AN ARTIFICIAL NEURAL NETWORK MODEL FOR PREDICTING ATTAINMENT OF THE 50:50 GENDER RATIO IN STEM COURSES IN KENYA

BY

CYNTHIA NABWIRE KIBET

MASTER OF SCIENCE IN DATA ANALYTICS

KCA UNIVERSITY

2020

AN ARTIFICIAL NEURAL NETWORK MODEL FOR PREDICTING ATTAINMENT OF THE 50:50 GENDER RATIO IN STEM COURSES IN KENYA

BY

CYNTHIA NABWIRE KIBET

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MSc. DATA ANALYTICS IN THE FACULTY OF COMPUTING AND INFORMATION MANAGEMENT AT KCA UNIVERSITY

OCTOBER, 2020

DECLARATION

Declaration by Candidate

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 the dissertation contains no material written or published by other people except where due reference is made and the author duly acknowledged.

Student Name: **Cynthia Nabwire Kibet** Registration No: **KCA 16/05106**

Sign: ____________________________ Date: **05/11/2020**

Declaration by Supervisor

I do hereby confirm that I have examined the master's dissertation of **Cynthia Nabwire Kibet** and approved it for examination.

Sign: Simon N. Mwendia

Digitally signed by Simon N. Mwendia
DN: cn=Simon N. Mwendia, o=KCA University, ou=College of Technology, email=smwendia@kca.ac.ke, c=KE
Date: 2020.11.10 13:45:54 +03'00'

Date: 10/11/2020

Dr. Simon N. Mwendia, PhD

ABSTRACT

In spite of the existing educational policies on gender and several other interventions that are aimed at empowering the girl child, education is not globally available and gender inequality is still a major problem world wide. Many nations are now concerned that fewer girls are going to school in comparison to their male counterparts, and also that males have higher participation and learning achievements than girls ,more particularly in Science, Technology, Engineering and Mathematics (STEM) subjects and courses.

STEM education is one of the pillars behind Kenya’s Vision 2030, which aims to turn the country into a newly industrializing, middle-income country providing a high quality life to all its citizens by the year 2030, in a clean and secure environment. STEM education is expected to provide learners with the knowledge, skills, attitudes and behavior required for inclusive and sustainable societies.

Graduation trends from the Commission for University Education (CUE) show that more than 30 % of graduating students each year are awarded commerce degrees or one of its other hybrids in business studies, 20 % graduate in education arts and another 20 % in other non-STEM courses. In a study conducted by Dr. Eusebius Juma Mukhwana, (Mukhwana et al., 2016) a former deputy commission secretary in charge of planning and research development at CUE, 74 % of all university students are enrolled in business, education arts and humanities. This leaves only 26% of the students in STEM.To make a bad situation worse, gender disparity within STEM fields is in favor of males. Female students represent only 35% of all the students enrolled in STEM-related fields of study at higher learning levels according to a study conducted by UNESCO through the ‘STEM and Gender advancement’ project in 2015. This disparity in gender is startling, moreso since careers in the STEM fields are now being commonly cited as jobs of the future that are being used, and shall continue to be used to drive innovation, inclusive growth and sustainable development. The female gender is held back by societal norms, biases and prejudice, and expectations that influence the quality of education they receive and even the subjects they choose to study at higher learning levels.

Following the above findings, the Kenyan government and stakeholders in the education sector have put measures in place in a bid to bridge this gap in gender. The main aim of this study therefore was to develop a model that would predict when the ratio of males to females in STEM will be 50:50 and further determine what measures can be put in place by government or society, to promote the interest and engagement of girls in STEM.

An Artificial Neural Network (ANN) was applied as the predictive data mining method to come up with the model. Exploratory data analysis was performed on the data and a regression model was built inorder to achieve the main objective of the study.

The study utilized the data in the repositories of the Kenya Universities and Colleges Central Placement Services (KUCCPS) for the years 2014, 2015, 2016, 2017 and 2018. The method of data collection was ‘Use of existing data as a data collection method for machine learning’ (Yuji et al., 2019). After the model was built, it was evaluated to determine its accuracy.

Key Words: Predictive Data Mining, ANN, Python, Educational Data, STEM, Gender reforms in Education

ACKNOWLEDGEMENT

I would like to acknowledge and thank my supervisor Dr. Simon Mwendia for his professional and intellectual guidance. His input and academic contribution have mentored me to excellence. I would like to extend my gratitude towards all my course work lecturers at the Faculty of Computing and Information Management (FOCIM-KCAU) for their hard work and academic nurturing that has given me the knowledge base for this research. I truly appreciate you.

I would also like to acknowledge my employer, the Kenya National Examinations Council (KNEC), who made it possible for me to start and complete my Master’s program and my colleagues at work, who have also given me ample time to work on my Masters’ degree.

Special gratitude to my wonderful family who have supported, encouraged and prayed for me during the period of this project. This project is every bit yours as it is mine.

Above all, I would like to thank the almighty God, without whom no part of this project would have been possible.

DEDICATION

I dedicate this work to my loving parents Mr. & Mrs. Kibet, who have been exceptional in their guidance and support. I sincerely thank you for your incredible contribution in every aspect of my life and for giving me a headstart into scholarship.

ACRONYMS

ANN: Artificial Neural Network

CEMASTEA: Centre for Mathematics, Science & Technology Education

CUE: Commission for University Education

EDM: Educational Data Mining

GDP: Gross Domestic Product

KDD: Knowledge Discovery in Databases

KNBS: Kenya National Buraue of Statistics

KUCCPS: Kenya Universities and Colleges Central Placement Service

MLP: Multilayer Perceptron

NACOSTI: National Commision for Science, Technology and Innovation

NASA: National Aeronautics and Space Administration

NSF: National Science Foundation

SDGs: Sustainable Development Goals

SMET: Science, Mathematics, Engineering & Technology

STEM: Science Technology Engineering & Mathematics

OPERATIONAL DEFINITION OF TERMS

Algorithm: A process or set of rules to be followed in calculations of other problem solving operations especially by a computer

Artificial Neural Network (ANN): the piece of a computing system designed to simulate the way the human brain analyzes and processes information.They have self-learning capabilities that enable them to produce better results as more data becomes available.

Data Lake: A single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning.

Data Mining: Analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Educational reform : A transformation plan and movement, which tries to bring about a systematic change in educational theory and practice in the level of basic or higher education (HE) in a given community and society.

Gender Equality: The state of equal ease of access to resources and opportunities regardless of gender, including economic participation and decision-making; and the state of valuing different behaviors, aspirations and needs equally, regardless of gender

Model: A simplified representation used to explain the workings of a real world system or event

Multilayer Perceptron: A class of feedforward artificial neural network (ANN) that utilizes a supervised learning technique called backpropagation for training.

Prediction: A statement about the future. It’s a guess, sometimes based on facts or evidence.

Sustainable Development Goals: are a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity by 2030.

50:50 A gender ratio of one to one

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

This chapter covers the background to the study, which includes the origin and history of STEM education to its contemporary state, the importance of gender inclusivity in STEM education and careers, status of STEM in Kenya and overall trends in the same. The chapter also outlines the problem statement, objectives of the study, research questions, hypotheses, scope, limitations and assumptions of the study.

1.1 Background of the Study

1.1.1 History of STEM Education

STEM, which is anacronym for Science Technology Engineering and Mathematics was originally called SMET but later changed in order for the acronym to sound better on the tongue (Sanders, 2009). It was an initiative created by the National Science Foundation (NSF) to provide students with problem solving skills that would make them critical thinkers and eventually more marketable in the workforce ,in comparison to their counterparts who would not have taken STEM .It is perceived that a student who takes part in STEM education at lower learning levels would have an edge over their classmates, if they happened not to go past secondary school level or an even greater advantage if they decided to particularly pursue a course in STEM ,at college level. (Butz et al., 2004). STEM education was also propelled by a number of events in history including the Morrill Act of 1862 that was behind the rise of land grant universities whose initial centre of attention was training in agriculture but soon after, engineering base training programs were created. (Butz et al., 2004). The more land grant institutions got established, the more STEM Education was taught and in the end got incorporated into the workforce. Other historical events that led to further growth and flourishing of STEM Education include the Second World War (WWII) because of the creation of weapons which required a lot of Science, Technology & Engineering, and the launch of the Sputnik satellite into space by the Soviet Union. The Sputnik lauch later challenged the USA into the creation of NASA which has continued to do more research and sponsor more students into studying the science of space.

1.1.2 Contemporary Aspects of STEM Education

In contemporary, the STEM agenda has been propelled through the Sustainable Development Goals (SDGs). SDGs are an assemblage of 17 goals that are global and are designed to help the world to end poverty, to ensure a sustainable future for everyone by ensuring all people (regardless of race ,gender,socioeconomic class) enjoy prosperity and peace by the year 2030 and to protect the planet we live in. The SDGs were set in the year 2015 by the UN General assembly with the intention of being achieved by the year 2030. Through the SDGs, the world collectively addresses environmental challenges, social and economic concerns that people are facing in the present day and age. SDG number four is on education and it is to ensure equitable and inclusive quality education and promote lifelong learning for all. SDG five is on gender equality and it is aimed at achieving gender equality and empower all women and girls. SDG four and five are of key interest to this study. STEM education is also an instrument to achieving other SDGs like ending hunger and getting to grips with climate change.

1.1.3 The 50:50 by 2030 Agenda

The United Nations through UN Women came up with an initiative called "Planet 50-50 by 2030: Step It Up for Gender Equality”. The initiative is a call to governments to commit to dealing with the challenges that are holding the female gender (girls and women) back from reaching their full or unrealisesd potential. The initiative is essential to the achievement of the 2030 agenda for Sustainable Development in providing a thorough ground plan for the future of the world, its resources and the people. This is due to the fact that the empowerment of girls and women is fundamental to the achievemet of the SDGs. The Step It Up vision anticipates a high society and planet where all the women and girls have equal rights and opportunites by the year 2030. The initiative asks nations and governments to commit to closing the gap in gender that results to inequality, by means of putting necessary laws and policies in place, to taking national action plans and making adequate investments.Kenya as a nation is taking baby steps towards achieving this agenda by putting measures in place, aimed at creating gender equality in STEM.

1.1.4 STEM Education in Kenya

The Kenyan government has put policies in place to aid in propelling STEM courses and to also bridge the gender gap within STEM. Such policies include: The Kenya Vision 2030’s Second Medium Term Plan (2013-2017), The Science, Technology and Innovation (ST & I) policy and strategy (2008), among many others.The campain for STEM in general has existed for quite a while but the most recent and notable one followed the education reforms by government through the Ministry of Education and the (then) Cabinet Secretary (CS) for Education, Dr. Fred Matiangi. The reforms included introduction of STEM in public schools in 2016, where the CS launched launched 47 extra-county schools .The wide gap between demand and supply of STEM related skills in Kenya are the reasons behind the launch of the model schools. The reforms in education were also effected in national examinations (KCPE and KCSE) to curb cheating and irregularities, which led to a decline in the number of students who attained mean grades ‘A’ and ‘B’. The level of achievement of girls in Kenya Certificate of Secondary Examination (KCSE) determines whether or not they shall join university and further whether they shall be admitted into STEM courses which require generally higher scores in Sciences and Mathematics. There are 70 universities in Kenya, 38 being private and 32 public. (Mukhwana et al., 2016). The Education sector in Kenya has grown post-hastely in the previous five to ten years.

Despite the general increase in number of students enrolled into universities, there are notable weaknesses in the relevance and quality of courses offered especially at undergraduate level because the industry and employers are still expressing the need for more STEM graduates into the workforce. The need is apparent, for universities in Kenya to emulate global trends specifically in the labour market so as to maintain their relevance (Mukhwana et al., 2016). University education in Kenya is not giving rise to enough graduates in STEM courses since currently, only 13% the whole graduate population are holders of STEM based degrees while only 30 % of graduates in STEM only are female. There is a shortfall in human resource in the fields of medicine and engineering yet universities continue to train more than the required numbers of students in Arts, Business and Humanities. Students in other non-STEM courses comprise of 74 % of all students enrolled (Mukhwana et al., 2016). This observation should be a wake up call as we get closer to 2030 because the Kenya Vision 2030 blueprint highlights STEM based careers and fights to ensure that the nation of Kenya becomes a newly industrialized middle income nation by 2030.

1.1.5 Application of ANNs to the Prediction Problem

This research project uses an Artificial Neural Network (ANN) model to predict when the ratio of males to females in STEM in Kenya will be 50:50. An Artificial Neural Network(ANN)model is an intelligent system and is used to solve complicated problems in many applications such as optimization, prediction, modeling, clustering, pattern recognitions, simulation, and others. The ANN structure consists of three layers: the input layer which has collected data, an output layer which produces computed information, and one or more hidden layers suitable to connect the input and output layer. A neuron is a basic processing unit of a NN and performs two functions: the collecting of the inputs and producing of the output. Each input is multiplied by connection weights, and its products and biases are added and then passed through an activation function to produce an output as shown below:

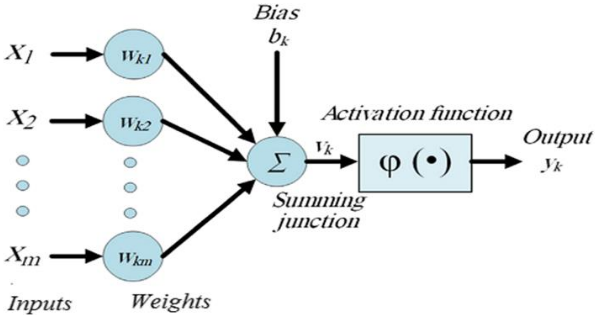


Figure 1: Structure of a Neural Network

Where in this case,X_1=Interest in STEM, X_2=Numeracy Skills,X_3=Linguistic Skills,X_4=Cognitive traits,X_5=Parental Beliefs,X_6=Socioeconomic status,X_7=Level of School and X_8=County of Origin. The inputs ($X_1 \ldots X_8$) are predictor variables which are derived the from the collected data and the Neural Network assigns weights to each, based on feature importance or how key the variable is in the achievement or attainment of the 50:50 ratio. The weights $Wk_1 \ldots Wk_8$ are determined by Neural Network. The activation function to be applied by this model is the Re Lu activation function because it allows for back propagation. It is also computatively efficient, thus allowing the network to converge very quickly, hence producing the output within very little time.The Back propagation algorithm is the most method used for training feedforward ANNs which is dependent on the gradient descent optimization technique. Back propagation is a technique based on supervised learning that is used for training Neural Networks, and it is processed to learn samples iteratively. Therefore, it compared the output predicted for each input with the actual value. To minimize the mean squared error between the network estimated and the measured data, the weights are adjusted for each training model. The network learns by :i.Initializing all weights and biases and normalizing the training data ii.setting the values of learning rate and momentum coefficients, iii commputing the output of neurons in the hidden layer and in the output layer, iv.Computing the error by comparing the actual and predicted values v.Updating all weights and biases vi.Repeat steps iii to v for all training data until the error converges to limit level, to produce the expected output of the year when the ratio is expected to be 50:50

ANNs are more valuable than traditional model-based methods in solving this research problem because they are data –driven and self –adaptive in the sense that there are very few apriori assumptions about the model for the problem under the study. It is for this reason also that ANNs are more attractive because of their capability to learn from experience, the experience in this case being the data available. (Zhang, 2009)

1.2 Statement of the Problem

Target number five of the Sustainable Development Goal (SDG) 4 is aimed at putting an end to gender inequality in education and ensuring equity in the access to all levels of education ie universities,colleges and technical training institutions for people from all walks of life; be they those living with disability, the vulnerable or indigenous people. The goal is to be achieved by the year 2030. SDGs were officially adopted in 2016 but a few years down the line, studies still continue to show persistence in the gender difference at all education levels and particularly within STEM (Migosi ,2018). Disparity in gender becomes clearer as learners get liberty to select subjects particularly in high school. The phenomenon gets worse even as learners thereafter proceed into university. All over the world, there is low access and participation of female learners in STEM disciplines. This is evidenced by research findings (Mbirianjau et al., 2011) that show that only 30% of leaners in STEM courses in Universities are female.

As a response measure to this problem, the government of Kenya and institutions of higher learning have established intervention measures with the aim of gender equity within STEM. Such measures include but are not limited to: formulation of policies that promote gender balance in education, follow up to ensure transformation of policy to action, affirmative action, putting up technical institutions in every subcounty, financial aid both to institutions and to learners in the form of HELB loans and burseries (Kapinga, 2010).

In spite of the intervention measures and policies on gender equity being in place since the year 2000, there is still corroboration of very few female learners getting admitted into universities to study STEM courses. The same institutions of higher learning paint the truest picture in terms of government’s achievement in the set intervention measures.

The institutions have also become very rich in data that can guide stakeholders better, in decision making concerning matters enrollment into and retention of the girl child in STEM. It is however notable that as much as this data exists in large numbers, higher learning institutions are still faced with the challenge of making smart managerial decisions or improving on the already existing ones. The process of decision making gets more and more complex with changing generations and dynamics.

Educational institutions are in need of efficient technology to aid in decision making and to assist in setting of new and better strategies in the matter of gender bias in STEM education.

Managerial systems in the universities can address the challenge of gender bias by being provided with new knowledge that relates to the processes of admission and retention of students.

A number of studies in educational data mining have been carried out in a bid to address the matter of enrollment of students. The studies have put to use different data mining techniques to gain knowledge that has aided the processes of admission. This knowledge has been particularly valuable in prediction of enrollment hence better guiding the selection and admission process thereof. However, there is lack of a prediction model that can be used to determine when the 50:50 gender ratio of enrollment in Higher Education Institution, particularly in STEM courses will be attained. (Wanjau et al., 2016). Latest Government reports from the Ministry of Education in Kenya also continue to show use of traditional statistical surveys to determine gender ratios in STEM education; further showing the apparent need of a scientific model for this case. (Basic Education Statistical Booklet, 2017)

The main objective of this study therefore, is to build a prediction model to determine when the 50:50 enrollment gender ratio in STEM will be achieved, using data mining. Based on the predictor variables used in the model, the study will further help to determine what additional intervention measures can be taken to increase the enrollment rate of the girl child into STEM courses.

1.3 Objectives of the Study

1.3.1 Main Objective

To develop a prediction model for determining when the 50:50 gender ratio of enrollment into STEM courses, among government sponsored students in Kenya will be attained.

1.3.2 Specific Objectives

i. To determine the factors that lead to low enrollment of females into STEM courses

ii. To establish appropriate data mining methods used to build the predictive model for enrollment into STEM ratios by gender

iii. To develop a model for predicting when the 50:50 gender ratio of enrollment of government sponsored students into STEM courses in Kenyan Universities will be attained.

iv. To evaluate the model for predicting attainment of the 50:50 gender ratio in STEM

1.4 The Research Questions

i. What are the factors that lead to low enrollment of females into STEM courses?

ii. What are the appropriate data mining method(s) that can be used to develop a model for predicting gender ratios of enrollment into STEM?

iii. What the appropriate ways of developing a model for predicting when the 50:50 ratio will be attained?

iv. How will the model for predicting the 50:50 gender ratio in STEM be evaluated?

1.5 Justification of the Study

The research findings from this study have the potential of providing useful insight to the Ministry of Education in Kenya, stakeholders in the education sector in Kenya, institutions of higher learning, domain experts and the personnel in charge of policy formulation. Such insight includes, knowing which measures of the ones put in place by government and stake holders to curb low enrollment of girls into STEM are the most effective, knowing the rate at which the gap of gender inclusion in STEM courses in Kenya is being bridged, and knowing the current gender ratios of students enrolled in STEM courses at universities in Kenya. The study at large is bound to increase enrollment of the girl child into STEM courses at higher learning levels which will in turn lead to a better economy. Stakeholders should also be able to put more emphasis on the importance of STEM education and inclusion of the girl child in the same.

Data mining technology has not been widely explored in Kenya in the field of education and therefore the findings of this study also have the potential of adding knowledge to that field of research. The data that is available yet not being fully utilised in institutions like the Kenya Universities and Colleges Central Placement Service (KUCCPS) will be put to better use in decision making on matters education. The findings will encourage more researchers to explore this area of knowledge and lead to new findings and innovations to make human life better.

1.6 Motivation of the Study

STEM education and gender inclusivity for economic development are under SDGs 4 and 5. STEM education is a big propeller of economic development in the world and therefore Kenya as well. Gender inclusion in the STEM fields is equally a matter of concern in the whole world. The deadline for achievement of vision 2030 and the SDGs is also fast approaching. Governments together with key stake holders are investing a lot of funds and resources in a bid to ensure the goals are met. A lot of other intervention measures to achieve this goal have also been put in place and there is need to assess the effectiveness of these measures through knowing when the goals will be achieved. To add to this, the increase in number of undergraduate students getting admitted into STEM courses does not contend with the demand for STEM professionals in the labour market yet, evidence points at decline in interest and preparedness of high school graduates for STEM courses.

In light of the increasing need to draw more girls who have graduated high school into this particular subject field of study, research dedicated to understanding the factors that influence academic choices of students in regard to STEM courses at university level , was crucial.

A lot of data has been collected with regard to existing differences in gender within STEM courses, in terms of enrollment and retention but less has been done to dig deep into this existing volumes to come up with patterns that will aid in achievement of SDG 4 and 5. The current existing methods of determining trends in enrollment ratios are the traditional statistical methods which are quite rigid and less accurate and dependable than this scientific modelling method.

This study came up with a predictive model that will increase efficiency in monitoring of enrollment by gender. The findings of this study are to hereafter help make the Kenyan economy globally more competitive while pushing for gender reforms in the STEM industry.

1.7 Scope of the Study

This study was focused on data that exists for all government sponsored students at college and university levels. This is because these groups of students studied and merited at their respective levels therefore they can be based on to paint a realistic predictable future. The data was also readily available for the years 2014,2015,2016,2017 and 2018. The study was conducted on data that has already been collected on this target group and exists in the repository of the Kenya Universities and Colleges Central Placement Service (KUCCPS). The research examined trends in intake into STEM courses by gender over the years 2014-2018. The research design that was used is exploratory analysis, through data mining processes.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This study was aimed at building an ANN prediction model that would determine when the 50:50 gender enrollment ratio among government sponsored students in STEM will be attained in Kenya. Chapter 2 of the study primarily focused on the available literature in order to establish what areas in this domain have been covered thus far and how they relate to this study area. The chapter also revealed the gaps in knowledge that were picked from the review of literature, and how this study filled some of the gaps.

2.2 Theoretical Review on the Gender Gap in STEM

Careers and subjects in STEM have shown evidence of supporting the development of a cohort of people who are problem solvers, thinkers and collaborators. Despite all pointers of research towards the importance of STEM in terms of social and academic aspects, discrepancies in gender wthin STEM still persist. Athough research findings show that the gender gap has been in existence since the introduction of STEM, this gap has overally decreased across the generations. To be more specific, there has been an increase in the number of women who have earned graduate and post graduate degrees in STEM fields from the year 1990. (Hill et al., 2010). It is however a surprising fact that in as much as girls may perform better than boys at the same learning level in STEM, girl children tend to lose interest at a faster rate and consequently do not pursue STEM at higher learning levels such as in highschool or at degree level and even at career level. Further to this, studies show that in the year 2008, 41% of students who were joining college and had plans to major in engineering and science were men yet only 30% were women. The question of interest to many professionals and researchers in the field of STEM is 'Why is there a gender gap?'

Findings in research further show that the gap in gender doesn't only exist at the level of students choosing their majors at higher learning levels but the same gap is relayed to the places of work. Studies also show that of all the employed members of faculty in STEM education, only one in five members are female. In workplaces that are specific to STEM, the male gender still represent a higher percentage of the employees, than females. For example, the Society of Women Engineers states that in 2003 approximately 20% (~12,000) of new engineers were women, compared with about 80% of men (~49,000), however this is an increase from past generations (Crawford, 2012).

Pegged on this findings, there are patterns in society that advance men in STEM and STEM-related fields while the same patterns discourage or leave behind women. Whereas this occurrence has been observed and studied for a long time, many theories exist to understand the gender gap in STEM.

Scientists, professionals and other relevant experts that have questioned this phenomenon have found that since historical times, men have been given incetives to excel in STEM and STEM-related subjects wheras women have been given incentives to perform well in other non-STEM related areas like languages and literature. (Burton, 1986). To uderpin this research, is the certitude that outside characteristics like teachers,society and parents also influence likes and dislikes for STEM based on gender, which is called gender socialization (Leaper et al., 2014).

Through out all the generations there has been the socialization aspect of women not being motivated and encouraged to take up interests that are STEM-centered. After the obvious recognition of the phenomenon of having significantly lower numbers of women in STEM than men,the following section looks deeper into the justification of this phenomenon and how educators can increase the interest and curiosity of girls in STEM.

2.2.1 Theories

The low representation of females in the fields of STEM is traceable to many causes. There are people who believe that it is a direct outcome of the socialization practices and stereotypes that prevail in many societies across the world, centered on the submissiveness of the female gender and dominance of males. To back up the idea of stereotypes being engrained in society and the issue of socialization, there are also some notions that are propelled in childhood like boys excelling in maths while girls being perceived as good in the kitchen and home keeping activities. (Gunderson et al., 2011 and Regner et al., 2014). These practices of socialization feed into the concept of stereotypes that in turn threaten the performace of girls in STEM. (Shapiro et al., 2012)

There are other groups of researchers who believe that the gap in gender in STEM is not really as a result of socialization and stereotype threat practices; and that it is instead linked directly to the role that is played by peer groups in a student's academic experience. (Crosnoe et al., 2008). According to this idea, students like to be part of a peer group and as a result, they like to engage in activities that their peers are also engaging in or if at all the activity is different, then they prefer that the activities be approved by their peers as well.

The last theory that attempts to explain the gap in gender within STEM also bases on stereotypes but more of stereotypes at work, in STEM careers (Cheryan et al., 2015). The theory goes ahead to focus on characterstics and personalities that have been labelled to belong to people in the fields of technology and engineering. Some of the characteristics include introverted personality and social awkwardness. (Cheryan et al., 2015). Since most women are naturally more social and outgoing, they tend to shy away from STEM and STEM related proffesions because of these stereotypes that have infiltrated into the views of society.

The First Theory; Gendered Socialization

There is a big difference between how girls and boys socialize in many societies in Kenya. The difference in socialization and behavior is largely brought about by the ideas that are preconceived about gender roles. Gender roles are sets of "behaviors, attitudes, and personality characteristics expected and encouraged of a person based on their sex." Studies show that boy children are normally brought up to conform to the roles that are profiled for males while girls are equally brought up to conform to roles that are profiled for females (Spark Notes, 2006). Some experts may not agree that these gender stereotypes are based on differences in genetics but the socialization differences are clear from an early age.

Researchers have continually come up with evidence to show direct relationship between gender stereotypes and STEM proffesions therefore implying that the theory of gender roles and socialization can be partly used to explain the gap in geder within STEM. According to (Dasgupta et al., 2014) women have been observed to leave the STEM pipeline even before they officially enter the STEM carreers. Through the pipeline, some women are lost, who could have been part of the next generation of scientists, technologists, engineers and matheticians. Research has futher shown that one of the contributing factors to women dropping off along the STEM pipeline is women being bombarded with negative stereotypes and socialized ideas especially about the subpar women's abilities in mathematics. (Gunderson et al., 2011)

It is an ugly phenomenon because it has been observed that the stereotypes and mentalities are communicated and inculcated in the minds of girl children when they are very young; and this is done by their teachers and parents yet sometimes, they do so unconsciously. It does not really matter whether the mentalities and stereotypes are communicated consciously or unconsciously because these same stereotypes still shape the attitude of the girl children towards mathematics and in the end, causes their interest in STEM fields to decline. It can therefore be strongly argued that stereotypes and stereotype threats are the major reason for underrepresentation of women in the field of STEM.

Similarly, socialization also occurs in the contexts of family. Parents and gurdians and the people that children are raised around do have a lot of influence on motivation and achievement of the child in school and in similar settings like those structured around topics related to STEM. (Partridge et al., 2008). Eccles (2014) also analyses and describes the influence of family on gender differeces within STEM. This research describes how the perceptions and beliefs of parents influence the activity choices of children and their outcomes. There are many ways through which parents or gurdians can influence their children, such ways include but are not limited to; the toys they make available to their children, the experieces they expose their children to such as visiting engineering plants and the likes and even the nature of television programmes that the children are allowed to watch. The research futher found that "parents may make causal attributions and these differences help to explain children's abilities and interests"

Overall, the research findings of Eccles (2014), alongside other studies have displayed the theory of gender socialization which sidelines women and girls in STEM. The sidelining diminishes the voice of women and ther legitimacy in the classroom and worforce. (Regner et al., 2014).

The Second Theory; Peer Groups

Whereas socialization plays a role in leading many girls to fall out of the STEM pipeline, other separate studies are centered on a different theory on peer relationships and pressure that is usually felt during the years of adolescence. It has been observed that a girl was more likely to take up mathematics or a course in mathematics because their close friend took it up or excelled in it, rather than because their course mate took it up. This relationships and associations appeared to be stronger as one got closer to graduating high school yet weaker amongst adolescets who had a prior record of failing in school. However girl children continued to consistently show all these observed patterns. (Crosnoe et al., 2008)

According to this research, most students in middle school are usually at the adolescent stage which is generally a difficult stage. At this stage, students want to perform like their peers and this trait could be linked to the group acceptance which is very important to adolescents. This research is also supported by others of similar nature such as You (2011) that found “that peers have an important influence on the behavior and development of adolescents” (p. 829). Essentially, “the child’s acceptance within the peer group is one of the key measures of positive/negative school experiences. Perceived support from peers can give students a sense of motivation and help students see the importance of pursuing academic success such as STEM related courses.”

In addition to this, a research by (Hoorn et al., 2014) examined the influence that peers have on the behavior of adolescents. In as much as this study mainly focused on antisocial and prosocial behavior, its findings back up the idea that at the adolescent stage, peers easily influence each other therefore causing themselves to be be vulnerable to peer feedback.   The behavior of peers was observed to be more prosocial when peers provide positive behavior. In the contrary though, when peers do not give feedback or even when they give antisocial feedback, positive behavior is observed to decrease. When children attain the stage of adolescence they have been observed to be more depedent on the judgement of their peers to decide what to do and to know how to engage at school and in the community which they live. They tend to engage more in what is deemed as cool by their peers rather than what is not. As a result of this, when few girls get enrolled into and stay in STEM courses the peer feedback through words or inaction can be perceived to be negative. To further back this idea up, (Leapers et al., 2011) found that the motivation of girls in STEM subjects such as mathematics and science, in the course of their adolescent years is positively associated with peer support. Overally there is no doubt that peer groups do actually have an impact on whether or not one will succeed. It is in the same breathe that peer groups will also cause many girls to turn away from STEM, if their peers do not approve of or succeed in STEM. In conclusion, peers have an important role to play in the engagement or disengagement of one, in STEM.

The third Theory: Stereotypes of STEM Professionals

To add on to the two preceding theories of stereotypes and gender socialization, there is another group of researchers from a different school of thought that attempt to explain the phenomenon of fewer women in STEM. This theory focuses on the represetation of views of the culture that surrounds careers in STEM. One of the studies that supports this theory gave a proposal “that students’ stereotypes about the culture of these fields—including the kind of people, the work involved, and the values of the field—steer girls away from choosing to enter them (Cheryan et al.,2015). ” These reserchers examined the stereotypes that are associated with professionals that are in the STEM fields. From their findings, they concluded that the STEM umbrella is more male-oriented and the proffesionals in the field also embody the characteristics of social isolation.

The phenomenon of social isolation as a characteristic of people in STEM is linked directly to the first theory of gender socialization. Most girls are brought up with the ideology that to be antisocial, or self isolation is not a desirable quality therefore not valued within the female gender. Girls are therefore raised to be social, to be interactive and even to be pleasers. Further to this, it was observed that even from historical times, girls and women have not been taught that they are born brilliant whereas their male counterparts are raised with the understanding that they are born very brilliant, and even more brilliant than women in most instances. (Cheryan et al., 2015). Therefore, girls are bound to be more interested in carreers related to STEM, if stereotypes associated with proffessionals working in STEM is changed.

Further to this, Steinke (2017) conducted research on how girls at adolescent stage grow identities that are either positive or negative, around STEM that could later have an influence on the choices they make later in life with regard to their careers and proffesions. The researcher suggested several ways of changing the perceptions of girls towards STEM. One such way was through incorporation of images of women who have succeeded in STEM, in several media avenues such that girls are aware that these women do actually exist. “Popular media have played a crucial role in the construction, representation, reproduction, and transmission of stereotypes of STEM professionals” (p. 716). Her research described the kind of images that embody gender based stereotypes, which are being portrayed to girl children through the different media.

To add on this, research that is more recent has found that if women fight to change the stereotypes that are associated with the female gender and develop the confidence of mind, the gap in gender is likely to be bridged at a faster pace. Dr. Anna Powers who is a leader in the domain of STEM provides women with access to fields that foster innovation studies, lectures and empower girls and wome to focus on STEM subjects. Dr. Anna advocates for girl children to embrace STEM through encouraging confidence, goal setting and fighting stereotypes that are rooted in history. (Kerpen, 2017).

2.2.2 Factors influencing the gender gap in STEM

The above theories show that there are many different factors that are complex and overlapping, which influence the participation, progression and achievement of girls in STEM subjects and courses. In a bid to explain and better understand these factors and how they are interrelated, an ecological framework has been adopted that presents the factors at four different levels namely ; individual, family, institutional and societal levels.

Individual level: They include biological factors which according to psychology are called a persons nature. Examples are; brain structure, genetics, hormones, cognitive traits like linguistic and spatial skills, individual behavior, abilities and skills. Individual factors also include psychological aspects like self-efficacy, interest and motivation in these subjects.

Family and peer level: These maily comprise of the close people and environment in which one is brought up. In psychology they are mainly factors that comprise of nurture. Such include the beliefs of ones parents or guardians, their socioeconomic status and other household factors and the peer influences on one.

School level: These are factors in the environment of learning a teacher’s beliefs, expectations and profile, the school curricula, learning resources and materials, the methods and strategies of teaching and student-teacher interactions, the practices of assessment and the overall environment at school.

Societal level: These comprise of outside factors that mostly the student themselves or their parents have little to do with. Most of the factors here are can be controlled by people at very high levels like the government and the media. Such factors include; gender equality and gender stereotypes.

2.2.3 Current STEM gender ratio in Kenya

All over the world, the number of women who are undertaking degrees in STEM is very small. The numbers are even more dismall in Africa given the economic level of many nations in the continent. The imbalance is greatest in engineering courses with the lowest percentage of women. As of the year 2010, only one in every four students was female. The nation of Guinea had the least percentage of women at 5.8 %. The percentage translates to one woman in every seventeen students. Lesotho and Cape Verde did very well with 55.7 % and 52.3 % of the students being women, respectively. (Mbirianjau et al., 2016)

In the country of Kenya, there is a clear gender disparity in STEM courses with women being only 30% of the students in STEM (Wandiri, 2009). The resultant ratio of males to females in STEM is therefore 70:30. Very few women participate and even fewer women complete their studies to graduation level. Even for the few women who complete their studies to graduation level, their scores are generally lower compared to those of males. The same situation holds true across many other countries; both developing and developed. The number of female students who graduate from STEM continues to decline as one moves higher up the education ladder, to higher levels of learning. The phenomenon of females reducing even more as one goes higher in learning levels is called “the leaky pipeline” This is attributable to all the factors discussed in the section before.

In the country of Kenya, the rate of female participation in public universities is less than 30%, despite the educational gender policies and interventions that have been put in place. According to Wandiri’s study that sought to document the participation of females in STEM courses in Kenyan universities, there are many factors that are at play. Such factors include family responsbility, sexual harrasemet and gender stereotyping. Gender biases were revealed, not only in enrollmet and completion but also in policies that tend to favour males in STEM discplines.

The current ratio of males to females in STEM is 70:30. The ideal ratio is supposed to be 50: 50 accordig to the United Nations 50:50 by 2030 step it up agenda. This study therefore seeks to know when the 50:50 ratio will be achieved in Kenya given the current rate of ratio growth.

2.3 Emperical Review

2.3.1 Data Mining Process

Data mining is a process that involves various steps in order to end up with the knowledge generation. Knowledge Discovery in Databases (KDD) is described as a nontrivial process of recognizing valid, novel, potentially useful and finally understandable patterns in data (Fayyad et al., 1996).

To expound on the definition, data were various sets of valid facts which are accessible in electronic disposition. Patterns are a group or series of data that recur in a way that is recognizable. Usually, the patterns are represented in the form of models that are exhibited or conveyed by a given language as subsets of the data. These patterns should have sound basis in logic by being factual such that, any data that is new can be modeled to produce similar patterns. The process of data mining involves several steps from the preparation of data to the enhancement of knowledge, all recurrently used untill the results that are being looked for are obtained or achieved. Nontrivial indicates that there ought to be a sort of inference computation so as to differentiate it from traditional methods of computation of values. (Mittal et al., 2016)
This procedure is pooled and iterative in nature and also comprises of several steps together with decisions reached by the user in attempts made at each step to finish or complete a given specific task of discovery, each one realized or accomplished by the application of a discovery method.

The terms Data Mining and Knowledge Discovery in Databases (KDD) are at times interchangeably used yet conversely, many people consider data mining to be a vital step of Knowledge Discovery in Databases (KDD) that results to beneficial models or patterns for data.

The various processes for data mining are displayed in figure 1 that follows.

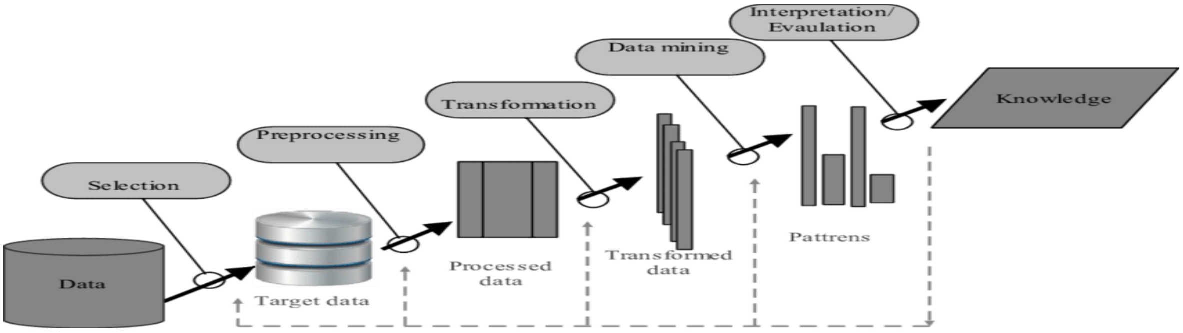


Figure 2: Data Mining Processes (Yongjian Fu, 2011)

Selection: Selecting pertinent data for the process of mining, from sources that are distinct.

Preprocessing: Since data is collected from many various sources, it tends to have inconsistencies. In order to get rid of these inconsistencies, various activities are conducted at the preprocessing phase whereby data that has errrors is removed or corrected. Noisy data or discrepancies also get removed and the data is combined since it is initially from individual distinct sources.

Transformation: Here data is transmited into appropriate form for mining. Feature selection, sampling, aggregation may be used.

Data Mining: At this stage, patterns in the data are extracted. This step is also very significant because algorithms used in the entire mining process are selected based on the patterns observed in the data.

Interpretation and Evaluation: To recognize and infer mining results or patterns into knowledge by elimination of redundancies and irrelevant patterns. Here, an assortment of visualization and GUI stratagems are used for transforming the advantageous patterns into the human understandable terms.

2.3.2 Tasks in Data Mining

The tasks of data mining are classified into two: Descriptive and Predictive. The two categories are also regarded as main objectives or goals of data mining. Further to this, there are six primary functions of data mining which are: Classification, Clustering, Dependency modeling, Deviation detection, Dependency modeling and Summarization (Fayyad et al., 1996).

Under the predictive data mining category, lies the functions of regression, anomaly detection and classification. Likewise, under descriptive data mining, lie the functions of dependency modeling and clustering.

Predictive models used for making forecasts usually utilize given variables whose values are known in a dataset in order to come up with predictions of values that are unknown for another given variable of interest whereas models used in descriptive mining usually group patterns and relationships and incorporate trends in data that are commonly known and therefore understandable to humans across. (Gorunescu et al., 2011)

Classification:

Classification is one of the long established and traditional techniques of mining data that is founded in machine learning. Classification identifies common characterists among a given group of objects or items in a database then thereafter follows the given metric of the model that shall be used in classification, to categorise these given items into diverse classes. The main ojective of classification is to scrutinize the training data and come up with or develop a model that is very accurate for each and every class by use of the features that are available in the data.

Classification makes use of mathematical and arithmetic approaches such as statistics, Decision Trees and Neural Networks. (Ming-Syan et al., 2006). Techniques that are more complex like Neural Networks or Decision Trees could also be used to forecast. However, all these techniques can be combined to attain way better results.

Clustering:

This is a technique of mining data that organises objects into categories depending on how similar they are. Differences between objects in the same cluster are minimised while dfferences between a cluster and the next are maximized. Objects within a cluster have similarities that are foreordained. Clustering is a form of unsupervised learning, in the terminology of machine learning because no label exists for the groups into which the objects shall be placed ; they are just grouped based on similarity and differences, not based on some checklist of characteristics that must be met to make one (an object) a member of a given class.

Dependency Modelling or Association Rule Mining:

This is one of the most acknowledged techniques of data mining and is classified as unsupervised data mining. Dependency modelling is aimed at finding relationships between records that belong to a large dataset and labels significant dependencies among variables. Association rule mining is commonly applied in market basket analysis in an attempt to analyse customer purchasing behavior. From the modelling, one is able to determine the kind of goods that are usually purchased together e.g. a customer who purchases bread is very likely to purchase milk as well or a cutomer who purchases construction materials is also very likely to purchase furniture too. This in turn helps to improve marketing strategies of items.

Anomaly detection:

Synonymous to its name it deals with the unearthing of most substantial changes or aberrations from the standard behavior. Anomaly detection is commonly used in the banking industry to detect fraudulent activities. Anomaly detection can aslo be applied to a given dataset in order to identify and get rid of outliers so as to come up with more accurate study findings.

Summarization:

Summarization is not part of the techniques of data mining but it results from the above techniques. It deals with determination of a depiction that is compact for a given data subset. This technique is also referred to as generalization or description.

2.3.3 Application of Data Mining in Educational Systems:

There is a growing interest among researchers towards the field of data mining in education because of how much the field has evolved. Interest in research in this field is further fuelled by the fact that very many students are admitted annualy into various institutions therefore adding very large volumes of data to the already exsting ones. The techniques of data mining can be applied to aid in bridging or filling gaps in knowledge in education systems. This is achieved by knowledge discovery, unveiling hidden patterns, variances and implications or connotations. Through these processes, stakeholders are able to effectively make informed decsions that lead to refining and clarification of the educational systems (Romero et al., 2005). In this light, this study sought to solve a social problem of gender balance in STEM education but the problem is termed as a predictive data mining problem in the context of machine learning. A regression model was also built, to come up with the prediction because there were no finite or fixed number of classes into which the outcome was supposed to fit, thereby making regression a more desirable technique as opposed to classification. A Multilayer Pereptron (MLP) type of Artifcial Neural Network was used to construct the regression model. This process was conducted using Python, in Tensorflow open-source library alongside keras libraries since they have the capacity to run alogsde each other.

The following section shall review Artifcial Neural Networks.

2.3.4 ANN Literature Review

The application of Artificial Neural Networks (ANNs) is traceable back to the field of finance, though as time has gone by, their application has spread into other fields like education.Since Artficial Neural Networks are self-adaptive and data-driven they do not necessarily require given assumptions with regard to the underlying model. Generally, there exists five categories of networks that are utilized as forecasting tools. The five are as follows: Feedforward Networks such as the Multilayer Perceptron (MLP), Polynomial Networks, Support Vector Machine, Recurrent Networks and ModularNetworks.

The following are some of the advantages or potential benefits of Artificial Neural Networks. (ANNs)

i) Non-linearity, which is as a result of the neural processor having many different nodes that work together to eventually fire the output. The aspect of non linearlity makes the ANN have a processing capacity which is as close to human reasoning as possible and even better.

ii) Data –driven learning which results from mapping of the output to the input. The network is therefore able to learn by examples through supervised machine learning.

iii) Adaptability. This characteristic gives ANNs the ability to change their own synaptic weights based on the given output verses desired output, to reduce and eliminate error; the adjustment occurs in real time which makes ANNs outstanding.

iv) ANNs have a huge response capacity in that they not only provide a pattern classification but also aid in very reliable decision making.

v) They are very fault tolerant as a result of the the massive interconnected nodes. In the event that one breaks down, the rest are still able to continue working and provide dependable results.

vi) Parallelism which makes Artificial Neural Networks to be integrated on a very large scale. Due to this aspect, they are able to accomplish given tasks very fast, and to also capture very complex behaviors.

vii) ANNs use the same notations across all fields that are engaged with networks. The uniformity in design and analysis makes them desirable.

viii) Finally, ANNs apply the analogy of Neurobiology (Haykin, 1994) such that they adapt all by themselves and are non-linear. They therefore do not require theoretical constructs with regard to the given model. Because of this given characteristics and advantages, ANNs are able to establish very complex relationships given very little. As little as the data only.

This study applies Multilinear Perceptrons in solving the stated problem and therefore the following section seeks to asses the performance of all the other types of networks relative to Multilayer Perceptrons. Several other studies have portrayed MLP to be the most accurate in forecasting of time series, which also guided the choice to apply them in this study.

Types of ANNs

There are many different types of networks that are used for the purpose of classification and/or regression. The following section categorizes the networks based on their general characteristics.

The first category comprises of Feedforwad Networks (FFNs). A common example of the Feedforward Networks is the Multilayer Perceptron (MLP). This is the simplest type of ANN since all the information simply moves forward from the input layer to the output layer, through the hidden layer if there is any. There are no loops or the output is not taken back to the input. Since all the output is forward, there is no established backward connection between a node and the one preceeding it. Other examples of neural networks that are Feedforward include: Radical Basic Fuction (RBF) (Bildirici et al., 2010), Generalized Regression Neural Network (GRNN) (Mostafa, 2010), Dynamc Neural Network (DNN) (Guresen et al., 2010) among many others.

The second category comprises of Recurrent Networks (RCNs) whose major characterstic is the dynamic nature of their connectivity therefore they are able to store information that will be used in future. Examples of networks that belong to the Recurrent Networks Category include the following; Elman Network (ELN) (Yumlu et al., 2005); the modification to the Elman Network (Perez-Rodriguez et al., 2005) Partially Recurrent Networks (PRN) (Perez-Rodriguez et al., 2005) and Autoregressive Networks (ARN) (Kodogi et al., 2002).

The third category comprises of the Polynomial Networks (PLNs). These networks offer methodical processing of input variables that are polynomial. However, it can still be exhaustive if one applies sigmoidal or gaussian functions in the training. Examples of networks that are polynomial in nature include: Pi-sigma networks such as Ridge Polynomial Networks (RPN) and its dynamic version (Ghazali et al., 2011), as well as the Function Link Network (FLN) (Hussain et al., 2008).

The fourth category comprises of Modular Networks (MNs). Modular Networks are comprised of many different modules in the networks which give provision for tasks to be solved separately then the aswers are thereafter combined in a manner that is logical. One way through which this is made possible is by use of various network architechtures (Zhang et al., 2001) and the other way is by use of application of different initialization weights while leaving the same architecture of networks. (Adeodato et al., 2011).

Finally, the fifth category comprises of Support Vector Machies (SVM). The network is a member of the kernel base model or nucleus. The idea behind SVMs is to plot every data item as a point on an n-dimensioal space where n is the number of features that are occurring, with the value of each feature being the value of a particular coordinate. Thereafter, the task of classification is performed by finding the hyperplane that clearly draws a difference between the two or more classes. (Carpinteiro et al., 2011)The hyperplane maximizes the margin of separation.

The Multilayer Perceptron (MLP)

As mentioned earlier, a Multilayer Perceptron is a type of Feedforward network. It comprises of an input layer whose work is only to feed the values therefore does not perform any processing, another layer called the hidden layer and a last layer called the output layer. A neural network also has a basic unit known as the neuron. The hidden layer and output layer consist of a set of nodes that pass the values from the previous layer to the next i.e. from input to hidden, from hidden to the next hidden and lastly to the output layer. The learning algorithm is usually implemented during the training phase whereby the various layers indicate the direction in which the information is flowing. The Multilayer Perceptron usually attains competency by means of the back propagation algorithm which is a gradient technique. The weights are not altered unless there is an error and the training is performed until the MLP consistently produces the correct output, and that is when one can say that the algorithm has learned. After the traning process is complete, the weights of the network are frozen then they are subsequently used to calculate the outputs for iputs that are new.

Back propagation in summary works in the following manner;

- Computes the error by calculating the difference between the output given by the model and the actual expected output.
- Based on the activation function, the error is made minimum– The second step is therefore checking whether the error has been minimized or not.
- After the error has been corrected through minimisation, the parameters are updated. This process is called the training phase .Now basing on the new updated weights, a new output is computed and fired and thereafter compared again with the actual output. If there is a difference between the two, the process is repeated untill when the model will fire the actual output that is expected. That is when the error will have become zero.
- The model is then considered accurate enough to make predictions. New inputs of the same dataset can be fed into the model which should then produce outputs that do not vary from those that are expected. This is now known as the testing phase.

The following is the Justification for using MLPs in this Classification Problem

1. It is very efficient in computation due to the aspect of parallelism.

2. MLPs are universally known and used for similar nature of computations therefore they made a good choice

3. They posses the ability to learn or gain competence through adaptation such that, the data that is fed into them for the training phase then becomes their experience, and future data from similar datasets can be used to achieve similar results as long as the model is accurate.

4. They are quite straight forward in the sence of not requiring theoretical constructs of the underlying model. Simply training the model alone is able to bring forth the decision fuction that is required.

5.MLPs have proven to be universal approximators when they have sufficient hidden nodes and two layers with a back propagation network.

2.3.5 Other Studies on Application of Data Mining in STEM Education

In the recent past, there has been a number of emerging studies in the field of educational data mining. A lot of research has been carried out to explain or find out and establish student enrollment trends by use of data mining techniques. Some of the research work has been reviewed in the following section.

In their work, (Fong et al., 2009), made use C4.5 and back-propogation algorithms for the process of student admission. In their study, they proposed a model that is a hybrid of decision tree classifier and neural networks. This model was aimed at predicting the likelihood of a given student joining a given university. The model would achieve this by analyzing the academic merits of this given student, their background and further, the admission criteria of the university from past records.

In his study, Kovačić carried out a case study to identify the extent to which data on enrollment can be utilized to predict the success of a student. The research applied CHAID and CART algorithms on the data for enrollment of students that were admitted to study Information Systems at the open polytechnic of New Zealand. The output was a two decision tree classfier that would categorize students into two clusters, the two being successful and unsuccessful. The CHAID algorithm obtained an accuracy level of 59.4 while the CHART algorithm accuracy was 60.5. (Kovačić, 2010)

(Moucary et al., 2011) conducted a research study to find a reliable and accurate prediction tool that would help administrators and course instructors to make decisions about enrolling students who had graduated from Bachelors of Engineering, into the Masters of Engineering program. Objective number one of the study was to discover the relationship between the most affecting factors. Objective number two was to come up with a prediction model that would equip the administrators and other people that control the master’s admission process, with an efficient and accurate decision-making tool. The research study employed the use of Matlab Neural Networks Pattern Recognition tool, together with Classification and Regression Trees (CART) with important cross validation testing.

There is also another research that made use of techniques in data mining, to forecast the admissibility of female students into higher education. This study was centred on coming up with models of data mining to make this prediction by employing the Naive Bayes Classifier and Decision Tree algorithm. The study utilized actual data of approximately 690 students of undergraduate courses from Government Arts College of Pudukkottai in India. The testing and training data comprised of the students personal data, their academic performance before college and at graduate level characteristcs. (Padmapriya et al., 2012)

In another study by Yadav et al., 2013, methodologies for data mining were utilized to select students to be enrolled into a given course. This study applied the classification task to evaluate previous performance of the students. The Decision tree algorithm was used to perform the classification. In order to determine the student that was legible for admission into a given course, the Decision Tree classfier used data from the students management system that contained records like the stream to which the student belonged, their pre-college performance and grades scored at high school graduation. From the study, the findings revealed that the past academic performance of a student could be used to create a model by use of ID3 decision tree algorithm which can be utilized for prediction of a student's enrollment into the course of Master of Coumputer Applications (MCA)

There is also a study by (Gupta et al., 2013), which explored sociological and demographic variables such as gender, age, education, ethnicity, disability and work status, and the evironment in which students study that could influence the dropping out of the students. The study scrutinized the extent to which these factors could help in pre-identification of students who shall succeed through school and those who shall not. Techniques of data mining like classification trees, feature selection and logistic regression were used to come up with the most importat factors leading to the success of a student and the profile of a typical successful and unsuccessful students were constructed. Emperical study findings pointed at the following as the most important factors that separate a successful student from one who is unsuccessful; course block, ethnicity and course programme. Of all the methods of growing a classification tree, Classification and Regression Tree (CART) was the most successful, with an overall percentage of 60.5%; correct classification. The risk approximated by both the cross-validation and gain diagram shows that all trees are not accurate in separating the successful students from those that are not successful, if they use enrollment data only. A similar conclusion was reached by utilizing logistic regression. This case study was to come up with a datawarehouse for a data mining system which would be used for the prediction of student enrollment into the university. The data warehouse would posses the ability to come up with summary reports as input data files for the data mining system which would then perform the predictions.

Another research study sought to find out whether a student's past academic performace could be used to come up with a classification model by use of a decision tree algorithm (ID3 and J48 decision tree algorithm). The outcome of the study is to help students to select the courses to be admitted to, based on their academic performance and skills. The study was able to predict the outcome of the students from their interactions with the assessment's system. This study developed a prediction model that was able determine whether a given student who attends college will get admitted into a STEM major or not. The study developed a logistic regression model predicting STEM major enrollment from combinations of features. (Priyanka et al., 2013)

The above studies have utilized the existing data that is available in the education sector. However, from the literature little has been done to deal with the existing gender differences in STEM. The question that consequently comes up as a gap from the literature reviewed is how data mining can be applied to the existing large datasets to come up with better solutions of bridging the existing gender gap in STEM in Kenya. To further make use of the research that helps predict whether or not a given student will join STEM courses, there is need to come up with a prediction model that will determine when the number of female students getting enrolled into STEM will be equal to that of males. There is a lot of data in the repositories of institutions like KUCCPS which can be put to this use. This study made use of this data.

2.4 Conceptual Framework

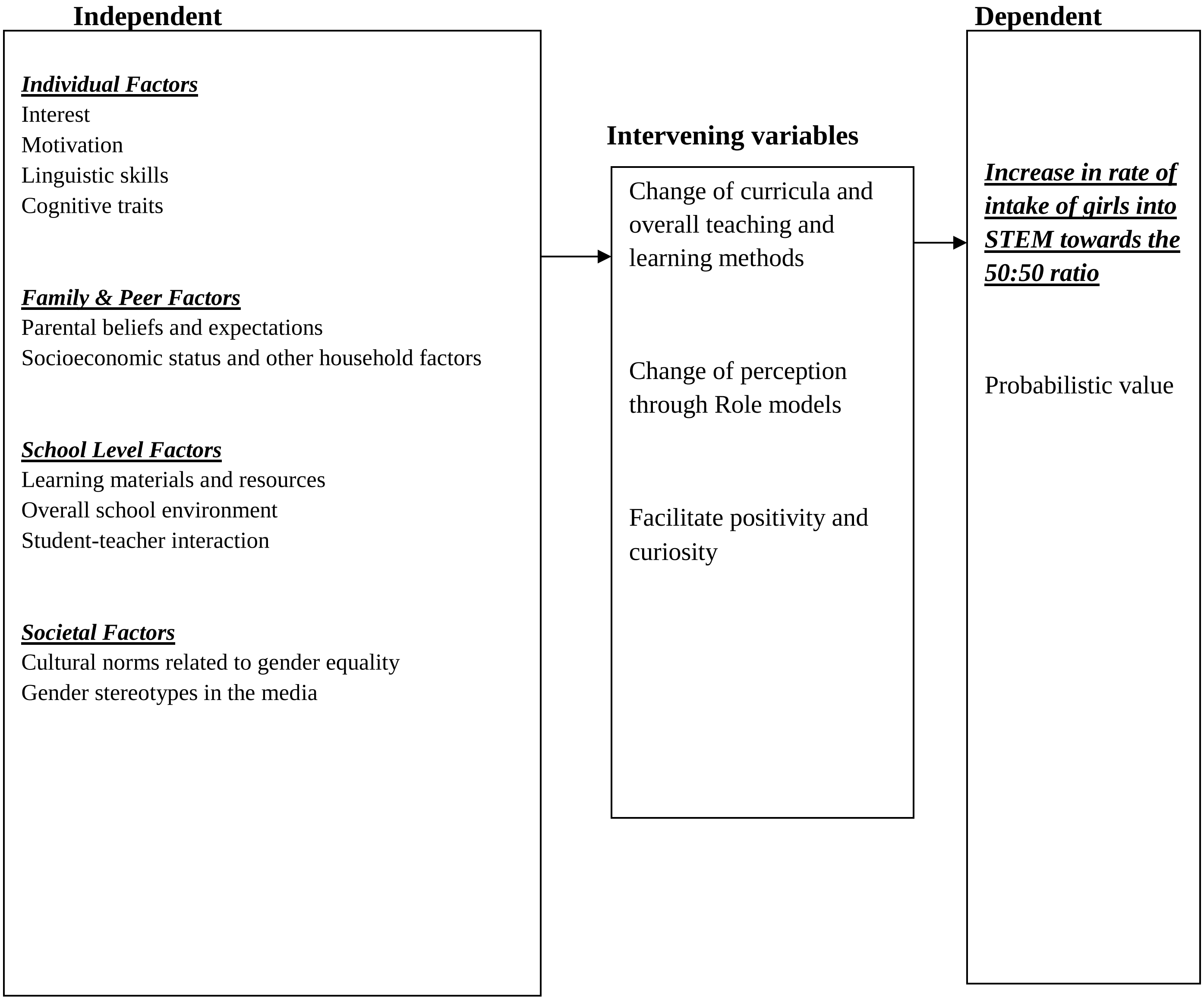

Independent
Individual Factors
Interest
Motivation
Linguistic skills
Cognitive traits
Family & Peer Factors
Parental beliefs and expectations
Socioeconomic status and other household factors
School Level Factors
Learning materials and resources
Overall school environment
Student-teacher interaction
Societal Factors
Cultural norms related to gender equality
Gender stereotypes in the media
Intervening variables
Change of curricula and overall teaching and learning methods
Change of perception through Role models
Facilitate positivity and curiosity
Dependent
Increase in rate of intake of girls into STEM towards the 50:50 ratio
Probabilistic value

Figure 3: Conceptual Model for Attaunment of the 50:50 Ratio in STEM

2.5 Operationalization of Variables

Variables	**Sub-variables**	**Indicators**	**Values(Data)**
Individual	Interest Motivation	Choice of all the three Sciences	Yes No
	Numeracy Skills	Grade in Mathematics	A,B,C,D,E
	Linguistic Skills	Grade in languages	A,B,C,D,E
	Cognitive traits	Overall mean grade	A,B,C,D,E
Family & Peers	Parental beliefs and expectations,	Religion	Christian Other religions Muslim
	parental education and socioeconomic status, and other household factors	Level of education of parents	Upper Upper middle Lower Middle Poor
School	Learning materials and resources, teaching strategies and student-teacher interactions, assessment practices and the overall school environment.	Level and location of school attended	National, County, Extra county .
Societal	Gender equality	County of Origin	One of the 47 counties in Kenya

Table 1: Operationalization of variables

CHAPTER THREE: RESEARCH METHODOLOGY

3.0 Introduction

This chapter was aimed at discussing the research design, the target population, sample size, sampling techniques and their justifications, describe the methods of data collection and how the data was analyzed.

3.1 Method for Achieving Objective 1 and 2

Objective one was to determine the factors that lead to low enrollment of females into STEM courses. To achieve this object, a review of the study data was performed. The predictor variables were used to determine the most important factors that affect enrollment of the girl child into STEM.

Objective number two was to establish appropriate data mining methods used to build the predictive model for enrollment into STEM ratios by gender. To achieve this objective, a survey of literature was conducted and the most appropriate method of constructing the Neural Network was established.(Amrouche et al., 2014), (Hassan et al., 2018) The input sets were sourced from the study data. The training and testing set were also established in the 80: 20 ratio as per the standard industry practice.

3.2Method to Achieve Objective 3 and 4

3.2.1 Research Design

Objective three was to develop a model for predicting when the 50:50 gender ratio of enrollment of government sponsored students into STEM courses in Kenyan Universities will be attained and objective four was to evaluate the model for predicting attainment of the 50:50 gender ratio in STEM. Data mining methodology was used to achieve objective 3 and 4.

Data mining methodology can be defined as a system that consists of rules, procedures, methods and principles that guide the data mining process (Berry & Linoff, 2009). Data mining methodology is an 11 step methodology designed to form a successful data mining project.. The 10 step data mining methodology was applied as follows:

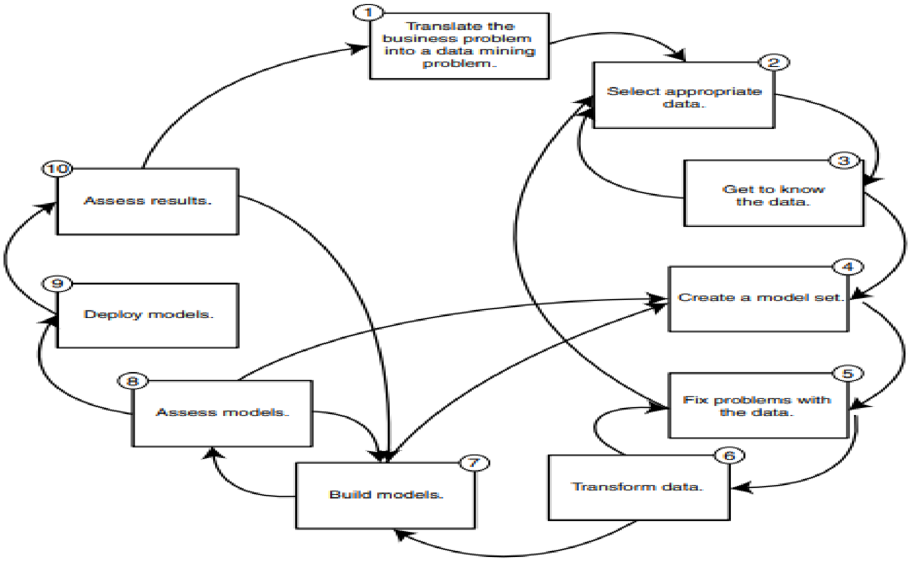


Figure 4 :Data Mining Methodology(Berry & Linoff 2009)

Step 1: Define the business problem

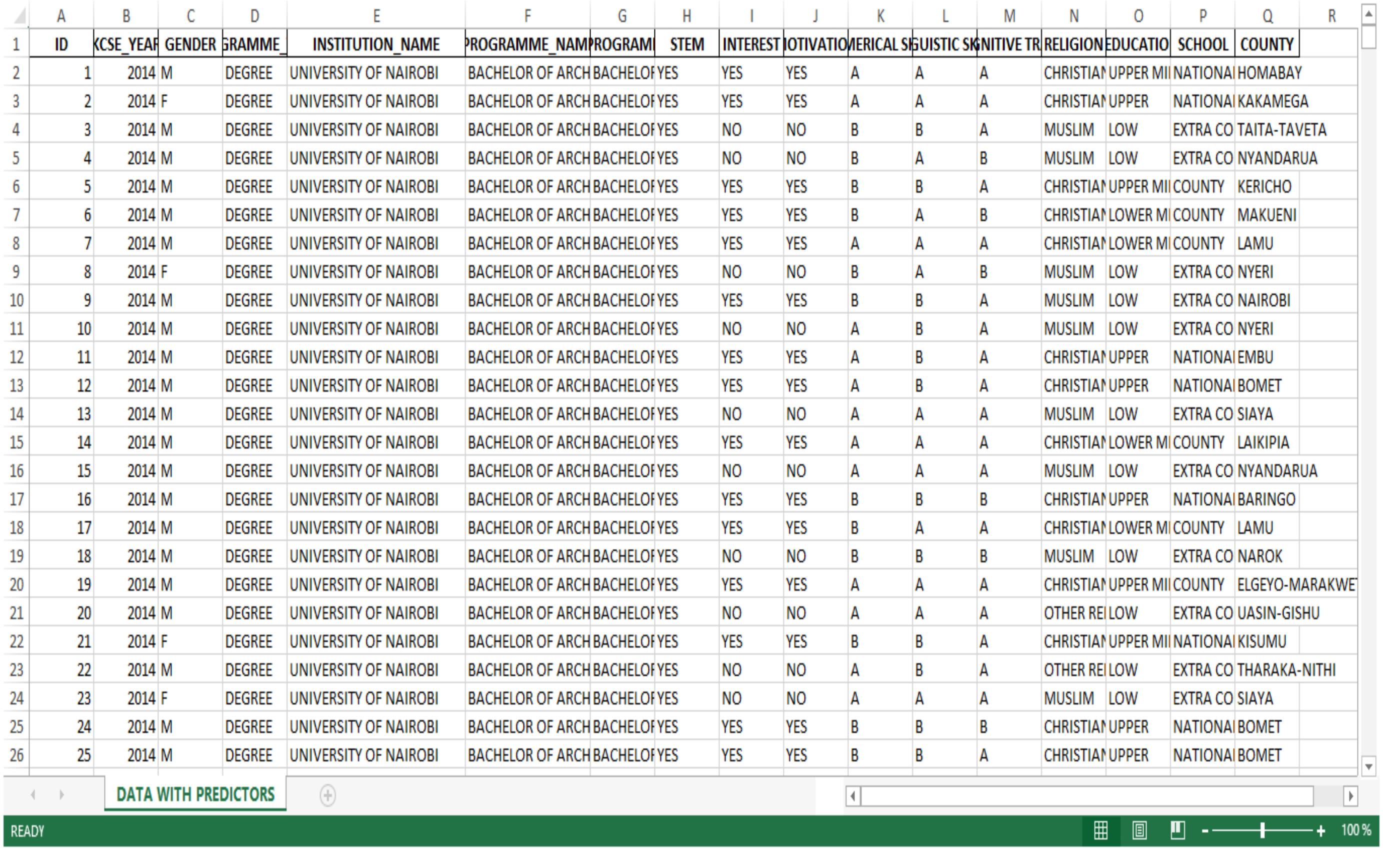
A well-defined business problem leads to a better design of the data mining model that will help solve the problem. This research has been able to clearly define what the problem at hand is through the statement of the problem and further aims at explaining how data mining can be applied on the existing data to try and come up with a solution to the problem. The general societal problem is lack of gender parity in STEM courses hence resulting to the specific research problem of: lack of a model to predict attainment of gender parity in STEM courses in Kenya.

Step 2: Translate business problem into a data mining problem.

A successful translation of a business problem to a data mining problem requires the problem to be formulated in one of the data mining tasks. This has been accomplished by looking at the problem and since the problem requires a predictive model, the best technique to use was regression. Regression was more desirable because there was no fixed number of classes into which the data could be categorized, yet the problem would still be classified under supervised learning. One of the deliverables of the model are "time" which tells us "when" the ratio will actually be attained. This time is interms of years, based on the nature of data available.

Step 3: Select appropriate data

Data mining requires data. In the best of all possible worlds, the required data would already be resident in a corporate data warehouse, cleansed, available, historically accurate, and frequently updated. The data required for this study was available at the repositories of KUCCPS and was accessible upon official request and authorization from the NACOSTI. Convenient sampling was applied to come up with the required sample to achieve the set objectives of the study. Convenience sampling is a non-probability sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher. The researcher therefore made use of data available in the KUCCPS repository for students who got admitted into STEM courses for the years 2014, 2015,2016,2017 and 2018. The data had 591,348 records.The figure below displays a smaple of the data in an Excel spreadsheet.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	ID	KCSE_YEAR	GENDER	GRAMME_	INSTITUTION_NAME	PROGRAMME_NAM	PROGRAMI	STEM	INTEREST	IOTIVATIO	IERICAL SI	GUISTIC SK	NITIVE TR	RELIGION	EDUCATIO	SCHOOL	COUNTY	
2	1	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	A	A	A	CHRISTIAN	UPPER MI	NATIONAI	HOMABAY	
3	2	2014	F	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	A	A	A	CHRISTIAN	UPPER	NATIONAI	KAKAMEGA	
4	3	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	B	B	A	MUSLIM	LOW	EXTRA CO	TAITA-TAVETA	
5	4	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	B	A	B	MUSLIM	LOW	EXTRA CO	NYANDARUA	
6	5	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	B	B	A	CHRISTIAN	UPPER MI	COUNTY	KERICHO	
7	6	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	B	A	B	CHRISTIAN	LOWER MI	COUNTY	MAKUENI	
8	7	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	A	A	A	CHRISTIAN	LOWER MI	COUNTY	LAMU	
9	8	2014	F	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	B	A	B	MUSLIM	LOW	EXTRA CO	NYERI	
10	9	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	B	B	A	MUSLIM	LOW	EXTRA CO	NAIROBI	
11	10	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	A	B	A	MUSLIM	LOW	EXTRA CO	NYERI	
12	11	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	A	B	A	CHRISTIAN	UPPER	NATIONAI	EMBU	
13	12	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	A	B	A	CHRISTIAN	UPPER	NATIONAI	BOMET	
14	13	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	A	A	A	MUSLIM	LOW	EXTRA CO	SIAYA	
15	14	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	A	A	A	CHRISTIAN	LOWER MI	COUNTY	LAIKIPIA	
16	15	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	A	A	A	MUSLIM	LOW	EXTRA CO	NYANDARUA	
17	16	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	B	B	B	CHRISTIAN	UPPER	NATIONAI	BARINGO	
18	17	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	B	A	A	CHRISTIAN	LOWER MI	COUNTY	LAMU	
19	18	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	B	B	B	MUSLIM	LOW	EXTRA CO	NAROK	
20	19	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	A	A	A	CHRISTIAN	UPPER MI	COUNTY	ELGEYO-MARAKWET	
21	20	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	A	A	A	OTHER REI	LOW	EXTRA CO	UASIN-GISHU	
22	21	2014	F	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	B	B	A	CHRISTIAN	UPPER MI	NATIONAI	KISUMU	
23	22	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	A	B	A	OTHER REI	LOW	EXTRA CO	THARAKA-NITHI	
24	23	2014	F	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	NO	NO	A	A	A	MUSLIM	LOW	EXTRA CO	SIAYA	
25	24	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	B	B	B	CHRISTIAN	UPPER	NATIONAI	BOMET	
26	25	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCH	BACHELOF	YES	YES	YES	B	B	A	CHRISTIAN	UPPER	NATIONAI	BOMET	

Figure 5: Data Sample in Excel Spreadsheet

Step 4: Get to know the data

Before building models, it is important to go through the selected data just to familiarize oneself with it. The researcher examined the data and used it to derive 11 predictor variables which would be of significance to the study. At this stage, anomalies in the data were aslo detected.

Step 5: Fix problems with the data

Problems in the selected dataset were dealt with. Normally all data that one intends to use to build models will most likely be dirty and hence have problems or a lot of noise. The data used in this study however, was fairly clean. The only problem was with the column PROGRAMME_TYPE, which had a total of 4 categories, some represented both in upper and lower case. This was resolved by changing all the string values in the column to uppercase.

Step 6. Transform the data

After fixing problems within the data, the data will have to be converted to a format that is appropriate for mining. At this stage, the reseatcher added some derived varivbles to make the data more meaningful and understandable. The column of interest was derived from the question of whether a student took all the three sciences or not.

Step 7: Building the model

Building models as a data mining process takes up less time in a data mining project since modern data mining software have automated the process. However for this model, coding for the model was performed from scratch in the Tensorflow environment using Python programming language.The code is attached in Appendix III

Step 8: Assess the models

This step determines if the model is working as expected. At this point there will be questions of accuracy of the model and confidence of the predictions from the model. The model being a regression one, it was evaluated using the R^2 metric , achieving a apredictiion accuracy level of 78.8% as shown in Appendix III.

Step 9: Deploy the models

This is the step where the model is moved from the mining environment to the scoring environment. Depending on the development environment, the process may be hard or easy. The hard part comes from the fact that some development may have been done in a special environment where the software can't run anywhere else and the miner has to move it and recode it in another programming language to make it possible for it to run. However in this study, the model will not require to be further deployed.

Step 10: Assess the results

In this step, actual results are compared against expected results. The actual measure of data mining is the value of actions taken as a result of the data mining. Having a measure of lift can help one choose a model and use of these models can help one choose how to apply the results from the models. In addition to using lift, it is also important to do a measure at the field.

Step 11: Begin Again

The completion of a data mining project raises more questions without answers. This is true since there comes new relationships that were not thought of before and this creates a new question to answer hence starting of data mining process again.

3.2.2 Target population

The target population are the government sponsored students that qualified fro STEM courses for the years 2014,2015,2016,2017 and 2018. Approximately 85,000 students are admitted into university and college annually, out of which 30% join STEM courses.

3.2.3 Sampling and sampling procedure

The process of creating an annotated sample is initiated by selecting a subset of data; the question that the researcher asked is: *what should the size of the training subset be, in order to reach a certain target classification performance?* The targeted classification performance was above 75%. Consequently, as much data as possible was required in order to achieve the targeted performance level. The sampling method applied was convenience sampling as the researcher retrieved the records for the available period at KUUCPS. Classification models have proven to perform better with as much data as possible hence the researcher obtained all the available data in the repositories of KUCCPS for students who were admitted into STEM since the year 2014,(inception of KUCCPS) upto the year 2018, which was the latest available data. The data had 591,348 records.

Further, the ***train_test_split()*** method was used to choose training and test samples. The size of test sample was 0.2(20% of the entire data set) and the remainder (0.8 or 80%) was the size of training sample, to achieve the targeted accuracy level of performance and also as per the standard industry practice.

3.2.4 Data collection methods

This is a machine learning research therefore the methods of data collection to be employed was the *machine learning methods of data* acquisition . Data discovery is necessary when one wants to share or search for new datasets and the datasets are available on the Web and or corporate data lakes. . (Yuji Roh et al., 2016). This study therefore utilized data discovery as the method of data collection since the data to be used was readily available on the Kenya Universities and Colleges Central Placement Service (KUCCPS) data lake. The KUCCPS requires one to have a permit to conduct research from National Commission for Science Technology and Innovation (NACOSTI) after which the researcher requests in writing and pays for the data. The request includes the period for which the researcher would like the data for and the scope i.e. per county or nationally. From their data lakes. KUCCPS was able to give the data in a hard disk.The data is also historical in nature making it perfect for the required data mining processes needed to achieve the objectives of this research.

3.3 Assumptions of the study

In the course of the study, the researcher assumed as follows:

i. That government sponsored students give the truest picture in the achievement of the 50: 50 ratio because they make the majority of the population of student that join university, and also because it is assumed that they are motivated to study STEM from high school level and work hard enough to achieve and get admitted into the same.

ii. That there is no multicollinearity, to imply that the independent or predictor variables are not too correlated to each other.

iii. That there is multivariate normality to therefore imply that each predictor variable is normally distributed, to produce an estimate that is not biased.

iv. That there are no influencial ouliers in the data which will affect the quality of the neural network model.

3.4 Data Analysis

The data used for the study was available in raw format in an Excel spreadsheet. It was imported into the Tensorflow environment for preprocessing and analysis using Python programming language. The input/ predictor varibles used to construct the ANN model were the following 11: Gender, Interest, Motivation. Numerical Skills, Linguistic Skills, Cognitive Traits, Religion, Parents Educational Level, School Type, and County of origin. The variables are fed into the input layer which then feeds into the three hidden layers. Each hidden layer with 3 neurons feeds into a single output neuron that carries the decision of the variable, which is the probability of a given student joiing a STEM course or not. A value closer to 0 implies one has a lower likelihood of ending up in STEM while a value closer to 1 implies a high likelihood of one ending up in STEM. The ReLU activation function is used. Decisions must be taken to divide the dataset into training, and test ratio. Data samples of 591,348 students are randomly mixed and 80% of them are used for training while 20% are used for testing. The training epoch is set to 250 with a batch size of 128. The ANN training performs continuously and terminates when the validation error failed to decrease during the validation process.The model was evaluated using the R^2 score evalution metric since it was a regression model, and it reported an accuracy score of 0.78. The accuracy score was expected to be above 0.75 to prove to be accurate and reliable enough to solve the given nature of problem.

CHAPTER FOUR: RESEARCH FINDINGS AND DISCUSSIONS

4.1 Introduction

This chapter discusses the findings of the study. The researcher sought to develop a model that would predict attainment of the 50:50 STEM gender ratio among government sponsored students in Kenya. To achieve this, the researcher sought to understand the factors that ,lead to under-enrollment of female students into STEM courses,establish appropriate data mining methods for predicting when the ratio will be attained, develop a model for predicting when the ratio will be attained and to further evaluate the effectiveness of the model.

4.2 Research findings

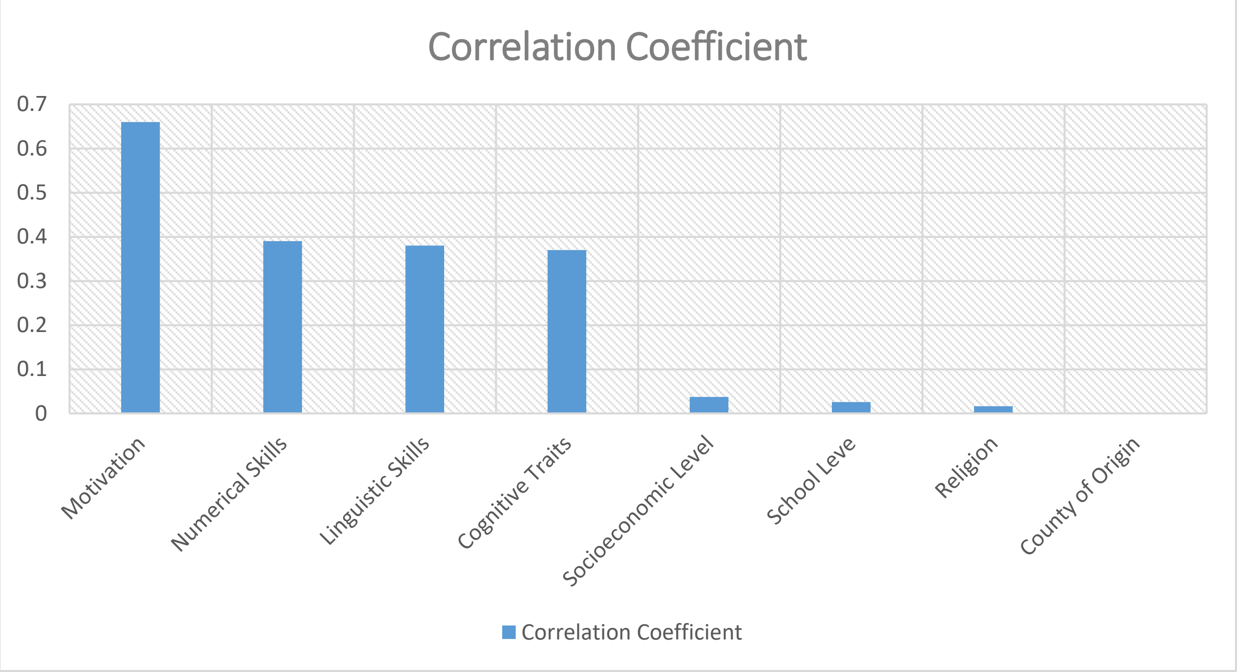
4.2.1 Objective One Findings

Factors leading to low enrollment of females into STEM courses. From the data that was collected, the researcher was able to come up with a number of factors leading to low enrollment of girl children into STEM courses. The variables were compared, relative to their correlation-coefficient towards the achievement of the 50:50 ratio and thereafter ranked from the most significant to the least significant as shown in the table below :

Table 2: Correlation between predictor variables and STEM

Variable	Correlation co-efficient	Level of Significance
Motivation	0.66	Very high
Numerical Skills	0.39	High
Linguistic Skills	0.38	High
Cognitive Traits	0.37	High
Parents Socioeconomic Level	0.037	Low
School Level	0.026	Low
Religion	0.016	Very Low
County of Origin	0.005	Very Low

Table 3: Graphical Representation of Feature Importance of the Predictor Variables



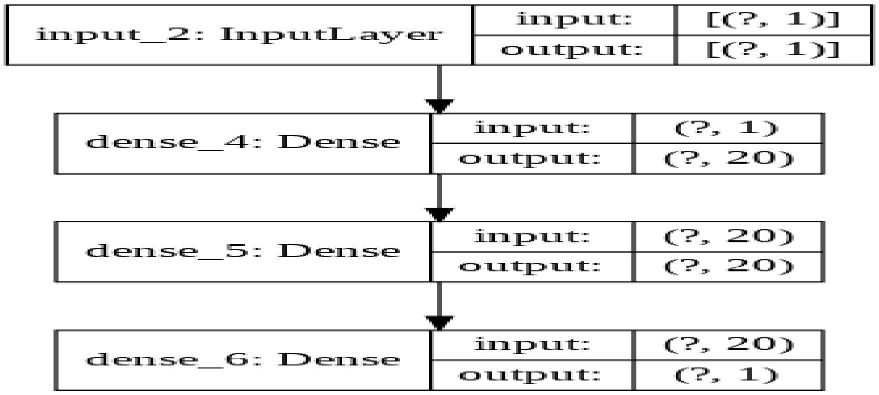
4.2.2. Objective Two Findings

Establishing appropriate data mining methods used to build the predictive model for enrollment into STEM courses ratios by gender . To achieve the second objective of the study, The model was visualized by calling ***plot_model()*** method that then creates a plot of the network model.

- ***model:*** Defines the model to be be plotted
- ***to_file*****:** Defines the name of the file to be created and save the plot.
- ***show_shapes***: defines whether or not to show the output shapes of each layer.
- ***show_layer_names*****:** (optional) defines whether or not to show the name for each layer.

The output for the above code was as shown in the following figure:

Table 4: ANN Model Structure



Layer		
input_2: InputLayer	input:	[(?, 1)]
	output:	[(?, 1)]
dense_4: Dense	input:	(?, 1)
	output:	(?, 20)
dense_5: Dense	input:	(?, 20)
	output:	(?, 20)
dense_6: Dense	input:	(?, 20)
	output:	(?, 1)

The model structure can also be visualized in the format below:

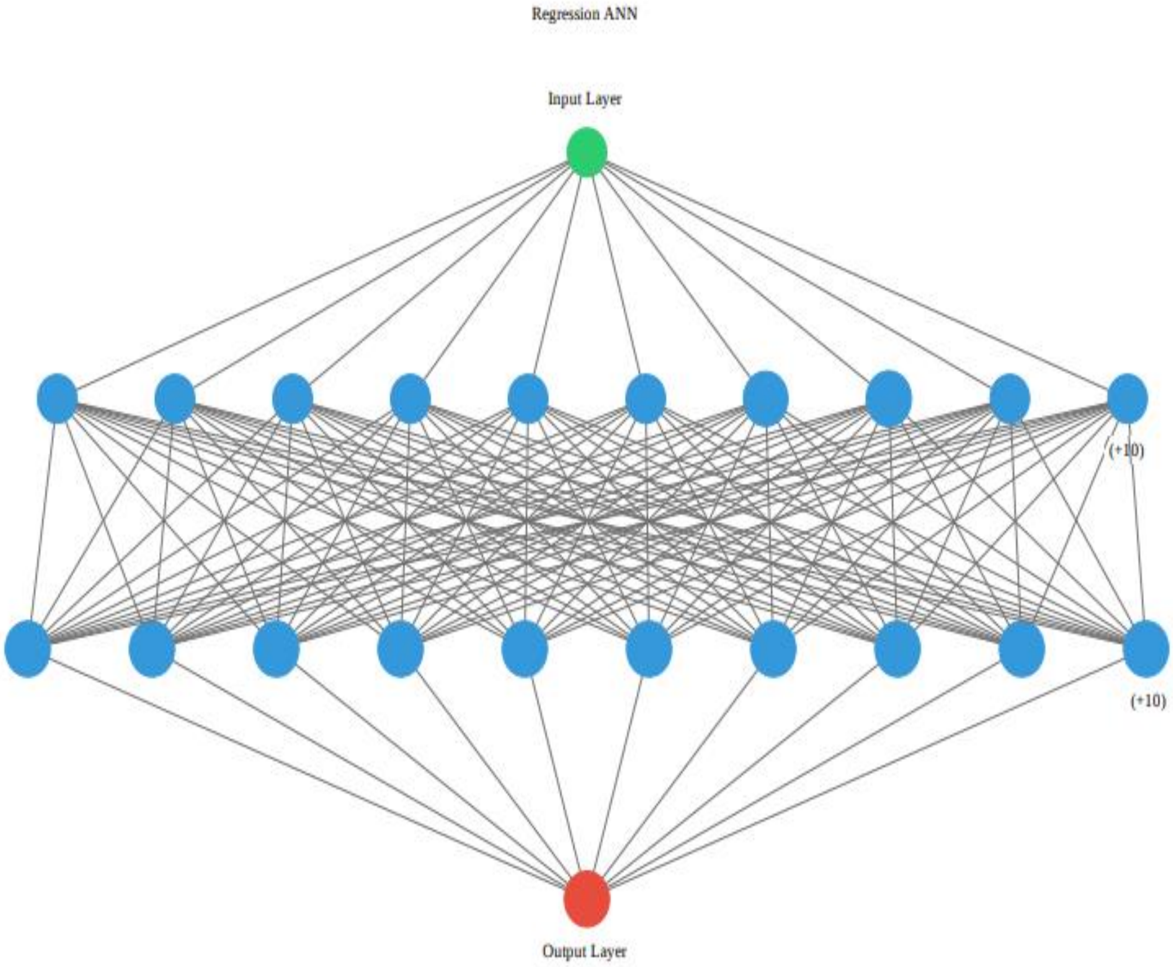


Figure 6: Visualized representation of the ANN structure

4.2.3 Objective Three Findings

Developing a model for predicting attainment of the 50:50 gender ratio in STEM among government sponsored students in Kenya.The following graph that shows the rate of enrollment for the available data was obtained:

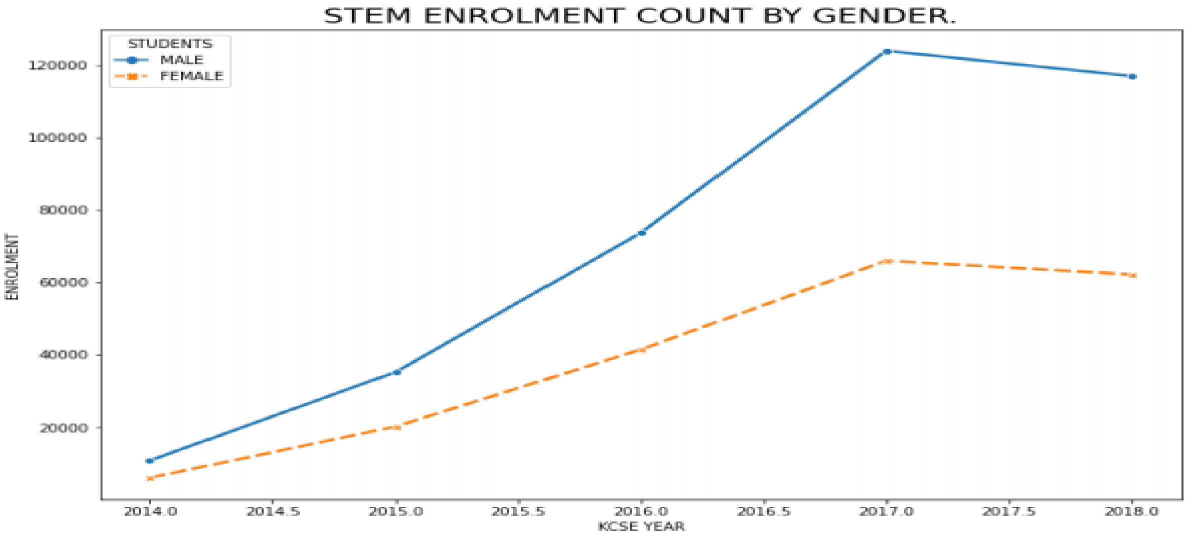


Figure 7 :Current Trend in Enrollment By Gender

The following model was obtained:

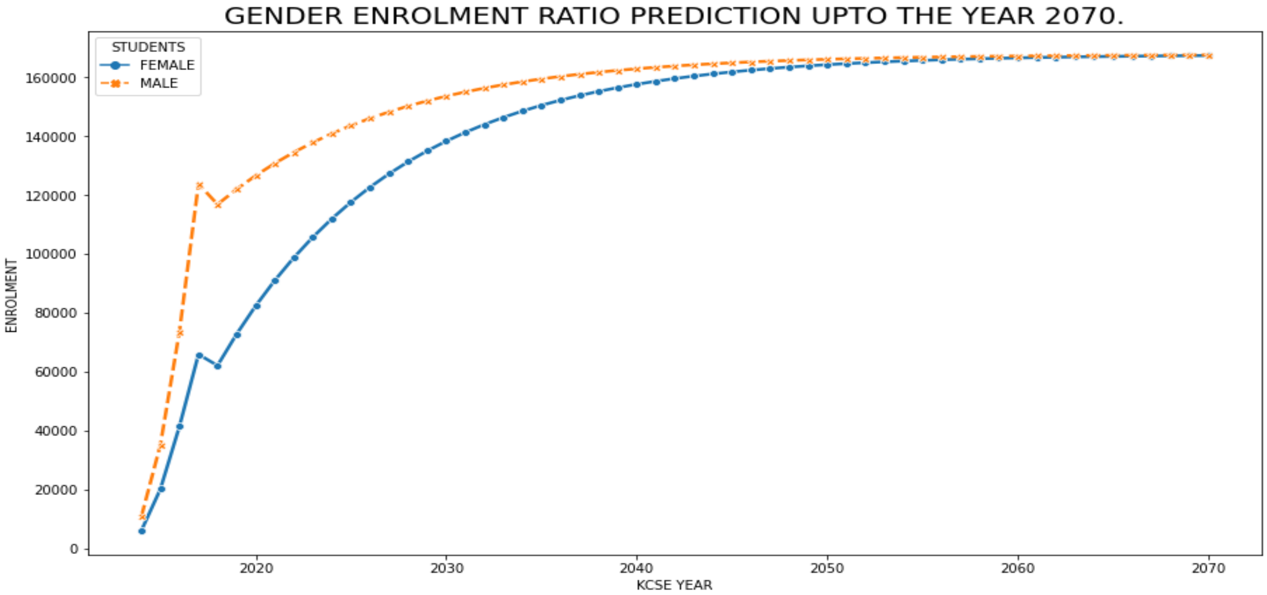


Figure 8: Model For Attainment of the 50:50 gender ratio

From the above visualization, the rate of enrollment of both males and females rises steadily at different rates. The 50:50 ratio is likely to be achieved in the year 2070. The ratio will however not be achieved by the year 2030 based on the data used in this study.

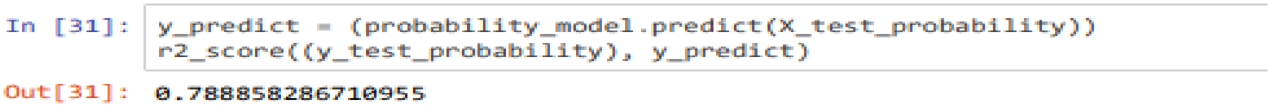
4.2.4 Objective Four Findings

The model was evaluated in order to validate its performance. It was tested on sections of the data during the training process. The R2 metric was used to evaluate the model as it is a regression model, making the R2 metric the most suitable one to use. The model reported an accuracy of 78.9% . The formula for the R2 metric is as shown below:

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

$$\frac{\text{MSE(model)}}{\text{MSE(baseline)}} \qquad \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (\overline{y}_i - \hat{y}_i)^2}$$

The R^2 computation is inbuilt for the ANN, and the evaluation results are as shown in the figure below:



```
In [31]:  y_predict = (probability_model.predict(X_test_probability))
          r2_score((y_test_probability), y_predict)

Out[31]:  0.788858286710955
```

Figure 9: Model evaluation

A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system is as follows:

0.90-1 = excellent (A)
0.80-0.90 = good (B)
0.70-0.80 = fair (C)
0.60-0.70 = poor (D)
0.50-0.60 = fail (F)
A value near 0.5 means the lack of any statistical dependence.

Since the model accuracy level is at 0.78, it is fair and therefore deemed reliable enough to solve the given problem.

4.3 Discussion of the Study Findings

This section discusses the results of this research, as presented in chapter 4. The findings are discussed in relation to the existing literature. Moreover, the limitations of the current research are addressed and possible solutions are offered.

This study constructed and utilized an Artificial Neural Network to come up with a predictive model to show when the 50: 50 gender ratio in STEM courses is likely to be achieved in Kenya. Due to the nature of the expected outcome being probabilistic, the study used a regression algorithm which was then evaluated using the R^2 metric and an accuracy level of 78.8% was achieved.

This findfings of this study corroborate with those of the (Yente et al. 2017) research, which proposed the use of ANN (Artificial Neural Network) via Tensorflow using Multilayer Perceptron (MLP) function as a data mining technique for prediction. The main objective of the research was to determine the enrollment of students based on attendance of marketing events. The model was built and evaluated usin the R^2 metric, achieving a prediction accuracy level of 74.1 %.

This study also established the following as the factors that lead to low enrollment of girls into STEM courses: Lack of interest in STEM, lack of Motivation, poor performance in numeracy skills; which results to low scores in mathematics, lower cognitive ability, insufficient awareness on STEM at county level. The factors were ranked based on their predictive power, towards achievement of the ratio.

In their study (Cheryan et al. 2015), it was observed that even from historical times, girls and women have not been taught that they are born brilliant whereas their male counterparts are raised with the understanding that they are born very brilliant, and even more brilliant than women in most instances. The matter of cognitive ability is however still a question of whether they perform poorly because society has made them believe boys will always be better, or whether they perform poorly because they simply can't perform better. The matter of cognitive ability is easily arguable since despite the fact that males have been observed to be generally performing better than females, there are still females who perfom better than males. Therefore were it a question of natural disposition, then it means that no measures whatsoever that are put in plac would be able to help get more females into STEM. However the findings of the study are still in tandem with the theory of gendered socialization.

Poor performance is also another factor among girls that leads to a lower rate of enrollment. This is also linked to the idea of stereotypes being engrained in society and the issue of socialization. There are also some notions that are propelled in childhood like boys excelling in maths while girls being perceived as good in the kitchen and home keeping activities. (Gunderson et al., 2011 and Regner et al., 2014). Research has futher shown that one of the contributing factors to women dropping off along the STEM pipeline is women being bombarded with negative stereotypes and socialized ideas especially about the subpar women's abilities in mathematics. (Gunderson et al., 2011) .These practices of socialization feed into the concept of stereotypes that in turn threaten the performace of girls in mathematics and consequently also in STEM. (Shapiro et al., 2012). The study findings of objective one are hereby seen to be in line with the underlying theories and other studies that explain the gender gap in STEM.

In a study by (Wanjau et el.,2016)to Predict Student Enrolment in STEM Courses in Higher Education Institutions, it is clear that the ratio of males in STEM is higher than that of females, as also seen earlier from the literature review. The trend shall continue like so upto the year 2070 when the number of girls in STEM shall be equal to that of boys. Attainment of the 50: 50 ratio is made more challenging by other secondary factors like the level of education of ones parents which is quite a difficult factor to control or intervene.

Other factors such as religion also play a role in the achievement of the ratio yet they are equally as hard to intervene through. The general cognitive levels of boys have also been perceived to be higher than those of girls which pauses a question of nature, of whether boys will naturally continue to outdo girls in technical matters such as STEM education. However, it can aslo be observed that there are several other factors that can lead to higher rates of enrollment of girl children, which the government and society can easily intervene in. Such include creating more awareness about STEM at county levels, working to improve the overall performance and attitude of girl children towards mathematics and sciences and developing the interests of girls in STEM while it is still early enough. It is however generally encouraging that the overall trend in enrollment into STEM courses is upward, thereby creating the impression that in as much as the 50:50 by 2030 may not be attained in time, the measures put in place are still yielding fruit but at a slower rate than the intended. More effort will be required inorder to achieve Kenya Vision 2030 alongside with the global SDG 4 and 5.

4.4 Limitations of the Study

The study was only limited to government sponsored students and not self-sponsored ones. This was based on the assumption that students who qualified for and got admitted into university and college by direct entry, had the initial will and attitude to take courses in STEM and they worked hard enough to qualify for them. This cohort is therefore bound to give a more realistic picture of the projection.

It was also limited to the only available data for the years 2014-2018 because the Kenya Universities and Colleges Central Placement Service (KUCCPS) only has data from the year 2014,as it was founded in 2013.

The study was further limited to KUCCPS as the source of data becauase it is the only institution that deals with admission of students into Universities and Colleges in Kenya, and therefore the only one that would have data that would be sufficient enough to fulfill the objectives of this study.

The performance analysis of the model was also simply limited to the R2 metric as the study was not using the same data to compare the performance of ANNs ,relative to other tools and methods used to perform similar predictions,such as support vector machine (SVM).

CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter presents the conclusions from the study findings and the recommendations for policy and future research.

5.2 Conclusions

The study achieved its objectives as discussed below:

5.2.1. Objective one

The following conclusion can be drawn to address the objective: to determine the factors that lead to low enrollment of girl children into STEM courses. The research data was studied and the predictor variables were used to determine the key factors that lead to low enrollment of girl children children into STEM. The factors were ranked from the most key to the least key and the following were established as key factors in the attainment of the 50:50 ratio : individual factors such as lack of interest, lack of motivation, poor numeracy skills, lower cognitive ability and lack of sufficient awareness at county level. It can therefore be concluded that these are the factors that need to be further looked into, in order to help hasten the attainment of the 50: 50 gender ratio in STEM in Kenya.

5.2.2. Objective two

Objective number two was to establish the most appropriate method of building the Neural Network.Typically, the ANN model was simply expressed as a mathematical function:Y=f(X,W) where Y and Xare the output and input vectors. W is a vector of weight parameters representing the connections within the ANN. The problem was established to be a regression problem therefore the model was based on a regression algorithm. The model was made up a 3 hidden layer multi perception regression model, with 3 neurons at each layer. The model takes in the pre-processed features discussed and predicts a probable estimation of an individual to end up in a STEM course or not. The models predictions are set to be on a scale of 0 to 1. The model was then visualized and concluded to be functional.

5.2.3 Objective Three

The third objective was to build the predictive model that would determine when the 50: 50 ratio would be attained. This was achieved using the 10- step data mining guiding principle by (Berry & Linoff, 2009). It can also be concluded that the input variables used and the data set were sufficient enough to come up with the predictive model.

5.2.4 Objective Four

The last objective was to perform the evaluation of the prediction model. This was achieved by use of the R2 metric which was determined as the most appropriate metric due to the outcome being probabilistic. The model was established to be 78% accurate and therefore concluded to be accurate enough to perform the prediction problem at hand.

5.3 Contributions of the study

1. The study has been able to show that data mining can be used on existing data to develop models that can aid in improvement of gender ratios of enrollment into STEM.
2. The study has identified the profiles of girls that are less likely to be admitted into STEM, this is likely to improve intervention measures aimed at improving intake of girls into STEM, as the measures shall be more target focused and not general.
3. The results from this study alongside with the study findings of other researchers can help policy implementors to come up with strategies to increase enrollment of girl children into STEM courses.

5.4 Recommendations for future research

Further research should be conducted on self sponsored students to observe the trend and see if they are a vital part in achievent of the 50: 50 ratio attainment.

A similar study can be conducted but on the females who are already enrolled into STEM to determine the retention ratios at the professional work place.

Since the study results have shown the profiles of females who are less likely to be admitted into STEM, policy makers should come up with specific measures that are specifically aimed at such nature of girls inorder to improve the enrollment ratio

The study recommends construction of a similar model using different tools like R to compare with the performance of the ANN regression mode; used in this study.

The model could be implemented at each secondary school level so that based on each of the predictor variables, teachers and parents are able to provide personalized guidance that should see to it that all females capable of joining STEM courses are able to do so.

Use of the same data to perform different operations like association rules, clustering and classification patterns within the data can be performed

REFERENCES

Archer, L., DeWitt, J., & Dillon, J. (2014). ‘It didn’t really change my opinion’: exploring what works, what doesn’t and why in a school science, technology, engineering and mathematics careers intervention. Research in Science & Technological Education, 32(1), 35-55.

Adams, C. (2014). Campaign aims to Recruit 1 Million STEM Mentors for Girls Education Week, 34(11), 8.

ASQ. 2012. U.S. youth reluctant to pursue STEM careers, ASQ surveys says. Milwaukee, ASQ. http://asq.org/newsroom/news-releases/2012/20120131-stemcareers-survey.html

AAI. 2015. State of Education in Africa Report: A report card on the progress, opportunities and challenges confronting the Higher Education Sector. Africa - America University

Buschor, C. B., Berweger, S., Keck Frei, a. and Kappler, C. 2014. Majoring in STEM - What accounts for women’s career decision making? a mixed methods study. The Journal of Educational Research, Vol. 107, No. 3, pp. 167-176. DOI: 10.1080/00220671.2013.788989.

Basheka, B. C. (2008): Value for Money and Efficiency in Higher Education: Resources Management and Management of Higher Education in Uganda and its Implications for Quality Education Outcomes. Uganda Management Institute, Kampala: OECD.

Baron-Cohen, S. (2003). The essential difference: The truth about the male and female brain. New York: Basic Books.

Dasgupta, N., & Stout, J. G. (2014). Girls and women in science, technology, engineering, and mathematics: STEMing the tide and broadening participation in STEM careers. Policy Insights from the Behavioral and Brain Sciences, 1(1), 21 29. doi:10.1177/2372732214549471

Girls, Inc. (2016). Science, Math, and Relevant Technology. Retrieved from http://www.girlsinc.org/resources/programs/girls-inc-operation-smart.html

Gudo, C. O., Olel, M. a. and Oanda, I. O. 2011. University expansion in Kenya and issues of quality education: Challenges and opportunities. International Journal of Business and Social Science 2 (20): 203 - 214.

Haussler, P. & Hoffman, L. (2002). an intervention study to enhance girls’ interest, self-concept, and achievement in physics classes. Journal of Research in Science Teaching, 39, 870888.

Hayden, K., Ouyang, Y., Scinski, L., Olszewski, B., & Bielefeldt, T. (2011). Increasing student interest and attitudes in STEM: Professional development and activities to engage and inspire learners. Contemporary Issues in Technology and Teacher Education, 11(1), 47-69.

Hill, C., Corbett, C., & St. Rose, a. (2010). Why so few? Women in science, technology, engineering, and mathematics. AAUW. Retrieved from https://www.aauw.org/files/2013/02/Why-So-Few-Women-in-Science-Technology-Engineering-and-Mathematics.pdf

Holmes, S., Redmond, a., Thomas, J., & High, K. (2012). Girls Helping Girls: Assessing the Influence of College Student Mentors in an afterschool Engineering Program. Mentoring & Tutoring: Partnership In Learning, 20(1), 137-150.

Hughes, R., Nzekwe, B., & Molyneaux, K. (2013). The Single Sex Debate for Girls in Science: A Comparison Between Two Informal Science Programs on Middle School Students' STEM Identity Formation. Research In Science Education, 43(5), 1979-2007.

Kenya National Bureau of Statistics (KNBS). 2016. Economic Survey. Kenya: KNBS, Nairobi.

Kolstad, R.K. & Briggs, L.D. (1995). Better Teaching of Science Through Integration. Journal of Instructional Psychology, 22(2), 130.

Kommer, D. (2002). Boys and girls learn differently: A guide for teachers and parents. American Secondary Education, 30(3), 88.

Kinanjui, K. and Mburugu, E. (Eds). 2004. African perspectives on development. Washington DC.

K. Hornik, M. Stinchcombe, and H. White: Multilayer feedforward networks are universal approximators. Neural Networks 2 (1989), 359-366.

Lamont, T. (2010). John Maeda: Innovation is when Art meets science. The Observer, November 14, 2010. Retrieved from www.guardian.co.uk/technology/2010/nov/14/my-brightideajohn-maeda

Lewin, T. (2008, July). Math scores show no gap for girls, study finds. New York Times, 157(54). p. 16

Maeda, J. (2013, July). artists and scientists: more alike than different. Scientific American.

Maloney, E., Waechter, S., Risko, E., Fugelsang, Jonathan. (2012, June). Reducing the sex difference in math anxiety: The role of spatial processing ability. Learning and Individual Differences, 22(4). p380-384. 5p.

Martin, P. (2014, September). MakeHERSpaces: STEM, girls and the Maker Movement. CRB Short Subjects. California Research Bureau, California State Libraries. www.library.ca.gov/CRB.

McLaughlin, C. (2011, September). Editorial: art and Technology. Childrens' Technology and Engineering. International Technology and Engineering Editors' association. P2

Moir. a., & Jessel, D. (1989). Brain sex: The real difference between men and women. Dell Publishing, New York, NY.

Ministry of Devolution and Planning. 2015. Sector Performance Standards. 2nd Ed.

Milgram, D. (2011). How to recruit women and girls to the science, technology, engineering, and math (STEM) classroom. Technology and Engineering Teacher, 71(3), 4-11.

Mosatche, H. S., Matloff-Nieves, S., Kekelis, L., & Lawner, E. K. (2013). Effective STEM programs for adolescent girls: Three approaches and many lessons learned. afterschool Matters, 17, 17-25.

Mukhwana, E., Oure, S., Kande, A., Njue, R., Too, J., Njeru, M. and Some, D. (2016). State of Univesity Education in Kenya. Commission for University Education Nairobi, Kenya

Mukhwana, E., Oure, S., Too, J. and Some, D. 2016. State of Post Graduate Research and Training in Kenya. Commission for University Education. Nairobi, Kenya

Nakayiwa, H. 2016. Higher Education and development: Prospects for transforming agricultural education in Uganda. african Journal of Rural Development 1 (2): 193 - 204.

O'Hanley, H. (2015). The steam initiative. Arts & Activities, 157(1), 11.

Piro, Joseph. (2010, March). Going from STEM to STEaM: The arts have a role in america's future too. Education Week. 29:24. PP28-29.

Riegle-Crumb, C.; Moore, C., & Ramos-Wada, a. (2010, October). Who wants to have a career in Science or Math? Exploring adolescents' future aspirations by gender and race/ethnicity. Wiley Online Library. www.wileyonlinelibrary.com

Roberts, a. (2012). a Justification for STEM Education. Technology and Engineering Teachers. Retrieved from https://www.iteea.org/File.aspx?id=86478&v=5409fe8e

R. P. Lippmann: an introduction to computing with neural nets. IEEE ASSP Magazine (1997), 4-22.

Sanders, J., & Peterson, K. (1999). Close the gap for girls in math-related careers. Education Digest, 65(4), 47.

Shaffer, L. (2014, Spring). Full STEAM Ahead: 10 tips for incorporating art into your STEM curriculum. Scholastic Instructor. Scholastic, Inc. New York, NY.PP32-25.

Sikora, J. & Pokropek a. (2012). Gender Segregation of adolescent Science Career Plans in 50 Countries. Science Education, Wiley Periodicals, Inc.

Wimmer, H. & Powell, L.M., 2015. a Comparison of Open Source Tools for Data Science. 2015 Proceedings of the Conference on Information Systems applied Research, pp.1–9. available at: http://iscap.info.

Witten, I.H., Frank, E. & Hall, M. A. (2011): Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, available at: http://www.cs.waikato.ac.nz/~ml/weka/book.html%5Cnhttp://www.amazon.com/DataMining-Practical-Techniques-Management/dp/0123748569.

APPENDICES

Appendix I: Research Schedule

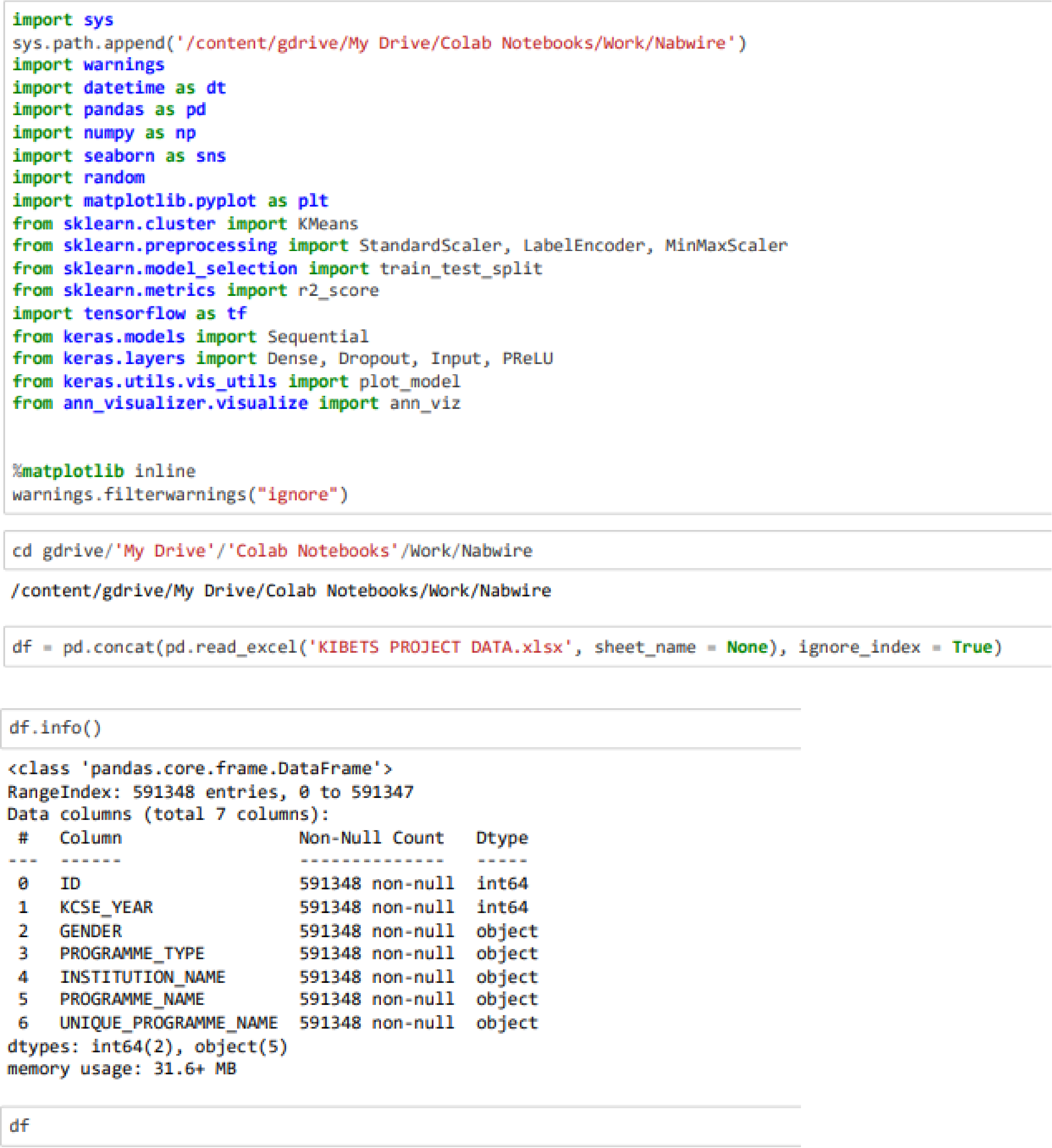
ACTIVITY	START	FINISH
Proposal drafting and writing	1st January 2020	29th January 2020
Proposal submission and (preparation for) Presentation	30th January 2020	7th February 2020
Data Collection and Analysis	8th April 2020	27th July 2020
Work in Progress 1 writing and submission	1st August 2020	30th September 2020
Thesis defense	October 2020	24th October 2020
Final report submission	October 2020	October 2020

Appendix II: Research Budget

Cost estimates for the project are as follows:

ITEM/ACTIVITY	**COST (KES)**	**PURPOSE**
Laptop	85,000	Design and Development of the model
Reference resources	5,000	Reference
Payment to NACOSTI	1,500	To seek permission to conduct research
Payment to KUCCPS	3,000	To be given the required data for the study
Hard disk	7,000	Storage of the data to be used in the research
Internet connectivity	5,000	Internet access
Miscellaneous expenses	5,000	Feasibility study, transport to and from KCAU and KUCCPS, Printing and binding of the proposal and thesis documents.
TOTAL	**111,500**	

Appendix III :Neural Network Code:



```
import sys
sys.path.append('/content/gdrive/My Drive/Colab Notebooks/Work/Nabwire')
import warnings
import datetime as dt
import pandas as pd
import numpy as np
import seaborn as sns
import random
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler, LabelEncoder, MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
import tensorflow as tf
from keras.models import Sequential
from keras.layers import Dense, Dropout, Input, PReLU
from keras.utils.vis_utils import plot_model
from ann_visualizer.visualize import ann_viz


%matplotlib inline
warnings.filterwarnings("ignore")
```

```
cd gdrive/'My Drive'/'Colab Notebooks'/Work/Nabwire
```

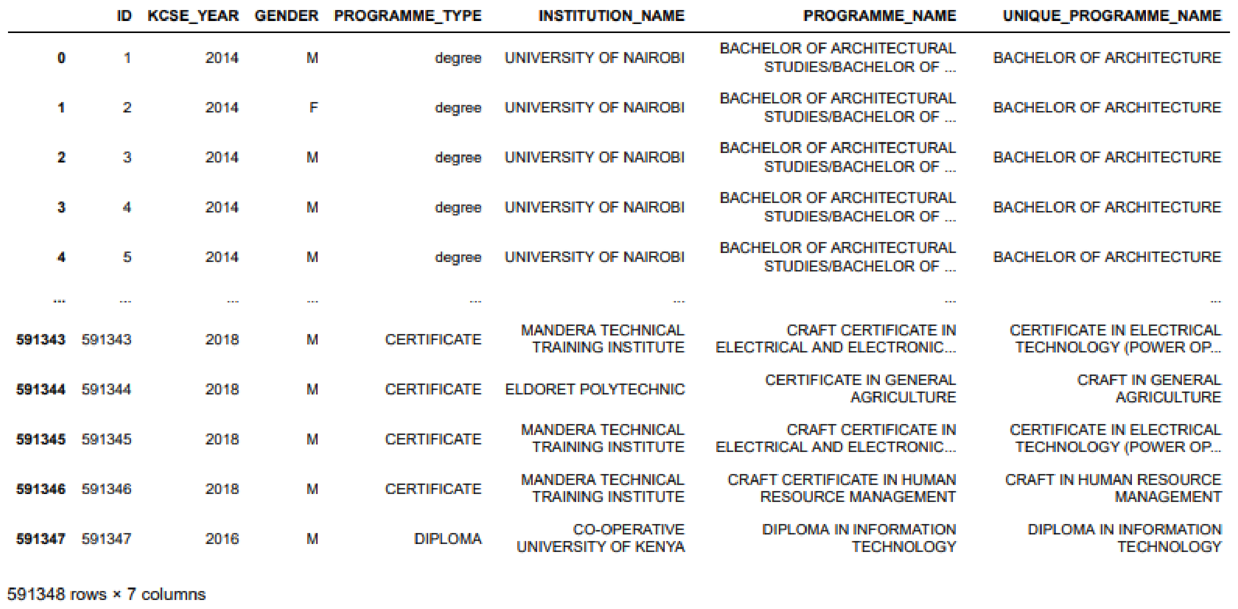
```
/content/gdrive/My Drive/Colab Notebooks/Work/Nabwire
```

```
df = pd.concat(pd.read_excel('KIBETS PROJECT DATA.xlsx', sheet_name = None), ignore_index = True)
```

```
df.info()
```

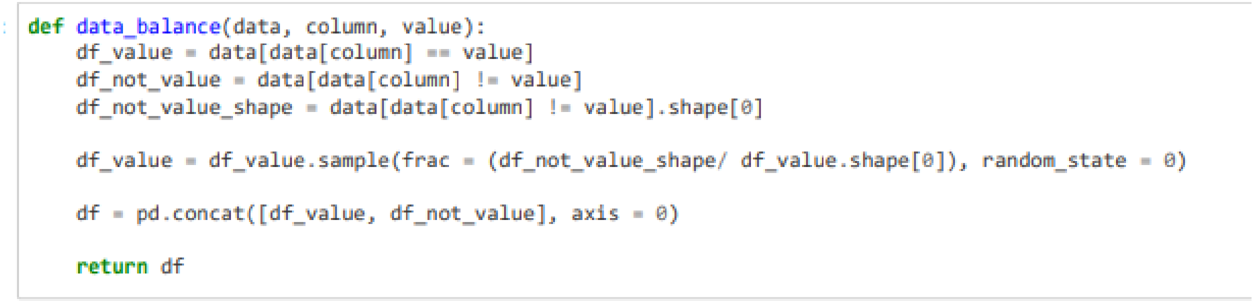
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 591348 entries, 0 to 591347
Data columns (total 7 columns):
 #   Column                Non-Null Count   Dtype 
---  ------                --------------   ----- 
 0   ID                    591348 non-null  int64 
 1   KCSE_YEAR             591348 non-null  int64 
 2   GENDER                591348 non-null  object
 3   PROGRAMME_TYPE        591348 non-null  object
 4   INSTITUTION_NAME      591348 non-null  object
 5   PROGRAMME_NAME        591348 non-null  object
 6   UNIQUE_PROGRAMME_NAME 591348 non-null  object
dtypes: int64(2), object(5)
memory usage: 31.6+ MB
```

```
df
```

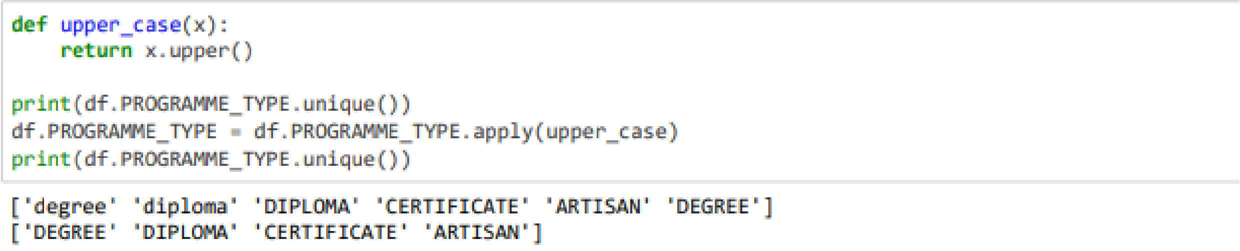


	ID	KCSE_YEAR	GENDER	PROGRAMME_TYPE	INSTITUTION_NAME	PROGRAMME_NAME	UNIQUE_PROGRAMME_NAME
0	1	2014	M	degree	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE
1	2	2014	F	degree	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE
2	3	2014	M	degree	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE
3	4	2014	M	degree	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE
4	5	2014	M	degree	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE
...	...	...	...	...	...	...	...
591343	591343	2018	M	CERTIFICATE	MANDERA TECHNICAL TRAINING INSTITUTE	CRAFT CERTIFICATE IN ELECTRICAL AND ELECTRONIC...	CERTIFICATE IN ELECTRICAL TECHNOLOGY (POWER OP...
591344	591344	2018	M	CERTIFICATE	ELDORET POLYTECHNIC	CERTIFICATE IN GENERAL AGRICULTURE	CRAFT IN GENERAL AGRICULTURE
591345	591345	2018	M	CERTIFICATE	MANDERA TECHNICAL TRAINING INSTITUTE	CRAFT CERTIFICATE IN ELECTRICAL AND ELECTRONIC...	CERTIFICATE IN ELECTRICAL TECHNOLOGY (POWER OP...
591346	591346	2018	M	CERTIFICATE	MANDERA TECHNICAL TRAINING INSTITUTE	CRAFT CERTIFICATE IN HUMAN RESOURCE MANAGEMENT	CRAFT IN HUMAN RESOURCE MANAGEMENT
591347	591347	2016	M	DIPLOMA	CO-OPERATIVE UNIVERSITY OF KENYA	DIPLOMA IN INFORMATION TECHNOLOGY	DIPLOMA IN INFORMATION TECHNOLOGY

591348 rows × 7 columns

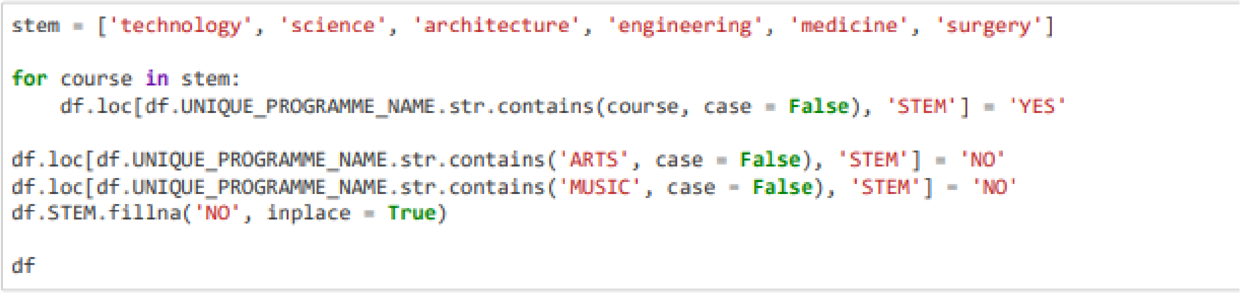


```
def data_balance(data, column, value):
    df_value = data[data[column] == value]
    df_not_value = data[data[column] != value]
    df_not_value_shape = data[data[column] != value].shape[0]

    df_value = df_value.sample(frac = (df_not_value_shape/ df_value.shape[0]), random_state = 0)

    df = pd.concat([df_value, df_not_value], axis = 0)

    return df
```

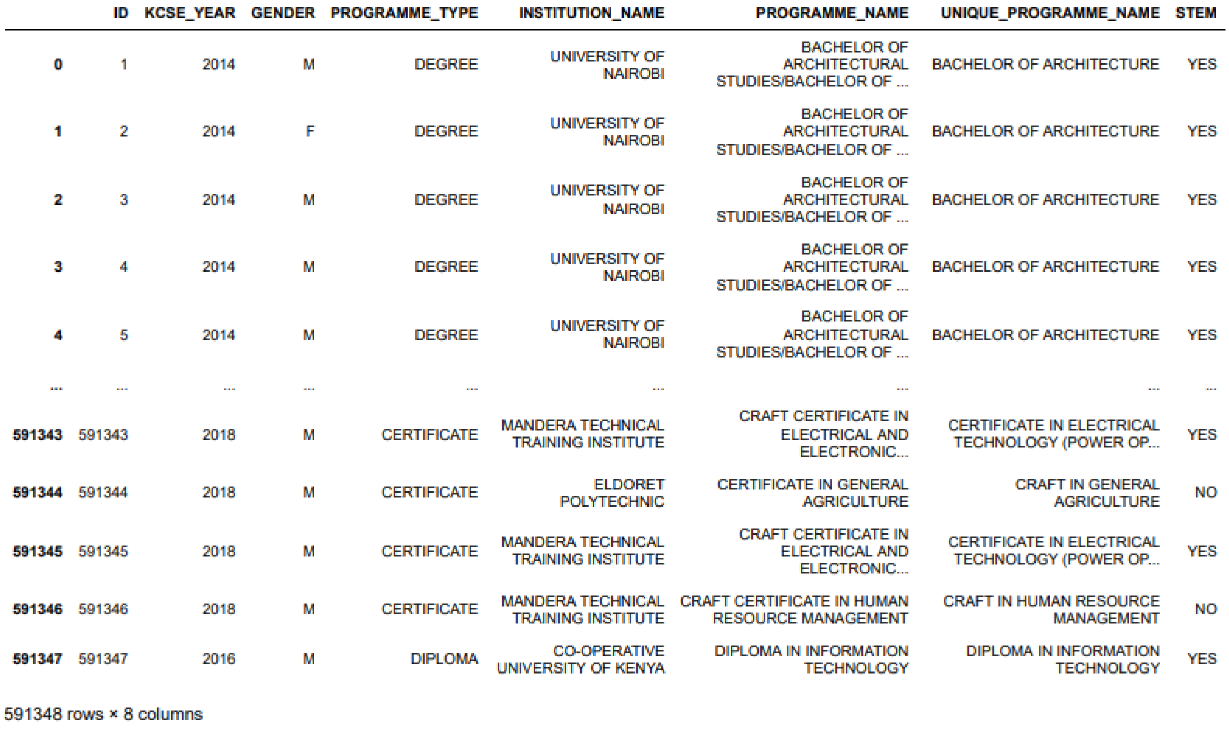


```
def upper_case(x):
    return x.upper()

print(df.PROGRAMME_TYPE.unique())
df.PROGRAMME_TYPE = df.PROGRAMME_TYPE.apply(upper_case)
print(df.PROGRAMME_TYPE.unique())
```

```
['degree' 'diploma' 'DIPLOMA' 'CERTIFICATE' 'ARTISAN' 'DEGREE']
['DEGREE' 'DIPLOMA' 'CERTIFICATE' 'ARTISAN']
```

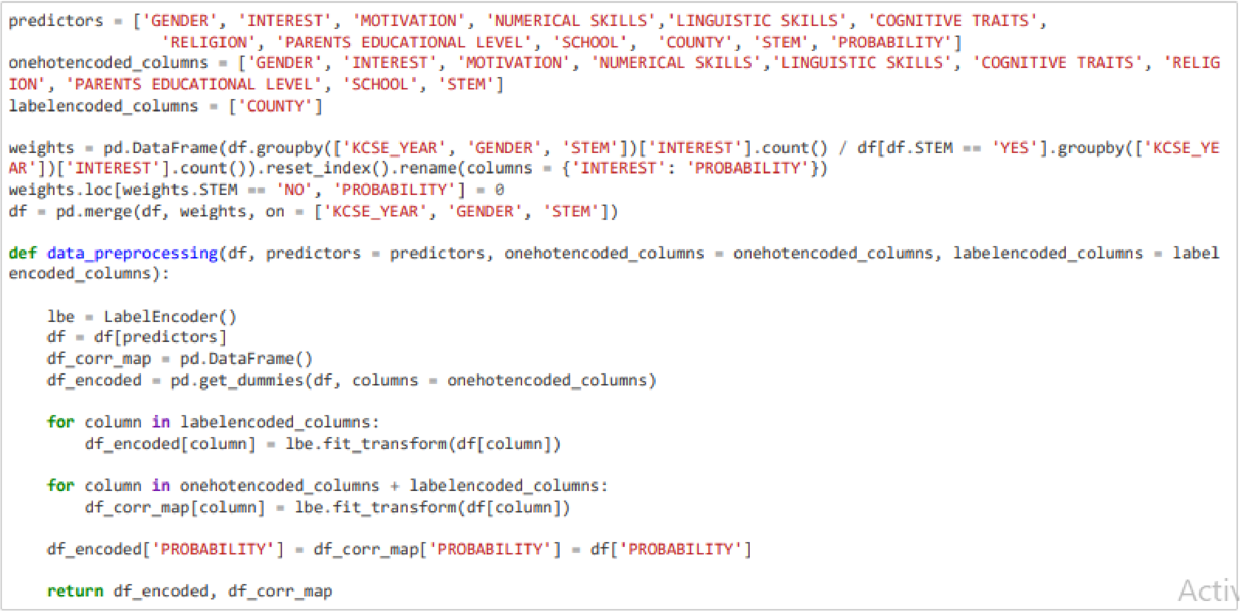


```
stem = ['technology', 'science', 'architecture', 'engineering', 'medicine', 'surgery']

for course in stem:
    df.loc[df.UNIQUE_PROGRAMME_NAME.str.contains(course, case = False), 'STEM'] = 'YES'

df.loc[df.UNIQUE_PROGRAMME_NAME.str.contains('ARTS', case = False), 'STEM'] = 'NO'
df.loc[df.UNIQUE_PROGRAMME_NAME.str.contains('MUSIC', case = False), 'STEM'] = 'NO'
df.STEM.fillna('NO', inplace = True)

df
```

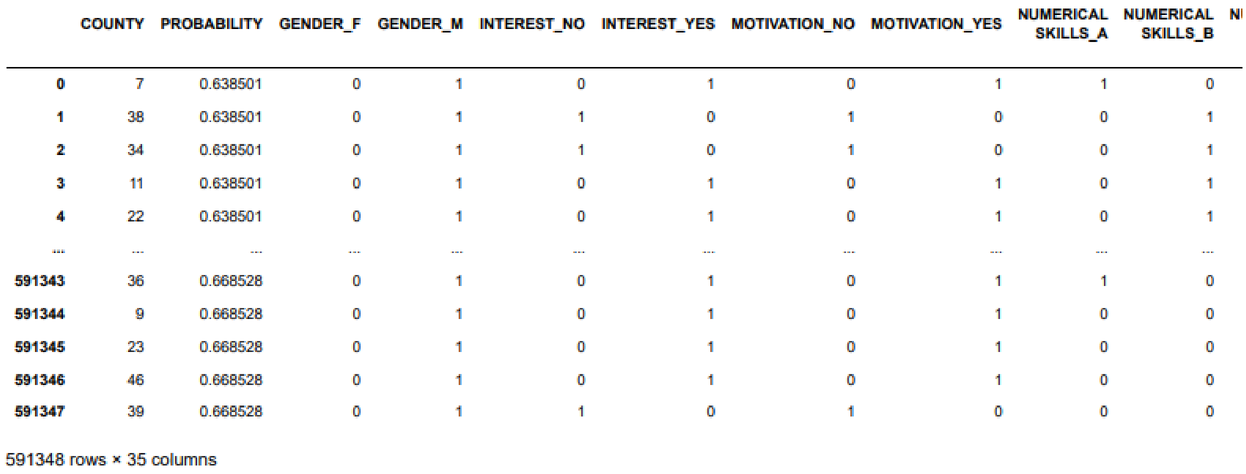


	ID	KCSE_YEAR	GENDER	PROGRAMME_TYPE	INSTITUTION_NAME	PROGRAMME_NAME	UNIQUE_PROGRAMME_NAME	STEM
0	1	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE	YES
1	2	2014	F	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE	YES
2	3	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE	YES
3	4	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE	YES
4	5	2014	M	DEGREE	UNIVERSITY OF NAIROBI	BACHELOR OF ARCHITECTURAL STUDIES/BACHELOR OF ...	BACHELOR OF ARCHITECTURE	YES
...	...	...	...	...	...	...	...	...
591343	591343	2018	M	CERTIFICATE	MANDERA TECHNICAL TRAINING INSTITUTE	CRAFT CERTIFICATE IN ELECTRICAL AND ELECTRONIC...	CERTIFICATE IN ELECTRICAL TECHNOLOGY (POWER OP...	YES
591344	591344	2018	M	CERTIFICATE	ELDORET POLYTECHNIC	CERTIFICATE IN GENERAL AGRICULTURE	CRAFT IN GENERAL AGRICULTURE	NO
591345	591345	2018	M	CERTIFICATE	MANDERA TECHNICAL TRAINING INSTITUTE	CRAFT CERTIFICATE IN ELECTRICAL AND ELECTRONIC...	CERTIFICATE IN ELECTRICAL TECHNOLOGY (POWER OP...	YES
591346	591346	2018	M	CERTIFICATE	MANDERA TECHNICAL TRAINING INSTITUTE	CRAFT CERTIFICATE IN HUMAN RESOURCE MANAGEMENT	CRAFT IN HUMAN RESOURCE MANAGEMENT	NO
591347	591347	2016	M	DIPLOMA	CO-OPERATIVE UNIVERSITY OF KENYA	DIPLOMA IN INFORMATION TECHNOLOGY	DIPLOMA IN INFORMATION TECHNOLOGY	YES

591348 rows × 8 columns

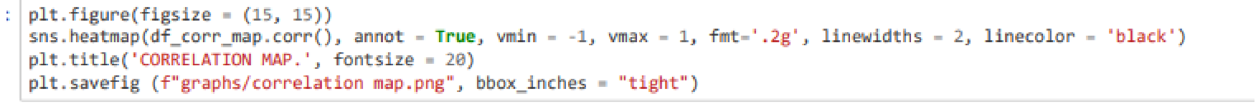


```
predictors = ['GENDER', 'INTEREST', 'MOTIVATION', 'NUMERICAL SKILLS','LINGUISTIC SKILLS', 'COGNITIVE TRAITS',
                'RELIGION', 'PARENTS EDUCATIONAL LEVEL', 'SCHOOL',  'COUNTY', 'STEM', 'PROBABILITY']
onehotencoded_columns = ['GENDER', 'INTEREST', 'MOTIVATION', 'NUMERICAL SKILLS','LINGUISTIC SKILLS', 'COGNITIVE TRAITS', 'RELIG
ION', 'PARENTS EDUCATIONAL LEVEL', 'SCHOOL', 'STEM']
labelencoded_columns = ['COUNTY']

weights = pd.DataFrame(df.groupby(['KCSE_YEAR', 'GENDER', 'STEM'])['INTEREST'].count() / df[df.STEM == 'YES'].groupby(['KCSE_YE
AR'])['INTEREST'].count()).reset_index().rename(columns = {'INTEREST': 'PROBABILITY'})
weights.loc[weights.STEM == 'NO', 'PROBABILITY'] = 0
df = pd.merge(df, weights, on = ['KCSE_YEAR', 'GENDER', 'STEM'])

def data_preprocessing(df, predictors = predictors, onehotencoded_columns = onehotencoded_columns, labelencoded_columns = label
encoded_columns):

    lbe = LabelEncoder()
    df = df[predictors]
    df_corr_map = pd.DataFrame()
    df_encoded = pd.get_dummies(df, columns = onehotencoded_columns)

    for column in labelencoded_columns:
        df_encoded[column] = lbe.fit_transform(df[column])

    for column in onehotencoded_columns + labelencoded_columns:
        df_corr_map[column] = lbe.fit_transform(df[column])

    df_encoded['PROBABILITY'] = df_corr_map['PROBABILITY'] = df['PROBABILITY']

    return df_encoded, df_corr_map
```

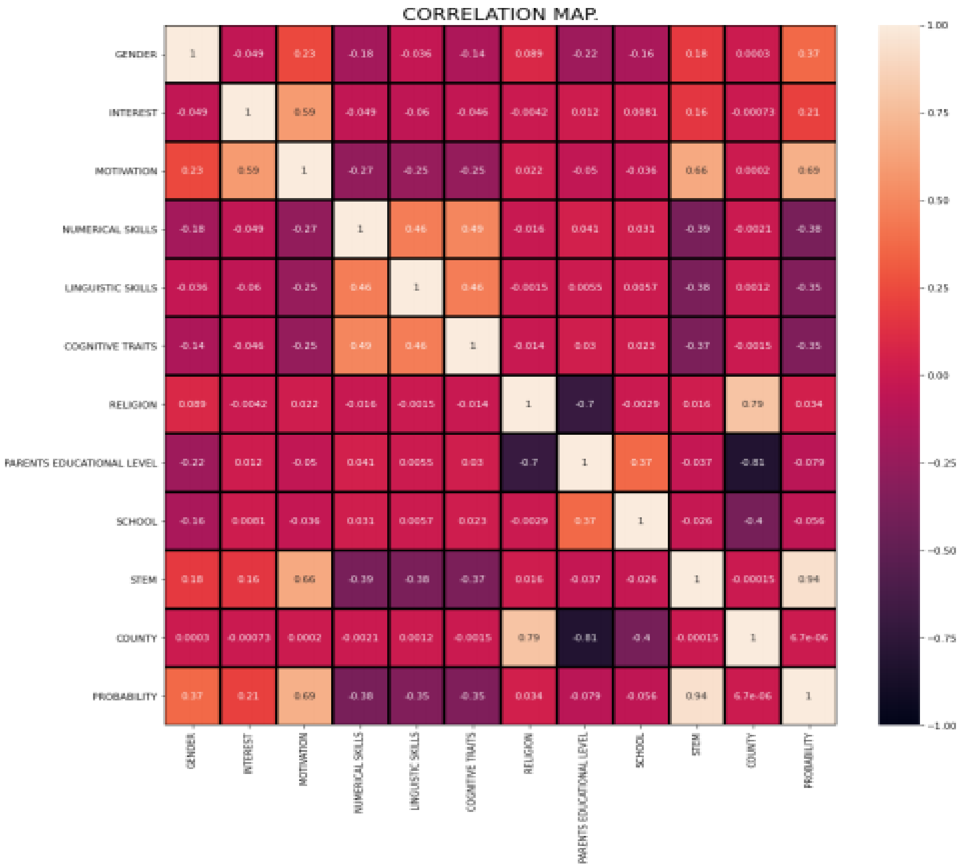


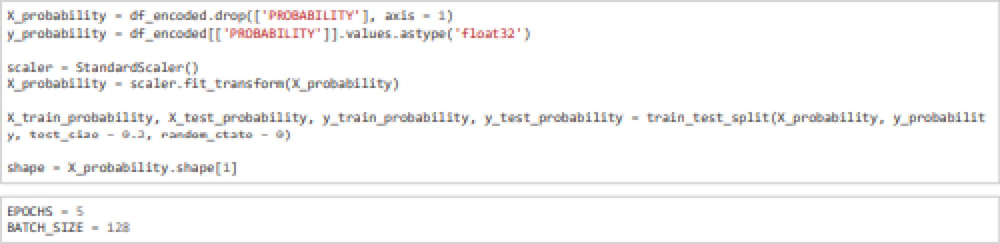
	COUNTY	PROBABILITY	GENDER_F	GENDER_M	INTEREST_NO	INTEREST_YES	MOTIVATION_NO	MOTIVATION_YES	NUMERICAL SKILLS_A	NUMERICAL SKILLS_B	NI
0	7	0.638501	0	1	0	1	0	1	1	0	
1	38	0.638501	0	1	1	0	1	0	0	1	
2	34	0.638501	0	1	1	0	1	0	0	1	
3	11	0.638501	0	1	0	1	0	1	0	1	
4	22	0.638501	0	1	0	1	0	1	0	1	
...	...	...	...	...	...	...	...	...	...	...	
591343	36	0.668528	0	1	0	1	0	1	1	0	
591344	9	0.668528	0	1	0	1	0	1	0	0	
591345	23	0.668528	0	1	0	1	0	1	0	0	
591346	46	0.668528	0	1	0	1	0	1	0	0	
591347	39	0.668528	0	1	1	0	1	0	0	0	

591348 rows × 35 columns



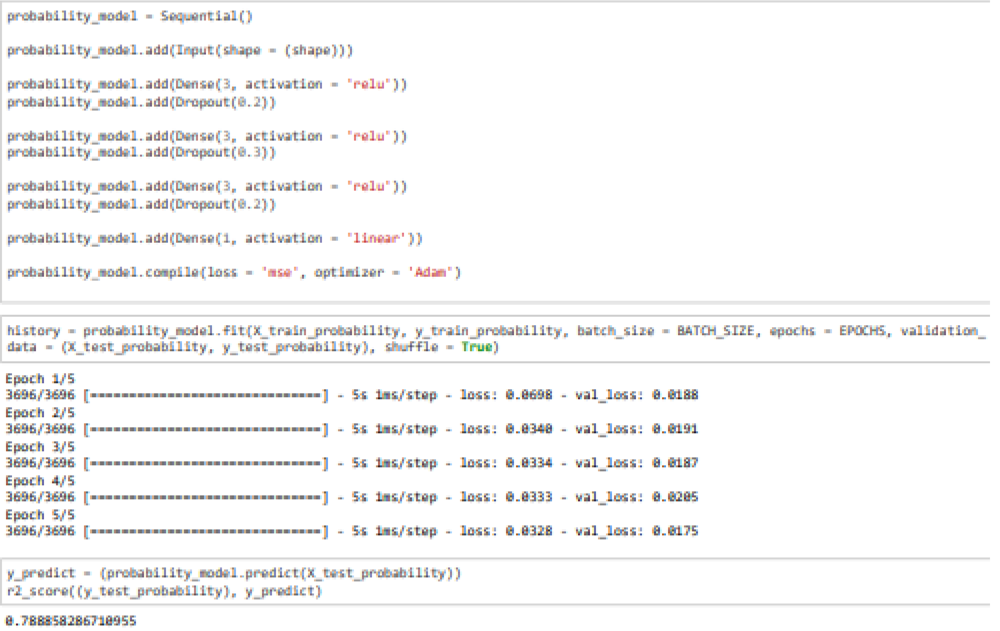
```
:  plt.figure(figsize = (15, 15))
   sns.heatmap(df_corr_map.corr(), annot = True, vmin = -1, vmax = 1, fmt='.2g', linewidths = 2, linecolor = 'black')
   plt.title('CORRELATION MAP.', fontsize = 20)
   plt.savefig (f"graphs/correlation map.png", bbox_inches = "tight")
```





```
X_probability = df_encoded.drop(['PROBABILITY'], axis = 1)
y_probability = df_encoded[['PROBABILITY']].values.astype('float32')

scaler = StandardScaler()
X_probability = scaler.fit_transform(X_probability)

X_train_probability, X_test_probability, y_train_probability, y_test_probability = train_test_split(X_probability, y_probabilit
y, test_size = 0.2, random_state = 0)

shape = X_probability.shape[1]
```

```
EPOCHS = 5
BATCH_SIZE = 128
```



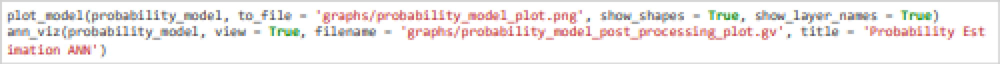
```
probability_model = Sequential()

probability_model.add(Input(shape = (shape)))

probability_model.add(Dense(3, activation = 'relu'))
probability_model.add(Dropout(0.2))

probability_model.add(Dense(3, activation = 'relu'))
probability_model.add(Dropout(0.3))

probability_model.add(Dense(3, activation = 'relu'))
probability_model.add(Dropout(0.2))

probability_model.add(Dense(1, activation = 'linear'))

probability_model.compile(loss = 'mse', optimizer = 'Adam')
```

```
history = probability_model.fit(X_train_probability, y_train_probability, batch_size = BATCH_SIZE, epochs = EPOCHS, validation_
data = (X_test_probability, y_test_probability), shuffle = True)
```

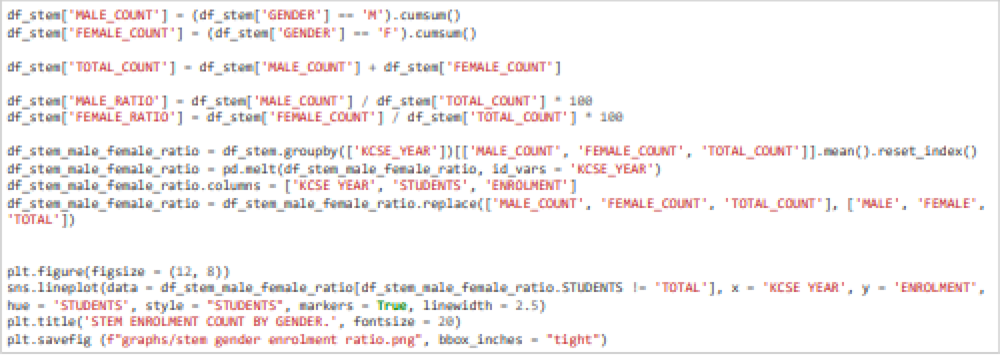
```
Epoch 1/5
3696/3696 [==============================] - 5s 1ms/step - loss: 0.0698 - val_loss: 0.0188
Epoch 2/5
3696/3696 [==============================] - 5s 1ms/step - loss: 0.0340 - val_loss: 0.0191
Epoch 3/5
3696/3696 [==============================] - 5s 1ms/step - loss: 0.0334 - val_loss: 0.0187
Epoch 4/5
3696/3696 [==============================] - 5s 1ms/step - loss: 0.0333 - val_loss: 0.0205
Epoch 5/5
3696/3696 [==============================] - 5s 1ms/step - loss: 0.0328 - val_loss: 0.0175
```

```
y_predict = (probability_model.predict(X_test_probability))
r2_score((y_test_probability), y_predict)
```

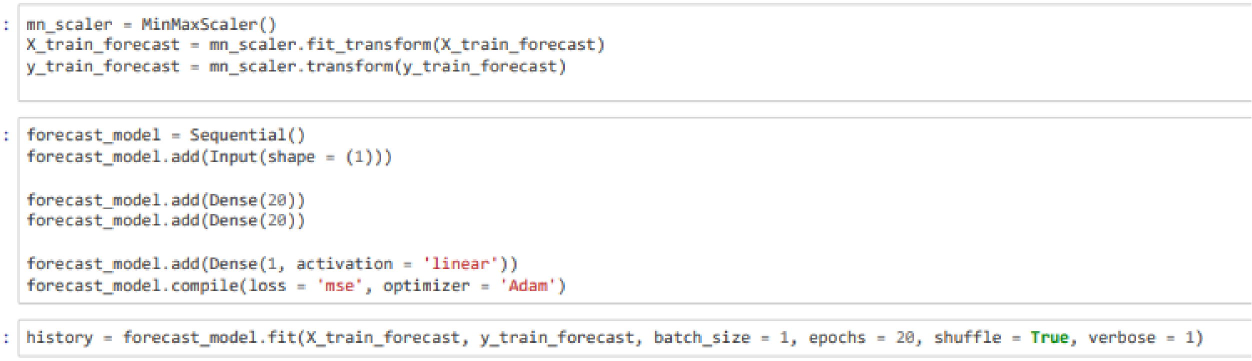
```
0.788858286710955
```



```
plot_model(probability_model, to_file = 'graphs/probability_model_plot.png', show_shapes = True, show_layer_names = True)
ann_viz(probability_model, view = True, filename = 'graphs/probability_model_post_processing_plot.gv', title = 'Probability Est
imation ANN')
```



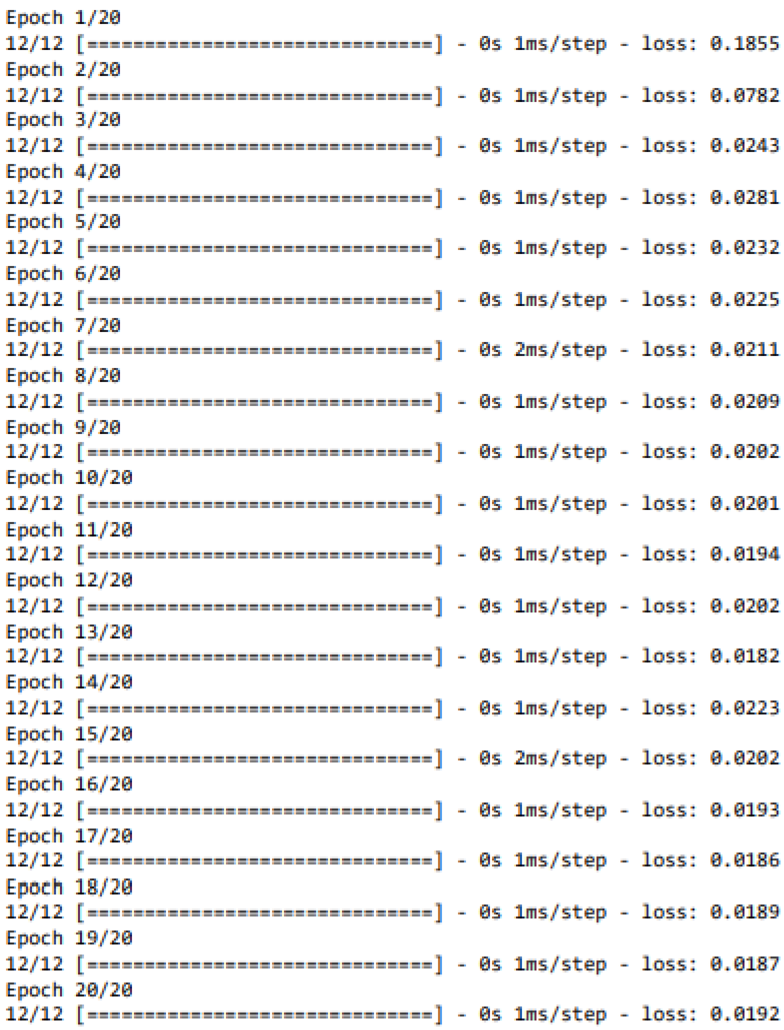
```
df_stem['MALE_COUNT'] = (df_stem['GENDER'] == 'M').cumsum()
df_stem['FEMALE_COUNT'] = (df_stem['GENDER'] == 'F').cumsum()

df_stem['TOTAL_COUNT'] = df_stem['MALE_COUNT'] + df_stem['FEMALE_COUNT']

df_stem['MALE_RATIO'] = df_stem['MALE_COUNT'] / df_stem['TOTAL_COUNT'] * 100
df_stem['FEMALE_RATIO'] = df_stem['FEMALE_COUNT'] / df_stem['TOTAL_COUNT'] * 100

df_stem_male_female_ratio = df_stem.groupby(['KCSE_YEAR'])[['MALE_COUNT', 'FEMALE_COUNT', 'TOTAL_COUNT']].mean().reset_index()
df_stem_male_female_ratio = pd.melt(df_stem_male_female_ratio, id_vars = 'KCSE_YEAR')
df_stem_male_female_ratio.columns = ['KCSE YEAR', 'STUDENTS', 'ENROLMENT']
df_stem_male_female_ratio = df_stem_male_female_ratio.replace(['MALE_COUNT', 'FEMALE_COUNT', 'TOTAL_COUNT'], ['MALE', 'FEMALE',
'TOTAL'])


plt.figure(figsize = (12, 8))
sns.lineplot(data = df_stem_male_female_ratio[df_stem_male_female_ratio.STUDENTS != 'TOTAL'], x = 'KCSE YEAR', y = 'ENROLMENT',
hue = 'STUDENTS', style = "STUDENTS", markers = True, linewidth = 2.5)
plt.title('STEM ENROLMENT COUNT BY GENDER.', fontsize = 20)
plt.savefig (f"graphs/stem gender enrolment ratio.png", bbox_inches = "tight")
```



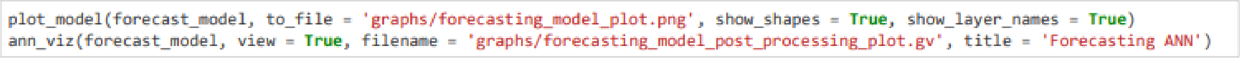
```
mn_scaler = MinMaxScaler()
X_train_forecast = mn_scaler.fit_transform(X_train_forecast)
y_train_forecast = mn_scaler.transform(y_train_forecast)
```

```
forecast_model = Sequential()
forecast_model.add(Input(shape = (1)))

forecast_model.add(Dense(20))
forecast_model.add(Dense(20))

forecast_model.add(Dense(1, activation = 'linear'))
forecast_model.compile(loss = 'mse', optimizer = 'Adam')
```

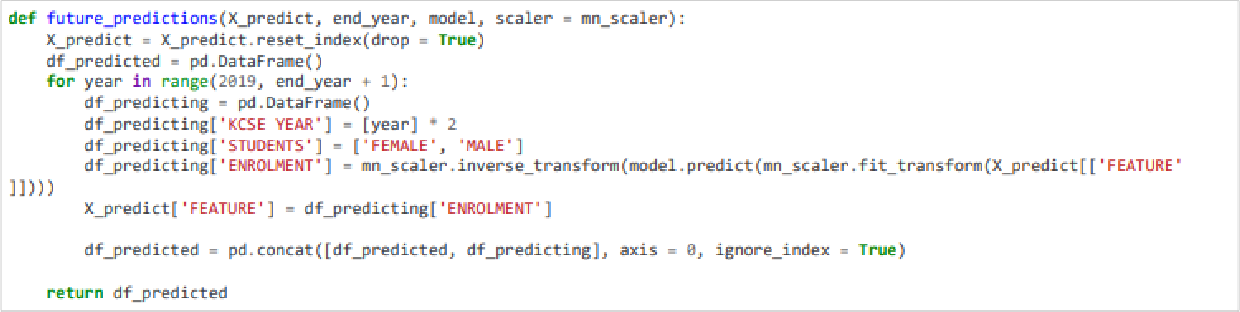
```
history = forecast_model.fit(X_train_forecast, y_train_forecast, batch_size = 1, epochs = 20, shuffle = True, verbose = 1)
```



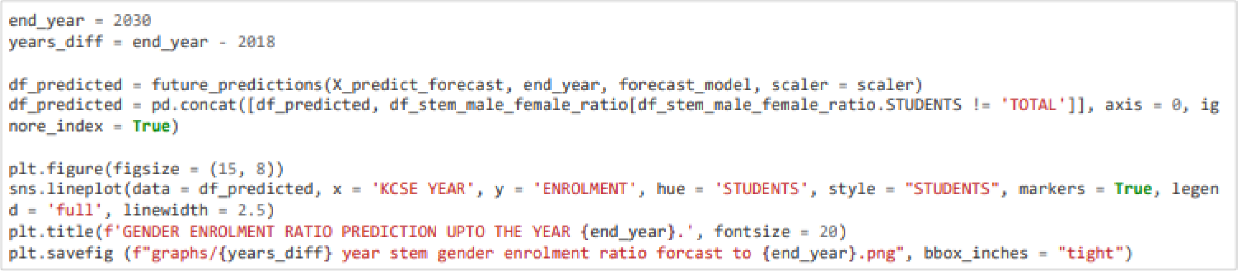
```
Epoch 1/20
12/12 [==============================] - 0s 1ms/step - loss: 0.1855
Epoch 2/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0782
Epoch 3/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0243
Epoch 4/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0281
Epoch 5/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0232
Epoch 6/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0225
Epoch 7/20
12/12 [==============================] - 0s 2ms/step - loss: 0.0211
Epoch 8/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0209
Epoch 9/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0202
Epoch 10/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0201
Epoch 11/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0194
Epoch 12/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0202
Epoch 13/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0182
Epoch 14/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0223
Epoch 15/20
12/12 [==============================] - 0s 2ms/step - loss: 0.0202
Epoch 16/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0193
Epoch 17/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0186
Epoch 18/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0189
Epoch 19/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0187
Epoch 20/20
12/12 [==============================] - 0s 1ms/step - loss: 0.0192
```



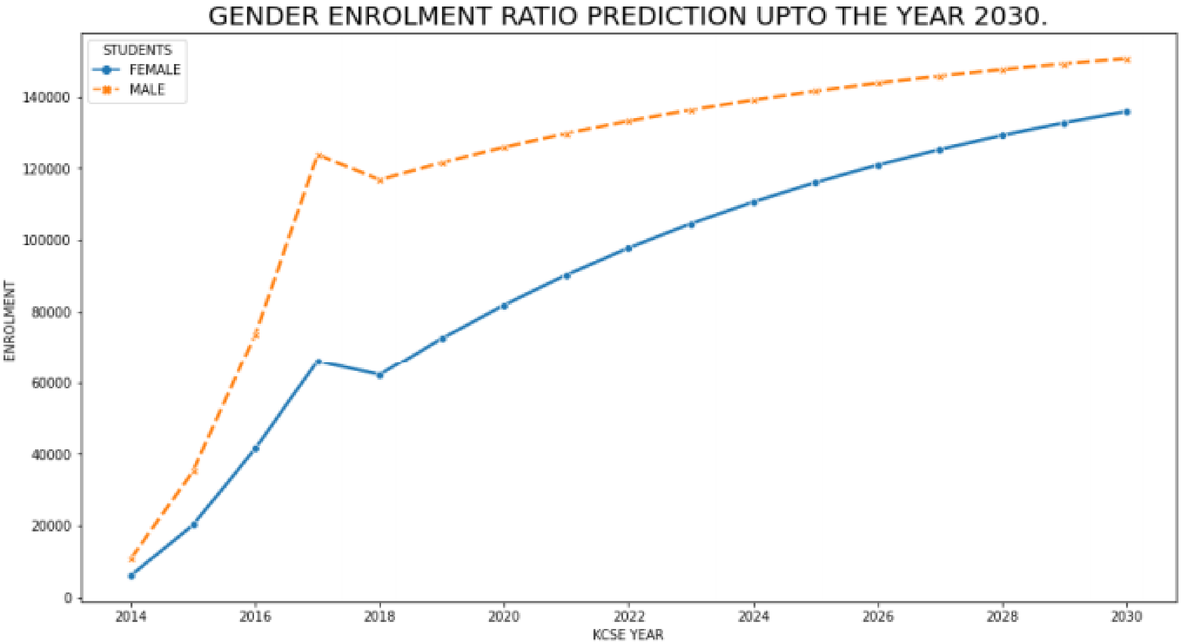
```
plot_model(forecast_model, to_file = 'graphs/forecasting_model_plot.png', show_shapes = True, show_layer_names = True)
ann_viz(forecast_model, view = True, filename = 'graphs/forecasting_model_post_processing_plot.gv', title = 'Forecasting ANN')
```

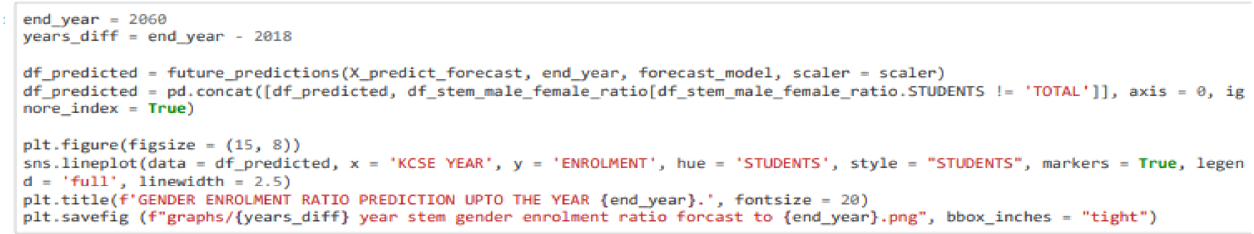


```
def future_predictions(X_predict, end_year, model, scaler = mn_scaler):
    X_predict = X_predict.reset_index(drop = True)
    df_predicted = pd.DataFrame()
    for year in range(2019, end_year + 1):
        df_predicting = pd.DataFrame()
        df_predicting['KCSE YEAR'] = [year] * 2
        df_predicting['STUDENTS'] = ['FEMALE', 'MALE']
        df_predicting['ENROLMENT'] = mn_scaler.inverse_transform(model.predict(mn_scaler.fit_transform(X_predict[['FEATURE'
]])))
        X_predict['FEATURE'] = df_predicting['ENROLMENT']

        df_predicted = pd.concat([df_predicted, df_predicting], axis = 0, ignore_index = True)

    return df_predicted
```

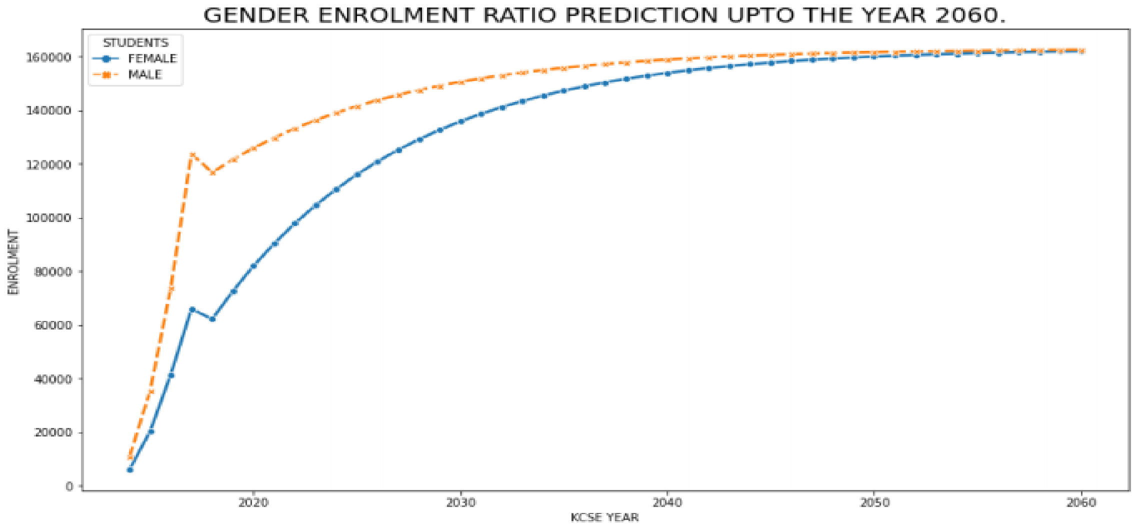


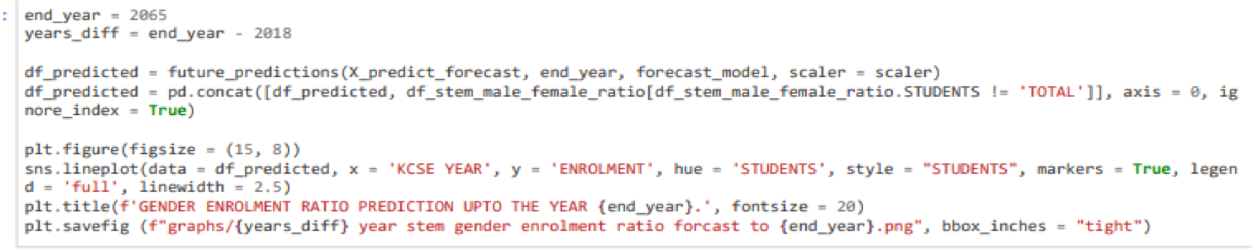
```
end_year = 2030
years_diff = end_year - 2018

df_predicted = future_predictions(X_predict_forecast, end_year, forecast_model, scaler = scaler)
df_predicted = pd.concat([df_predicted, df_stem_male_female_ratio[df_stem_male_female_ratio.STUDENTS != 'TOTAL']], axis = 0, ig
nore_index = True)

plt.figure(figsize = (15, 8))
sns.lineplot(data = df_predicted, x = 'KCSE YEAR', y = 'ENROLMENT', hue = 'STUDENTS', style = "STUDENTS", markers = True, legen
d = 'full', linewidth = 2.5)
plt.title(f'GENDER ENROLMENT RATIO PREDICTION UPTO THE YEAR {end_year}.', fontsize = 20)
plt.savefig (f"graphs/{years_diff} year stem gender enrolment ratio forcast to {end_year}.png", bbox_inches = "tight")
```



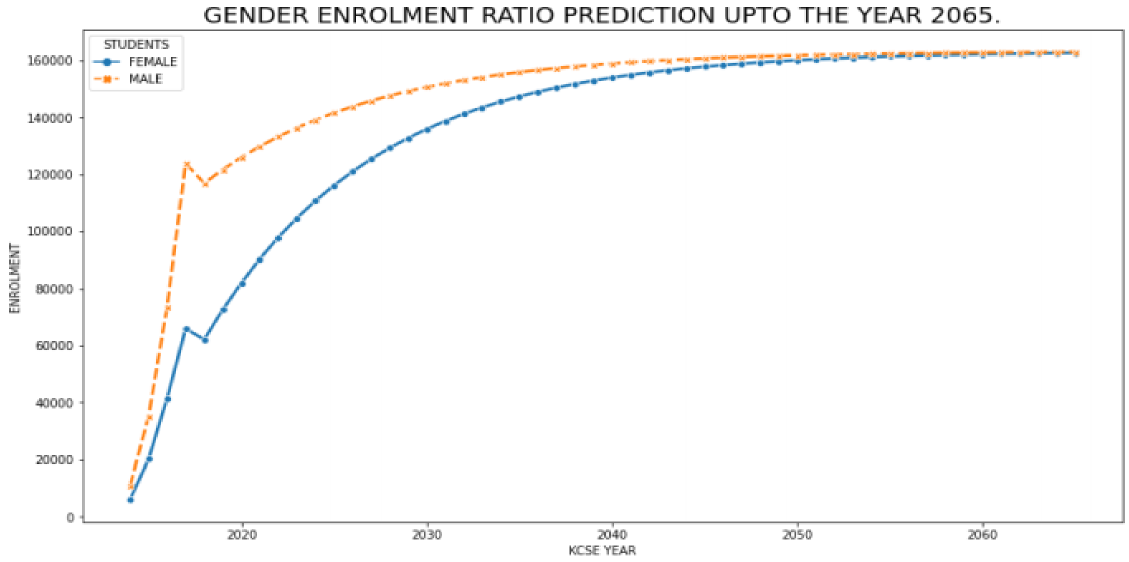


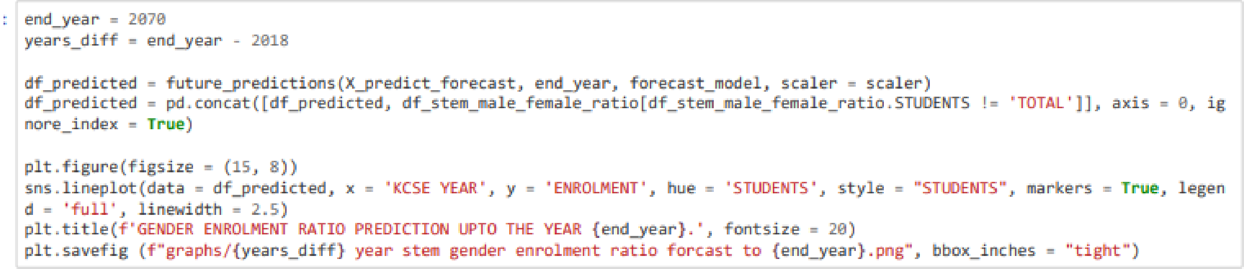
```
end_year = 2060
years_diff = end_year - 2018

df_predicted = future_predictions(X_predict_forecast, end_year, forecast_model, scaler = scaler)
df_predicted = pd.concat([df_predicted, df_stem_male_female_ratio[df_stem_male_female_ratio.STUDENTS != 'TOTAL']], axis = 0, ig
nore_index = True)

plt.figure(figsize = (15, 8))
sns.lineplot(data = df_predicted, x = 'KCSE YEAR', y = 'ENROLMENT', hue = 'STUDENTS', style = "STUDENTS", markers = True, legen
d = 'full', linewidth = 2.5)
plt.title(f'GENDER ENROLMENT RATIO PREDICTION UPTO THE YEAR {end_year}.', fontsize = 20)
plt.savefig (f"graphs/{years_diff} year stem gender enrolment ratio forcast to {end_year}.png", bbox_inches = "tight")
```



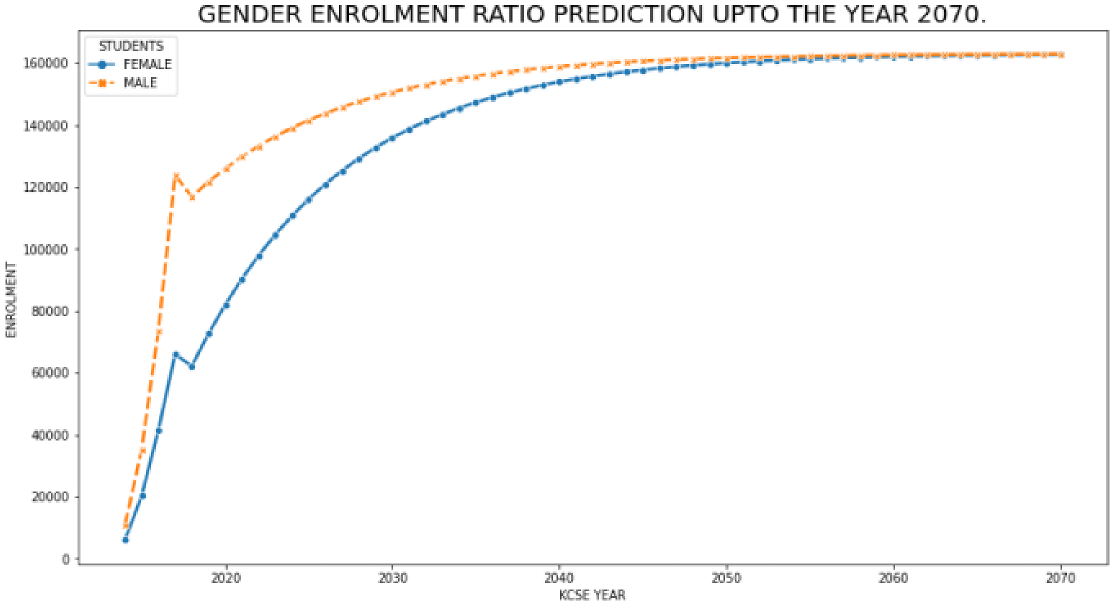


```
end_year = 2065
years_diff = end_year - 2018

df_predicted = future_predictions(X_predict_forecast, end_year, forecast_model, scaler = scaler)
df_predicted = pd.concat([df_predicted, df_stem_male_female_ratio[df_stem_male_female_ratio.STUDENTS != 'TOTAL']], axis = 0, ig
nore_index = True)

plt.figure(figsize = (15, 8))
sns.lineplot(data = df_predicted, x = 'KCSE YEAR', y = 'ENROLMENT', hue = 'STUDENTS', style = "STUDENTS", markers = True, legen
d = 'full', linewidth = 2.5)
plt.title(f'GENDER ENROLMENT RATIO PREDICTION UPTO THE YEAR {end_year}.', fontsize = 20)
plt.savefig (f"graphs/{years_diff} year stem gender enrolment ratio forcast to {end_year}.png", bbox_inches = "tight")
```





```
: end_year = 2070
  years_diff = end_year - 2018

  df_predicted = future_predictions(X_predict_forecast, end_year, forecast_model, scaler = scaler)
  df_predicted = pd.concat([df_predicted, df_stem_male_female_ratio[df_stem_male_female_ratio.STUDENTS != 'TOTAL']], axis = 0, ig
  nore_index = True)

  plt.figure(figsize = (15, 8))
  sns.lineplot(data = df_predicted, x = 'KCSE YEAR', y = 'ENROLMENT', hue = 'STUDENTS', style = "STUDENTS", markers = True, legen
  d = 'full', linewidth = 2.5)
  plt.title(f'GENDER ENROLMENT RATIO PREDICTION UPTO THE YEAR {end_year}.', fontsize = 20)
  plt.savefig (f"graphs/{years_diff} year stem gender enrolment ratio forcast to {end_year}.png", bbox_inches = "tight")
```


GENDER ENROLMENT RATIO PREDICTION UPTO THE YEAR 2070.
STUDENTS
FEMALE
MALE
ENROLMENT
KCSE YEAR