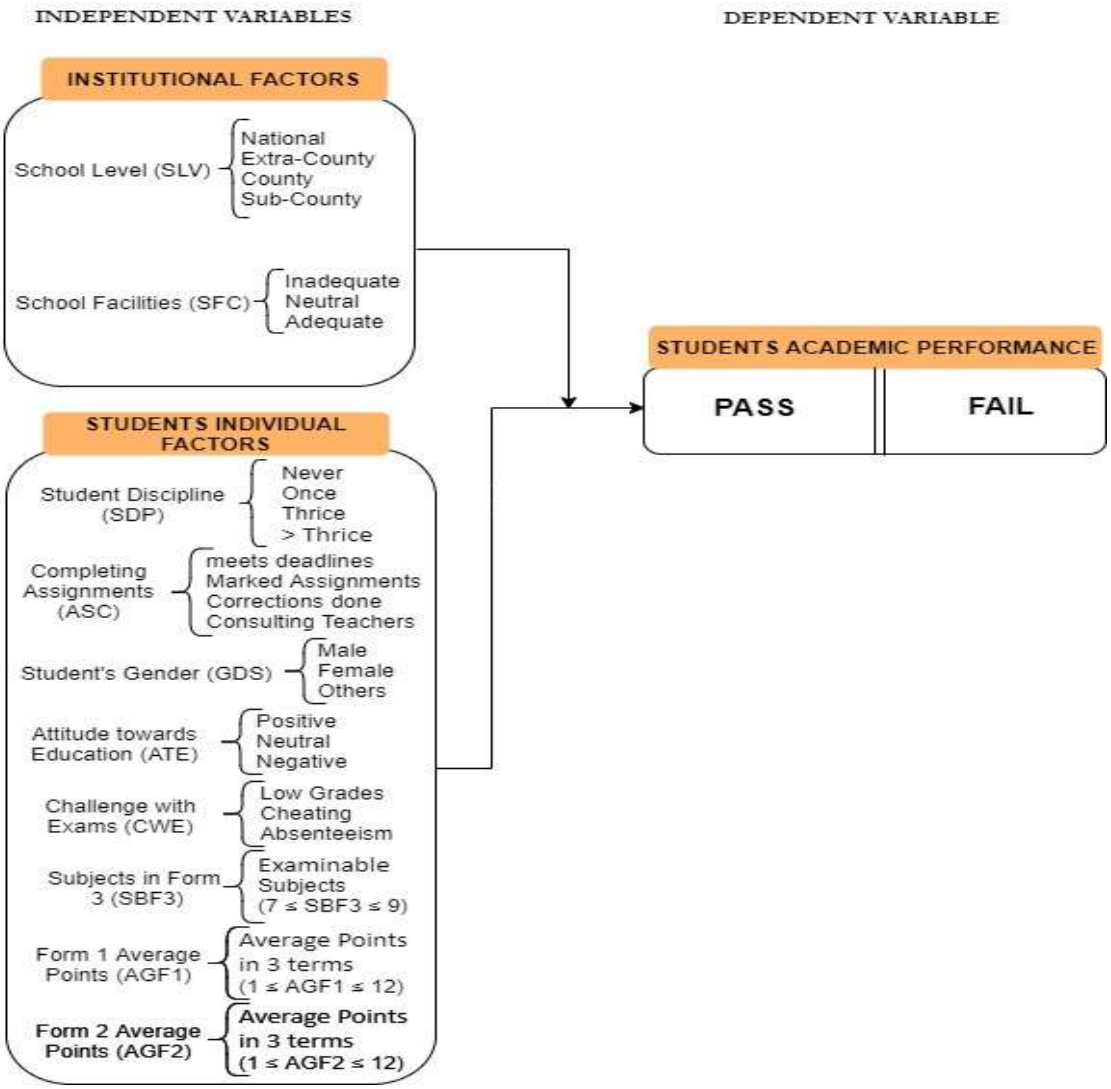
to address hence develop a model efficient in terms of computational cost and hopefully of higher accuracy.

2. The urgent need to arouse initiation of strategic intervention from various stakeholders of education at the secondary level using data from public secondary schools in Kitui west constituency. These interventions could be aimed at; reducing the big number of students who fail in their exams due to varied reasons, lack of learning materials etc., reducing school dropouts by understanding students and addressing their needs to the individual level in an effort to improve the learning environment.
3. There is lack of a comprehensive research in this topic at the secondary level in Kenya. Thus the researcher sought to contribute to the body of research on developing models for predicting students' academic performance using dataset obtained from public secondary schools in Kitui west constituency. This is after it was clearly established from the empirical review that, research on this topic at secondary level has been left out despite its data being very different from the other levels i.e. the primary and higher levels of education. Also in Kitui county research in this topic has not been carried out hence, this becomes the first of its kind in the whole County. The secondary level of education is very crucial in one's career as it marks the early specialization stage for various career options in higher levels of education. In Kenya if a student fails to score marks equal either to or above a defined threshold, which in this case is a C+, the student is not considered for a chance in the universities. These students are encouraged to try their lack in colleges and TIVET institution for diploma or certificates in various skillsets.

2.4 Conceptual Framework

FIGURE 3

Conceptual Framework



2.5 Operationalization of Variables

This refers to a process of defining a variable/feature or even a concept in a measurable way. Fuzzy concepts are defined for measurement or expressing them qualitatively/quantitatively. This process involves finding a measurable, quantifiable and valid index for both independent and dependent variables enhancing their manipulation at different levels. Operationalization of variables allows subjective variables that are difficult to measure be easily quantified. It brings out the exact meaning of a variable improving its quality in a given research hence improving the efficiency of the research design. Lastly, the process standardizes the variables making hypothesis strong and clear (Blau, 1962). The table below illustrate operationalization of the variables used in this study.

TABLE 4

Operationalization of variables

Variable Category	Sub-Category	Variable	Symbol	Value
Student	Student Identity		SID	001 - 5500
Institutional Variables	School Variables	- School level - School facilities	SLV SFC	Categorical “
Student Individual Variables	Student behavior	- Student discipline - Completing Assignment	SDP ASC	“ “
	Student background	- Student gender	GDS	“
	Student Attitude	- Attitude towards education	ATE	“
	Student Performance	- Challenges with exams - Subjects in Form Three - Form 1 average Points - Form 2 average Points	CWE SBF3 AGF1 AGF2	“ Numeric “ “
Student Academic Performance	KCSE Results	- (≥ 7 points) PASS (1) - (≤ 6 points) FAIL (0)	SAP	Categorical

2.6 Summary

The chapter has reviewed theories, concepts and empirical studies related to this study. It has also identified different features that influence students' academic performance and presented their relationships using a conceptual framework. In addition, feature selection methods used have been discussed together with various machine-learning techniques applied in EDM.

Research gaps to be addressed by this study have also been identified from the empirical review. Lastly, the chapter terminates with the operationalization of variables identified earlier on, in the study, which was used to provide answers to the research questions in this study.

CHAPTER 3

METHODOLOGY.

3.1 Introduction.

A description of the procedures and strategies employed during the implementation of the study is provided in this chapter. The main content outlined in the section includes; research design used in the study, research process, sampling and sampling procedure, target population, data collection procedure and data processing and analysis.

3.2 Research design.

A research design is a logical structure of enquiry or a framework of methods and techniques applied and used by the researcher to identify suitable research methods for the subject matter under study (Creswell, 2003). From the main objective of the study, which is to develop a model for predicting students' academic performance using supervised machine learning techniques, diagnostic, secondary (desk) and experimental research designs best suites this study. The combination of the three research designs is the most suitable for this study because, diagnostic research design under case history method that is concerned with any significant historical information concerning the research problem (case), was used to examine the underlying cause (factors/variables) of the problem. Therefore, this design led to achievement of specific objective one, through the empirical review contacted in the previous chapter.

Underlying causes (factors/variables) identified in the previous section guided the collection of historical data that best suites the study. This is because supervised machine learning algorithms require data accumulated over time from which to learn, reveal patterns, sequences, associations etc. that will be used for the prediction process. Thus secondary (desk) research design (Stewart & Kamins 1993, Verschuren, et al 2010) under quantitative framework, was employed in this study to extract secondary data relevant to the objectives of this study from various department in public secondary schools in Kitui west constituency. These departments include; the exams departments, administration, class teachers, Guidance and counseling departments etc. Data which was stored either in hardcopy or softcopy files was extracted to an excel document.

From the empirical review in the previous chapter, features that influence students' academic performance were identified as mentioned earlier. With appropriate feature selection method, the main features that influence student academic performance were identified as the optimal subset of features, which on its part can inform some decisions to stakeholders on designing intervention measures to address dismal performance among concerned learners. The optimal subset provides answer to the first research question in this study "What are the main features that affect students' academic performance in public secondary schools in Kitui west constituency?" Then experimental research design under quantitative framework will be used to achieve the remaining specific objectives (two and three) with the optimal subset of features and various supervised machine learning algorithms identified from the previous chapter, empirical review section.

3.3 Research process

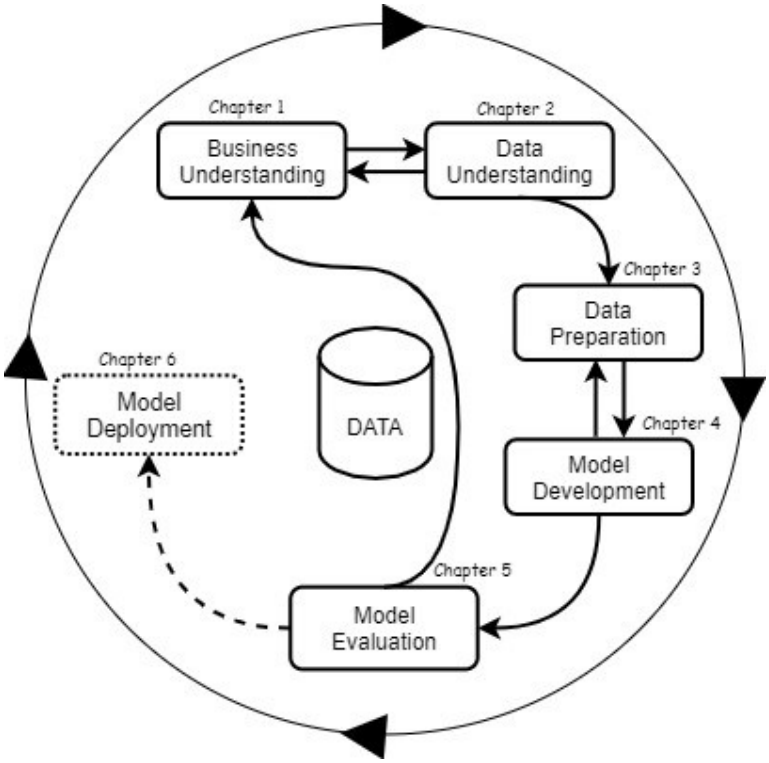
The study will adopt the Cross-Industry-Standard Process for Data Mining (CRISP-DM) by CRISP-DM, 2000) model to develop a classification model that will be used to predict students' academic performance in public secondary schools in Kitui west constituency. Binary supervised machine learning approach is the one that suites the study. This is because historical data obtained from the above section was used to train supervised machine learning algorithms, which will place students into two classes (Pass or Fail). Any new data record will be placed in any of the two classes during the models prediction process. CRISP-DM data mining model was conceived in 1996 and its development initiated in 1997 under a European Union project funded by ESPRIT. Since then, the model has evolved to become the technology neutral industry standard for the data mining process.

The model advances the following stages illustrated in the figure below as the most appropriate in knowledge discovery process. The models fit very well with the chapters in this study hence it is the most appropriate knowledge discovery model to base the study on. Figure 2 below illustrates a graphical representation of CRISP-DM data mining model, modified to suite the study and a discussion of its various stages thereafter.

FIGURE 4

Cross-Industry-Standard Process for Data Mining (CRISP-DM)

.



Source: CRISP-DM, 2000

3.3.1 Step 1: Business Understanding

This phase encourages the researcher to get a clear understanding of the business requirements of the data mining process. The stage provides a clear definition of the research problem being addressed by the study. For example, in the case of this study the problem of dismal performance among public secondary schools in Kitui west constituency. Diagnostic research design under case history method was adopted by reviewing several studies carried out earlier that concern the objectives of the study. This led to a deep understanding of the research problem and revealed factors that contribute to dismal performance among students in public secondary schools in Kitui west constituency. Thirty-eight factors/variables identified influence student academic performance as illustrated on table 2: features that generally influence students' academic performance. This laid down the research foundation, as the factors/variables identified guided the next stage of data understanding. This stage coincides

with chapter one of this study as illustrated in the diagram and provides the first remarkable step towards the achievement of research objective one.

3.3.2 Step 2: Data Understanding

This stage involves understanding data sources, collection methods and tools, the type of data to be collected in terms of its structure, quality and the challenges experienced during its collection or extraction. For this study, secondary(desk) research design (Stewart & Kamins 1993, Verschuren et al, 2010) under quantitative framework, was employed to extract secondary data relevant to the objectives of this study from various departments in public secondary schools in Kitui west constituency. The historical data that is objective as opposed to subjective accumulated over the research period and concerned the target population of the study. According to the main objective of the study, this type of data best suites the study as it accumulated over time during the normal operation of the schools, so there is no possibility of bias towards the research objectives.

FIGURE 5

Data Extracted from Various School's Repositories.

SID	SST	SLV	TAS	TCM	TAB	SFC	TTN	MED	FED	PGE	PHM	FTP	FSZ	FIN	DFP	PLP	CSR	CSP	AGS	RGS	GDS	STS	STH	ATE	CWE	SDP	SEC	ABS	TCS	ASC	SBF1	SBF2	SBF3	ENT	AGF1	AGF2	AGF3	SAP
001	3	3	1	3	2	3	3	3	4	1	1	1	1	3	2	1	1	1	2	2	1	2	3	1	1	3	2	2	2	2	12	12	8	2	7	6	6	0
002	4	4	3	3	1	2	3	2	2	1	1	1	1	2	1	1	2	1	2	2	2	2	2	1	2	1	3	1	3	1	11	8	8	10	8	6	2	1
003	1	2	3	3	1	3	3	2	3	1	1	1	1	2	2	1	1	1	1	2	2	3	3	1	1	4	2	1	3	1	12	8	8	4	8	9	5	1
004	4	4	2	3	1	3	3	3	3	1	1	3	1	3	1	1	1	1	2	2	2	3	3	2	1	2	3	1	2	1	9	8	7	12	5	7	5	0
005	1	2	2	2	2	3	3	2	2	1	3	1	1	2	1	2	1	1	1	2	2	3	2	2	1	3	2	1	2	2	12	8	8	11	6	6	5	1
006	1	2	3	3	2	2	3	5	5	1	1	1	1	2	2	2	1	2	1	2	2	3	2	1	2	4	3	1	3	1	12	8	8	1	7	8	7	1
007	2	3	3	2	2	2	2	4	5	1	1	1	1	3	2	1	2	1	1	2	1	2	2	1	1	2	3	1	2	2	12	12	8	5	7	6	3	1
008	3	3	3	3	1	3	3	2	5	1	1	1	2	3	2	1	1	3	2	2	1	2	2	1	1	1	2	1	1	2	12	12	8	9	5	4	3	0
009	1	2	3	3	1	3	2	2	3	1	1	1	2	2	2	2	1	1	2	2	2	3	2	1	1	2	3	2	3	1	12	8	8	8	8	8	7	1
010	2	2	2	2	2	3	2	3	3	1	1	1	1	3	2	1	2	3	2	2	1	2	2	1	2	3	3	1	2	2	11	8	8	7	2	4	5	0
011	3	4	2	3	2	3	2	5	2	1	2	2	2	2	1	1	2	1	2	2	1	1	2	1	1	2	3	1	2	1	12	12	8	4	10	7	7	1
012	1	2	3	3	1	2	2	3	5	1	1	1	1	3	1	2	1	1	2	2	2	3	2	1	1	1	2	1	2	2	12	12	8	6	8	7	5	1
013	1	2	3	3	1	2	3	4	4	1	2	2	1	2	1	1	1	1	1	2	2	3	3	1	2	1	3	1	2	2	12	8	8	2	8	6	4	1
014	2	3	3	3		1	2	3		1	1	1	1	3	2	1	2	1		2	1	2	2	1	2	2	2		2	1	12	12	8					
015	4	4	2	3	1	3	3	4	3	2	2	1	1	3	1	2	3	3	2	2	1	1	2	2	1	2	3	3	2	2	11	11	8	3	9	6	6	0
016	1	2	2	3	2	2	2	2	3	1	1	2	1	2	1	1	2	3	1	2	2	3	3	1	1	1	3	1	2	1	12	12	8	10	6	4	3	0
017	4	4	3	3	1	1	1	1	1	3	1	1	2	1	1	1	3	3	2	2	2	3	2	1	1	1	3	3	1	1	11	11	8	6	4	4	6	0
018	1	2	3	3	1	3	3	3	5	1	1	1	1	3	1	1	1	1	1	2	2	3	2	1	2	2	3	1	3	1	12	8	8	7	9	9	9	1
019	1	2	3	3	2	3	3	2	2	1	1	1	1	2	1	1	1	1	1	2	2	3	1	1	2	3	2	3	2	2	12	8	8	8	8	7	7	1
020	4	4	3	3	1	2	1	3	3	1	1	1	1	3	1	1	1	1	2	2	2	3	2	1	1	1	3	1	2	1	11	8	8	3	4	4	6	0
021	2	2	3	3	2	1	1	2	5	1	1	1	3	2	1	1	2	1	2	2	1	2	2	1	1	2	3	2	3	2	12	12	8	12	10	8	3	1
022	4	4	2	2	2	2	2	5	5	2	1	4	1	3	2	2	3	3	2	2	2	3	2	1	2	4	3	4	1	2	11	8	8	1	6	6	10	0
023	1	2	3	3	1	3	3	2	2	1	1	1	2	3	2	1	1	1	2	2	2	3	3	1	2	2	3	1	2	1	12	8	8	10	8	7	7	1
024	4	4	3	3	1	2	2	2	2	1	1	1	1	2	1	1	1	1	1	2	2	3	3	1	1	1	3	1	3	1	11	11	8	12	10	7	4	1
025	3	3	1	3	2	3	3	3	4	1	1	1	1	3	2	1	1	1	2	2	1	2	3	1	1	3	2	2	2	2	12	12	8	1	7	6	5	0

Sheet1

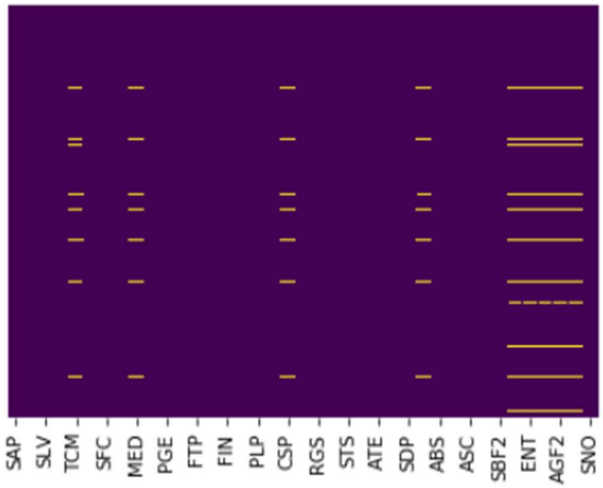
3.3.3 Step 3: Data Preparation

This involves preparing final dataset for the next phase of model development. Various data pre-processing techniques were utilized such as sampling, data normalization, feature extraction and engineering and other dimensionality reduction techniques that suites the dataset. The output of this stage is an optimal subset of features, which will be used in the next phase of model development. In this study, most of data preparation procedures like dealing with inconsistences, outliers, mismatch, etc. were dealt with during data entry into the excel document using appropriate excel tools. Data entry process was keenly contacted to avoid any data entry mistakes and crosschecking the document severally minimized chances of such errors. To further minimize such errors, data preprocessing was carried out in python where missing values discovered were dealt with, superfluous variables were dropped and feature selection was carried out using *minimum Redundancy Maximum Relevance* (mRMR), a filter feature selection method, to obtain the optimal subset of features. This was done by implementing the Mutual Information Quotient variant (MIQ) on the cleaned dataset. This variant was preferred due to its flexibility, effectiveness, easy to implement and low computational cost. The resultant is a final well-refined dataset composed only of those features that are more predictive to the target class.

FIGURE 6

Heat maps Displaying Missing Values in the Uncleaned and Cleaned Dataset

```
In [418]: sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')
Out[418]: <matplotlib.axes._subplots.AxesSubplot at 0x1f805b86248>
```



```
In [420]: sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')
Out[420]: <matplotlib.axes._subplots.AxesSubplot at 0x1f814bc4f48>
```

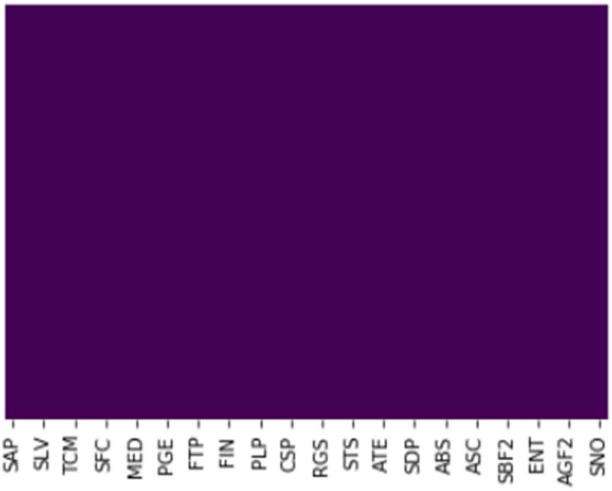


Figure 6 above is a visualization of missing values before and after the dataset was cleaned. The yellowish color show how blanks are distributed within the dataset. The uniform dark purple coloration in the second figure indicates absence of missing values after data cleaning process.

FIGURE 7

Dropping of Superfluous Variable(S); Students' Serial Number (SNO)

```
In [94]: df.drop(["SNO"], axis = 1, inplace=True)
         df.head()
```

Out[94]:

	SAP	SST	SLV	TAS	TCM	TAB	SFC	TTN	MED	FED	...	ABS	TCS	ASC	SBF1	SBF2	SBF3	ENT	AGF1	AGF2	AGF3
0	0	4	4.0	1.0	2.0	2	2	2.0	4.0	4	...	1.0	3	1	10	7	7.0	11.0	7.0	7.0	6.0
1	0	1	2.0	3.0	3.0	1	3	3.0	3.0	5	...	1.0	2	1	12	8	8.0	9.0	5.0	5.0	6.0
2	1	4	4.0	3.0	3.0	1	2	2.0	3.0	4	...	3.0	2	2	11	8	8.0	7.0	10.0	8.0	3.0
3	1	1	2.0	2.0	3.0	1	3	2.0	3.0	3	...	1.0	2	1	12	8	8.0	3.0	8.0	7.0	7.0
4	0	2	2.0	3.0	3.0	2	3	2.0	3.0	3	...	1.0	2	1	11	11	8.0	9.0	7.0	5.0	3.0

5 rows × 38 columns

Feature selection was implemented, immediately after the dropping of superfluous variables on the cleaned dataset to obtain the optimal subset of features that consist of ten features that are more predictive to the target class.

FIGURE 8

Dataset Composed of Optimal Subset Of Features and the Target Class.

```
In [422]: pymrmr.mRMR(df, 'MIQ',10)
Out[422]: ['AGF2', 'SBF3', 'ASC', 'SFC', 'GDS', 'SDP', 'ATE', 'SLV', 'AGF1', 'CWE']
```

```
In [423]: Data=df[['AGF2', 'SBF3', 'ASC', 'SFC', 'GDS', 'SDP', 'ATE', 'SLV', 'AGF1', 'CWE','SAP']]
          Data.head(5)
```

Out[423]:

	AGF2	SBF3	ASC	SFC	GDS	SDP	ATE	SLV	AGF1	CWE	SAP
0	7.0	7.0	1	2	1	4	1.0	4.0	7.0	1.0	0
1	5.0	8.0	1	3	2	4	1.0	2.0	5.0	2.0	0
2	8.0	8.0	2	2	1	3	1.0	4.0	10.0	1.0	1
3	7.0	8.0	1	3	2	1	1.0	2.0	8.0	2.0	1
4	5.0	8.0	1	3	1	3	1.0	2.0	7.0	1.0	0

This implies that, the main features that influence students' academic performance include; Form 2 average points, Subjects in Form Three, Completing Assignment, School facilities, Student gender, Student discipline, Attitude towards education, School level, Form 1 average points and Challenges with exams. This provides the answer to the first research question.

3.3.4 Step 4: Model development

This stage involves subjecting the dataset composed of optimal subset of features against various machine-learning models that have a history of good performance and are efficient in the training due to computing resource challenges. A random sample of 5000 subjects was selected from the dataset consisting of optimal subset of features. It was then split into 70% training sample and the remaining 30% test sample. The training sample was used to train; Gradient boost classifier, Random forest classifier, Decision tree classifier and Deep Neural Network classifier. Once the development and training stage was over, developed models were ready for evaluation using the test data set in the next stage. This provided the solution to research question two.

FIGURE 9

Developed and Trained Models

```
#Define Keras model
model = Sequential()
model.add(Dense(12, input_dim=10, activation='relu'))
model.add(Dense(8, activation = 'relu'))
model.add(Dense(1, activation = 'sigmoid'))
print(model.summary())
```

```
Model: "sequential_7"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_19 (Dense)             (None, 12)                132
_________________________________________________________________
dense_20 (Dense)             (None, 8)                 104
_________________________________________________________________
dense_21 (Dense)             (None, 1)                 9
=================================================================
Total params: 245
Trainable params: 245
Non-trainable params: 0
_________________________________________________________________
None
```

```
#Execute the model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'] )
```

```
#Train the model on the Training sample
Trained_model = model.fit(X_Training_sample, Y_Training_sample,epochs=100, batch_size=10)
```

```
In [429]: from sklearn.tree import DecisionTreeClassifier
          dtree = DecisionTreeClassifier()
          dtree.fit(X_train,y_train)

Out[429]: DecisionTreeClassifier()
```

```
In [434]: from sklearn.ensemble import RandomForestClassifier
          rfc = RandomForestClassifier(n_estimators=100)
          rfc.fit(X_train, y_train)

Out[434]: RandomForestClassifier()
```

```
In [437]: from sklearn.ensemble import GradientBoostingClassifier
          gradient_boost = GradientBoostingClassifier()
          gradient_boost.fit(X_train, y_train)

Out[437]: GradientBoostingClassifier()
```

3.3.5 Step 5: Model evaluation

This stage involves subjecting the developed model to various metrics of evaluation discussed in the previous chapter to find out the extent to which the model has contributed towards addressing the business requirements or solutions of the research problem. Developed and trained models were evaluated using test sample on, Accuracy, Precision and Recall. Decision

tree achieved the best overall accuracy followed closely by Random forest, thus recommended for implementation for predicting students' academic performance in public secondary schools in Kitui west constituency. This stage produced a validated model ready for deployment, and this provides the solution to the third and the last research question.

FIGURE 10

Evaluation of the developed Models.

i) Decision Tree Classifier

```
print("Accuracy on training set: {:.3f}".format(dtree.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(dtree.score(X_test, y_test)))
```

```
Accuracy on training set: 0.974
Accuracy on test set: 0.966
```

```
predictions = dtree.predict(X_test)
from sklearn.metrics import classification_report,confusion_matrix
print(classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

           0       0.98      0.95      0.96       729
           1       0.95      0.98      0.97       771

    accuracy                           0.97      1500
   macro avg       0.97      0.97      0.97      1500
weighted avg       0.97      0.97      0.97      1500
```

```
print(confusion_matrix(y_test,predictions))
```

```
[[690  39]
 [ 12 759]]
```

ii) Random Forest Classifier

```
print("Accuracy on training set: {:.3f}".format(rfc.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(rfc.score(X_test, y_test)))
```

```
Accuracy on training set: 0.974
Accuracy on test set: 0.966
```

```
rfc_pred = rfc.predict(X_test)
print(confusion_matrix(y_test,rfc_pred))
```

```
[[688  41]
 [ 10 761]]
```

```
print(classification_report(y_test,rfc_pred))
```

```
              precision    recall  f1-score   support

           0       0.99      0.94      0.96       729
           1       0.95      0.99      0.97       771

    accuracy                           0.97      1500
   macro avg       0.97      0.97      0.97      1500
weighted avg       0.97      0.97      0.97      1500
```

iii) Deep Neural Network

```
In [78]: #Evalaute the Trained Model with test data
         _, accuracy = model.evaluate(X_Test_sample, Y_Test_sample)
         print('Accuracy: %.2f' % (accuracy*100))

         1500/1500 [==============================] - 0s 183us/step
         Accuracy: 82.87

In [79]: from sklearn.metrics import classification_report,confusion_matrix
         print(confusion_matrix(Y_Test_sample,predictions))

         [[590 148]
          [109 653]]

In [80]: print(classification_report(Y_Test_sample,predictions))

                       precision    recall  f1-score   support

                  0.0       0.84      0.80      0.82       738
                  1.0       0.82      0.86      0.84       762

             accuracy                           0.83      1500
            macro avg       0.83      0.83      0.83      1500
         weighted avg       0.83      0.83      0.83      1500
```

iv) Gradient Boost Classifier

```
gradient_boost_pred = gradient_boost.predict(X_test)
print(confusion_matrix(y_test,gradient_boost_pred))
```

```
[[628 101]
 [106 665]]
```

```
print(classification_report(y_test,gradient_boost_pred))
```

```
              precision    recall  f1-score   support

           0       0.86      0.86      0.86       729
           1       0.87      0.86      0.87       771

    accuracy                           0.86      1500
   macro avg       0.86      0.86      0.86      1500
weighted avg       0.86      0.86      0.86      1500
```

3.3.6 Step 6: Model Deployment

This stage is out of the studies scope as illustrated with dash line in the figure 4 above. It involves deployment of the best overall model after the validation stage for solving the research

problem. In the case of this study, Decision tree was the best overall with an accuracy of 97% that proved stable after validation, but was followed closely by Random Forest with an accuracy of 97% which on validation went down to 93% as a result of high associated variance. Therefore, this study recommends Decision Tree for deployment in predicting student academic performance to arouse appropriate intervention (s) from any interested and concerned stakeholder.

3.4 Target Population

Also known as, universe population refers to a set of elements such as people, items and /or objects from which representative samples are obtained for the study process. (Mugenda &Mugenda, 2003) defines target population as the number of subjects, used by the researcher to generalize the results of the study. That is members of a given group targeted by the researcher for an investigation related to the study being carried out (Pole and Lampard 2020) or an entire aggregation of respondents that meet the designated set of criteria (Burns & Grove 1997). Kitui west constituency consists of two neighboring sub counties: Matinyani and Kitui west. Information obtained from sub-county education offices from the two sub counties indicate that there are 43 public secondary schools in the sub-county with a total, form 1 enrolment of 6,500 learners on average per year. The study targets a group of students admitted to public secondary schools within Kitui west constituency in the year 2017 up to the time they sat for their exit exam in the year 2020, from all 43 public secondary schools. This study therefore, adopts a census approach where all students that form the target populations from all public secondary schools in Matinyani and Kitui west sub counties were included. This provided sufficient historical data that was used for training supervised machine learning algorithms mentioned earlier in the process of developing the model.

3.5 Sampling and Sampling Procedure

Census method was adopted to gather secondary data that concerns students' academic performance from all targeted subjects in the 43 public secondary schools within the stated research period. Then a proportion of resultant dataset randomly selected, (70% - training sample) was used for training the models while the remaining (30% - test sample) for testing the models. Due to unavoidable circumstances emanating from transfers and drop out, data for 5,500 subjects was obtained out of the projected 6,500 subjects. This dataset was sufficient for the study and was used as illustrated earlier in this chapter.

3.6 Data collection procedure

This refers to a systematic way of gathering data that is relevant to the solution of the research problem (Burns & Grove, 1997). Upon being cleared by the School of Graduate Studies & Research (SGS) of KCA University and acquiring the necessary document (research orientation letter) that grants authority to proceed to the data collection phase, researcher visited schools within the research area, approached administrators and sought permission to collect data in their schools. The researcher then approached various department heads in the schools to assist in the extraction of relevant data from their repositories. As mentioned earlier, internal desk research approach was used where secondary data accumulated over time (historical) generated within the institution that's relevant to the study's objective was extracted to an excel document. During this exercise, guidance was drawn from the conceptual framework or the table of variables in the operationalization of variables. At this stage, the researcher ensured that only historical data that is relevant to the study's objectives and has accumulated within the research period was collected. The resultant dataset was found to be consistent with similar datasets from other sources.

3.7 Data Processing and analysis

Data analysis is a process that involves systematic organization and synthesis, categorizing, ordering, manipulating and summarizing the data and describing them in meaningful terms (Weisberg, 1989). Data extracted and loaded to excel document in the previous stage was subjected to data preprocessing techniques like dealing with missing values, data inconsistences, outliers and noisy data values as illustrated earlier in this chapter. Once data-cleaning process was over, the resultant cleaned dataset was subjected to data validation procedures such as data type check, range check, uniqueness check with respect to students' ID numbers, format check etc. Then the final refined dataset was converted to *.csv* document and loaded to a Data lake in Hadoop. This data was then queried into python using Pyspark where the rest of data processing and analysis took place.

FIGURE 11

Extracted Dataset Loaded Into Hadoop Data Lake.

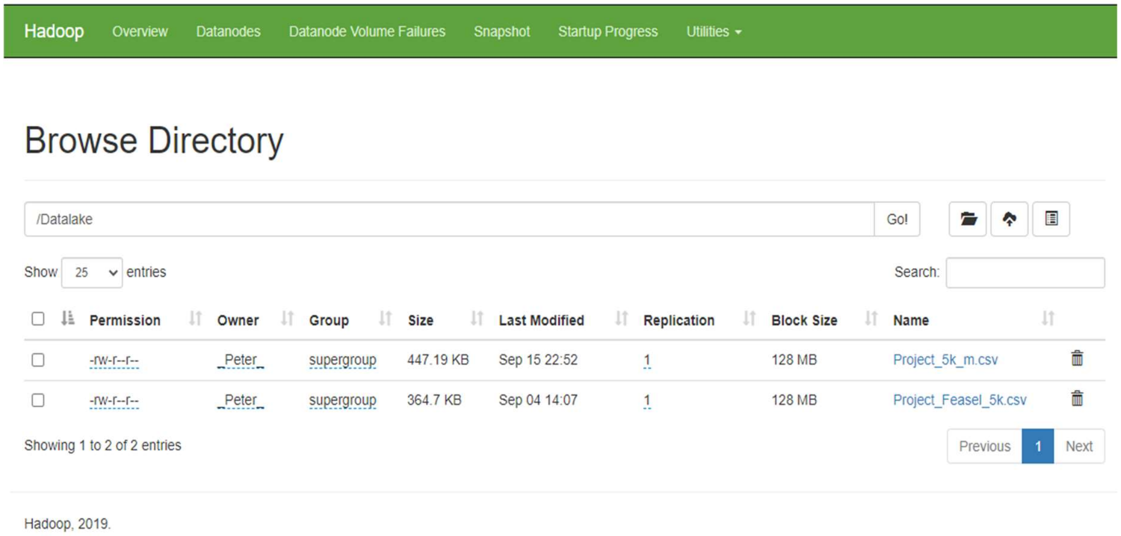


FIGURE 12

Data Loaded to Python from Hadoop Data Lake Using Pyspark.

```
In [88]: # 4.2 read spark data frame
         spark_data = spark.read.csv("hdfs://localhost:9000/Datalake/Project_5k_m.csv",header='true',inferSchema='true')
         spark_data.dtypes
         spark_data.show(5)

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+--
--+----+----+---+----+----+----+---+
|SAP|SST|SLV|TAS|TCM|TAB|SFC|TTN|MED|FED|PGE|PHM|FTP|FSZ|FIN|DFP|PLP|CSR|CSP|AGS|RGS|GDS|STS|STH|ATE|CWE|SDP|SEC|ABS|TCS|ASC|SB
F1|SBF2|SBF3|ENT|AGF1|AGF2|AGF3|SNO|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+--
--+----+----+---+----+----+----+---+
|  0|  4|  4|  1|  2|  2|  2|  2|  4|  4|  1|  1|  1|  1|  3|  2|  1|  1|  1|  3|  2|  1|  2|  3|  1|  1|  4|  3|  1|  3|  1|
10|   7|   7| 11|   7|   7|   6|  1|
|  0|  1|  2|  3|  3|  1|  3|  3|  3|  5|  1|  1|  1|  1|  3|  2|  3|  1|  1|  2|  2|  2|  3|  3|  1|  2|  4|  3|  1|  2|  1|
12|   8|   8|  9|   5|   5|   6|  2|
|  1|  4|  4|  3|  3|  1|  2|  2|  3|  4|  1|  2|  3|  2|  2|  1|  1|  2|  2|  2|  2|  1|  2|  2|  1|  1|  3|  3|  3|  2|  2|
11|   8|   8|  7|  10|   8|   3|  3|
|  1|  1|  2|  2|  3|  1|  3|  2|  3|  3|  1|  1|  1|  1|  1|  1|  1|  1|  3|  2|  2|  2|  3|  3|  1|  2|  1|  3|  1|  2|  1|
12|   8|   8|  3|   8|   7|   7|  4|
|  0|  2|  2|  3|  3|  2|  3|  2|  3|  3|  1|  1|  4|  2|  2|  1|  1|  1|  1|  2|  2|  1|  2|  2|  1|  1|  3|  3|  1|  2|  1|
```

The final technique applied at the data preprocessing stage was feature selection, where any of the feature selection technique discussed in the previous chapter was used to produce an optimal subset of features. This provided answer to the first research question. The optimal subset of features was used with machine learning algorithms identified from empirical review to have previously registered good performance for model development. The model that registered the highest accuracy was recommended for predicting student academic performance. This provided answer to the second research question. These predictions can

arouse initiation of strategic interventions that can reduce, if not eliminate dismal performance among the students in public schools in Kitui west constituency. Apart from accuracy, other evaluation metrics such as Recall, Precision, ROC (Receiver Operating Characteristics) etc. were used which provides answer to the third and the last research question. The chosen model can be deployed in any interested secondary school offering 8:4:4 system of education in either Kitui west constituency or other counties within the country, which is out of scope of this study.

CHAPTER FOUR

DATA ANALYSIS, FINDINGS AND DISCUSSION

4.1 Introduction

Results obtained from the study area presented in this chapter. The main objective of the study being, to develop a model for predicting students' academic performance in public secondary schools in Kitui west constituency. Diagnostic research design under case history method that is concerned with any significant historical information concerning the research problem (case) was used to examine the underlying cause (factors/variables) of the problem. Therefore, this design led to achievement of specific objective one, through the empirical review contacted in the previous chapter. Historical data extracted from public secondary schools in Kitui west constituency was used to develop a supervised machine learning classification model for predicting students' academic performance. The developed models were evaluated on various metrics to establish their effectiveness in addressing the research problem.

4.2 Demographics of collected data.

This represents distribution of the subjects in terms of; age, gender, pass and fail groups etc. For the dataset used in this study, these distributions were as displayed in the visualizations below.

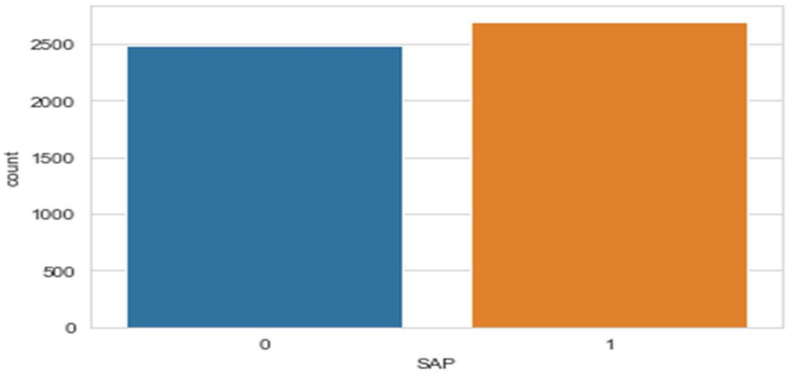
FIGURE 13

Proportion of Pass (1), Fail (0) – (a) and the Distribution across Gender – (b)

(a)

```
In [35]: sns.set_style('whitegrid')
         sns.countplot(x='SAP',data=Data)

Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x28b93b4dc08>
```



(b)

```
In [34]: sns.set_style('whitegrid')
         sns.countplot(x='SAP',hue='GDS',data=Data)

Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x28b92b04ec8>
```

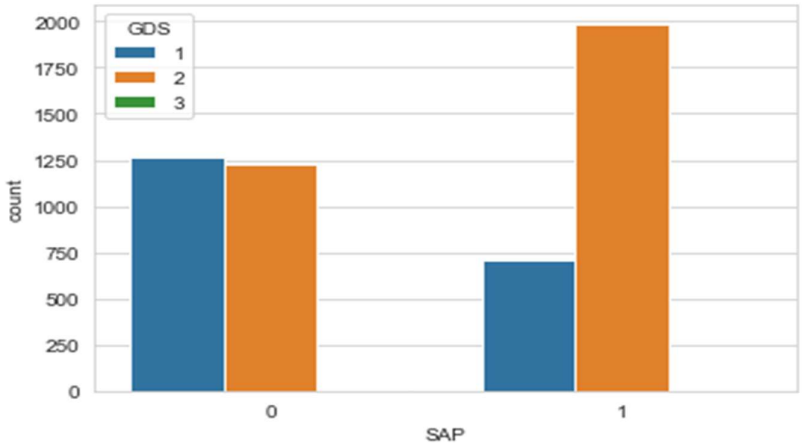


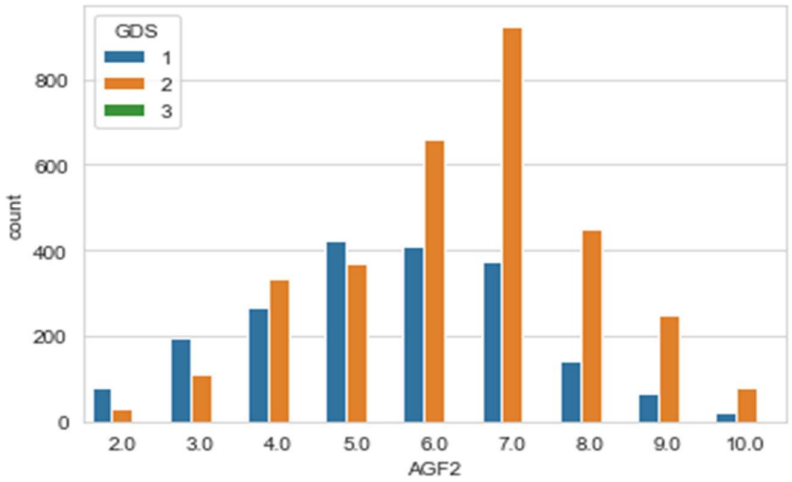
Fig 13 (a) shows that the dataset is slightly imbalanced towards the pass group (1), (b) shows distribution of students' performance across gender, where almost equal number of boys and girls failed while majority of students who passed were boys (2). The third category (3) are insignificant.

FIGURE 14

Distribution of Average Grade in Form 2 across Gender.

```
In [39]: sns.set_style('whitegrid')
         sns.countplot(x='AGF2',hue='GDS',data=Data)

Out[39]: <matplotlib.axes._subplots.AxesSubplot at 0x28b93cfcd48>
```



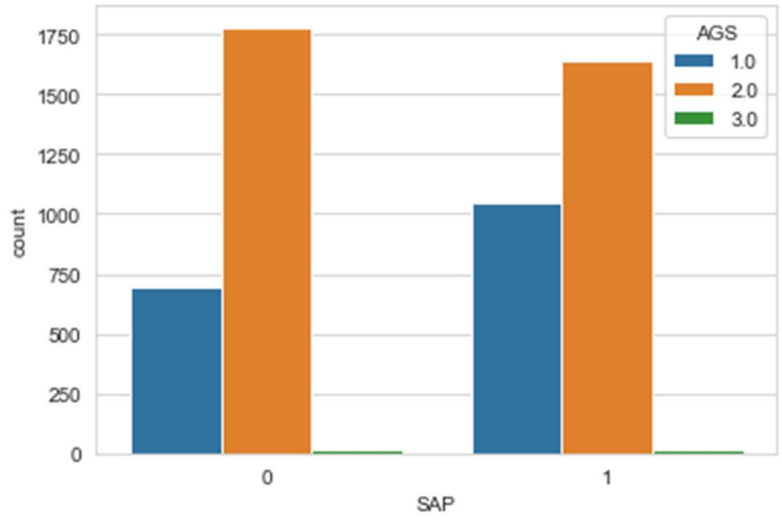
It is clear that girls (1) performance was positively skewed while that of boys (2), was negatively skewed. More boys attained pass mark compared to girls with mode grade being a B – (7 points) for boys and C- (5 points) for girls. Average grade in form two is the highest ranked attribute with a rank index of (0.638677). As mentioned earlier category (3 others) is insignificant.

FIGURE 15

Distribution across Students' Age

```
In [42]: sns.set_style('whitegrid')
         sns.countplot(x='SAP',hue='AGS',data=df)

Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x28b93ef21c8>
```



From the figure 12 above, most students were above 18 years of age (1) by the time, they sat for their exit exam. More students passed under the category 17 years and below (2) as compared to the other categories. Category (3) 'Not sure', refers to students whose records could not be traced at the time this research was being carried out.

4.3 Objective one results.

In order to achieve the objective one of this study, the researcher employed diagnostic research design, case history method where information in the history concerning the research problem was obtained through empirical review. From this review, features that influence students' academic performance were identified. Then using mRMR, feature selection method

mentioned in the previous chapter, optimal subset of features was selected. This subset consisted of features that are more predictive to the target class and represented the main features that influence students' academic performance. Pearson correlation coefficient (Pearson r) was used in this study to examine the strength and direction of independent variables influence on the dependent variable. This method best suite the study because, it evaluates the linear relationship between variables based on raw data. The figure below shows feature ranks among the optimal subset of features.

FIGURE 16

Feature Ranking Using Pearson Correlation Method (Pearson's R).

```
In [98]: Correlation=Data.corr(method='pearson')
         print("\nCorrelation:\n",Correlation)

Correlation:
           AGF2      SBF3       ASC       SFC       GDS       SDP       ATE  \
AGF2   1.000000  0.070654 -0.122466  0.070543  0.246738 -0.044031 -0.162292
SBF3   0.070654  1.000000 -0.051900  0.079510  0.120099 -0.004642  0.002289
ASC   -0.122466 -0.051900  1.000000 -0.021721 -0.045651  0.144900  0.064728
SFC    0.070543  0.079510 -0.021721  1.000000  0.100276  0.008145 -0.162577
GDS    0.246738  0.120099 -0.045651  0.100276  1.000000  0.001524  0.007739
SDP   -0.044031 -0.004642  0.144900  0.008145  0.001524  1.000000  0.007570
ATE   -0.162292  0.002289  0.064728 -0.162577  0.007739  0.007570  1.000000
SLV   -0.363408 -0.184408 -0.064812 -0.325430 -0.355263 -0.085132  0.133416
AGF1   0.744715  0.020560 -0.053996  0.039409  0.127251 -0.035440 -0.130665
CWE    0.140259  0.059727 -0.104207  0.145248  0.122338 -0.066026 -0.085136
SAP    0.638677  0.128417 -0.163859  0.137777  0.247329 -0.108669 -0.159274

           SLV      AGF1       CWE       SAP
AGF2 -0.363408  0.744715  0.140259  0.638677
SBF3 -0.184408  0.020560  0.059727  0.128417
ASC  -0.064812 -0.053996 -0.104207 -0.163859
SFC  -0.325430  0.039409  0.145248  0.137777
GDS  -0.355263  0.127251  0.122338  0.247329
SDP  -0.085132 -0.035440 -0.066026 -0.108669
ATE   0.133416 -0.130665 -0.085136 -0.159274
SLV   1.000000 -0.261014 -0.185884 -0.370668
AGF1 -0.261014  1.000000  0.095895  0.471674
CWE  -0.185884  0.095895  1.000000  0.160238
SAP  -0.370668  0.471674  0.160238  1.000000
```

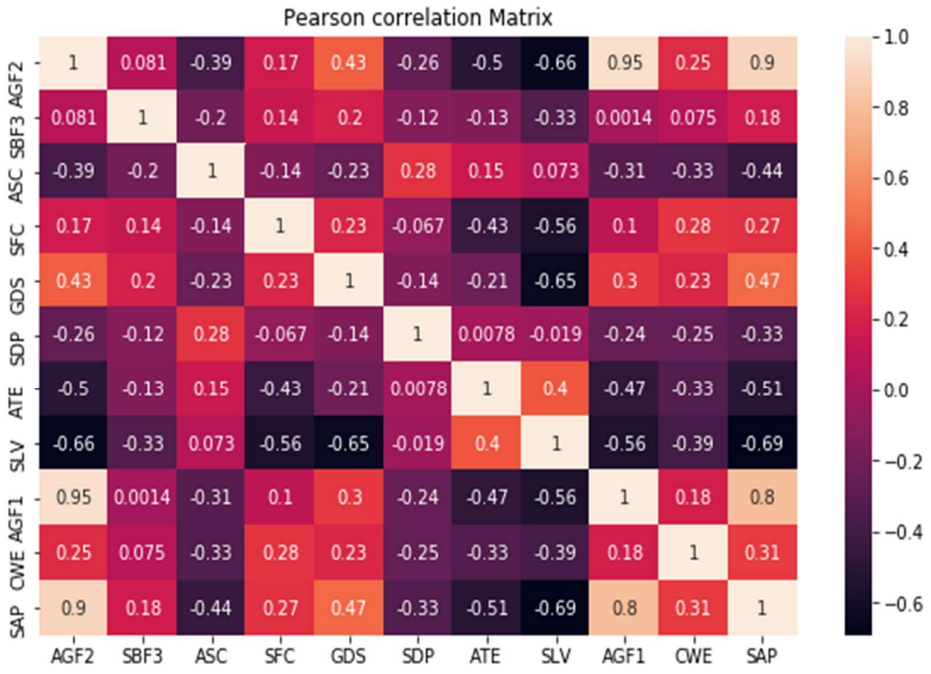
The correlation matrix below, presents a summary of relationships among variables used in this study in terms of strength and direction.

FIGURE 17

Correlation Matrix for Features Used in the Study.

```
In [99]: plt.figure(figsize = (10,6))
         plt.title('Pearson correlation Matrix')
         sns.heatmap(Correlation.corr(),annot=True, fmt='.2g')

Out[99]: <matplotlib.axes._subplots.AxesSubplot at 0x13afaf74b08>
```



This matrix represents relationships strength and their direction among variables. Values towards one (1) represents a strong relationship while those towards zero (0) represents weak relationship. Some features were not included in the optimal subset of features even after being ranked high by the Pearson method because, mRMR employs a two-step selection strategy strictly implemented in the order depicted by its name, minimum Redundancy Maximum Relevance. Results of the feature selection process using mRMR and the correlation matrix are as displayed in the table 2 below.

TABLE 5

Ranking for both Selected and Dropped Features.

Serial No	Feature/ Variable	Rank Index	Selected	Dropped

1	Students Academic Performance (SAP)	1.0000	Target	Target
2	School Type (SST)	-0.407466		✓
3	School Level (SLV)	-0.370668	✓	
4	Teacher's Attitude towards Students (TAS)	0.158721		✓
5	Teacher's Commitment (TCM)	0.094344		✓
6	Teacher's Absenteeism (TAB)	-0.053124		✓
7	Trained teachers (TTN)	0.146641		✓
8	School facilities (SFC)	0.137777	✓	
9	Family type (FTP)	-0.064319		✓
10	Mother's education (MED)	0.143582		✓
11	Father's education (FED)	0.170154		✓
12	Parent/Guardian encouragement (PGE)	0.067202		✓
13	Parental level of participation (PLP)	-0.071947		✓
14	Parents state of harmony (PHM)	-0.018771		✓
15	Family income (FIN)	0.057093		✓
16	Family size (FSZ)	0.033718		✓
17	Difficulties in fees payment (DFP)	0.138025		✓
18	Community security (CSR)	-0.084050		✓
19	Community support (CSP)	-0.078334		✓
20	Student discipline (SDP)	-0.108669	✓	
21	Student English Command (SEC)	0.009754		✓
22	Days Absent (ABS)	-0.146095		✓
23	Consulting teachers (TCS)	0.201909		✓
24	Completing Assignment (ASC)	-0.163859	✓	
25	Student age (AGS)	-0.112568		✓
26	Student religion (RGS)	-0.000736		✓
27	Student gender (GDS)	0.247329	✓	
28	Student status (STS)	0.188279		✓
29	Study time at Home (STH)	0.072724		✓
30	Attitude towards education (ATE)	-0.159274	✓	
31	Challenges with exams (CWE)	0.160238	✓	
32	Subjects in Form One (SBF1)	0.187638		✓
33	Subjects in Form Two (SBF2)	0.173544		✓
34	Subjects in Form Three (SBF3)	0.128417	✓	
35	KCPE Points (Entry Status) (ENT)	0.002895		✓
36	Form 1 average Points (AGF1)	0.471674	✓	
37	Form 2 average Points (AGF2)	0.638677	✓	
38	Form 3 average Points (AGF3)	-0.012495		✓

Therefore, the optimal subset of features constitutes the main features that influence students' academic performance in public secondary schools in Kitui west constituency. These features include; School level (SLV), School facilities (SFC), Student discipline (SDP), Completing assignments (ASC), Student gender (GDS), Attitude towards education (ATE), Challenges with exams (CWE), Subjects in form three (SBF3), Average points in form 1(AGF1) and the last but not the least the Average points in form 2 (AGF2). Compared to the other features, this subset consisted of features that have maximum relevance and minimum redundancy to the target class; hence, they were used in the model development stage.

4.4 Objective Two Results

Models were developed then trained using the training sample with four supervised machine learning algorithm as discussed earlier at the model development stage in the previous chapter. The four algorithms used include; Decision tree classifier, Random forest classifier, Gradient boost classifier and Deep Neural Network classifier. The figure below represents the visualizations of the developed and trained models

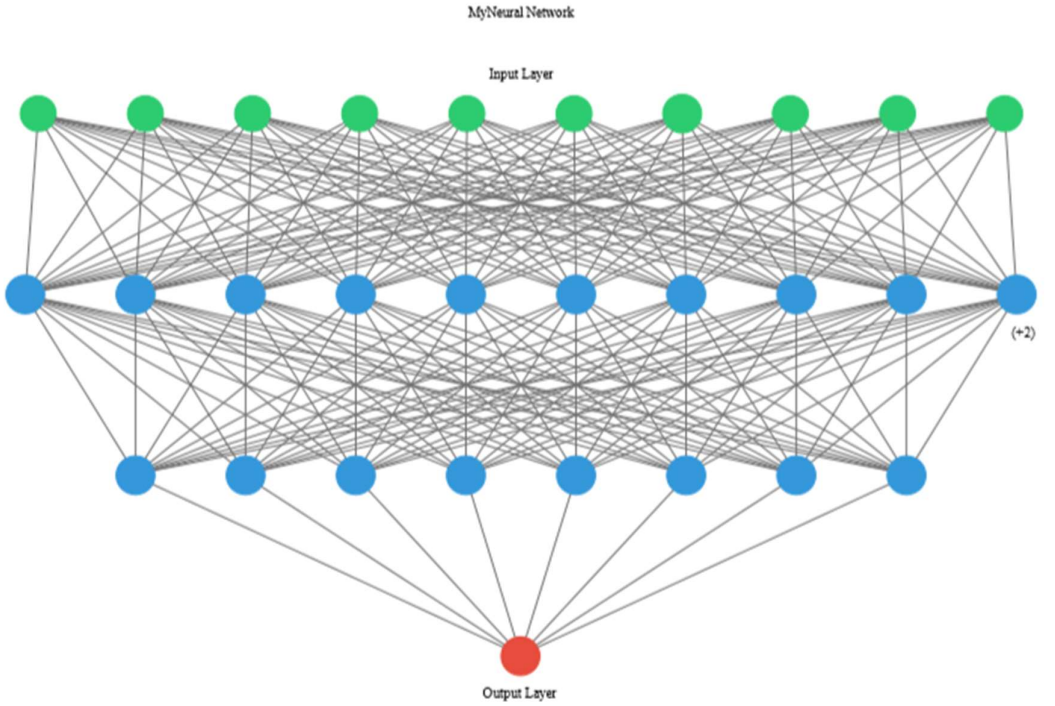
FIGURE 18

Visualizations of Decision Tree and Deep Neural Network Models

```
dot_data = StringIO()
export_graphviz(dtree, out_file=dot_data,feature_names=features,filled=True,rounded=True, max_depth=3)

graph = pydot.graph_from_dot_data(dot_data.getvalue())
Image(graph[0].create_png())
```

```
import os
os.environ['PATH'] = os.environ['PATH']+';'+os.environ['CONDA_PREFIX']+r"\Library\bin\graphviz"
ann_viz(model, view=True, filename="C:/pyfilesdata/network.gv", title="MyNeural Network")
```



4.5 Objective Three Results

Developed models were evaluated on accuracy, Recall and Precision using the test sample. Variation on performance was registered across the developed models as illustrated in fig. 14 below. Decision tree was the best overall with an accuracy of 97% on both evaluation and validation, followed closely by Random forest with an accuracy of 97% on evaluation, which dropped, to 93% on validation due to associated variance. Figure 14 below shows results of the developed models and their interpretation thereafter. A confusion matrix provides a measure of reliability and accuracy of the model. i.e.

		Predicted	
		Fail	**Pass**
Actual	**Fail**	**True Fail (High Intervention)**	False Pass (Type I error)
	Pass	False Fail (Type II error)	**True Pass (Low Intervention)**

In order to evaluate how effective, the developed and trained models are, predicted values must be compared with the actual values. A confusion matrix facilitates such exercise by displaying the predicted values against the actual values for each of the developed models as shown below. Various evaluation metrics can be computed from the confusion matrix, for example in this study, the following metrics were used to compare models performance.

i. Accuracy (A) – which is the ratio of correctly predicted cases either (positives or negatives) to the total number of cases. $Accuracy = \frac{TF+TP}{TP+FP+TF+FF}$ (Powers, 2011)

- Accuracy does not indicate how cases in a minority class are classified especially in an imbalanced dataset.

ii. Two more other metrics were used in this study, which include; Precision and Recall. Which both seek to compensate the weakness of Accuracy by presenting the ratios of successfully predicted positives (ratio of accurately predicted positives to all cases predicted as positive) and negatives (ratio of correctly predicted negatives to all cases predicted as negative) respectively.

- $Precison\ (P) = \frac{TP}{TP+}$ $Recall = \frac{TF}{TF+F} = (Sensitivity)$

iii. The last but not the least metric considered in this study is the F – measure. This metric takes into account both precision and recall as defined by (Fawcett, 2005) $F = 2.\frac{Precison\ .\ Recall}{Pecision+R}$. This metric determines the model's effectiveness through combining Precision and recall to attain a balanced value.

FIGURE 19

Performance Evaluation for the Developed Models

1. Decision tree (DT)

```
print("Accuracy on training set: {:.3f}".format(dtree.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(dtree.score(X_test, y_test)))
```

```
Accuracy on training set: 0.974
Accuracy on test set: 0.966
```

```
predictions = dtree.predict(X_test)
from sklearn.metrics import classification_report,confusion_matrix
print(classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

           0       0.98      0.95      0.96       729
           1       0.95      0.98      0.97       771

    accuracy                           0.97      1500
   macro avg       0.97      0.97      0.97      1500
weighted avg       0.97      0.97      0.97      1500
```

```
print(confusion_matrix(y_test,predictions))
```

```
[[690  39]
 [ 12 759]]
```

		Predicted	
		Fail	**Pass**
Actual	**Fail**	**690**	39 (Type I error)
	Pass	12 (Type II error)	759

2. Random Forest (RF)

```
print("Accuracy on training set: {:.3f}".format(rfc.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(rfc.score(X_test, y_test)))
```

```
Accuracy on training set: 0.974
Accuracy on test set: 0.966
```

```
rfc_pred = rfc.predict(X_test)
print(confusion_matrix(y_test,rfc_pred))
```

```
[[688  41]
 [ 10 761]]
```

```
print(classification_report(y_test,rfc_pred))
```

```
              precision    recall  f1-score   support

           0       0.99      0.94      0.96       729
           1       0.95      0.99      0.97       771

    accuracy                           0.97      1500
   macro avg       0.97      0.97      0.97      1500
weighted avg       0.97      0.97      0.97      1500
```

		Predicted	
		Fail	**Pass**
Actual	**Fail**	**688**	41 (Type I error)
	Pass	10 (Type II error)	761

3. Gradient Boost(GB)

```
gradient_boost_pred = gradient_boost.predict(X_test)
print(confusion_matrix(y_test,gradient_boost_pred))
```

```
[[628 101]
 [106 665]]
```

```
print(classification_report(y_test,gradient_boost_pred))
```

```
              precision    recall  f1-score   support

           0       0.86      0.86      0.86       729
           1       0.87      0.86      0.87       771

    accuracy                           0.86      1500
   macro avg       0.86      0.86      0.86      1500
weighted avg       0.86      0.86      0.86      1500
```

		Predicted	
		Fail	**Pass**
Actual	**Fail**	**628**	101 (Type I error)
	Pass	106 (Type II error)	665

4. Deep Neural Network

```
In [78]: #Evalaute the Trained Model with test data
         _, accuracy = model.evaluate(X_Test_sample, Y_Test_sample)
         print('Accuracy: %.2f' % (accuracy*100))

         1500/1500 [==============================] - 0s 183us/step
         Accuracy: 82.87

In [79]: from sklearn.metrics import classification_report,confusion_matrix
         print(confusion_matrix(Y_Test_sample,predictions))

         [[590 148]
          [109 653]]

In [80]: print(classification_report(Y_Test_sample,predictions))

                       precision    recall  f1-score   support

                  0.0       0.84      0.80      0.82       738
                  1.0       0.82      0.86      0.84       762

             accuracy                           0.83      1500
            macro avg       0.83      0.83      0.83      1500
         weighted avg       0.83      0.83      0.83      1500
```

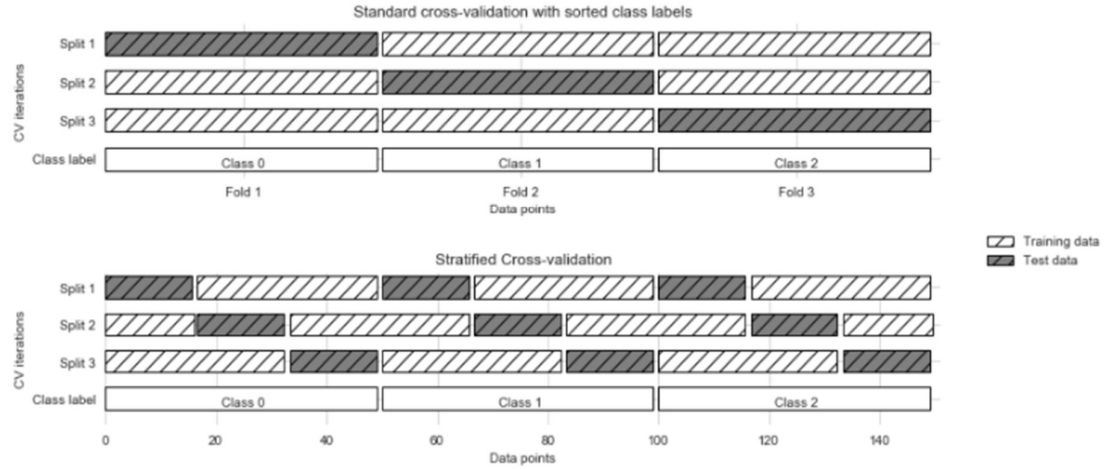
		Predicted	
		Fail	**Pass**
Actual	**Fail**	**590**	148 (Type I error)
	Pass	109 (Type II error)	653

As illustrated earlier in this study (4.2 Demographics of collected data**),** the dataset used in this study was slightly imbalanced, hence stratified k-fold cross validation was the best that suited the study. Under this validation strategy, the dataset is split in to a given number of fold i.e. K, in such a manner that the proportions between classes in each fold is the same as they are in the whole dataset. This ensures a more reliable estimate of the classifier's performance on generalization. The figure below presents a comparison between, stratified k-fold cross validation and the standard cross validation.

FIGURE 20

A Comparison between Standard Cross Validation and Stratified K-Fold Cross Validation.

```
mglearn.plots.plot_stratified_cross_validation()
```



Stratified k-fold cross validation was carried on developed models with Decision tree classifier and Random forest classifier. Results were as illustrated in the figure below.

FIGURE 21

Validation of the Decision Tree and Random Forest Based Models

```
from sklearn.model_selection import StratifiedKFold
kfold = StratifiedKFold(n_splits=10)
scores = cross_val_score(dtree, X, y, cv=kfold)
print("Cross-validation scores:\n{}".format(scores))
print("Mean accuracy: {:.2f}".format(scores.mean()))
```

```
Cross-validation scores:
[0.972 0.964 0.956 0.968 0.964 0.974 0.972 0.97  0.956 0.964]
Mean accuracy: 0.97
```

```
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import cross_val_score
kfold = StratifiedKFold(n_splits=10)
scores = cross_val_score(rfc, X, y, cv=kfold)
print("Cross-validation scores:\n{}".format(scores))
print("Mean accuracy: {:.2f}".format(scores.mean()))
```

```
Cross-validation scores:
[0.96 0.89 0.92 0.93 0.97 0.91 0.97 0.9  0.94 0.9 ]
Mean accuracy: 0.93
```

Although both models had the same accuracy (97%), the variance on accuracy registered by ten Random forest models under stratified cross validation strategy had higher variance compared to corresponding decision tree models. Thus, the performance of decision tree model was more stable and thus reliable, hence recommended for implementation for predicting students' academic performance.

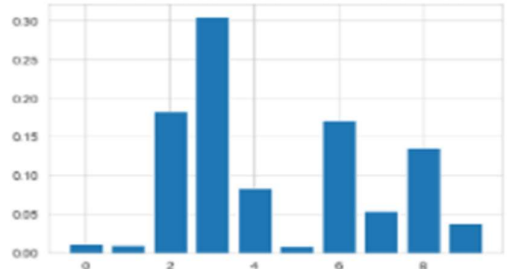
The last stage in this process of model development is extraction of feature importance from the model selected for deployment based on evaluation and validation results. A fitted linear machine learning based model uses weighted sum of the input values (features) for prediction. The weighted sum is composed of a set of coefficients obtained by the fitted model that vary across the input values (features). These coefficients can be extracted to directly represent feature importance scores (Brownlee, J. 2020). The extracted feature importance scores, represents the contribution of each feature used in the model development towards the model's performance on generalization. For the Decision Tree model, which this study recommended for deployment based on the results registered on evaluation and validation, feature importance scored were extracted and visualized using a bar graph as shown in the figure below.

FIGURE 22

Feature Importance Scores on the Developed Model

```
from sklearn.datasets import make_classification
from sklearn.tree import DecisionTreeClassifier
from matplotlib import pyplot
# define dataset
X, y = make_classification(n_samples=1000, n_features=10, n_informative=5, n_redundant=5, random_state=1)
# define the model
model = DecisionTreeClassifier()
# fit the model
model.fit(X, y)
# get importance
importance = model.feature_importances_
# summarize feature importance
for i,v in enumerate(importance):
    print('Feature: %0d, Score: %.5f' % (i,v))
# plot feature importance
pyplot.bar([x for x in range(len(importance))], importance)
pyplot.show()
```

```
Feature: 0, Score: 0.01143
Feature: 1, Score: 0.00976
Feature: 2, Score: 0.18194
Feature: 3, Score: 0.30545
Feature: 4, Score: 0.08390
Feature: 5, Score: 0.00833
Feature: 6, Score: 0.17130
Feature: 7, Score: 0.05381
Feature: 8, Score: 0.13597
Feature: 9, Score: 0.03810
```



Features were indexed in the same order they were fed in the model i.e. AGF2, SBF3, ASC, SFC, GDS, SDP, ATE, SLV, AGF1, CWE. It is clear from fig. 20 above that feature 3 (SFC – school facilities), feature 2 (ASC – completing assignments on time), feature 6 (ATE – attitude towards education) and feature 8 (AGF1 – average point in form 1) contributed significantly to the weighted sum and hence the model's performance on generalization. The more the school is equipped in terms of teaching/learning facilities the lower the probability of a student admitted in such school to register dismal performance. In Kenya, based on school facilities public secondary schools are categorized in to National, Extra county, County and Sub-county schools. Attempting all assignments given and completing them on time, enhance learning and hence reduces dismal performance among learners. Vision, goals and target determines the internal driving force towards learning. Students who are committed to their vision in life through well-set goals and targets always perform well in their exams. Thus, the attitude a student has towards learning significantly influence their academic performance. Form one being the entry class in secondary schools; performance registered by the end of that

year determines students' academic performance in the other classes. A good performance at this level implies that the students has understood all pre-requisites concepts for the later content which enhances the probability of such a student performing well in the higher levels. The remaining six feature did not contribute much, thus education administrators should concentrate more on the four features described above to address dismal performance in their institutions.

4.6. Discussion of Results

The finding of this research were meant to arouse initiation of intervention from various stakeholders that will be aimed at reducing dismal performance among learners in public secondary schools. Four models were developed and trained with training sample and later evaluated on various metrics with the test sample to determine their effectiveness in solving the research problem. Comparison among the models indicate that Decision tree and Random Forest classifiers performed the best overall as presented in the table 3 below. However, on validation in which stratified K-fold cross validation was used, Decision tree classifier's performance proved far much stable with an accuracy of 97% on average at K = 10, compared to Random Forest classifier's accuracy of 93% on average. This implies that for the 10 iteration conducted, the variance on accuracy obtained was very low for a Decision tree classifier compared to the Random Forest classifier. This indicated that on predicting students' academic performance, Decision tree performance was more stable than the other three classifiers considered in this study hence recommended for deployment.

It is worthy to note that single Decision tree classifiers are prone to overfitting especially if pre-pruning has not been done. Pre-pruning is supposed to be done once the classifiers begins to over fit i.e. the accuracy on training set is equal to 100%. For the results registered in this study, Decision tree classifier had not begun to over fit hence the results are reliable i.e. accuracy on training set of 0.974 verses accuracy on test set of 0.966.

The results obtained were in agreement with those registered by earlier models with some remarkable improvement across the evaluation metrics e.g. accuracy as illustrated in the table below. This is due to the research designs adopted, method used for data collection and feature selection method employed. Comparison of the models results with earlier models was as shown in table 3 below.

TABLE 6

Comparison of Developed Models Results with Earlier Models.

Authors	Level/ Country	Sample size	Feature Selection	Optimal subset	ML Model	Performance (Accuracy)
Current study	Kenya (Public Secondary Schools)	5000	mRMR	10/38 features	**Decision Tree (DT)**	**97.0%**
					Random Forest (RF)	97.0%
					Gradient Boost (GB)	86.0%
					Deep Neural Network (DNN)	83.0%
Saa, A. et. al, 2019	UAE (University)	56,000	Information Gain	15/34 features	Decision Tree (DT)	68. 49
					Random Forest (RF)	75.52%
					Gradient Boost (GB)	72.45%
Yousafzai, B. et. al, 2020	Pakistan (Secondary)	80,000	genetic algorithm (GA)	29/106 features	Decision Tree (DT)	96.64%
Amrieh, E. et al 2016	University of Jordan	500	Empirical review	16 features	Decision Tree (DT)	75.8%
					Artificial Neural Network (ANN)	80.0%
					Random Forest (RF)	75.6%
Ha, D. T. et al, 2020	University of Vietnam	561	Simple Pearson correlation coefficient	42 features	Decision Tree (DT)	73.48%
					Random Forest (RF)	80.7%

4.7 Summary of results

The chapter presented the finding of the current study from the data analysis phase. Ten main factors were identified using mRMR, feature selection method that are more predictive to the target class based on minimum redundancy and maximum relevance. These features constituted the optimal subset of features that were used to develop four models discussed above; Decision Tree, Random Forest, Gradient Boost and Deep Neural Networks. Developed models were evaluated using four metrics extracted from respective confusion matrices.

Decision Tree classifier registered the best overall accuracy hence recommended for deployment in predicting students' academic performance in public secondary schools in Kitui west constituency.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter presents conclusions inferred from the research findings and recommendations to offer guidance to researchers interested in this topic. Contributions of the study are also highlighted in this chapter.

5.2 Conclusions

From the research findings, it can be deduced that, ten factors significantly influence students' academic performance in public secondary schools in Kitui west constituency. This optimal subset of features (10 features selected) represents features with minimal redundancy within and maximum relevance with the target class. They include; Form 2 average points (AGF2), Subjects in Form Three (SBF3), Completing Assignment (ASC), School facilities (SFC), Student gender(GDS), Student discipline (SDP), Attitude towards education(ATE), School level (SLV), Form 1 average points (AGF1) and Challenges with exams (CWE).

The relationship and direction of these factors within and with the target class was determined using Pearson correlation matrix that shows the variable's influence on the target class. For example, Form two average grade ranked the best among the optimal subset of features with rank index of 0.638677 and on the other side, school level with a rank index -0.370688. From this kind of information, the insight to educational administrators is that they should initiate strategies to reduce attitude the school level has on learners as well as promoting aspects that will lead to an improvement of learners' performance in form two.

Four features were found to contribute significantly towards the recommended Decision tree model generalization performance. These include; feature 3 (SFC – school facilities), feature 2 (ASC – completing assignments on time), feature 6 (ATE – attitude towards education) and feature 8 (AGF1 – average points in form 1). Hence, educational stakeholders are encouraged to focus on the four features during their initiation of strategic intervention to address dismal performance in their schools.

Among the four developed models, decision tree registered the best performance overall with an accuracy of 97% both on evaluation and validation for predicting student academic performance. Random forest followed closely with an accuracy of 97%, which

dropped to 93% on validation due to associated variance. Thus, Decision tree was declared the best machine-learning algorithm for developing the predictive model.

The strengths of the recommended model include; model's results are more stable hence reliable since it is not affected by scale of the dataset and it can easily be visualized especially for small trees or up to a certain depth as in the case in this study. This enhances the model understanding by even the non-experts. The major weakness of Decision Tree models is that they are prone to overfitting even with pre-pruning. Overfitting should always be tested by checking whether the accuracy on the training set has hit 100%, this is the time the model begins to over fit.

5.3 Contributions of the study

The study has contributed both empirically and theoretically towards the research problem. On the empirical contribution, the study has clearly demonstrated how feature selection methods can improve the efficiency and accuracy of the model. The study has recommended a model that can identify high intervention students with an accuracy of 97%, who need measures to rescue them from registering dismal performance prior to their main exam. This will guide the initiation of strategic interventions meant to address the issue of dismal performance among public secondary schools in Kitui west constituency. By using data generated during the normal operations of the institutions with data mining techniques, this study has demonstrated that a predictive model can be developed that could be useful in facilitating decision making in various learning institutions. From the recommended model feature, importance scores extracted revealed that four features contributed significantly towards the models prediction, this can provide insights to education administrators especially in an environment of limited resources on which variables of performance to respond to first. On theoretical contributions, the study has contributed towards learning practice by illuminating aspects that concern student academic performance prediction i.e. main features (optimal subset of features) that influence student's academic performance, feature selection methods, machine-learning methods in EDM etc. The findings of this study combined with earlier studies findings in the topic and domain knowledge can provide guidance on strategies to minimize dismal performance in various public secondary schools in Kitui west constituency.

5.4 Recommendations for Future Research

This study was limited to Kitui west constituency; future studies on the topic should focus on other constituencies and compare the results with the current study. The study used only secondary data obtained through desk research methodology from public secondary schools within the target constituency, future studies to appropriately combine both secondary and primary data and compare their results with the current study. In this study python was used as the main data analysis tool, future studies should employ other data analysis tools like WEKA, Hub Spot, Rapid miner, Tableau Public etc. and compare the results. Only four machine-learning algorithms were used in this study for models development, future studies to use different algorithm among many at their disposal to develop similar models and compare the results. Factors so far not captured in this study like syllabus coverage, subject teacher mastery of content, administrative strategies employed etc. to be included in future studies on this topic.

REFERENCES

Alnoukari, M., & El Sheikh, A. (2012). Knowledge discovery process models: from traditional to agile modeling. In *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications* (pp. 72-100). IGI Global.

Al-Rahmi, A. M. (2020). Constructivism Theory: The Factors Affecting Students' Academic Performance in Higher Education.

Amra, I. A. A., & Maghari, A. Y. (2017, May). Students' performance prediction using KNN and Naïve Bayesian. In 2017 *8th International Conference on Information Technology (ICIT)* (pp. 909-913). IEEE.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*(8), 119-136.

Ashraf, A., Answer, S., & Khan, M. G. (2018). A Comparative study of predicting student's performance by use of data mining techniques. *American Scientific Research Journal for Engineering, Technology, and Sciences* (ASRJETS), 44(1), 122-136.

Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education*, 7(3), 112-118.

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in higher education*, *12*(2), 155-187.

Billah, M., & Waheed, S. (2020). Minimum redundancy maximum relevance (mRMR) based feature selection from endoscopic images for automatic gastrointestinal polyp detection. *Multimedia Tools and Applications*, 79(33), 23633-23643.

Bin Mat, U., Buniyamin, N., Arsad, P. M., & Kassim, R. (2013, December). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. In *2013 IEEE 5th conference on engineering education (ICEED)* (pp. 126-130). IEEE.

Blau, P. M. (1962). Operationalizing a conceptual scheme: The universalism-particularism pattern variable. *American Sociological Review*, 159-169.

Bornstein, M. H., & Bradley, R. H. (2014). Socioeconomic status, parenting, and child development: Routledge, (147 – 189).

Brownlee, J. (2020). *Data preparation for machine learning: data cleaning feature selection, and data transforms in Python*. Machine Learning Mastery.

Burns, N., & Groves, K. (1997). *Practice of nursing research*. Philadelphia, PA: WB Saunders Company. Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of educational psychology, 104*(1), 32.

Cresswell, J. W. (2003). Qualitative, quantitative, and mixed methods approaches. *Research Design,* 3-26.

Echegaray-Calderon, O. A., & Barrios-Aranibar, D. (2015, October). Optimal selection of factors using Genetic Algorithms and Neural Networks for the prediction of students' academic performance. In 2015 *Latin America Congress on Computational Intelligence* (LA-CCI) (pp. 1-6). IEEE.

Evans, G.E., & Simkin, M.G. (1989). What best predicts computer proficiency? *Communications of the ACM*, *32*(11), 1322-1327.

Gajwani, J., & Chakraborty, P. (2021). Students' Performance Prediction Using Feature Selection and Supervised Machine Learning Algorithms. *In International Conference on Innovative Computing and Communications* (pp. 347-354). Springer, Singapore.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Ha, D. T., Loan, P. T. T., Giap, C. N., & Huong, N. T. L. (2020). An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques. *International Journal of Computer Science and Information Security (IJCSIS)*, *18*(3).

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T. ... & Liao, S. N. (2018, July). Predicting academic performance: a systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175-199).

Hirokawa, S. (2018, October). Key attribute for predicting student academic performance. In *Proceedings of the 10th International Conference on Education Technology and Computers* (pp. 308-313).

Jalota, C., & Agrawal, R. (2021). Feature Selection Algorithms and Student Academic Performance: A Study. In *International Conference on Innovative Computing and Communications* (pp. 317-328). Springer, Singapore.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Kiess, H. O. and Bloomquist, D. W. (1985): Psychological Research Methods: A Conceptual Approach. Boston: Allyn and Bacon.

Kieti, J. M. (2018). *An investigation into factors influencing students' academic performance in public secondary schools in Matungulu sub-county, Machakos County* (Doctoral dissertation).

Leavy, P. (2017). Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches (124 – 164)

Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. (2019). Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of educational computing research*, 57(2), 448-470

Madhavan, S. (2015). *Mastering Python for Data Science*. Packt Publishing Ltd.

Mark Brundrett (2011). The global challenge for primary schools: education in a world of 7 billion people, Education 3-13, 39:5, 451-453.

Mark Brundrett. (2014) Education for all: the challenges of achieving universal early childhood care and primary education. *Education 3-13* 42:3, pages 233-236.

Mgala, M. (2016). Investigating prediction modelling of academic performance for students in rural schools in Kenya.

Mgala, M., & Mbogho, A. (2014, November). Selecting relevant features for classifier optimization. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 211-222). Springer, Cham.

Mills, G. E., & Gay, L. R. (2019). *Educational research: Competencies for analysis and applications*. (Pp.84 -85). Pearson. One Lake Street, Upper Saddle River, New Jersey 07458.

Ministry of Devolution and Planning. (2013). Millennium Development Goals. Status report for Kenya 2013.

Montague, M., Reynolds, M. M., & Washburn, M. F. (1918). A Further Study of freshmen. *The American Journal of Psychology*, *29*(3), 327-330.

Moreira, J., de Leon Ferreira, A. C. P., & Horváth, T. (2019). *A general introduction to data analytics.* Wiley.

Motala, S., Ngandu, S., Mti, S., Arends, F., Winnaar, L., Khalema, E., & Martin, P. (2015). Millennium development goals: Country report 2015.

Mugenda, O. M. and Mugenda, A.G. (2003). Research methods: Quantitative and qualitative Approaches. Nairobi. Acts Press.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. "O'Reilly Media, Inc.".

Obadiah M., Kelvin O. & Raphael A. (2019, October). Towards Prediction of Students' Academic Performance in Secondary School Using Decision Trees. In *International Journal of Research and Innovation in Applied Science (IJRIAS) | Volume IV, Issue X, October 2019|ISSN 2454-6194.*

Ogwoka, T. M., Cheruiyot, W., & Okeyo, G. (2015). A Model for predicting Students 'Academic Performance using a Hybrid K-means and Decision tree Algorithms. *International Journal of Computer Applications Technology and Research,* 4(9), 693-697.

Orodho J. A. (2002). Techniques of Writing Research Proposals and Reports in Education and Social Sciences. Nairobi: Masola Publishers.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, *27*(8), 1226-1238.

Pole, C. J., & Lampard, R. (2002). Practical social investigation: Qualitative and quantitative methods in social research. Pearson Education.

Polit, DF, & Hungler, BP (1995). Fundamentals of nursing research. In *Fundamentals of Nursing Research* (pp. 391-391).

Punlumjeak, W., & Rachburee, N. (2015, October). A comparative study of feature selection techniques for classify student performance. In 2015 *7th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 425-429). IEEE.

Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students' performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies*, 24(6), 3577-3589.

Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. *arXiv preprint arXiv:0912.3924.*

Ramphele, F. (2018). *Predicting grade progression within the Limpopo Education System* (Master's thesis, Faculty of Science).

Rugutt, J. K., & Chemosit, C. C. (2005). A Study of Factors that Influence College Academic Achievement: A Structural Equation Modeling Approach. *Journal of Educational Research & Policy Studies,* 5(1), 66-90.

Saa, A. A., Al-Emran, M., & Shaalan, K. (2019, March). Mining student information system records to predict students' academic performance. In *International conference on advanced machine learning technologies and applications* (pp. 229-239). Springer, Cham.

Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019, March). Student performance prediction and classification using machine-learning algorithms. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 7-11).

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414-422.

Sharma, D., & Aggarwal, D. (2021). A Predictive Approach to Academic Performance Analysis of Students Based on Parental Influence. *In International Conference on Innovative Computing and Communications* (pp. 75-84). Springer, Singapore.

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254).

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, *1*(1), 64-85.

Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *How to test the validation of a questionnaire/survey in a research (August 10, 2016)*.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, *45*(1), 89-125.

Vercellis, C. (2009). *Business intelligence: data mining and optimization for decision-making* (pp. 1-18). New York: Wiley.

Verschuren, P., Doorewaard, H., & Mellion, M. (2010). *Designing a research project* (Vol. 2). The Hague: Eleven International Publishing, pg 160 – 201.

Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximizing Classification Accuracy. *Pertanika Journal of Science & Technology*, 26(1).

Wang, M., Zhou, S., & Yang, Z. (2019). A brief introduction to deep learning techniques. *Autom. Technol. Appl*, *38*(05), 51-57.

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (1989). *An introduction to survey research and data analysis*. Scott, Foresman & Co.

Yousafzai, B. K., Hayat, M., & Afzal, S. (2020). Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. *Education and Information Technologies,* 25(6), 4677-4697.

Zaffar, M., Hashmani, M. A., & Savita, K. S. (2017, November). Performance analysis of feature selection algorithm for educational data mining. In 2017 *IEEE Conference on Big Data and Analytics (ICBDA)* (pp. 7-12). IEEE.

Zaffar, M., Savita, K. S., Hashmani, M. A., & Rizvi, S. S. H. (2018). A study of feature selection algorithms for predicting students' academic performance. *Int. J. Adv. Comput. Sci. Appl*, 9(5), 541-549.

Zhao, Z., Anand, R., & Wang, M. (2019, October). Maximum relevance and minimum redundancy feature selection methods for a marketing machine-learning platform. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 442-452). IEEE.