MASTERS IN INFORMATION SYSTEM MANAGEMENT

A NEURAL NETWORK PREDICTION MODEL FOR DIPLOMA AND CERTIFICATE STUDENT'S PROGRESSION IN UNIVERSITIES

BY:

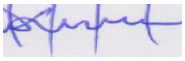
NGIGI STANLEY MUNGA - 19/07153

Supervisor:

Dr. Simon Mwendia

DECLARATION

This research project is my original work and has not been presented for a degree in any other University or any other award.

Signature: ……  … Date: 13/09/2021

Stanley Munga Ngigi

19/07153

I confirm that the work reported in this research project was carried out by the candidate under my supervision.

Signature …… Date ……

Dr. Simon Mwendia

Department of Information Management

School of Information Technology

KCA University

DEDICATION

I dedicate this research project to my wife, dad, and mum; Mercy, Gibson, and Cecilia Ngigi, for the support you have given me. God bless you.

ACKNOWLEDGEMENT

I sincerely thank the Almighty God for guiding me and providing the means for my postgraduate studies and the far He has brought me.

Secondly, I want to express my love and gratitude to my family; my parents, who have been with me emotionally, spiritually, and financially. Thank you for trusting and investing in my future. My siblings, classmates, and friends, thank you for the guidance and support.

Insincerity, I wish to thank Dr. Simon Mwendia (my supervisor) for his support, guidance, and leadership over the period. Also, I pass my gratitude to Dr. Lucy Mburu, Dr. Njenga for their positive contribution towards the understanding and development of this research project.

Table of Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

ANN Artificial Neural Network

ARIMA-Autoregressive Integrated Model

HELB –Higher Education Loans Board

SPR-Student progression rate

ST- Students progression

MO-Motivation

Demo- demographics

SO – Social and family

GLOSSARY

Students progression

This involves the transition from one level to another in a university i.e. from certificate to diploma and diploma to degree

Student progression rate

This refers to the number of students that progress in a given level and which level has a high number of student progressions in a given semester or in a given academic calendar.

Artificial neural networks

This is a Mathematical model that tries to imitate the behavior of the human brain which involves nonlinear relationships among different datasets that cannot always be fully identified using conventional analyses.

ABSTRACT

The most important priority of a private academic university is financial stability, which is determined by 100% student progression to the next level of study. Poor student advancement will result in the university's demise, as student progression is the primary source of revenue for a private university. The institution can plan adequately for the next semester and determine the number of workers required without undue stress as a result of the students' progress. Students in Kenyan colleges make predictions using linear forecasting models, which presume that data is linear. As a result, the progression of students based on linear models may be erroneous. Nonlinear models have been used to predict student outcomes with great effectiveness. As evidenced by the literature study, the artificial neural network stands out. The suitability of several models of student advancement to predict student progression in Kenya will be investigated. A literature review will be used to investigate the viability and performance of various models. The study's particular goals were to evaluate students' dropout and deferment rates, create an appropriate artificial neural network model that employs the identified elements to forecast progression rate, and validate the model. The data for this project will be gathered via the Zetech University database system. The report included information on students who were enrolled from 2007 to 2019. The study included a total of 5000 pupils. The artificial neural network model was validated using the sigmoid activation function after the data was separated into training and test sets. The rate of advancement was discovered to be 78.5 percent. Universities should establish intervention programs for students who are on the verge of deferral or dropping out, according to the report.

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Diploma and Certificate students' progression is a major concern to guardians, sponsors, and university administrators, the majority of the students enroll in the universities for their diploma or certificate courses intending to progress until to the degree level. However, in many instances, this has not been the case. Majority of these students after completing their diploma or certificate courses they do not progress from certificate to diploma and diploma to degree (Gibbs, 2004). Students encounter a variety of challenges, including a lack of government funding by the Higher education loans board (HELB), lack of finances to further their academics until completion, missing marks from the respective department, geographical factors such as distance making it difficult for students to afford transport cost, job vows and health matters are challenges are some of the challenges or difficulties that students face among many others (Hayden, 2012). The most important mission of universities as academic institutions is to ensure that there is a one hundred percent progression of all the students enrolled for diploma and certificate courses because this will help the university maintain its population and generate income since the students are the major source of income.

The study will take part in analyzing the student progression in Kenyan private universities using the data of 2007 to 2019 to find the accuracy level of student progression.

The data mining methodology will aid university officials in calculating the number of students who are likely to progress through the university, as well as offering the appropriate predictive model and assist them. It will provide university administrators and managers with enough data

to make informed judgments. Data mining techniques will help in identifying the proper patterns and classifications for determining the number of students who progress from certificate to diploma and diploma to degree in the timeline set by the academic calendar and which gender progress in a high number.

Artificial neural networks will enable us to understand the trend of the student progression by determining the accuracy level of the student progression from the year 2007 to 2019.

Lau (2019), ANN extends the capability of analyzing the complicated amount of data that are easily classified through the use of conventional statistical techniques.

Siri, D. (2015) defined ANN as a mathematical model that tries to imitate the behavior of the human brain which involves nonlinear relationships among different datasets that cannot always be fully identified using conventional analyses. ANN has been applied in solving problems in areas like engineering, agriculture, education, medicinal chemistry, and pharmaceutical research. Linear models have been applied in the area of inflation forecast in Kenya which are autoregressive (AR), error correction, and calibrated macroeconomic models but these types of time-series models are limited because they assume linearity which is not consistent with the nonlinear nature.

Some institutions employ Artificial Neural Networks (ANNS) to anticipate progression to circumvent the linear nature of autoregressive models. This is because ANNs are good at capturing nonlinearities. Have the ability to imitate the majority of functional requirements, are immune to outliers, and are unaffected by disruption. Some studies have compared the prediction ability of Neural Network models to that of linear models. The neural network-based models were shown to be more precise in prediction in these investigations. As a result, ANNs according to (Hil'ovska et al, 2012) will be used by numerous intuitions.

According to a review of relevant literature, several research on student progression have been undertaken in Kenya, but no study has compared the performance of AAN models to that of other models in predicting student progression in Kenya. As a result, this study generated an ANN model that will be used to forecast student development in Kenya and compared it to an autoregressive moving average model.

1.2 Statement of the problem

The advancement of diploma and certificate students is a significant source of worry in educational policymaking sectors (Demetriou & SchmitzSciborski, 2011; Tinto, 2006). Around 30% of students enrolled in diploma and certificate programs around the world do not continue to pursue their degrees. In Kenyan universities, a similar problem of low diploma student development exists.

Private institutions entirely depend on funds from clients. Their operations are affected if the client's contribution pool is low, the pool depends on the number of students admitted as well as the number of students who progress. With a predicted number of intakes, progression is the major stream of income that distinguishes private institutions. Private institutions with higher progression rates are in better positions compared to ones experiencing low progression.

The stability of private institutions depends on the number of their students. They should strive to ensure a high number of diploma students' progress. This can be done by advertising their courses and making an accurate prediction on the number of students expected to progress. Linear and nonlinear models can be used to predict student improvement. In most cases, linear models have been used to predict student progression.

When linear models are used to forecast a student's progress, accurate predictions are obtained. Binners et al.(2005), data linearity is assumed for linear models, however, evidence suggests that

the output contains nonlinearities. Because of this assumption, linear models are intrinsically limited in their ability to forecast student development.

Haider et al. (2007) found that the root mean squared error of student progression prediction based on nonlinear models is lower than that of progression prediction based on linear models in a study of student progression forecasting.

Binner et al. (2005) compared the linear ARIMA and VAR models to the non-linear ANN model in terms of student progression accuracy. The linear models ARIMA and VAR were utilized, while the non-linear model's ANN was used. Nonlinear models, they concluded, provide more accurate forecasts.

Binner et al (2006) used quarterly data in the United States to compare the performance of ANN and linear models, and found that the ANN model offered better forecasts.

1.3 RESEARCH OBJECTIVES

1.3.1 MAIN OBJECTIVES

The primary goal of this study is to design an artificial neural network data mining model for predicting Diploma and Certificate students' progression in a university.

1.3.2 Specific Objectives

i) To examine factors influencing diploma and certificate student's progression in university

ii) To develop a neural network model for predicting diploma and certificate student's progression

iii) To validate the developed model.

1.4 Research Questions

i. What factors influence universities diploma and certificate student progression?

ii. Which is the appropriate neural network model for predicting diploma and certificate student progression?

iii. What is the validated model?

1.5 Significance of the Study

The major outcome of this study will be the development of a diploma student's progression model in Kenya. Increased student progression in private universities in Kenya.

The study will compare students' progression performance of the ANN models with the ARIMA model will make it possible for Kenyan stakeholders in academia to choose the best prediction model.

From the literature on the study conducted in Kenya ARIMA model has been the powerhouse in predicting student progression rates. However, the accuracy level of the ARIMA model is low.

Not many studies conducted in Kenya have employed ANN, however globally ANN is nonlinear but its accuracy level is higher, (Shahiri Mohamed,2015) conducted a study predicting student's dropout rate using ANN and the prediction accuracy was at 98%.

1.6 Motivation of the Study

The primary motivation for this research is to ensure that diploma and certificate students' progress efficiently. An artificial neural network, a data mining approach, will assist university academic registrars in quickly obtaining essential information such as student progress in terms of gender and various faculties. This will guarantee that lectures and university registrars plan for the necessary resources in the future, as well as determining the number of students progressing. The motivation to carry out this study using an artificial neural network as opposed to other data

mining algorithms is because ANNs are good at capturing nonlinearities and can imitate the majority of functional requirements, are immune to outliers, and are unaffected by disruption.

Hil'ovska et al (2012) compared the prediction ability of Artificial Neural Network models to that of linear models. The neural network-based models were shown to be more precise in prediction in these investigations also Artificial Neural Network is a nonlinear model which is easy to understand and use as compared to other statistical methods. Artificial Neural Network is a non-parametric model while most statistical methods are the parametric model that needs a higher background of statistics.

Artificial Neural Networks compared to the other data mining algorithms can handle large amounts of datasets and ANNs can detect the complex nonlinear relationship between the independent and independent variables and also it can detect possible interactions between predator variables.

1.7 Scope of the study

The survey was limited to Kenyan universities, with the student fraternity serving as the primary respondent. To demonstrate certificate and diploma student progression rates, data was gathered from the university administration system database at the registrar of academics. The study used the data for all the semesters from January to April, May to August, and September to December in the year 2007 to 2019.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter comprises a review of the literature that exists. There are seven core sections in this chapter.2.2 Theoretical review 2.2.1Factors influencing student progression rates. Section 2.3 introduces the concept of Artificial Neural Networks, how it works and Artificial neural network techniques. Section 2.4 discusses the Empirical review. 2.5 describes the conceptual framework for the research. Section 2.6 Operationalization of variables 2.7 gives a synopsis of the chapter

2.2 Theoretical review

2.2.1 Overview of student's progression rates

Shahiri, Husain, and Rashid discovered that cumulative grade point average (CGPA) is the most influential variable because it influences future educational and professional mobility in a literature analysis on data mining approaches used for forecasting student success. According to their findings, neural networks have the best prediction accuracy (98%) while decision trees have the second-best prediction accuracy (80%). Both support vector machines and k-nearest neighbor have the same prediction accuracy as naive Bayes (83 percent), however naive Bayes had lower prediction accuracy (76 percent).

Thai Nghe, Janecek, & Haddawy (2007) examined two classifiers, the Decision Tree and the At two different institutes, a Bayesian Network was used to predict students' GPA at the end of their third year of undergraduate studies and the end of their first year of postgraduate study. Each data set contains 20,492 and 936 complete student records, respectively. According to the findings, the Decision Tree outperformed the Bayesian Network in all classes. In all cases of classes, the accuracy was significantly enhanced by applying the resampling technique,

especially for the Decision Tree. At the same time, because the Decision Tree method tends to focus on local optimal, resampling was utilized to reduce misclassification, especially for minority classes of imbalanced datasets. Ian & Eibe (2005) presented a case study in which educational data mining was utilized to identify failing students' behavior and warn students at risk before the final exam. Another case study was presented by Romero, Ventura, & Garcia (2008). They mined e-learning data using each phase of the data mining procedure.

Polpinij (2002) used educational data mining to predict students' final grades using data acquired from a Web-based system. Beikzadeh & Delavari (2005) employed educational data mining to identify and then improve educational processes in higher education systems, allowing them to make better decisions.

Waiyamai (2003) employed data mining to help with the development of new curricula and the selection of an acceptable major for engineering students.

In other studies, Kotsiantis, Pierrakeas, & Pintelas (2003) compared six classification approaches to predict drop-outs amid a course (Naive Bayes, Decision Tree, Feed-forward Neural Network, Support Vector Machine, 3-nearest Neighbor, and Logistic Regression). Demographic information, the results of the first writing assignments, and attendance at group meetings were all included in the data collection. The data set contained records of 350 students. Their best classifiers, Naive Bayes and Neural Network, we're able to predict about 80 percent of drop-outs. The findings also revealed that a simple model like Nave Bayes may generalize effectively on tiny data sets, whereas other methods like Decision Tree and Nearest Neighbor require considerably bigger datasets.

Zulu (2008) identifies student attrition as a framework that might be used in various research that uses pre-enrolment and post-enrollment indicators to predict student success and failure. This framework is used in part in my research. Ramsden (1992) and Laurillard (1993) both argue that a variety of factors influence students' achievement in higher education, according to (Ditcher & Tetley, 1999). The learning context, teaching tactics, student motivation, and students' grasp of course requirements are all provided as examples.

The issue is that none of these research analyzed data using a neural network. They also employed primary data in the form of questionnaires, which may have been skewed during the data collection process. Family, institutional, and lecture-related issues are the underlying components in this study.

2.2.2 The important factors that influence student's progression rates

Motivation - In studies, motivation is important because they rely on self-direction and self-learning. Motivation, or a lack thereof, might result in a student's slow growth. (2016, Bawa)

(Eric, 2010) student achievement is closely linked to accountability and motivation. He explains that interest in classes is strongly related to student aptitude and attitude toward learning. (K. L. Smart & J. J. Cappell,2006) Time required to complete a module, a lack of real-world examples, problems accessing a resource, and support systems are all shown to be key sources of motivating limitations.

Economics

Face to face owing to the need for physical class attendance, which necessitates the use of time and travel expenses, among other things. According to research performed in Pakistan, the most

common reason for dropout was a failure to pay fees on time (Darakhshan Muslim, Syed Muhhammad, Syeda Aneeqa, 2017).

Learner dropout is heavily influenced by the financial difficulties of some families and the scarcity of financial aid in eLearning courses. Most governments do not subsidize diploma and certificate programs, and thus do not provide financial aid to students at lower levels.

Social and family – Home is the location where the foundation of learning and education takes place. To obtain good academic outcomes that contribute to kids' advancement, parents, children, and other family members must promote a learning environment in their homes. Parents are responsible for aiding pupils who are having difficulties in certain classes, for example. This help could be provided in the form of private tutoring. They equip kids with technology and other learning materials to help them enhance their academic performance at home. Parents have an important role in their children's development and growth (Kudari, 2016). Children frequently talk to their parents about any problems they are having in school, whether academic or otherwise. Parents offer their children security, encouragement, and aid in resolving their problems.

Demographics

Learner dropout is influenced by a variety of personal factors, including age and gender. (Darakhshan Muslim, Syed Muhhammad, Syeda Aneeqa, 2017) According to the author, some familial considerations, such as marriage, have an impact on female students and may cause them to drop out of classes. The student's age may play a role in dropout since younger students are more familiar with technology than older students, allowing them to use the online platform more simply.

2.3 Time- series Prediction models

Any data can be dissected into a trend, a cyclical element, a seasonal factor, and an error term, according to these models. There are two types of time series models: univariate and multivariate. The most widely used multivariate model is the vector autoregressive (VAR) model. The autoregressive (AR) and autoregressive integrated moving averages (ARIMA) models are the most used univariate models.

2.3.1 ARIMA

It is made up of ARMA and integrated components. There are two types of ARIMA models: seasonal and non-seasonal. ARIMA (p, d, q) (P, D, Q) m is the non-seasonal model, while ARIMA (p, d, q) (P, D, Q) m is the seasonal model. Positive numbers make up the parameters p, d, and q. For the seasonal part of the model, P, D, and Q indicate the AR, differencing, and MA factors. Each season's number of periods is represented by b.

2.3 Artificial Neural Networks

Artificial neural networks are theoretical applications based on research into the brains and nervous systems of many animals. In the field of computer science, it is classified as artificial intelligence. Artificial Neural Networks (ANNs) were created in the 1950s to replicate the structure of the organic brain. To solve large issues, neural networks use a "divide and conquer" strategy (Gershenson, 2003). An Artificial Neural Network is made up of several nodes and the connections that connect them. Neural networks have been applied in many fields, including physics, computer science, finance, and many others

2.3.2 How Artificial Neural Networks works

Our brains and their interconnections serve as inspiration for Artificial Neural Networks. They're based on the natural neurons present in the brain. Synapses on the dendrites of natural neurons

transmit information. The neuron is triggered and a signal is emitted if the signal is stronger above a predetermined threshold. Nothing happens if this is not the case.

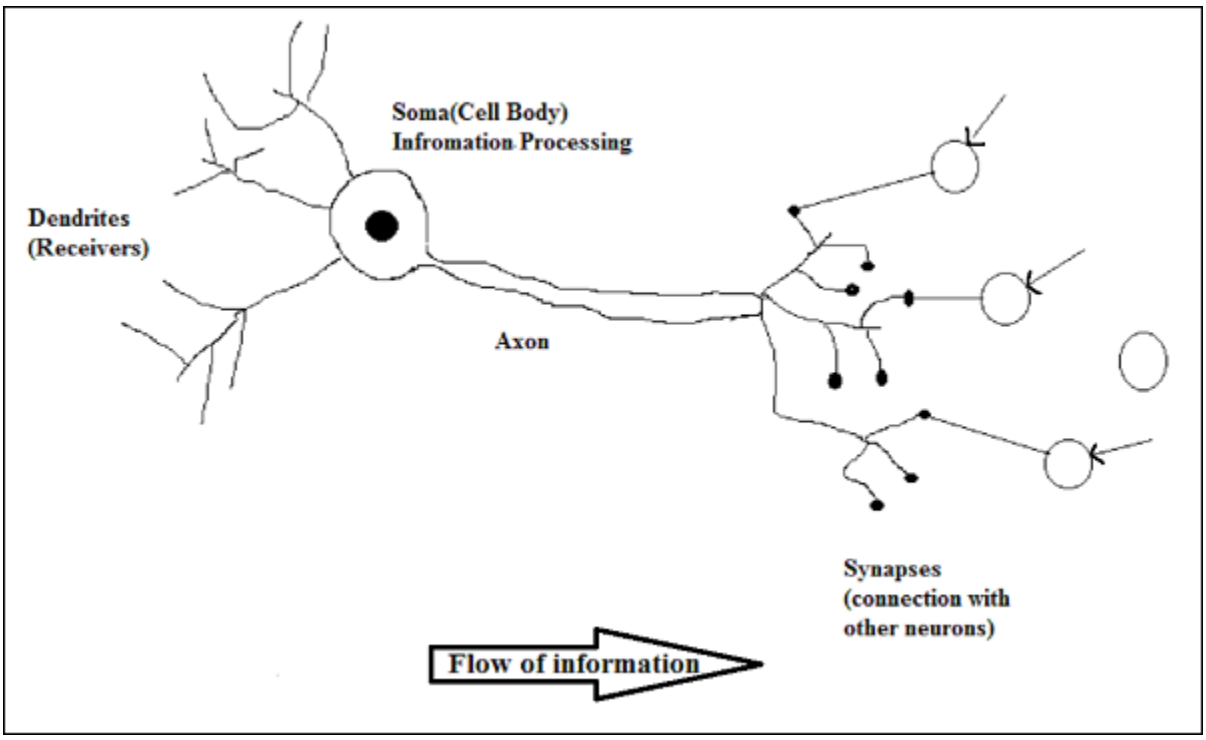


Figure 1 Natural neuron (source: Artificial Neural Networks for Beginners)

Because they accept some input, conduct some operations, and then provide an output, the nodes are computational units. These operations can be simple or complex at times. Information travels from one end to the other because they are connected. The signal flow may be unidirectional or bidirectional. Many thresholds and functions in the network combine to provide noticeable global activity. ANNs are algorithms that are used to find patterns in data over time and use that information to make better decisions. There are three levels in a basic ANN: input, output, and hidden layers. An ANN may learn and adjust its weights and biases over time to match the expected output.

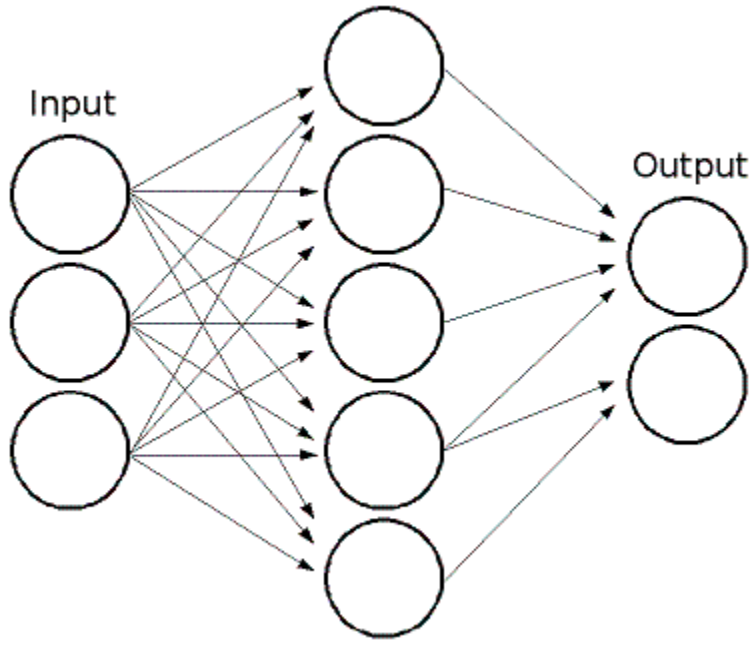


Figure 2 Artificial Neural Network (source: **http://neuralnetworksanddeeplearning.com/**)

A perceptron is a device with several inputs and a single output. Weights are modified for each input until the ANN produces the desired output.

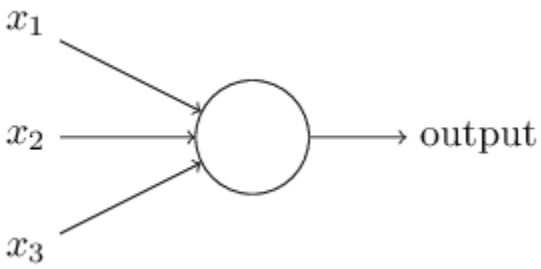


Figure 3: Perceptron (source: **http://neuralnetworksanddeeplearning.com/**)

Each neuron is given a numerical value. The greater the input must be to activate the threshold, the higher the weight. Negative weights are also possible. Weights are modified during training to get the desired output dependent on particular inputs. Learning is a part of the process of determining the best weights.

2.3.3Artificial Neural Network Techniques

According to Maxwell (2015), some popular neural networks are

1. Feedforward networks

2. Feedback/recurrent networks

3. Convolutional networks

The classification is based on the network's data flow.

2.3.1.1 Feedforward Neural Networks

In the development of feedforward networks, there are no loops. From the input layer to the output layer, the signal flows in a straight path with no backward movement. As a result, the weighted connections' activations are only communicated forward. Multilayered feedforward neural networks are the most often utilized feedforward neural networks. An input layer, an output layer, and a hidden layer are used to organize the neurons (s). The hidden layer of a network could have one or more layers. The connections between the layers are all unidirectional, going from input to output. This research employs a multilayer feedforward neural network. We also needed to know how many neurons should be in a hidden layer.

2.3.1.2 Recurrent Neural Networks

The connections of recurrent neural networks have loops. As a result, the weighted connections are employed to feed backward prior network activations. As a result, data travels in both directions.

2.3.1.3 Convolutional Neural Networks

Since its inception, convolutional neural networks have been almost exclusively connected with computer vision applications. This is because their architecture is well-suited to undertaking complicated visual analysis. Instead of a two-dimensional array, the convolutional neural network architecture is characterized by a three-dimensional arrangement of neurons. A

convolutional layer is an initial layer in such neural networks. Each convolutional layer neuron only analyzes data from a limited portion of the visual field. Following the convolutional layers are corrected layer units, or ReLU, which let the CNN handle more complex data.

CNN's are most commonly utilized in object detection applications such as machine vision and self-driving cars. While these artificial neural networks are the most popular in today's AI applications, many more are being developed to achieve a level of capability more like that of the human brain. Every new revelation about how our brains work leads to a new advance in AI, resulting in improved neural network models. As we learn more about our brains, it will only be a matter of time before we can replicate the entirety of our brain's functioning in computers.

2.4 Empirical review

2.4.1 Review of related work

In this part, we'll look at various university-based empirical research that has used neural networks. These studies are mentioned here to emphasize the importance of the current research. Poor admissions results are a serious concern in higher learning education. Students quit universities for a variety of reasons, including a lack of previous knowledge in the topic of study, extremely low marks and inability to pass a test, and a lack of financial resources. Predicting students' grades is a difficult task for university administrators who want to avoid the phenomena of early school exit.

To determine the factors that influence students' performance, (Oladokun et al., 2008) used a neural network. Students were divided into three groups based on their test scores. The prediction accuracy obtained by the authors of the research was around 74%. To predict student graduation rate, (Karamouzis et al., 2008) employed a three-layer perceptron network trained by backpropagation. The authors' network model was 70.27 percent accurate for successful graduates and 66.29 percent accurate for failing graduates.

An artificial neural network was employed in a study by Stamos (2008) to predict student graduation results. The ANN network is a three-layered perceptron that was trained using

backpropagation methods. The experiments used a sample of 1,407 student profiles during training and assessment. The chosen sample was divided into two groups and represented Waubonsee College students. The training set consisted of 1,100 profiles, whereas the testing set consisted of 307 profiles. The two sets' average prediction rates were 77 percent and 68 percent, respectively.

In Prediction of General High School Exam Result Level Using Multilayer Perceptron Neural Networks, Mohammed Awad (2018) employed an artificial neural network technique. The study's findings revealed that the ANN method was more precise than the logistic regression technique.

Blatchford (2013) noted that the student's progression is determined by how excellent the teacher is when he or she is teaching the units by motivating and inspiring teaching. This means that pupils or students will always love the units or subjects concerning how teachers make it more effective in class and thus passing the units and progressing to the next level. The use of this model will not predict the accuracy and effective transition from a certificate to diploma and diploma to degree.

Monitoring on student progression in class is determined by the regular formal and informal assessment during class time (Victoria, 2019). This model entails the collection of pre and post-tests which may need for the teacher to adjust their instructional strategies to better the student and meets the objective of the students to progress in the next level as being competent. For a student to progress, most public schools use universal screening, progress monitoring, and response intervention (RTI) to measure the student needs which is tested at least once a year and thus makes it not a good prediction model for student progression in class or level.

None of these studies focused on creating an artificial neural network to predict student progression rates.

2.5 Conceptual framework

The conceptual framework depicts the relationship between key factors found in the study's research concerns. The conceptual framework is vital in visualizing the relationship surrounding the study's core areas concept, according to (Sisimwo,2016).

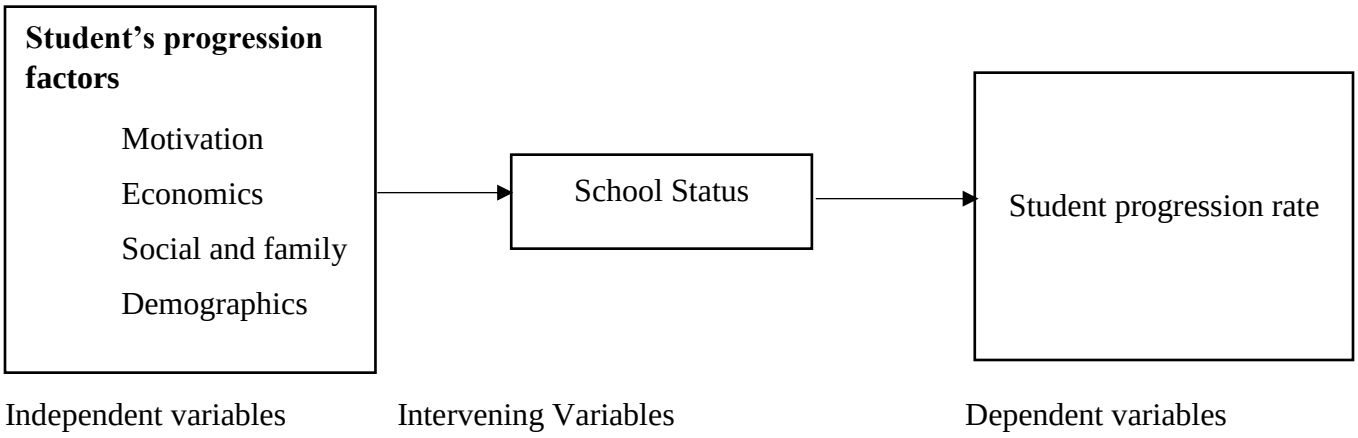


Figure 5 Conceptual framework Munga (2021)

2.5.1 Operationalization of Variables

2.5.1 Operationalization of Variables

The variables of the study were operationalized as per table

Table 2.5.1: Operationalization of variables

VARIABLE	SUB VARIABLE	INDICATOR	VALUES
Student factors	Motivation	• The length of time it takes to	• Numeric

		finish a module.	
	Economic	•Financial situation • Financial assistance	Scale 1-Parent 2-Guardian 3-scholarship 4-self sponsor
	Social and family	• Relationship status	1-single 2- Married 3-Divorced 4-Windowed/windowed
	Demographic	Sex	Female/male If female 1 and if male 0
	Students progression rate	• Number of students who have enrolled for post requisite of diploma and certificate	Number

2.6 Summary and the Research Gaps

This chapter examines the study variables and their relationships within a theoretical framework of choice. Dropouts, school dropouts, and deferments have all been demonstrated to have an impact on a student's advancement rate. Other explanations for the three elements investigated in the study have been presented. To stress the relationship between the study's elements, a conceptual framework was established. In literature, the gaps underline the necessity to construct a model for university students' progression rates.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

The methods used in the study were described in depth in this chapter. The research design, which is a blueprint for performing a study that outlines the design process, was discussed in this chapter. The research design also took into account the study's chosen research paradigm. The data source, which described the study's population, was discussed in this chapter. It also included data selection and sampling procedures. Data pre-processing was also discussed, which involved gathering data and ensuring that it is in the proper format for analysis. Data mining was also covered, which explains how to collect data and do descriptive and predictive analytics with the R statistical software or python programming language. The topic of data transformation was also explored, which refers to ensuring that the data is scaled in the reprocessing process. The Knowledge Gap is then presented to present the study's output results

3.2 Research Design

The word "research design" refers to a plan that directs each study and assists in data collection, analysis, and interpretation of results. It can also be utilized by researchers as a template for deciding on methods and instruments to utilize in data gathering and analysis to answer the research questions (Cooper & Schindle, 2014). The Fayaz model for data mining was used in this investigation, as illustrated in Figure 3.

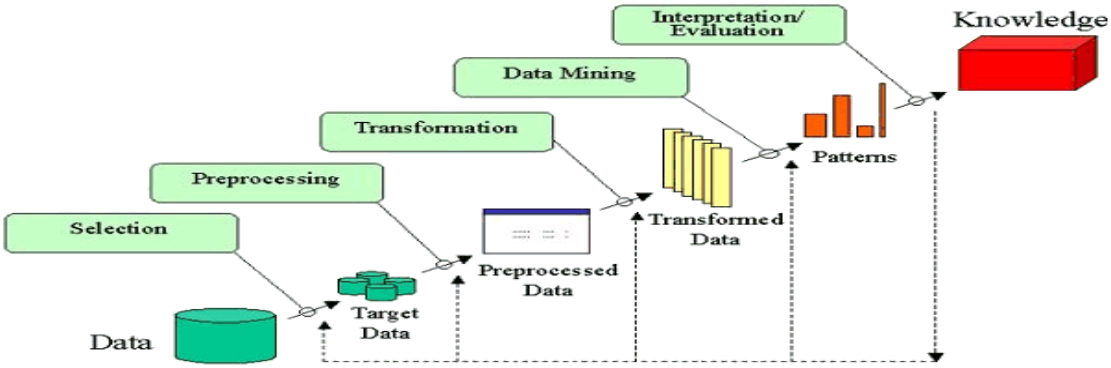


Figure 6: A model design approach for the study

Source: Fayyad et al. (1996)

3.2.1 Selection of Data

The entire student population was used in the study; it was necessary to conduct a census on the student population in the private universities in Kenya. The study worked with a representative dataset by doing data reduction after acquiring data from the university data source and saving it in an excel spreadsheet. As part of the numerosity reduction process, the data was replaced or estimated using alternative, smaller data representation methods such as means, clustering, sampling, and histograms. The data was then cleaned up by the researcher. The practice of discovering and eliminating faulty or erroneous records from a database is known as data cleansing. The binning technique was used to complete this operation. Binning methods were used to sort data value by checking its "neighborhood," or the values in its immediate vicinity. The sorted values were divided into several "buckets" or bins. The data was then divided into equal frequency bins before being smoothed using means. According to the research, several Private Universities in Kenya are divided into three faculties: ICT, hospitality, education, and business. The data from the faculties were used in the study using a convenience sampling methodology. This accounted for a third of the faculties and, as a result, 33% of the population.

Mugenda and Mugenda (2013) claim that 10% of the population is indicative of the entire population. According to the data in the database, there will be occasions where diploma and certificate students earned university placement but did not enroll in their studies. There were instances where students entered in the second year, indicating that some of the students had previously completed diploma courses and qualified to participate in second-year university degree programs. As a result, the study undertook data cleaning to protect the students by filling in the missing first-year records with first-year enrolment averages. Due to repeated registration numbers, the database also featured a mix of registration numbers, with student information appearing more than once in the database. Dimension reduction was used to deal with this as well. Data reduction was carried out to improve the efficiency with which the data was handled for analysis. After that, the data was divided into training and test groups. Pesaran, Pick, & Timmermann (2011), as well as Andic &Ogunc (2012), have successfully used this strategy of sample splitting (2015).

3.2.2 Data Pre-processing and Transformation

Identifying the data to be used in training the model, testing the model, and assessing the output error with training and validation data are all part of this process. The type of data that enters into each data set, as established by the research sampling technique, is critical. The population that was identified was represented by the sample taken. The data was translated into three independent comma-delimited (CSV) format files, each providing information on student enrolling, deferment, and dropout. The student ids were used to link all three of these files. Following the preparation of data input files, data cleaning were carried out as follows:

i. Conversion of data types (character to numeric, character to factor)

ii. Aggregating data down to the student id level

iii. Identifying and deleting outliers as needed

iv. Removing or cleaning up missing values

v. Combining multiple datasets into a single one

The following factors datasets were used to determine the model efficacy through preprocessing. They include:

Motivation, Economic factor, social and family factor, and demographics.

After reading the CSV files, as the first step of the data preparation process, all the character type data were transformed into numeric data.

Motivation<- read.csv (file. Choose (), header= TRUE)

Economic<-read.csv (file. Choose (), header=TRUE)

Social and family<-read.csv (file. Choose (), header=TRUE)

Demographics<-read.csv (file. Choose (), header=TRUE)

Myfile<-as.data. Frame (Motivation, Economic, Social and Family, Demographics)

Attach (Myfile)

Myfile Motivation, Economic, Social and Family, Demographics

3.3.3 Data mining

The study will feed motivation, economic, social and family, and demographics to the model for processing, the model will be trained on the type of input data and the projected output of the training session. Data will be saved in an excel data sheet, which was converted to a CSV file (comma delimited) and read into the WEKA analysis tool.

When using the WEKA analyzing tool, the data was changed to an attribute relation file format (ARFF) file for ease of analysis.

After reading the data into the WEKA Analyzing tool, the next step was to scale the data interval to ensure efficiency in working with the data. The data was then separated into training data and data sets to determine the efficiency in data collected, the efficacy on the model was determined in percentages of 100% preferred [70:30]. According to, seventy percent (70%) of the source data is utilized to train the model, whereas thirty percent (30%) was used to test the model (Minewiskan, 2018). The training data was then into the ANN model via the model neurons specified, and the process was repeated in several trials until the artificial neural network model converges on the correct efficacy.

3.3.4 Model Evaluation

Model evaluation is a process of ensuring that the system meets the user's requirement by observing the actual model output and the evaluated output model. It involved the use of several trials to find the accuracy of the model and validated to ascertain that it meets the user's requirement on the progression of the students in the university.

The model was evaluated by calculating the measure of performance through the use of root mean squared error and the mean absolute error. These measures were calculated against the actual recorded progression rate values in comparison with the progression rate values forecasted by the model over the same period. The model was run in several trials and the average accuracy of the model was determined. A confusion matrix was obtained to show the various performance measures and the accuracy of the model. The output of the results was displayed at the Knowledge gap. Several performance measures were conducted in this study.

3.3.5 Knowledge representation

Knowledge representation is a technique for visually representing data to the user. This included data that has been mined for information. Different strategies were used to generate the output of the model to ensure that the efficacy of the model is met.

3.4 DATA COLLECTION

3.4.1 Source of data

The data was collected from the private university database system which involved several private universities in Kenya ranging from 2007 to 2019 in all semesters. It involved the raw data which was cleaned to get the right trend of student progression involving all attributes defined by this study.

3.4.2 Sampling method

The study used a stratified sampling method since it involved creating a section or group of population. In this study, it was put into consideration a certificate group and a diploma group which was used to determine the progression rate of one level to another. When using the stratified sampling method, it is easy to quantify because there is a clear number of users in each group.

This method of sampling was adopted because it assisted in targeting a certain group of students from certificate to diploma and diploma to degree since in our Kenyan economy people who are dominating in the job market are degree holders thus bringing into the attention of making sure that there is a 100% students progression to reduce the gap of employment and job satisfaction to the community who need the service of the graduates. To get the sample it involved several private universities in Kenya from several faculties and to get the respondents the study used already collected and stored data in the database. Database administrators played a very great role in providing the raw data that assisted in sampling.

3.4.3 Data collection methods

This study used data that had already been stored in the database for all the sampled private universities in Kenya.

3.5 Discussion

Knowledge discovery in databases (KDD) model was considered the best and the cornerstone for all subsequent KDD process models by Fayyad et al., Piatetsky-Shapiro, & Smyth (1996), who provided a more comprehensive and detailed KDD process steps concerning data analysis, data selection, pre-processing, and transformation stage.

Because SEMMA and CRISP-DM are implementations of the KDD model, all of the phases that are implemented on KDD were also to be implemented on the model indicated, making KDD the best model to employ because it gives all of the important stages of an accurate system. It became the most effective way for conducting data mining projects on the examined data. The current study is on student progression, and it was conducted using the ARIMA model, which failed to produce an accurate prediction.

The ARIMA model could only predict students' progression, whereas the ANN model predicted students' progression for a large number of students at the university, helping management make the best decisions and even preparing for the future.

CHAPTER FOUR

RESULTS AND DISCUSSIONS

4.1 Introduction

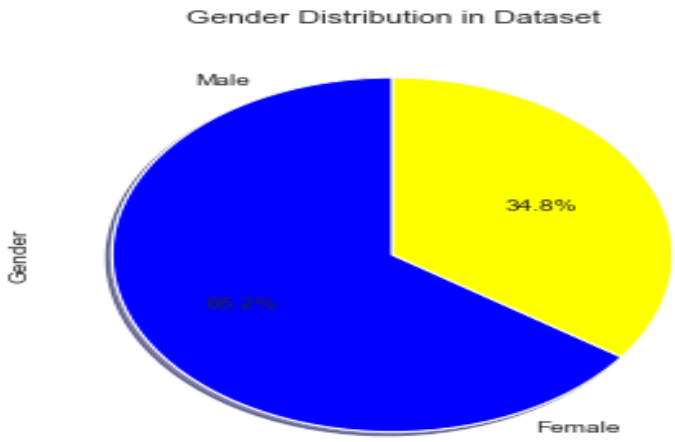
To test the proposed model, real data was collected from a multidisciplinary university. The collected data contains courses, cost implication, motivation, demographic factors, socio-family, and progression factors, from 2007 to 2019, with 5000 students records. These datasets describe data distribution with sample information, in addition to the training and testing sample number. Training ratio and total sample number are also considered.

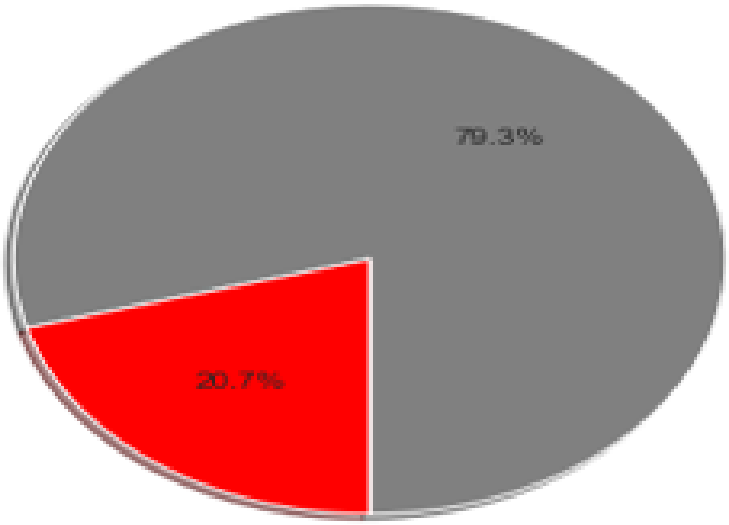
The dataset represents student progression. The data is divided into two unequal parts. The main part (data collected from 2007 to 2016) is used for training, while the remaining sample part (data collected from 2017 to 2019) is used for testing. Dataset of Economics Education represents the highest dataset percentage with a value of 18%, while the lowest dataset percentage belonged to Physics Education with 0.9% value.

4.2 Demographic Information on students Dataset

In the analysis, we focused mainly on the variables of the study during the period. Pre-processing was a crucial part to be done at the very beginning of any data science project. It included dealing with null values, detecting outliers, removing irrelevant columns through analysis, and cleaning the data in general. The data set used contained (4000) rows with four (4) columns. The four (4) attributes consisted of cost, motivation, social & family, demographics.

4.2.1participants in the study





The data was also normally distributed and hence appropriate for further analysis. This was evidenced by the histograms which showed balanced bins.

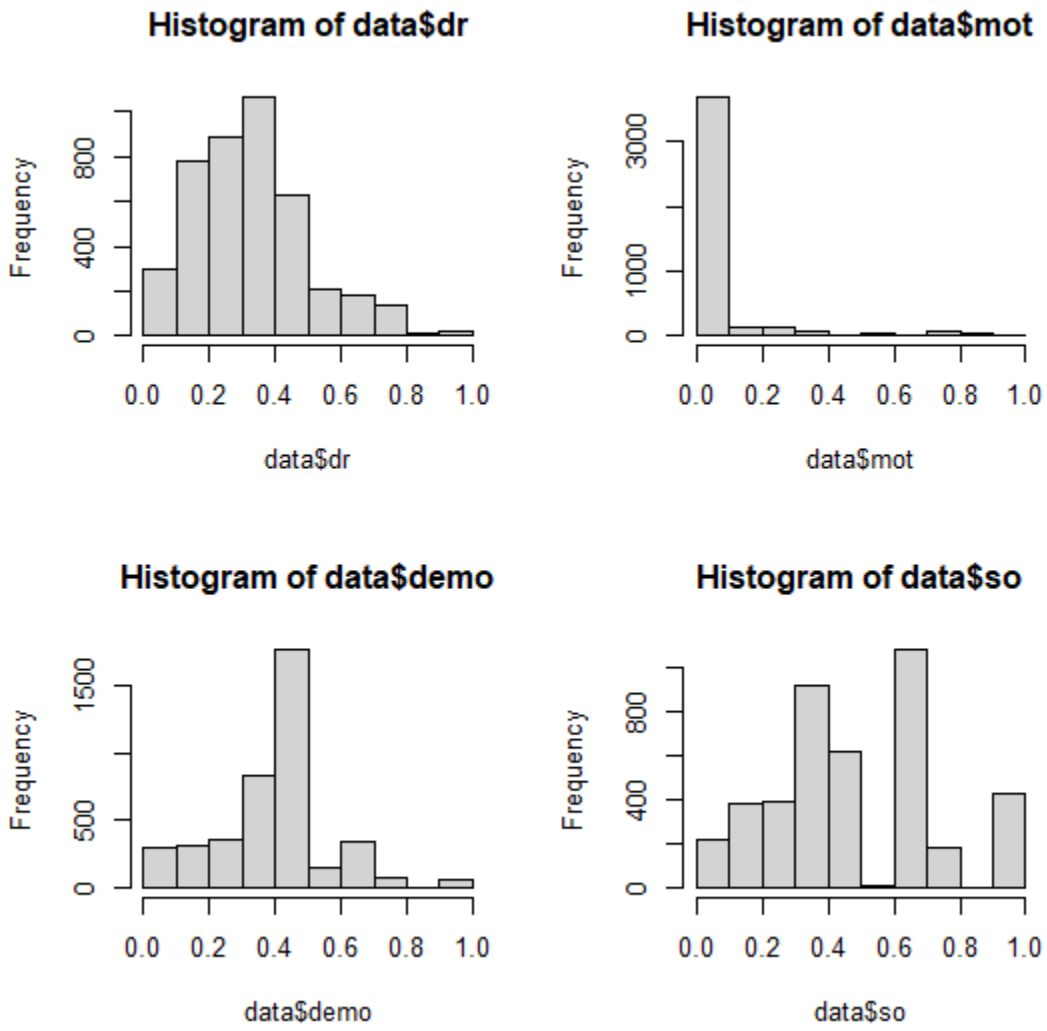


Figure 8:Histograms Results

Figure 4.1 Depiction of the Distribution of the Data

Where dr= Cost, mot=motivation, demo= demographics, so=social and family.

4.3 Factors influencing progression rate of students-Objective one

The researcher was interested in investigating the factors that influence the progression rate of students. The use of feature selection using Cramer's V Correlation was applied to the dataset to explore how the variables correlate with each other.

The use of Cramer's V Correlation to identify how the features or variables correlate with each other or if the variables increase at the same time they are said to correlate otherwise inversely if one variable increases while the other decreases, they anti-correlate. Cramer's V Correlation is similar to the Pearson Correlation coefficient. While the Pearson correlation is used to test the strength of linear relationships, Cramer's V correlation is used to calculate correlation in tables with more than 2 x 2 columns and rows. Cramer's V correlation varies between 0 and 1. A value close to 0 means that there is very little association between the variables. A Cramer's V correlation value of close to 1 indicates a very strong association.

Cramer's V	**Relationship**
0.25 or higher	Very strong relationship
0.15 to 0.25	Strong relationship
0.11 to 0.15	Moderate relationship
0.06 to 0.10	Weak relationship
0.01 to 0.05	No or negligible relationship

Table 4. 1Cramer's V Correlation

A co-efficient close to 1 means that there's a very strong positive correlation between the two variables. In our case, the blue shows very strong correlations. The diagonal line is the correlation of the variables to themselves thus, they will be 1. The correlation heat map of various variables is shown in Figure 4.5 below.

After checking the correlation matrix, it was observed that there are several attributes with high correlation.

Attributes	LEFT
Demographic factors	0.55
Social and family	0.41
Cost	0.36
Motivation	0.24

Table 4. 2Attributes Relationship

From the Correlation Matrix above, several attributes were more significant than others were. The revised framework is as in Figure 4.5 below. The key factors affecting the progression rate of students are motivation, cost, social and family, and demographic factors.

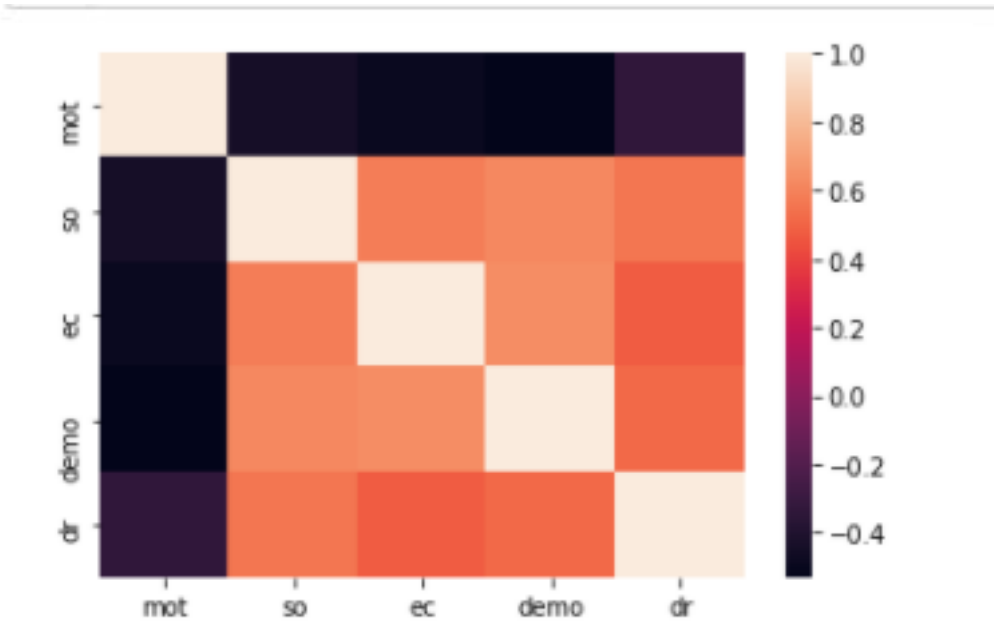


Figure 4. 1 Revised Framework

4.3.1Findings based on the objectives of the study

The objective was to investigate and identify factors that determine certificate and diploma university students' progression rates. Four factors were discovered to be relevant in determining student progression rates. These were the students' demographic, financial, motivational, and social aspects. Table 3 shows the outcomes for objective one.

Table 4.5 Objective one results

Variable	Beta coefficient	Significance	Remark
Intercept	−3.8461		
Demographic	0.5973	0.000	Significant
Social and family	0.938	0.000	Significant
Cost	-1.8019	0.000	Significant
Motivation	0.8928	0.001	Significant

The study used the relevant data to obtain regression coefficients for diploma and certificate student progression rates in Kenyan private universities. The artificial neural network was configured to obtain a combination of demographic, social, and family, cost, and motivation data from the students. The regression analysis showed that demographic achieved alpha of 0.5973 percent which was significant at the 5 percent level while social and family achieved 0.938%, cost achieved 1.8% and motivation achieved 0.89%.

4.4 Objective two results

The objective was to develop an appropriate artificial neural network model that uses the identified factors for predicting progression rate. The results of this objective were presented in figure 2

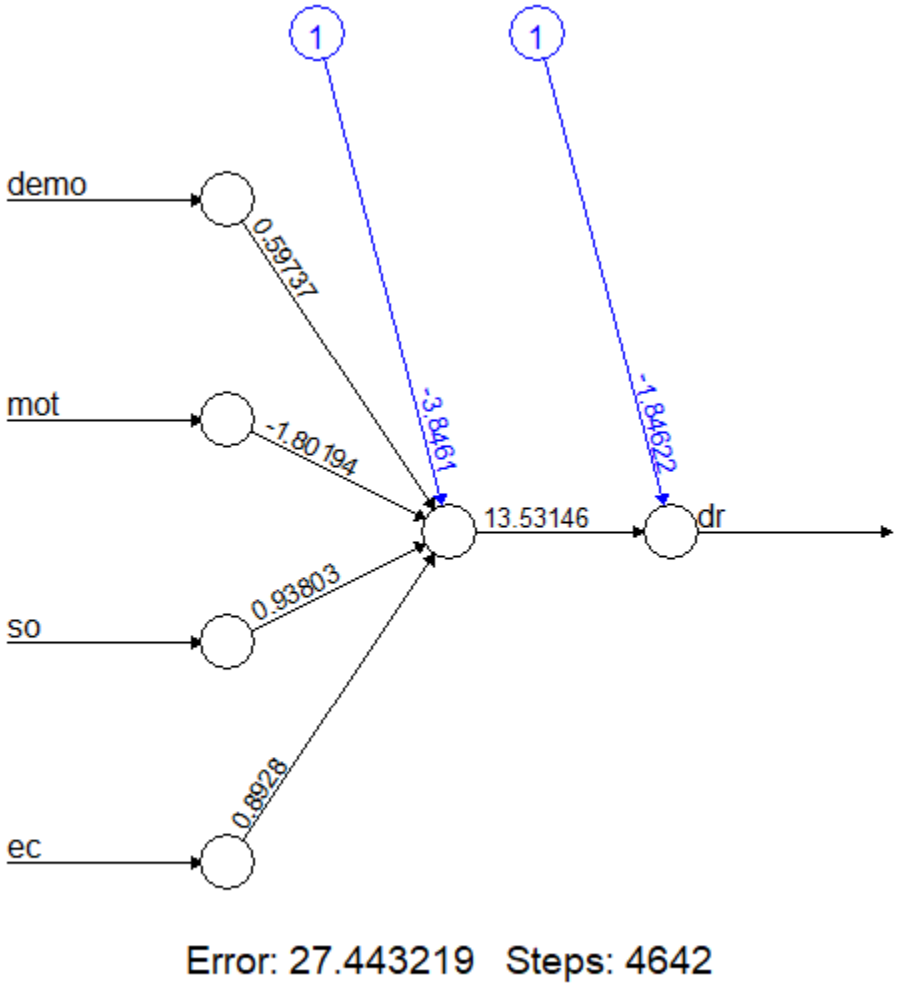
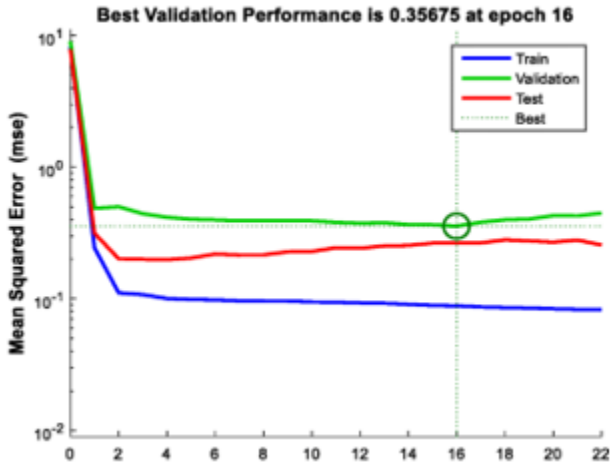


Figure 4.4. Artificial Neural Network Model for the study.



The results in figure 2 showed that demographic achieved alpha of 0.5973 percent which was significant at the 5 percent level while social and family achieved 0.938%, cost achieved 1.8% and motivation achieved 0.89%.

The study's artificial neural network model includes using a backpropagation algorithm to improve the network's ability to learn the inputs. The analysis stage was used to manage and monitor the learning process. There were various stages to the procedure. The input variables were first introduced into the artificial neural network via the neural network's input layer. The effect of introducing the inputs was carried over to the subsequent layers, namely the hidden and output layers. The outputs of the network were then calculated. Adjusting the weights of the neural network to reduce inaccuracy in-network computation as part of the learning process. The error generated was signaled back to the preceding layers and the weights were adjusted appropriately.

4.5 Objective three results

The third objective was to evaluate among the developed models to find out which best performs the prediction. After building the models, the resulting models were analyzed. The accuracy of the models and the insights gained from the resulting neural network was important considerations. Model accuracy is usually straightforward to measure; techniques such as k-fold cross-validation

(Tan, 2005). Cross-validation calculates the accuracy of the model by separating the data into two different populations: a training set and a testing set (70% train set and 30% test set). Stratified partitioning splits the data in such a way that the proportion of response class values remain the same in both train and test datasets. There are several metrics to evaluate models in data mining, like TPR, FPR, TNR, accuracy, precision, recall and F measure, and so on. Below are some of the performance metrics evaluated on the models. To estimate the model's performance, different evaluating measures were considered:

i) **Confusion Matrix**:

The average categorization results' confusion matrices. The anticipated class is shown in the rows of the confusion matrices, while the actual class is shown in the columns. True Positive, False Positive, and PPV are listed in the first column. The False Negative, True Negative and NPV values are displayed in the second column. Specificity, sensitivity, and total accuracy are shown in the last column.

Although it is not a performance metric in and of itself, practically all performance metrics are predicated on the confusion matrix's output.

Confusion matrix results are presented below

	Predicted(0 or Stayed)	Predicted(1 or Left)
Actual (0 or Stayed)	2000	160
Actual(1 or Left)	45	1895

The matrix above represents the actual and predicted outcome in the form of Stayed (0) and Left(1). There are four important terms:

- True Positives: Cases in which we predicted Stayed and the actual also Stayed.
- True Negatives: Cases in which we predicted Left and the actual output was Left.
- False Positives: Cases in which we predicted Stayed whereas the actual output was Left
- False Negatives: Cases in which we predicted Left whereas the actual output stayed.

ii) **Performance Metrics**: Accuracy, Precision, Recall, F1-Score, ROC_AUC Area and Support

Accuracy	**Precision**	**Recall (Sensitivity)**	**F1 - Score**	**ROC_AUC Area**	**Cross-Validation Accuracy**
0.779	**0.80**	**0.85**	**0.91**	**0.57**	**76.88%**

Table 4. 3Performance Metrics

From Table 4.5 above the model produced a good accuracy of **77.9%**

4.6 Discussion of the Results

Even though the model did not achieve a classification accuracy higher than 0.800, the model could be used as a guide to lead intervention policies so that the high rates of progression would increase (Marbouti et al., 2016).

The progression rate of students pursuing higher education at institutions was examined in this study. The study was crucial because there has been a lot of concern in Kenya about the future of university education. A rising number of students are failing to complete their courses and

graduate. This issue has been raised in public by major players in higher education institutions. The government, through the ministry of higher education, has gone so far as to introduce additional curriculum-based training, with a few exceptions, to encourage students to stay in school. Similarly, parents have visited universities to inquire about their children's progress. Many parents were surprised to learn that their children had been dropped from their classes during their early years. Other parents have complained to universities about their children not attending classes, only to discover that these kids had applied to delay their studies without their parents' knowledge. As a result, it was critical to conduct this research to have a better understanding of the state of university education in the country as a whole, so that relevant interventions might be implemented.

This investigation discovered that the university's student progression rate had declined dramatically, impacting the number of students who were making progress in their studies. This was in line with a study published in the University News (2018) report, which said that the rate of enrollment has been steadily declining since the government implemented radical education reforms. According to Julie (2018)'s research, dropout rates have risen dramatically, resulting in students' poor performance.

The study's predictive analytics revealed that the variables in the study had some degree of association. This implied that economic and social issues were linked substantially. As a result, it may be concluded that a student who varies his or her studies is more likely to finally drop out. This is a fundamental rule in association mining that shows that variables in a study are frequently closely related.

This research also resulted in the creation of an artificial neural network model. Figure 5 depicts the model, which was determined to be the best because it had the lowest absolute error. An

optimal model, according to Hasith (2016), produces the smallest absolute error as a measure of performance.

The study's third goal was to validate the model that had been created. This was also a success in this investigation since the accuracy was 0.7790 which translated to 77.9% for the first validation, after validating for the second time the model was 0.7567 which translated to 75.67 %, for the third validation the model was 0.7679 which translated to 76.79% then the model was validated for the four times which resulted to 0.7791 which translated to 77.91 thus the accurate prediction of the model was able to validate accurately at 77.9 % which was found to be consistent with earlier studies. Hasith (2016) conducted research to uncover the factors that cause students to drop out of higher education institutions.

Hasith (2016) found that the neural network model utilized in his study had an accuracy of about 80%, which was a difference of roughly 1% from the current study. Overall, this study was a success because the study's objectives were met.

4.7 Summary of findings

The study focused on developing a prediction model for diploma and certificate students' progression in universities using neural networks. The control variables for the study were demographic, gender, motivation, social, and family.

These variables were analyzed using the artificial neural network model. Data transformation and feature selection techniques were applied to the dataset to determine the impact on the performance of the classification algorithms. Lastly, we reduced the features and determined the appropriate features that can be used to predict the students' progression. Four performance metrics—accuracy, sensitivity, specificity, and F-Measure—were evaluated to check the performance of the model.

The effect of data transformation on classification performance was tested by converting the features into a categorical form using the techniques of equal width and equal frequency. Results indicated that models with categorical data performed better than those with continuous data. Observation showed that equal width data transformation performance results were better compared with equal frequency data transformation. The classification model correctly predicted 4000 students, with an accuracy of 77.9%.

The neural network's accuracy demonstrated a high predicting capacity. The network had a good predicting ability for the training set data compared to the test data in terms of accuracy. The quantity of social and family indicators has a significant impact on students' progression rates. This was backed up by a significant percentage of students differing their classes rather than dropping out entirely. Reduced student progression rates may, in the long run, result in an alarmingly low number of students being absorbed into the job market. As a result, there may be a reduction in the workforce, resulting in lower economic output.

CHAPTER FIVE

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Introduction

Predicting students' performance is mostly useful to help the educators and learners improving their learning and teaching process. The goal of the study was to use artificial neural networks to determine the rate at which certificate and diploma students progressed.

The study's summaries were derived from the analysis in Chapter four, and conclusions were drawn from the findings and recommendations based on the study's objectives. The recommendations were made for policy areas of interest as well as future research.

5.2 Conclusions

This study made important contributions to the knowledge base on certificate and diploma students' progression in the Kenyan private universities, which are currently facing a crisis of student numbers.

From objective one results it can be concluded that diploma and Certificate student's progression is a major concern to guardians, sponsors, and university administrators, majority of the students enroll to the universities for their diploma or certificate courses to progress until to the degree level. The first objective of this study was, therefore, to identify significant predictors that can be used to predict the progression of these students to the next level of the study among the factors studied were motivation, economics, social and family, and demographic factors. The study found that economic factors were most significant. Availability of financial resources should therefore be given a priority by stakeholders to ensure an improved progression of the students.

From objective two results, it can be concluded that an Artificial neural network is an important machine learning platform. It has been previously applied in wide-ranging areas including face recognition, imaging, weather forecasting. However, it has not been greatly applied in the prediction of the progression of students. This is one area this study sought to address. The study achieved this objective by developing an artificial neural network model. The model takes in four factors and has an accuracy of about 78%. The model had a good prediction accuracy that was in agreement with past studies as described in the discussion of results. The model was able to show the significance of the study variables.

From objective three results it can be concluded that a good prediction model is measured by several metrics. The metrics for this study included mean absolute error, F- measure, Receiver operating curve, specify, and sensitivity. The model for this study showed good performance measures and therefore is deemed appropriate to apply prediction of progression of diploma and Certificate students. A minimum absolute error was obtained which was comparable with related studies as presented in the discussion of results.

5.3 Study Contributions

Artificial Neural network is one of the computation models, the model is inspired by the human brain. It is one of the recent advancements in the artificial intelligence field. The model has a biologically inspired simulation performed on a computer to cluster, classify and pattern recognize data. The present study applied Artificial Neural Network because of its advantages, it can perform tasks where linear programs cannot, the network learns and no reprogramming is needed and it can be implemented in any application.

The study contributes to the education industry by assisting in the identification of the factors of certificate and diploma student's progression, which has afflicted several Kenyan private

universities. Students pursuing diplomas and certificates are diverse, and the study looks into the elements that influence their progression According to the related study, numerous factors contribute to certificate and diploma student progression, which are student factors. The students' factors include motivation, demographics, economics, social, and family. The study showed that economic factors were most significant and thus availability of financial resources should therefore be given priority by stakeholders to ensure an improved progression of the students.

This research contributes to the body of knowledge by demonstrating how data mining algorithms such as ANN can be used effectively at Kenyan private universities to enhance education efficacy and, as a result, lead to goal-oriented decisions and policymaking in the area of education.

On the other hand, the study found that the artificial neural network method can be used in higher education institutions to anticipate students' development for better resource planning because it is simple to comprehend and provides higher accuracy. This was also clear in a literature study, which revealed that many other researchers had discovered that decision trees gave higher retention prediction findings.

Despite its success in diploma and certificate students' progression and other areas, ANNs have significant flaws that will need to be addressed in the future, such as model robustness, transparency and knowledge extraction, extrapolation, and uncertainty.

5.4 Recommendations

This study brought interesting findings that can be believed to bring a positive change if implemented. First learning of students touches the heart of everyone, be it, the government, the

parents, stakeholders, and even managers of the universities. Therefore, interesting comments can be borrowed from this study.

The identification of student predictor characteristics that are crucial in predictability accuracy is an important follow-up to this work. This is useful for three reasons. First, the government can intervene directly if they know why private university certificate and diploma scholars fail to proceed to the next study year and at the stipulated period frame, in conjunction with the university administrators as well as the Ministry of education. Future research could be performed to improve learning behavior and improve attributions. Academic management boards can provide data on learning behavior to help with input attributions for the predictive model, allowing for faster progression.

The government should provide more resources to help increase the number of students enrolling in the university and reduce the rate of deferment of students due to lack of financial aid. University management boards should pay keen attention to the matters to help discover leading causes of deferment and drop out of students. The outcomes of this research showed deferment of students' certificate and diploma students is on increase. A separate study can be done to help identify the main problem that leads to the increasing deferment of students.

The voices of students are critical components in educating educators about the complicated topic of dropout. This study demonstrates that early progress monitoring, academic support, and a safe and welcoming learning environment are all necessary for transformation to occur. If educators pay attention to the words of dropouts, they will be better able to prevent pupils from leaving. Students' voices are a valuable resource for educators looking for solutions to the dropout epidemic since their responses may result in stronger supports for those on the verge of dropping out. The findings suggest that progress should be monitored frequently that early communication between

all stakeholders is established, that academic support is increased, and that safe and inviting learning settings be created. These findings suggest that other at-risk students will have a bright future.

Learner dropout is influenced by a variety of personal factors, including age and gender. some familial considerations, such as marriage, have an impact on female students and may cause them to drop out of classes and are therefore not able to proceed with their studies. The student's age may play a role in dropout since younger students are more familiar with technology than older students, allowing them to use the online platform more simply. Modern education is becoming more and more digital. Stakeholders should therefore invest more in the training of the disadvantaged group of students instead of assuming that they will learn on their own. In many instances, some groups of students are not able to make presentations via digital platforms such as KENET and zoom. This makes them fail in their assessments. The management should therefore consider revising their assessment methods.

Performance metrics of developed models play an important role in determining the goodness of a machine learning algorithm. Although this study was successful in applying an artificial neural network, other methods still exist. They include support vector machines, Arima models, and decision trees. Other studies can be undertaken to compare the performance of the different machine learning algorithms in predicting student progression.

References

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of educational data mining*, *1*(1), 3-17.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61-75). Springer, New York, NY.

Demetriou, C., & Schmitz-Sciborski, A. (2011, November). Integration, motivation, strengths, and optimism: Retention theories past, present and future. In *Proceedings of the 7th National Symposium on student retention* (Vol. 201).

Dissanayake, H. U. (2016). Predicting Student Retention: A Comparative Study of Predictive Models for Predicting Student Retention at St. Cloud State University.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-37.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD* (Vol. 96, pp. 82-88).

Ian, H. W., & Eibe, F. (2005). Data Mining: Practical machine learning tools and techniques.

Jung, H. W., Varkoi, T., & McBride, T. (2014, November). Constructing process measurement scales using the ISO/IEC 330xx family of standards. In *International Conference on Software Process Improvement and Capability Determination* (pp. 1-11). Springer, Cham.

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003, September). Preventing student dropout in distance learning using machine learning techniques. In *International*

conference on knowledge-based and intelligent information and engineering systems (pp. 267-274). Springer, Berlin, Heidelberg.

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1-15.

Nghe, N. T., Janecek, P., & Haddawy, P. (2007, October). A comparative analysis of techniques for predicting academic performance. In *2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports* (pp. T2G-7). IEEE.

Niazi, AK, Kamran, M., Tariq, Z., Malik, M., Ilyas, K., & Zaman, MT Statistics Industrial Perceptive Sales Analysis using Data mining.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368-384.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, *46*(5), 30.

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).

Sisimwo, J. (2016). *Electronic resources and its application in collection development practices in academic libraries: the case of United States International University* (Doctoral dissertation, University of Nairobi).

Tinto, V. (2006). Research and practice of student retention: What next?. *Journal of college student retention: Research, Theory & Practice*, *8*(1), 1-19.

Sisimwo, J. (2016). *Electronic resources and its application in collection development practices in academic libraries: the case of United States International University* (Doctoral dissertation, University of Nairobi), 1(1),5-20.

Zheng, G., Dahl, J. A., Niu, Y., Fedorcsak, P., Huang, C. M., Li, C. J., ... & He, C. (2013). ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Molecular cell*, *49*(1), 18-29.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368-384.

APPENDICES

Research Schedule

Table 1

No	Activity	Duration (In hours)	Start date (proposed)	Proposed deadline	start date (Actual)	End date	Deliverables
1	Concept paper writing	11	02/05/2021	07/05/2021	14/05/2021	26/05/2021	A well-thought-out project concept
2	Writing a Proposal	31	16/05/2021	19 /05/2021	2/06/2021	21/6/2021	A proposal with documentation
3	Presentation of a Proposal	16min	31/7/2021	31/07/2021	31/07/2021		Presentation of the project proposal
4	Data Collection	35	16/08/2021	19/08/2021	19/08/2021		A list of users' criteria
5	Data Analysis	20	20/08/2021	20/08/2021	22/08/2021		needs that have been examined
6	Research Reporting	30	24/08/2021	24/08/2021	30/08/2021		Research that is well-presented
7	Research Project Submission	30	13/9/2021	13/9/2021	25/9/2021		A Clear documented project

4.2 Budget and Justification for Budget

Table 2: Budget

No	Item	Quantity	Description	Cost per unit (ksh)	Total Price (ksh)
1	Internet data				12000
2	Stationaries				4000
3	Laptops	1	Good specifications		75000
4	Flash disk	1	32GB	2000	2500
5	printer	1	desk jet(Hp)	6500	7000
6	Transport fee				5000
7	Proposal Binding	12			3500
8	Unexpected expenditures			3500	3000
Total					**102500**