PERFORMANCE PREDICTION IN MATHEMATICS USING EDUCATIONAL DATA MINING TECHNIQUES: A CASE OF MZUMBE UNIVERSITY IN TANZANIA.

PAUL KAVISHE MUSHI

MASTER OF SCIENCE IN INFORMATION SYSTEMS

THE UNIVERSITY OF DODOMA

SEPTEMBER, 2020

PERFORMANCE PREDICTION IN MATHEMATICS USING EDUCATIONAL DATA MINING TECHNIQUES: A CASE OF MZUMBE UNIVERSITY IN TANZANIA.

BY

PAUL KAVISHE MUSHI

A STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SYSTEMS

THE UNIVERSITY OF DODOMA

SEPTEMBER, 2020

DECLARATION AND COPYRIGHT

I, **Paul Kavishe Mushi,** declare that this **study** is my own original work and that it has not been presented and will not be presented to any other University for a similar or any other degree award.

Signature of the candidate: **Date**: 27/11/2020

CERTIFICATION

The undersigned certify that he has read and hereby recommend for acceptance by the University of Dodoma, This study entitled**, "Performance Prediction In Mathematics Using Educational Data Mining Techniques: A Case of Mzumbe University In Tanzania"** in partial fulfillments of the requirements for award of the Master's Degree of Science in Information Systems at the University of Dodoma.

………………………………………………………

DR. DANIEL NGONDYA

(SUPERVISOR)

Date: 27/11/2020

ACKNOWLEDGEMENT

First of all, thanks and honor be to Almighty God for the gift of life and healthy to pursue my master’s degree, I know it’s you Lord making things work righteous.

Second, many thanks may go to the sponsor of my studies Mzumbe University that made me study comfortably since coursework and as a result perform this study on time.

Third, great thanks to the Mzumbe University management for allowing me to collect data as a result marked the successful of the study on time. Thanks to staffs from admission office, loan office, accommodation office, ARIS office, Department of mathematics and statistics and others who in one way or another helped me to accomplish this work.

Fourth, I would like to thanks my supervisor, Dr. Daniel Ngondya for the moral support in this study and guidance as well as encouragement to work hard so that to finish on time, May God bless him abundantly together with his family.

Fifth, I would like to appreciate all the academic staffs and management at CIVE for their academic and moral supports and knowledge sharing since coursework up to the dissertation.

Sixth, special thanks to my classmates of 2018/2019 at CIVE to mention few, Mr. Yusuph Rajabu and Miss. Mevis Steven Mwaluanda for their team work spirit since coursework up to the dissertation.

Seventh, thanks to the colleagues Ibrahim Mtandu and Eliya Masesa for their technical cooperation.

Lastly but not least, thanks to my lovely family for their support, love, prayers and encouragement to go for further studies, God bless you all.

DEDICATION

I would like to dedicate my work to my late father Project Gasper Mushi, my lovely mother Rafaela Michael, my lovely spouse Neisha Daud Msangi and my beloved son Ethan Paul Mushi, this dissertation is a reason why I didn't spend much time with you.

ABSTRACT

Nowadays, Higher Learning Institutions (HLIs) store a large amount of students' data. However, those data are not widely used to solve the academic problems of the students that are available at the HLIs such as poor performance of the students in some of the courses. Educational Data Mining (EDM) is the technology that can be applied to predict the performance of the students on the available dataset at the HLIs. This study intended to solve the problem of poor performance in Mathematics for degree management students at HLIs using EDM techniques. The purpose of the study was to predict Management degree Students' performance in Mathematics using EDM taking Mzumbe University (MU) as a case study. The quantitative research approach was applied in this study basing on the design science steps. Secondary data were collected to create the dataset through document review from examination office (final examination (FE), course work (CW) and Remarks), admission office (age, gender, entry category and ordinary level mathematics grades), accounts office (sponsorship details), department of mathematics and statistics (number of instructors) and accommodation office (living location) at MU including Main campus Morogoro and Mbeya Campus. Different Machine Learning (ML) algorithms were applied on training set (60%) such as K-Nearest Neighbor (K-NN), Random Forest (RF), Decision Tree (DT), Support Vector Classification (SVC) and Multilayer Perceptron (MLP). ML algorithms were validated using 10-fold cross-validation and validation dataset (20%) and the best algorithms were RF, DT and K-NN. During the evaluation of the three best ML algorithms using 20% of the dataset, RF ML algorithm was found to be the best for model development in mathematics performance prediction in this study with the accuracy of 99% and F1-scores of 99% and 100% for fail and pass class respectively. Moreover, DT was able to generate rules that were applied to recommend the minimum grade of D for ordinary level mathematics in admission to degree management students to reduce the failure rate at HLIs.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ARIS	Academic Records Information System
BHSM	Bachelor of Health System Management
BHRM	Bachelor of Human Resource Management
BHRM-MCC	Bachelor of Human Resource Management at Mbeya Campus
BHRM-MU	Bachelor of Human Resource Management at Main Campus
BLGM	Bachelor of Local Government Management
BPA-RAM	Bachelor of Public Administration in Records and Archives Management
CART	Classification And Regression Trees
CW	Course Work
DM	Data Mining
DT	Decision Tree
EDM	Educational Data Mining
FE	Final Examination
FN	False Negative
FP	False Positive
FPR	False Positive Rate
HESLB	Higher Education Students' Loans Board
HLIs	Higher Learning Institution
ICT	Information Communication Technology
ID3	Iterative Dichotomiser 3
IDM	Institute of Development and Management

KDD	Knowledge Discovery in Database
K-NN	K-Nearest Neighbor
LR	Logistic Regression
MCC	Mbeya Campus College
ML	Machine Learning
MLP	Multilayer Perceptron
MU	Mzumbe University
NB	Naïve Bayes
RF	Random Forest
RML	Reinforcement Machine Learning
RPL	Recognition of Prior Learning
SML	Supervised Machine Learning
SMO	Sequential Minimal Optimization
SOPAM	School of Public Administration and Management
SVC	Support Vector Classification
SVM	Support Vector Machine
TCU	Tanzania Commission for Universities
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UDOM	University of Dodoma
UDSM	University of Dar es Salaam
UML	Unsupervised Machine Learning

CHAPTER ONE

INTRODUCTION

1.1. General Introduction

Higher Learning Institutions (HLIs) in the world are doing a better job of educating people to increase knowledge and innovations. Private sectors and governments are investing a lot of funds to sponsor students worldwide for improving economic growth, bring change to society and reduce unemployment problems through self-employment skills after University life (Vista College, 2019). Even parents have high expectations of good performance and completion of studies successfully from their children. Hence, different stakeholders in HLIs such as parents or guardians, sponsors, government, students, lecturers and decision-makers have obligations to fulfil the good academic performance to scholars. Also, students in HLIs devote their time to stay at University for acquiring knowledge and skills. However, students may pass or fail in various programmes at the Universities and Colleges in the developed and developing countries. This might be due to various reasons such as demographic factors (age, gender, and marital status), economic factors of the parents, living location and sponsorship (Osmanbegovic & Suljic, 2012). Hence there was a need to address the problem of mass failure in HLIs.

There have been numerous studies conducted outside Africa on mass failure in computer science and engineering disciplines in HLIs (Ashenafi, 2017). Also, there was another study concerning the performance predictions in Mathematics at the secondary level (Cortez et al., 2003). One of the research was done by Sivasakthi (2018) using Data Mining (DM) Classification Algorithms to predict Students' Performance in Introductory programming which was the Computer Science discipline. First-year students have to study that course but the performance was not good hence the research was done using DM techniques to predict earlier the students who are likely to pass or fail (Sivasakthi, 2018).

Likewise, in developing countries such as Nigeria, the researcher Olatunji et al. (2017) conducted a study at the Federal University of Technology Akure. The researcher considered the students pursuing Quantity Surveying, Estate Management, Architecture and Industrial Design both belonging to the construction discipline. The study was concerned with the assessment of academic (continuous assessment, Grade

Point Average and entry standards) and non-academic (economic issues and family matters) factors. The researcher assessed factors affecting the performance of undergraduate students to improve the performance of the students in HLIs (Olatunji et al., 2017).

Furthermore, in Tanzania, the researcher Fussy (2018) conducted a study at Mkwawa University College of Education concerning the academic performance of the students in Higher Learning Institutions. The study focused on the quality of academic advising given to students when they join Universities at the earlier stages. In that study, the researcher found that during the first semester of the first year, science-based education students in the 2015/2016 academic year, almost half of the students (46% of 338 students) failed in one course (Fussy, 2018). The researcher concluded that there is a need to revisit and improve student advisory services to obtain higher retention and graduation rates.

However, those are not the only disciplines in HLIs with mass failure. The students pursuing Management programmes also face the problem of failing in their courses massively. Also, courses like Mathematics may have a high failure rate to students due to various factors that have been identified through literature reviews and from the field such as mathematics grades at ordinary level. Hence this study could inform the decision-makers such as Tanzania Commission for Universities (TCU), Higher Education Students' Loans Board (HESLB) and HLIs' Management team on the appropriate minimum entry qualifications in Mathematics to Management programmes in HLIs. In Tanzania, there are HLIs that offer degree programmes in management discipline which include art and science programmes. Those HLIs are like Mzumbe University (MU), University of Dar es salaam (UDSM), The University of Dodoma (UDOM) and Institute of social work. Other HLIs are Jordan University College, Local Government Training Institute, Moshi Co-operative University and The Mwalimu Nyerere Memorial Academy (TCU, 2020).

Moreover, considering MU as a case study, these Management programmes differ in admission entry criteria such that others include those with science subjects' background and other art subjects background. Moreover, Mathematics entry criteria and courses are undertaken by students after being admitted differ. Hence, this study intends to focus on students in four Management programmes who have been admitted without Mathematics background consideration. The programmes are Bachelor of Local Government Management (BLGM), Bachelor of Health System Management (BHSM), Bachelor of Human Resource Management (BHRM) and Bachelor of Public Administration-Records and Archives Management (BPA-RAM). The programmes are provided at the School of Public Administration and Management (SOPAM) and MU Mbeya Campus College (MCC). The programmes have the same admission entry criteria and it is mandatory to study Mathematics course during the first year at the first semester.

In addition, various Educational data mining (EDM) techniques have been used in research for the prediction of the students' performance such as Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (K-NN), Artificial Neural Network (ANN), Random Forest (RF) and Decision Tree (DT). The study intended to apply some of the EDM techniques to predict students who are more likely to fail in Mathematics course at HLIs.

Therefore, there could be a trend of students' mass failure in Mathematics for Management programmes if the research was not going to be carried out. Hence, there was a need to predict the performance at the earlier stages for first-year students for reducing the failure rate in HLIs. The study based on using EDM techniques such as ANN, K-NN, and SVM, DT and RF and come up with the best ML algorithm for model development in mathematics performance prediction.

1.2. Statement of the Problem

HLIs are playing a great role to educate people that leads to knowledge and innovation among the societies as well as the development of the nation. Parents and sponsors such as government through HESLB are investing a lot of funds to students. Moreover, students themselves invest their time to achieve higher education at the Universities or colleges with much expectations of increasing economic growth of the nation. Also, students do expect employments, get better jobs and attain a healthier and happier life after attaining higher education.

However, worldwide HLIs are facing the problem of students' performance including high failure rate in courses at various disciplines such as Engineering, Computer Science, Management and Business. Failing may lead to supplementary, carryovers or discontinuation during the semester or at the end of each academic year. Due to this problem of mass failure at HLIs, it may lead to a high number of students who might not complete the studies hence may lack qualities of being employed. Also, loss of funds by parents and government through HESLB, unhealthier and unhappier life in the surrounding societies. At MU, students pursuing Management programmes, Mathematics was one of the courses with high failure rate in the first semester during the first year.

Therefore, most of the performance prediction researches have focused on the discipline of computer science and engineering. Hence, there is a need for researchers to look into other disciplines using EDM techniques such as ANN, DT, RF, SVM and K-NN. Therefore, this study intended to focus into other disciplines such as Management discipline by training various ML algorithms widely applied in EDM and come up with the best ML algorithm for model development in mathematics performance prediction.

1.3. Objectives

1.3.1. Main Objective

The main objective of this study was to develop a mathematics performance predictive model for Management degree Students using EDM Techniques.

1.3.2. Specific Objectives

1. To identify the requirements for training ML algorithms in Mathematics performance prediction.
2. To identify the best trained ML algorithms in Mathematics performance prediction.
3. To evaluate the best selected trained ML algorithms in mathematics performance prediction.

1.3.3. Research Questions

1. What are the requirements for training ML algorithms in Mathematics Performance Prediction?
2. What are the best trained ML algorithms in Mathematics performance prediction?
3. What are the performance evaluations for the best selected trained ML algorithms in Mathematics performance prediction?

1.4. Significance of the Study

This study will contribute awareness to different stakeholders such as students themselves, lecturers, University Management (decision-makers), government (HESLB and TCU) as well as researchers. They will be aware of the factors that highly lead to students fail in Mathematics course and take the measures at the earlier stages to come with good performance. As a result students avoid supplementary, carryovers or even discontinuation.

To the University (admission office) and TCU will be aware of the Mathematics background and decide on minimum grade for criteria on selection. Also, the Lecturers and decision-makers (University Management) will be aware on the factors for students to fail in Mathematics hence they will use the trained ML algorithm to predict the performance of the new incoming students.

To the government, HESLB as the main sponsor in Higher Learning Institutions will be aware on the needs of the students' sponsorship to reduce the failure rate in Mathematics if found to be the contributing factor in performance.

Moreover, to the researchers, the application of EDM in Tanzania is not much explored. Hence, researchers in Tanzania and the world at large will fill the gap that was not covered in this research area and hence solve more problems in the academic area.

Furthermore, to the students themselves, through earlier identification of the factors that highly lead to failing in Mathematics course, students may do quick decisions such as studying hard in Mathematics course to avoid supplementary or carryover at the University.

Therefore, with the help of EDM Techniques from the existing data at MU, knowledge was extracted with the data set available. Due to the application of the various EDM techniques such as K-NN, RF, DT, ANN and SVM, the best algorithm was identified for model development in prediction of mathematics performance.

1.5. Scope of the Study

The scope of this study was limited to Degree students pursuing Management Programmes which are BHRM, BHSM, BPA-RAM and BLGM from Main Campus Morogoro and Mbeya Campus at MU. Moreover, this study based in one course, that was Mathematics undertaken in the first semester during the first year from 2014/2015 to 2018/2019 academic year due to the financial and time constraints.

1.6. Limitations of the Study

The study had some challenges in the data collection such as time and financial constraints. Therefore, this led to the decrease of some of the features such as students' marital status, parent's educational status, and parents' economic status whose collection would need more time and financial support.

1.7. Organization of the Study

Apart from this chapter one of the introduction, the rest of this study was organized into chapter two, chapter three, chapter four and chapter five as follows. Chapter two describes the literature review that contains the concepts applied in this study such as DM, EDM, Machine Learning (ML), classification algorithms, student academic performance, HLIs, supplementary examinations and evaluation metrics for classification problems. Also, at chapter two the related works were reviewed in students' academic performance prediction and the review was conducted in predictor variables for mathematics performance prediction. ML algorithms such as K-NN, RF, DT, SVC and MLP were found in other studies with their accuracies and widely used in academic performance predictions. In this study, mentioned algorithms were reviewed to be included in the training of ML algorithms to find the best algorithm in mathematics performance prediction. Furthermore, some of the predictor variables were identified through this chapter such as age, gender, entry category, campus location, living location, sponsorship and academic background while other variables identified at the field. Similarly, the research gap and conceptual framework were addressed in the chapter two of this study.

Likewise, chapter three contains methodology which shows the research setting as to where the research was conducted and in this study, it was at MU Morogoro and Mbeya. The quantitative research approach was applied and steps for design science were followed that include problem identification, setting objectives, development, evaluation and conclusion. Secondary data were collected through document reviews such as excel files from admission office, accounts office, accommodation office and from Academic Records Information System (ARIS). Data were explored and prepared for training ML algorithms. Also, it described the distribution of dataset into training (60%), validation (20%) and testing 20%. Likewise, evaluation metrics considered were accuracy, precision, recall and F-measure. Ethical consideration, reliability and validity were considered such as not to disclose the students' information anywhere and not to produce the results to the other organization except UDOM.

Moreover, chapter four includes findings and discussion based on the three research questions from the specific objectives. In identifying the requirements for training ML algorithms, another predictor variable was identified at the field which was a number of instructors to teach mathematics course. Including other 9 predictor variables from related works made a total of 10 predictor variables to predict students' performance in mathematics. Dataset was created using all features and feature importance was calculated to see how they influence performance to students in a mathematics course. Furthermore, chi-square, Z test and Analysis of Variance (ANOVA) statistical tests were applied to find the relationships between predictor variables and to the target variable. K-NN, RF, DT, SVC and MLP algorithms were trained and validated using validation dataset whereby accuracies and F-scores were calculated to obtain best algorithms which were K-NN, RF and DT for evaluation to result into one best algorithm. After evaluation of the best three algorithms, RF was the best algorithm in mathematics performance prediction with an accuracy of 99% and F1-scores of 99% and 100% for fail and pass class of the target variable respectively.

Similarly, chapter five describes the conclusion, recommendations and research areas for further studies. It was concluded that the RF algorithm was the best in this study for mathematics performance prediction at HLIs to degree management students. The accuracy 99% and F1-scores of 99% and 100% for fail and pass class of the target variable respectively. Also, it was recommended that HLIs should apply ML algorithms to predict the students' performance earlier to advise them and reduce the failure rate to a particular course. Moreover, it was recommended minimum grade of D for mathematics at the ordinary level to TCU and admission offices of HLIs during admission to management degree. For future work, the other researcher may consider the online courses to predict the performance of the students such as in Moodle as an e-learning management system. Likewise, the researchers may consider the primary data collection to have more variables by involving current students at HLIs in predicting student's performance.

CHAPTER TWO
LITERATURE REVIEW

2.1. Introduction

This literature review chapter of the study describes the works of various researchers in EDM concerning the performance predictions of the students at HLIs. The chapter is organized as follows; concepts applied in this study, empirical literature review, research gap and conceptual framework.

2.2. Concepts Applied in the Study

2.2.1. Data Mining

In Data Science field, DM is the extraction of useful unknown information from the large databases with knowledge acquisition purposes. DM involves using different techniques such as Classification, association and clustering whereby normal queries and report from relational databases cannot provide the hidden information (Agarwal, 2013). DM includes seven phases which are data cleaning, data integration, data warehousing, data selection, data transformation, DM and last pattern evaluation phase. The first four phases are for data preprocessing and the last three are for retrieving the hidden information on the processed data (Agarwal, 2013). Therefore, this study was about the application of DM to solve the problem of students' performance in mathematics course a case of MU. Useful knowledge was extracted from the students' data available at the University by applying DM techniques.

2.2.2. Educational Data Mining

In DM, useful knowledge can be extracted from various fields such as medicine, marketing, real estate, engineering, education and customer relationship management. EDM comes when applying DM in data related to the education field to discover classification, clustering and association rules that can be applied in students' performance prediction (Bunkar, 2012). EDM can predict students' performance at Primary schools using various DM techniques such as DT, ANN, K-NN, RF, Linear regression and SVM (Innocentia et al, 2018). Therefore, this study based on the education sector at HLIs by applying selected DM techniques on predicting the students' performance in a mathematics course for management students a case of MU.

2.2.3. Machine Learning

ML is the science of enabling computers to learn and act like a human being after feeding data and information. Three types of ML emerge which are Supervised Machine Learning (SML), Unsupervised Machine Learning (UML) and Reinforcement Machine Learning (RML) (Mellor, 2013). DM employs the ML algorithms to identify the patterns from a large amount of data and draw knowledge which is useful information from the pattern (Nisbet, Miner, & Yale, 2018).

SML is the type of ML whereby both input and output data are known after being labelled for classification to enhance learning for future data processing (Mellor, 2013). SML involves target (dependent variable) which is the outcome to be predicted from the known set of predictors known as independent variables. There are various algorithms under SML which are Linear Regression, Logistic Regression (LR), SVM, NB, K-NN, DT and RF. These algorithms can be applied in bioinformatics, quantitative structure, database marketing, handwriting recognition, information retrieval, learning to rank, information extraction, object recognition in computer vision, optical character recognition, spam detection and pattern recognition (Singh & Sharma, 2016).

In UML, the information is neither classified nor labelled to train the algorithm that allows such algorithm to act on that information without guidance. There are two groups of problem-solving under UML which are clustering and association problems. Clustering problems are concerned with discovering inherent groupings in the data such as grouping customers according to purchasing behaviors. While association problems are for discovering the rules that describe the large part of the data such as customers who tend to buy product X may also buy product Y. Algorithms employed in Unsupervised ML are K-means algorithm, Apriori algorithm, Expectation-maximization algorithm and Principal component analysis algorithm. UML may be applied in Human Behavior analysis, social network analysis to define groups of friends, market segmentation of companies by location, organizing computing clusters based on similar event patterns and processes.

RML is the type of ML algorithm that allows the machine to automatically determine the ideal behavior of a specific context to maximize the performance. Algorithms that can be applied in RML are Q-learning algorithm, State-Action-Reward-state-action algorithm (SARSA), Deep-Q Network algorithm (DQN) and Deep Deterministic Policy Gradient algorithm (DDPG).

Therefore, the type of ML applied in this study was SML due to the nature of problem that was to be addressed in performance prediction of the students in mathematics course. The algorithms applied in SML classified the target variable (remarks) as either pass or fail class from the known values of the predictor variables.

2.2.4. Classification Algorithms

Classification is the process employed in SML to classify data into predefined categorical class labels in two steps which are training and testing (Mythili & Mohamed, 2014). The first step is to analyze the data tuples with attributes using training data set to understand each class label of the training data then classification algorithm is applied to train ML algorithms. Then, second step is to use the test data set to examine the accuracy of the trained ML algorithms whereby if the accuracy is appropriate, ML algorithm is used to classify the unknown data tuples. Classification techniques that may be applied in EDM are DT, Multilayer Perceptron (MLP), K-NN, SVM and RF (Anuradha & Velmurugan, 2015).

DT is one of the popular and powerful classifier in classification that are employed in EDM to predict the students' performance whether will pass or fail (Yadav & Pal, 2012). Normally, DT classifiers have a tree like structure that begin from root attributes and end at the leaf nodes while having several branches that contain different attributes to represent outcome of the test whereby the leaf node on each branch represent a class (Bhavsar & Ganatra, 2012). DT algorithms describes the relationship between attributes and their relative importance. DT classifiers have the advantages of representing the rules that can be easily understood and interpreted by the users, no need of complex data preparation and perform well for numerical and categorical variables (Anuradha & Velmurugan, 2015). Therefore, in this study DT algorithm was selected to be one of the trained ML algorithm on the training dataset. Then trained DT algorithm was evaluated on how it classified well the target variable as either pass or fail in prediction of students' performance in mathematics.

Similarly, MLP is a class of feedforward artificial neural network (ANN) that consists of at least three layers of nodes which are an input layer, a hidden layer and an output layer. Each node is a neuron that uses a nonlinear activation function except for the input node. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron that distinguishes data which is not linearly separable (Wikipedia, 2020). Hence, this study trained MLP algorithm and then tested it for predicting the students' performance in mathematics as either the student will pass or fail.

Also, the K-NN algorithm also is one of the best classification algorithms in supervised learning that classify the objects according to the closest training samples in the feature space (Anuradha & Velmurugan, 2015). K-NN is referred to as instance-based learning or lazy learning since the function is only approximated locally and all computation is delayed until classification. The K-NN algorithm decides among points from the training set that are similar enough to be considered when choosing the class to predict for a new observation. It is done by picking the k closest data points to the new observation, and to take the most common class among these that is why it is known as k Nearest Neighbors algorithm (Bhavsar & Ganatra, 2012). Therefore, the K-NN ML algorithm also was trained in this study for predicting the students' performance in mathematics. Later, the trained algorithm was tested in evaluation to see how it classified well the pass and fail class on the testing dataset.

Likewise, the RF algorithm is one of the classifiers that belong in an ensemble learning method that is used for classification, regression and other tasks that can be performed with the help of the DT. DT can be assembled at the training time and the output of the class can be either classification or regression. With the help of these RF, one can correct the habit of overfitting to the training set (Mythili & Mohamed, 2014). Therefore, this study also trained RF algorithm for students' performance prediction in mathematics. Later, it was tested if it could classify the target variable remarks into the pass and fail classes.

Furthermore, SVM is one of the SML algorithms that provide an analysis of data for classification, regression analysis and outliers' detection. While they can be used for regression, SVM is mostly used for classification that carries out plotting in the n-dimensional space. The value of each feature is also the value of the specified coordinate then ideal hyperplane is found that differentiates between the two classes (Muhammad & Yan, 2015). There are several benefits for using SVM such as its effectiveness in high dimensional space, it uses a subset of training points in the decision function (called support vectors), memory-efficient when applied, and it is versatile because holds different kernel functions can be specified for the decision function Kumar, 2012). In this study, SVM as one of the ML algorithm was trained on the training dataset for students' performance prediction in mathematics. In a classification problem, SVM in classification problems is termed as Support Vector Classification (SVC), and in this study, SVC was one of the applied algorithms to classify the target variable remarks into the pass and fail class during the training of the algorithms.

2.2.5. Evaluation Metrics for Classification Problems

One of the reviewed studies was about predicting students' success in introductory mathematics course as either pass or fail by applying ML algorithms whereby accuracy, precision, recall and F-measure were used as the evaluation metrics (Vihavainen, Luukkainen, & Kurhila, 2013). Confusion matrices normally are used in statistics, DM, ML models and other artificial intelligence (AI) applications that can also be called an error matrix (Hossin & Sulaiman, 2015). There are some terms used in the confusion matrix as shown in Figure 2.1 as follows;

- Accuracy refers to the proportion of the total number of predictions that were correct.
- Positive Predictive Value or Precision refers to the proportion of positive cases that were correctly identified.
- Negative Predictive Value refers to the proportion of negative cases that were correctly identified.
- Sensitivity or Recall refers to the proportion of actual positive cases that are correctly identified.
- Specificity refers to the proportion of actual negative cases that are correctly identified.

Predicted Class

		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN)	Sensitivity (TP/(TP+FN))
	Negative	False Positive (FP)	True Negative (TN)	Specificity (TN/(TN+FP))
		Precision (TP/(TP+FP))	Negative Predictive Value (TN/(TN+FN))	Accuracy ((TP+TN)/(TP+TN+FP+FN))

Figure 2.1: Confusion Matrix for Evaluation of Trained ML algorithms

Source: (Researcher, 2020)

2.2.6. Student Academic Performance Prediction

In EDM, student academic performance prediction is one of the burning research areas recently and around 60% of the scholars are doing research using DM techniques on the dataset available (Kumar & Salal, 2019). The dataset contains the attributes (for example age, gender, marital status, grade, GPA, coursework) and their historical data (instances). As a result, students' performance may be estimated on final grade (A, B, C, D, E or F) or give remarks as either pass or fail on a course in a semester or Grade Point Average (GPA) at the end of the academic year. Also, in performance prediction may be determined whether a student will graduate or not. Moreover, academic performance prediction may be applied with the purpose of finding the key factors for the student academic performance either at primary, secondary, college or HLIs (Ashenafi, 2017).

Therefore, this study based on the classification of the target variable (remarks) in the student's academic performance prediction as either fail or pass class from the predictor variables. This was accomplished by the training of the selected ML algorithms using the training dataset and then predicting using the testing dataset.

2.2.7. Higher Learning Institutions

According to TCU, HLIs offers a level of education and training into various disciplines (Branch of Knowledge, skills and competences acquired through teaching or learning at University or professional level through specific programmes). The disciplines such as Education, Law, ICT, Management and Engineering may lead to intermediate and full academic or professional qualification and competence. The qualifications awarded may be certificate, diploma and degree from HLIs, for example, MU, The University of Dodoma (UDOM) and (UDSM) (TCU, 2019). Therefore, this study selected one of the reputable HLIs in Tanzania that offers Management programmes which was MU as the case study.

2.2.8. Supplementary Examination

Supplementary examination refers to the examination which a student is allowed to sit after failing in the first sitting at the end of the semester (Mzumbe University, 2018). Therefore, in this study, fail class from the target variable (remarks) is the supplementary status of the student in performance prediction that implied the student would have to sit for supplementary examination when next offered.

2.3. Related Works

Different studies have been reviewed in this study to obtain the way to the performance prediction in mathematics using EDM techniques. The research papers reviewed were those that helped to meet the objectives of this study as described in the following paragraphs.

One of the studies conducted by Sivasakthi (2018) using DM Classification Algorithms to predict Students' Performance in Introductory programming course which was within the Computer Science discipline. First-year students have to study that course but the performance was not good hence the research was done using DM techniques to predict earlier the students who are likely to pass or fail (Sivasakthi, 2018). Classification algorithms applied were MLP, DT (J48), DT (REPTree), Sequential Minimal Optimization (SMO) and NB and their predictive accuracies were 93.23%, 92.3%, 91.03%, 90.03% and 84.46% respectively. Predictive variables that were considered in that study were gender, higher secondary education, the medium of instruction, private coaching, area of the school, grade in Course at college and target variable was grade in Course at test (Sivasakthi, 2018). Therefore, through this review, it was necessary to apply classification algorithms to predict students' performance in mathematics course under management discipline taken during the first year in the first semester.

Besides, Zhou (2018) conducted a study on predicting pass rates to students in Mathematics and Portuguese by applying SVM and DT. In that study, the researcher did the experiments to find out the accuracies of the trained ML algorithms when all features are involved. The results were 91.5% for DT, 92.6% for SVM, RF was 72.4% and ANN was 88.3% in mathematics prediction (Zhou, 2018). Therefore, other predictor variables depend on each other that need to be considered in training the ML algorithms to have higher accuracies. Nevertheless, when other predictor variables were not involved in mathematics performance prediction, the results were 83.9% for DT, 87.3 for SVM, 52.7% for RF and 81.3% for ANN (Zhou, 2018). Hence, in this study, all the predictor variables for the mathematics performance prediction were involved apart from finding the significance of the predictor variables to the outcome variable and found few of them had little influence as described in chapter four.

Additionally, there was another related study done by Ünal (2020) who compared the five-level grading that had an accuracy of 71.14% for RF algorithm and binary level grading in mathematics that had an accuracy of 91.39% (Ünal, 2020). Also, from that study, DT had an accuracy of 73.42% when five-level grading was applied and then an accuracy increased to 89.11% when binary level grading was used. Therefore, it was concluded that the accuracies of the trained algorithms increases as the level of the classes to the target variable decreases and that was why this study opted to base on the binary classification by applying the ML algorithms in mathematics prediction.

Furthermore, according to Ashenafi (2017) on comparative analysis of the selected studies, the studies for the binary outcome prediction such as the pass or fail were compared. Linear Regression and Support Vector Machine algorithms may be applied to predict numerical scores outcome such as predicting the student' performance in final examination (FE) (Widyahastuti & Tjhin, 2017). However, the researcher argued that actual prediction of scores have historically been less accurate than those that predict whether the student has passed or failed the particular course (Ashenafi, 2017).

That was why in this study the performance prediction in mathematics to students pursuing management degree was considered and classification of the target variable was involved as either pass or fail.

Moreover, it was necessary to review other studies that have involved students' performance prediction using different EDM techniques. One of the researchers Yadav & Pal (2012) conducted a study to predict the performance of engineering students and trained DT algorithm had an accuracy of 67.8% (Yadav & Pal, 2012). From the confusion matrix, it was clear that the true positive rate of the model for the FAIL class was 0.786 accurate for Iterative Dichotomiser 3 (ID3) and C4.5 DT that mean generated model was successfully identifying the students who are likely to fail. The researcher recommended that ML algorithms such as the C4.5 DT algorithm can learn effective predictive models from the students' data accumulated from the previous year (Yadav & Pal, 2012). Factors considered by that researcher were student's programme, admission type, gender, grades in high school and senior secondary, living location, the medium of teaching, family size, family annual income and family status, father's and mother's occupation and qualification. Therefore, this study classified the depended variable remarks as either pass or fail considering the predictor variables from five previous academic years (2014/2015 to 2018/2019) so as to have more data for training and testing ML algorithms. And one of the ML algorithms trained in this study was DT since in this review it has the highest accuracy to be considered for training in this study.

Likewise, another related study performed by Pandey & Taruna (2014) involved RF ML algorithm for students' performance prediction and correctly classified 701 instances and 297 instances were incorrectly classified and the accuracy was 72.04% (Pandey & Taruna, 2014). Therefore, in this study RF algorithm was one of the trained ML algorithms to predict the students' performance in mathematics.

Also, Nikam (2017) compared several selected studies in DM using classification algorithms such as DT, ANN, SVM, RF, K-NN and NB. The researcher found that mentioned algorithms are capable of processing a large amount of dataset and predict categorical class labels using the training set to classify data, as a result, classify newly available data (Nikam, 2017). Therefore, in this study that intended to predict

Mathematics performance, classification algorithms were applied. Trained ML algorithms were able to classify the remarks into the pass and fail class labels using the training data set then using testing data set to predict the students' performance.

Moreover, Estate Management programmes also face poor performance by students at the Federal University of Technology Akure in Nigeria (Olatunji et al., 2017). Hence, there was a need to perform a study in Management discipline to enhance better performance to students at University to have good managers in industries.

According to Ashenafi (2017) from the research on comparative analysis of selected studies in student performance prediction, the researcher concluded that most of the performance prediction researches have focused on the discipline of computer science and engineering. Hence, there is a need for researchers to look into other disciplines (Ashenafi, 2017). Therefore, this study intended to focus into other disciplines such as in Management to students pursuing Human Resource, Local Government, Public Administration and Records and Archives Management for mass failure predictions in a particular course.

2.3.1. Predictor Variables in Mathematics Performance Prediction

According to the reviewed kinds of literature, there were various predictor variables for the student to pass or fail in mathematics at the University such as mathematics background at ordinary level (Kumah et al., 2016) and continuous assessment or coursework (Okwute & Musa, 2018). Other factors that contribute to the students' academic performance prediction regardless of the specific course that may be employed in the mathematics performance were determined from the reviewed studies. The predictor variables determined were personal attributes such as gender, age, interest with the study, admission type, and study behavior (Abu et al., 2019).

Moreover, family attributes such as parents' qualification, parents' occupation, family income, family status, and family support for study (sponsorship) were also found to influence student academic performance prediction (Kumar, Singh, & Handa, 2017). Furthermore, academic attributes such as high school grade, student's previous semester marks, class test grade, seminar performance, assignment performance, attendance in class and lab work and previous schools marks also found to be the contributing factors for the student academic performance prediction (Kumar, Singh, & Handa, 2017). Therefore, some of the predictor variables identified from the similar studies were used due to the data available at the field and some of the predictor variables were added from the field to form dataset in order to train ML algorithms for students' performance prediction in mathematics course as elaborated in chapter four of this study.

2.4. Research Gap

There are various researches in EDM that have based on the students' academic performance improvements in science and engineering disciplines in University education on the courses such as introductory programming and Mathematics which have a mass failure during the first year. However, other disciplines such as Management also face the mass failure in Mathematics course at University education in Tanzania such as MU and no EDM have been applied to analyze such problems. Therefore, this study applied EDM Techniques to predict the students' performance in Mathematics course pursuing Management programmes at MU in Tanzania under the assumptions that it is traditional classroom-based learning and neither web-based courses nor learning management systems.

2.5. Conceptual Framework

A conceptual framework was designed to guide the study in mathematics performance prediction. Figure 2.2 depicts the research phenomenon from the problem awareness on the factors contributing to mathematics Performance to the Performance Prediction in Mathematics as either Pass or Fail. Being aware of the problem to be solved concerning the Mathematics performance prediction for Management degree students at MU was the first step. Then, identifying predictor and target variables that led to data collection from the admission office such as educational background data and examination office such as course work (CW) and FE in Mathematics. Thereafter, data were selected and integrated to get the target dataset. Data cleaning and preprocessing techniques were applied to remove outliers, noise and missing values to have necessary values in dataset. Moreover, data were transformed into the form that can be suitable for applying algorithms to train ML algorithms for Mathematics performance prediction. After trained ML algorithms, the results were evaluated based on the evaluation metrics. At last, the knowledge was extracted from the dataset used to train and test the ML algorithms for Mathematics performance prediction.

The trained ML algorithms identified the two classes for the students' performance in Mathematics course as either Pass or Fail. If then rules were generated according to the results obtained after prediction to relate the predictor variables and the target variables to conclude the Mathematics performance.

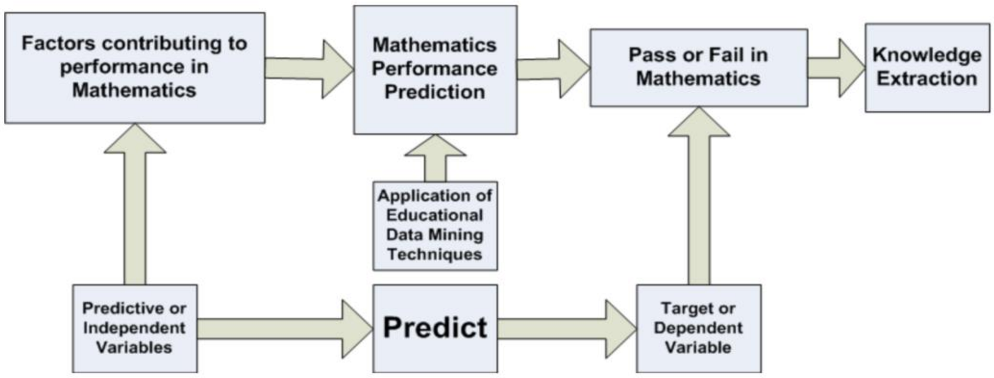

Figure 2.2: Conceptual Framework for the Research

Source: Researcher (2020)

2.6. Chapter Summary

This chapter two of literature reviews showed the key concepts of ML, DM and the related works that have been written by various scholars in students' performance prediction at primary, secondary, college and HLIs. From the three types of ML, this study employed SML since the problem solved based on classification where the target variable was labeled as either pass or fail

CHAPTER THREE
RESEARCH METHODOLOGY

3.1. Introduction

This chapter describes on how the research was conducted in achieving the objectives of this research and was organized as follows; research setting, research approach and design, data collection methods and tools, data set, EDM techniques, data analysis and trained ML algorithms evaluation, ethical considerations and reliability and validity of the research to be undertaken in EDM.

3.2. Research Setting

The research was conducted at MU as it is among the reputable University for Management programmes in Tanzania with many Management degree programmes compared to other universities (TCU, 2020). Under SOPAM, management programmes offered at MU are BLGM, BHSM, BPA-RAM and BHRM (Mzumbe University, 2016).

MU was established in 1953 when the British Colonial Administration established a Local Government School in the country. The school was aimed at training local Chiefs, Native Authority Staff and Councilors. The level of training was elevated after Tanzania (Tanganyika) independence to include training of Central Government Officials, Rural Development Officers and local Court Magistrates. In 1972, the Local Government School was merged with the Institute of Public Administration of the UDSM to form the Institute of Development Management (IDM-Mzumbe) thereafter full-fledged University in December 2006 (Mzumbe University, 2016).

Likewise, Management programmes provided by MU in bachelor level are BLGM, Economics in Project Planning Management (EPPM). Other programmes are BHSM, and Human Resources Management (at Main Campus and Mbeya Campus). Also, other programmes are Procurement and Supply Chain Management (BPSCM), Education in Language and Management (BELM). Moreover, Information and Communication Technology with Management (BSc. ICT-M), Production and Operation Management (BSc. POM). Furthermore, programmes in management are Industrial and Engineering Management (BSc. IEM), Library and Information Management (BSc. LIM). Similarly, other programmes are Business Administration in Entrepreneurship Innovation and Management (BBA-EIM) and BPA-RAM (Mzumbe University, 2019).

3.3. Research Approach and Design

This study employed the philosophical Worldview of Post positivism that deals with the cause and effect relationship type of research which is the quantitative research approach. (Creswell, 2014; Kumar, 2011). The quantitative research approach basing on the design science research was used with process steps that include awareness of the problem, suggestions by setting objectives, development, evaluation and then the conclusion (Peffers et al., 2007). Figure 3.1 shows the research method that was applied in this study from the data collection up to the knowledge obtained after applying ML algorithms on training dataset for the student's performance prediction in the Mathematics course.

3.4. Data Collection Methods and Tools

Document review was applied to gather data from MU in the examination office, accounts office and students' admission office from the academic year 2014/2015 to 2018/2019 for first-semester Management students in the Mathematics course. Documents reviewed were students' hostel allocation excel files that enabled to extract students who are at campus and who are outside campus. Likewise, lecturers' workload allocation excel files were reviewed to obtain number of instructors involved in teaching mathematics course. Moreover, students' admission excel files were collected to extract data such as age, gender, entry category, campus location, and ordinary level mathematics grades. Furthermore, students' loan allocation excel files were collected to know whether student is loan beneficiary or depends on parents/guardians. Also, course CW and FE results extracted from academic records information systems.

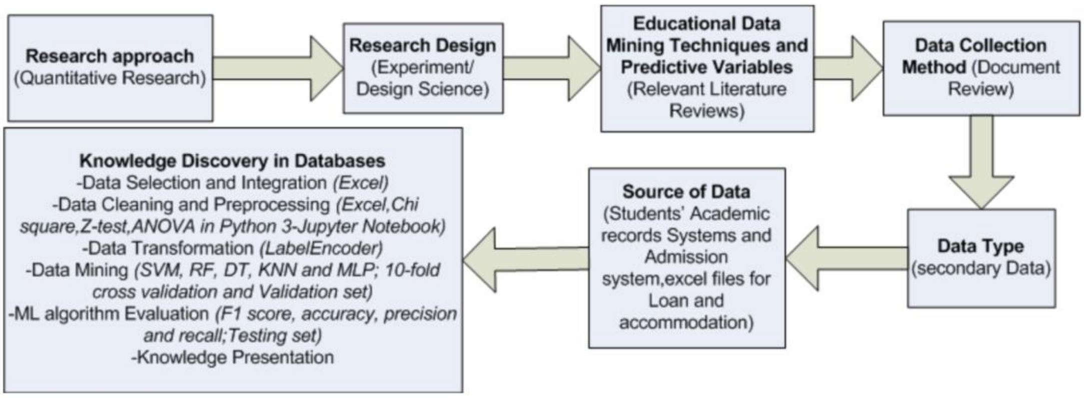

Figure 3.1: Research Methodology

Source: Researcher (2020)

3.5. Data Collection at the field for Dataset Preparation

The researcher was able to collect the students' secondary data from both campuses which were the main campus and Mbeya campus. Data were collected to form a total of 3347 instances before cleaning and preprocessing. Students' data were gathered from various offices at MU from 2014/2015 to 2018/2019 academic years such as the admission office, accommodation office, examination office, Department of Mathematics and Statistics and accounts office. The students had registration numbers that were unique for identification when data were collected for each student from different sections and departments. Data collected from the Department of Mathematics and Statistics was the number of instructors involved in teaching the mathematics course in different academic years. Also, Students' loan allocation data were obtained from the accounts office dealing with the students' affairs to know who has a loan from the government and those who get support from their parents and guardians. Likewise, students' living location data were obtained from the accommodation office as to whether the student was living on the campus or off-campus. Students' results which were CW, remarks and FE for mathematics course were collected from the examination office. Additionally, from the admission office, data collected were age, ordinary level mathematics grades, gender, and entry category.

3.6. Data Analysis, ML algorithms Training and Evaluation

After collection of the data from the field, it was necessary to analyze using Microsoft Excel (2016 version) and Jupyter Notebook (version 5.7.4) application programs. In excel, functions such as normsdist function was applied for finding p value after having z value in Z test. Likewise, find and replace functions in excel were applied in dataset creation as they were easy to access and make data easily visible during manipulation. Python version 3 was used as it was a powerful programming language that can manipulate data and perform feature engineering. Jupyter Notebook was used for code editing and displaying the analyzed data as it was easy to inspect data and features through graphs and tables (Slater et al., 2017). In data analysis, the mentioned tools were able to process 3259 instances of the dataset.

3.6.1. Data Exploration and Preparation

Before the application of ML algorithms on training dataset for performance prediction in Mathematics, it was needed to understand, clean and prepare the data as it was the most important part of this study. Table 3.1 indicates the students' performance pursuing four Management programmes in Mathematics course at MU from

2014/2015 up to 2018/2019 academic year during the first year in the first semester. In the 2014/2015 academic year, 115 out of 535 students had supplementary. In 2015/2016 academic year, 174 out of 602 students had supplementary. In academic year 2016/2017, 314 out of 695 had supplementary. In academic year 2017/2018, out of 773 students, 172 students had supplementary and 150 out of 742 students had supplementary in 2018/2019. Figure 3.2 and Figure 3.3 also show that there was a problem of mass failure in Mathematics to Management programmes and hence research was supposed to be carried out.

Table 3.1: Mathematics Performance in Management Degree Programmes at MU

Source: Researcher (2020)

	2014/2015		2015/2016		2016/2017		2017/2018		2018/2019				
Programmes	FAIL	PASS	FAIL	PASS	FAIL	PASS	FAIL	PASS	FAIL	PASS	TOTAL (FAIL IN 5 YEARS)	TOTAL STUDENTS IN 5 YEARS	FAIL (%)
BPA-RAM	27	53	30	49	39	42	1	59	46	69	143	415	34
BHSM	35	143	33	107	42	97	96	202	47	101	253	903	28
BHRM-MU	24	100	56	109	62	62	8	130	23	106	173	680	25
BHRM-MCC	11	90	42	127	123	141	61	156	7	272	244	1030	24
BLGM	18	34	13	36	48	39	6	54	27	44	112	319	35
Total	115	420	174	428	314	381	172	601	150	592	925	3347	28

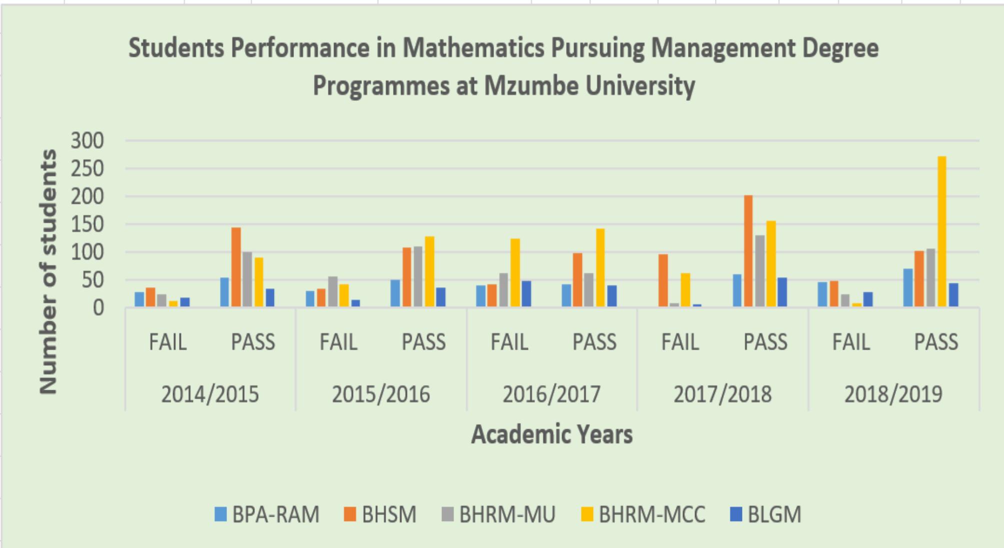

Figure 3.2: Students Performance in Mathematics Pursuing Management Degrees

Source: Researcher (2020)

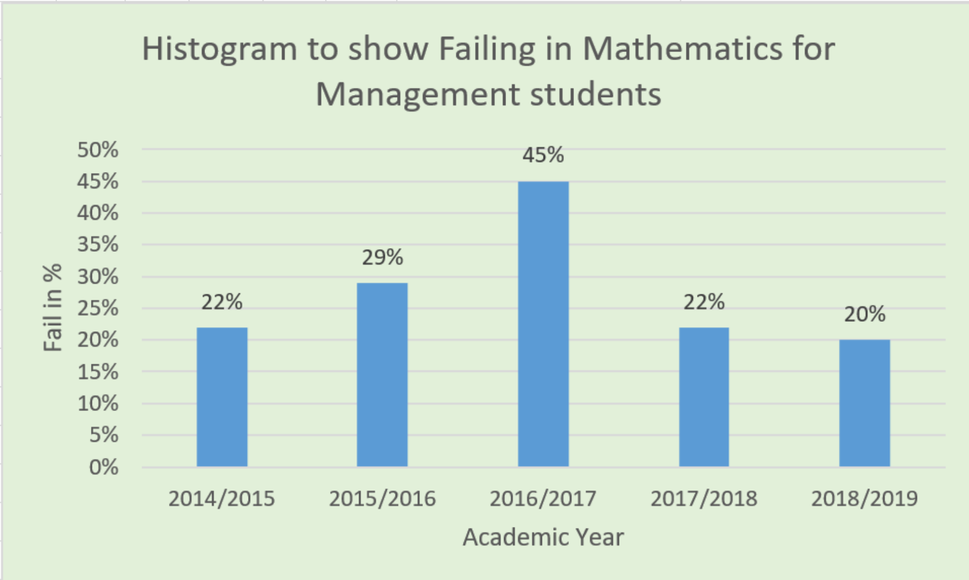

Figure 3.3: Management Students Failing in Mathematics from 2014 to 2018

Source: Researcher (2020)

In Variable Identification, it was necessary to identify the predictor (input) variables and the target variables (output) for performance prediction in Mathematics. Also, data types were identified on collected dataset such as numeric and character, and categories of variables as either categorical or continuous were identified.

In Univariate Analysis, the variables had to be explored individually using the methods and statistical measures that depended on the category of the variable as either categorical or continuous variable. In the case of continuous variables, it was needed to understand the central tendency (mean, median, mode, max, and min) and the spread of variables (Range, Quartile, Variance, and Standard Deviation) using statistical metrics visualization methods such as Histogram and Box plot through which the researcher was able to identify missing values and outliers. For categorical variables, a frequency table was used to understand the distribution of each category with the count and count percentage metrics for each category, and Bar chart was used to visualize the categorical variables.

In Bi-variate Analysis, it was essential to understand the relationship (association and disassociation) between two variables using Bi-variate analysis methods that depends on either the variables were both continuous (Scatter plot and correlation was applied), both categorical (Chi-square test was applied) or categorical and continuous variables where Z-test, T-test or ANOVA was applied to assess the significance between variables.

Moreover, the scatter plot and correlation were applied to find the relationship between numerical variables such as CW and FE. Similarly, the Z test was applied to find the association between categorical variable with two classes and the numerical variable such as CW and remarks (Pass or fail). Likewise, ANOVA was applied to find the relationship between the categorical variable with more than two classes such as mathematics ordinary level results in grades (A, B+, B, C, D, E and F) and numerical variable CW or FE in this study. The calculations of the chi-squares, Z test and ANOVA are in appendices C, D, E, F, G, H, I, J, K, L, M, N, P, and Q where by the results were presented and discussed in chapter four.

In missing values treatment, a better predictive model to be developed should have no missing values in the dataset. In this study, the researcher applied methods such as removing the entire row with the missing values. The method was applied because some of the students had neither CW nor FE, as well as remarks had no values. It was necessary to handle them to avoid wrong prediction or classification and biased model. In this study, 88 instances were removed out of 3347 instances as some of the students had postponed studies or examinations as a result 3259 instances remained to form a dataset.

In outlier treatment, any observation in the dataset that appeared far away and diverges from overall pattern was considered an outlier and needed to be treated using methods such as binning and mean imputation. An outlier may cause wrong representation on the dataset as it may affect the mean of the variable in the dataset. In this study, there were outliers in the age predictor variable that only four students had aged from 40 to 49 hence binning method was applied to group them in the range from 40 to 49. In variable transformation, it was one of the step in feature engineering that was the science of extracting more information from the existing data. In variable transformation, the binning method was applied to categorize the variables for instance categorical variable to numerical. The variable transformation was done to all categorical attributes using label encoder in SciKit-learn library in Python as some of them seen in appendices R, S and T. In variable creation, as one of the steps in feature engineering, it was done to obtain more variables from the existing variables by applying methods such as creating derived variables using functions and dummy variables when converting categorical to numerical variables.

In this study, the dataset was explored especially the distributions of the data on variables using bar charts for categorical variables and mean, variance, standard deviations, quartile and medians for continuous variables together with Histograms. Frequency distribution of the remarks variable which is categorical has 2403 students (73.7%) for pass class or Fail class that has 856 (26.3%) from 2014/2015 to 2018/2019 academic years during the first year. Figure 3.4 is the Bar chart which shows the performance in mathematics from 2014/2015 to 2018/2019 academic years with remarks pass or fail of degree students management students at MU in the first semester of the first year

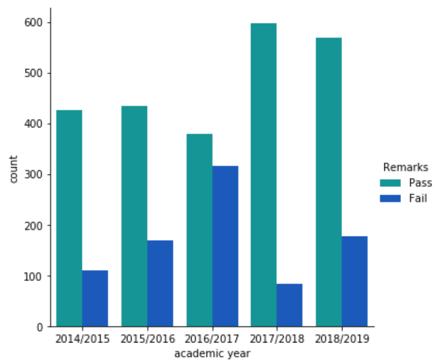

Figure 3.4: Remarks (Pass or Fail) from 2014/2015 to 20218/2019 Academic Years

Source: Researcher (2020)

Frequency distribution for gender attribute that was male and female in the dataset created, male were 1450 students (44.5%) and female were 1809 students (55.5%) from academic year 2014/2015 to 2018/2019 during the first year in the first semester as it was more depicted in Figure 3.5.

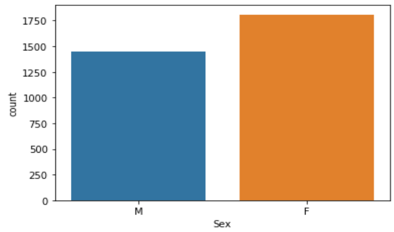

Figure 3.5: Gender Variable in Dataset

Source: Researcher (2020)

Moreover, it was explored on the dataset to understand the distribution of the sponsorship variable to students pursuing management degree programmes at MU. It was found that the number of students who were loan beneficiaries from the government by 44.5% in five years from 2014/2015 to 2018/2019 academic years which is 1450 students while students with private sponsorship were 1809 students which was 55.5% during the first year in the first semester. Figure 3.6 shows the distribution of the sponsorship as private and by HESLB to male and female first-year students of management degree programmes from MU during the first semester.

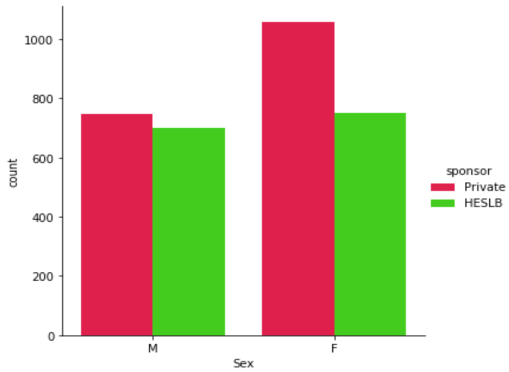

Figure 3.6: Sponsorship Distribution on Gender in Dataset

Source: Researcher (2020)

Furthermore, the living location of the students was explored using frequency distribution where 679 students (20.8%) were accommodated in University Hostels and 2580 students (79.2%) were accommodated in private hostels out of University as shown more in Figure 3.7.

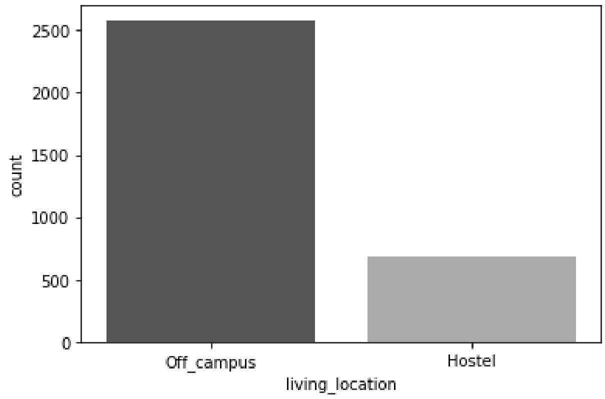

Figure 3.7: Living Location of the Students in Dataset

Source: Researcher (2020)

Likewise, the frequency distribution of the entry category was explored as either RPL, diploma or form six as the entry qualification to join the first year in degree management programmes at MU from 2014/2015 to 2018/2019 academic years. It was found that 719 students (22.1%) were from diploma and 2534 students (77.7%) were from form six and 6 students (0.2%) were from Recognition of Prior Learning (RPL) entry equivalent as Figure 3.8 depicts more on entry category distribution based on the gender (female and male).

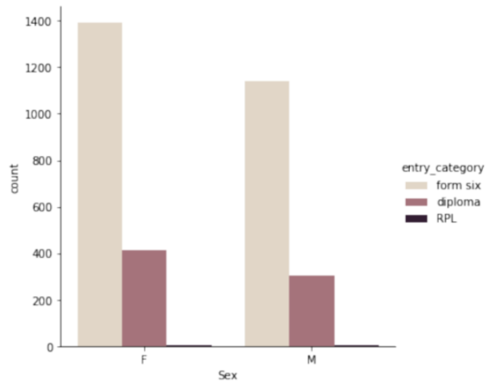

Figure 3.8: Entry Category Distribution in Dataset

Source: Researcher (2020)

Also, it was important to analyze the numerical variables which were FE, CW and age of the students by using univariate analysis metrics as shown in appendix B, C and D. Mean, standard deviation, minimum, first quartile, second quartile, third quartile and maximum values were applied as shown in Table 3.2 with their Histograms as shown in Figure 3.9.

Table 3.2: Univariate Analysis on Numerical Variables

Variable	Mean	STD	Min	25%	50%	75%	Max	Count
CW	27.21	6.07	0.5	23	27	31.65	47	3259
FE	23.77	10.02	0	18	23	30	50	3259
Age	21.73	3.15	17	20	21	23	49	3259

Source: Researcher (2020)

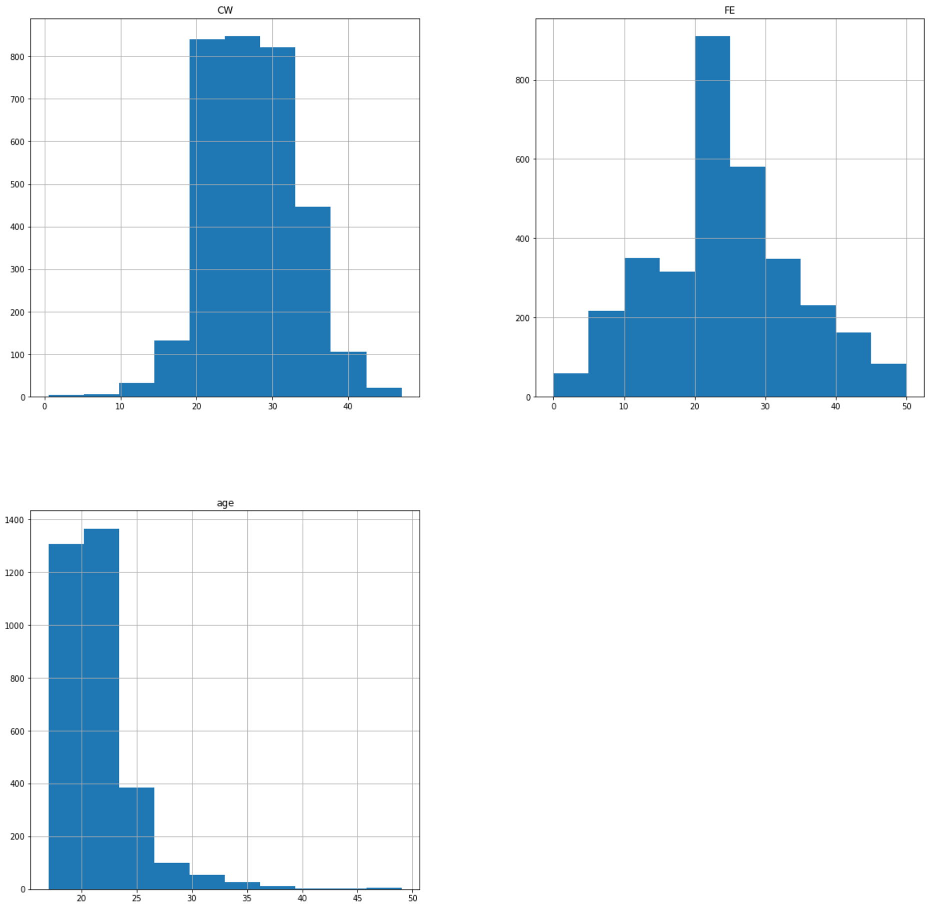

Figure 3.9: CW, FE and Age in Dataset

Source: Researcher (2020)

Mathematics ordinary level grades were explored as shown in Figure 3.10 whereby there were 1480 students (45.4%) with F grade, 87 students (2.7%) with E grade, 1309 students (40.2%) with D grade, 326 students (10%) with C grade, 38 students (1.2%)

with B grade, 9 students (0.2%) and 10 students (0.3%) with A grade from academic year 2014/2015 to 2018/2019 academic year.

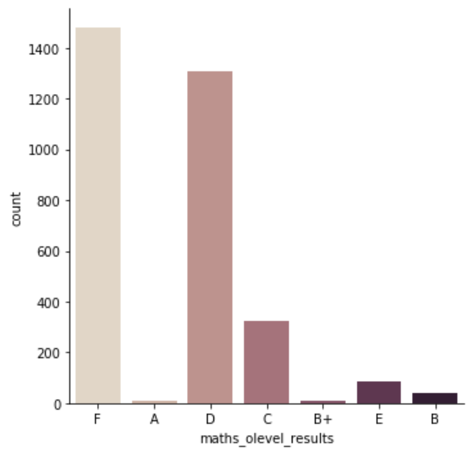

Figure 3.10: Mathematics Ordinary Level Grades in Dataset

Source: Researcher (2020)

3.6.2. The Training of ML Algorithms

The dataset was divided into 60% of the dataset in training of the ML algorithms, 20% of the dataset in the validation of the trained ML algorithms, and 20% of the dataset in testing the trained ML algorithms. Various EDM techniques have been used for the prediction of the students' performance such as SVM, K-NN, ANN and DT (Roy & Garg, 2017; Shahiri, Husain, & Rashid, 2015). Therefore, this study applied some of the ML algorithms using the data available at the University such as K-NN, SVC, RF, DT and MLP in mathematics performance prediction as they are widely used in EDM with good performance predictions.

The division of the dataset was considered after the review of the study for reducing dropout rates using ML approaches, the researcher had a training set (60%), validation set (20%) and testing set (20%) (Mduma, Kalegele, & Machuve, 2019). Since to train ML algorithms better, it was advised to train the ML algorithms with a dataset in the range of 60% to 80% and test the ML algorithms with 40% to 20% (Omid, 2013). The purpose was to avoid lowering the samples in the trained ML algorithms as it would result in trained ML algorithms with too few samples. Trained ML algorithms were evaluated based on the available metrics such as accuracy and correctly and incorrectly predicted instances by Confusion Matrix (Amjad Abu, 2016). Upon the training of the

ML algorithms, a validation technique known as K-fold cross-validation (10 fold cross-validation in this study) was applied as it was normally preferred method. It allowed the ML algorithms to be trained on multiple train-test splits and with a small dataset compared to Hold out validation technique hence it gave an indication that how well the trained ML algorithms performed on unseen data (Allibhai, 2018).

3.6.3. Evaluation of Trained ML Algorithms

After the application of ML algorithms on training dataset for mathematics performance prediction, it was better to evaluate each ML algorithm's performance. Therefore, classification evaluation metrics were used to check the performance of the trained ML algorithms to discriminate among the results (Hossin & Sulaiman, 2015). Since this study is solving the classification problem, a confusion matrix as a table was applied to outlines different predictions and test results then contrast them with real-world values. Evaluation metrics applied to the trained ML algorithms in this study were recall, precession, accuracy and F-measure.

In this study, evaluation metrics such as Sensitivity or Recall was applied. It was calculated as the ratio of positive class correctly detected that gave how good the trained ML algorithm was to recognize a positive class as shown in Equation 3.1 (Tiwari, 2019).

$$\text{Recall} = \text{TruePositives} / (\text{TruePositives} + \text{FalseNegatives}) \tag{3.1}$$

Moreover, precision as evaluation metric was calculated as the accuracy of positive class that computes how likely positive class prediction was correct as shown in Equation 3.2.

$$\text{Precision} = \text{TruePositives} / (\text{TruePositives} + \text{FalsePositives}) \tag{3.2}$$

Furthermore, Specificity was considered as the ratio of actual negatives from which trained ML algorithm predicted as a negative class or true negative as shown in Equation 3.3.

$$\text{Specificity} = \text{TrueNegatives} / (\text{TrueNegatives} + \text{FalsePositives}) \tag{3.3}$$

In some cases, researchers try to obtain the best precision and recall simultaneously by applying F1 Score which is the harmonic mean for precision and recall values. The formula for F1 score is as shown in Equation 3.4

$$\text{F-Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3.4)$$

The higher the F1 score, the more is the predictive power of the trained classification ML algorithm. Hence when the score was close to 1 it meant that a perfect trained ML algorithm for prediction. However, a score close to 0 meant decrement in the predictive capability of the trained ML algorithm (Tiwari, 2019).

Therefore, the evaluation metrics applied in this study were accuracies, recall, precision and F-measure due to the fact that the study involved classifying as either the student pass or fail in mathematics. One of the studies employed the mentioned evaluation metrics when compared the classification algorithms such as ANN, NB, DT, LR and SVM to find the best algorithm on predicting the students’ performance in board examination (Rustia et al., 2018).

3.7. Ethical Consideration

It was important in this research study to anticipate ethical issues before conducting the research, beginning of the research, collecting data, analyzing data and reporting, sharing and storing data (Creswell, 2014). Therefore, this study observed ethical issues such as not disclosing the students’ results and their information anywhere to anyone for other activities other than intended for this study. Also, to report the results and findings as they were obtained due to the given data without biases or favoring the organization that supplied data for this study.

3.8. Reliability and Validity

The study ensured the reliability and validity of the work by using a reasonable dataset for training the ML algorithms and testing the trained ML algorithms. Validation of the trained ML algorithms was done using the k-fold cross-validation technique (10-fold cross-validation in this study) to get better accuracy of the trained ML algorithms. Moreover, the preprocessing of the data was done such as data exploration, visualization and feature selection using statistical metrics such as Histograms, Bar charts, Chi-square test, Z-test, and ANOVA.

3.9. Chapter Summary

This chapter has explained the research settings as to where the research has been conducted, research approach and design, data collection methods and tools, dataset, EDM techniques to be applied in training ML algorithms, data analysis and evaluation, ethical consideration, reliability and validity of the research.

CHAPTER FOUR

FINDINGS AND DISCUSSION

4.1. Introduction

This chapter presents the findings and discussion of the study based on the three research objectives that led to the research questions to guide the study. The chapter is organized into findings and discussion based on the specific objectives of this study.

4.2. Identifying the Requirements for Training ML algorithms in Mathematics Performance Prediction

In this study, it was necessary to identify the requirements before training the ML algorithms for performance prediction in mathematics. The requirements included identification of the predictor variables, data set creation after data collection from the field and feature importance, extraction and selection.

4.2.1. Identification of Predictor Variables

The requirements for training the ML algorithms were identified that include predictor variables determination from the reviewed related works and through documentary review in the field for dataset creation. Predictor variables such as gender, age, ordinary level mathematics grades, living location, sponsorship, entry category, CW, campus location and FE were identified in related works. While one of the predictor variable was identified at the field named as the number of instructors involved in teaching the mathematics course at the particular academic year that has values as either one instructor or two instructors. From the field, it was found that the number of instructors may differ because the mathematics course has statistics topics that other instructors may be included to support teaching. The predictor variable was extracted from the workload allocation for each academic year from 2014/2015 to 2018/2019 academic year.

4.2.2. Data Collection and Dataset Creation

After having the predictor variables from the literature reviews and the field, data were extracted from various sources at MU to form a dataset. Collected data were from 2014/2015 to 2018/2019 academic years from both Main campus and Mbeya campus. The dataset after removing missing values, irrelevant rows and outliers that were 88 instances, new dataset had 3259 instances with eleven attributes which were date of birth that helped to obtain age, gender, entry category, mathematics grades at ordinary level, sponsorship status, living location, coursework, number of instructors, campus location and FE as well as remarks as either fail or pass as the outcome variable. Table 4.1 depicts the list of attributes that were used to form a dataset of 3259 instances for the students' performance prediction in mathematics as shown in appendix O.

Moreover, the instances collected were enough to form a dataset as it was compared from other studies such as Zohair in 2019 reerarched students' performance prediction with 273 instances (Zohair, 2019). Also, Abu in 2016 had research with 270 instances in performance prediction using classification algorithms (Abu, 2016).

Furthermore, in one of the studies, the researcher predicted the students' performance in mathematics with 279 instances from 2007 to 2010 academic years basing on the attributes such as oral grades, test grades and final grades in semester one and two (Livieris, Drakopoulou, & Pintelas, 2012). Compared to this study, more instances were involved which were 3259 instances and 10 attributes were considered after checking their significances to the output variable. Attributes such as coursework, FE, mathematics grades at the ordinary level, sponsorship, living location, age, gender, number of instructors, campus location and entry category were used as shown in Table 4.1.

4.2.3. Feature Importance, Extraction and Selection

In data collection to form the dataset, not all the features collected will be used to form the dataset for training ML algorithms. Some might be dropped if found not to have importance in improving the performance of the trained ML algorithms and others might be generated from other features.

In this study, the chi-square metric was applied to find the relationship between the two categorical variables between the independent variable and the depended variable such as between entry category and remarks, age group and remarks, mathematics ordinary level grades and remarks, living location and remarks and gender with remarks as one of the feature selection technique (Zaffar, Hashmani, Savita, & Rizvi, 2018).

Table 4.1: List of Attributes that Formed a Dataset

Sno	Variable name	Description	Possible Values
1	Age	Student's age	Quantitative data (0-60)
2	Gender	Sex of the student	Qualitative data (Male or Female)
3	Entry category	As whether student passed in form six, RPL or diploma and used either of them to enter at University.	Qualitative data (Diploma, Form six or RPL)
4	Ordinary level mathematics performance	Grades obtained at ordinary level	Qualitative data (A, B+,B , C, D, E or F)
5	Sponsorship	Who supports the students education at University	Qualitative data (Private or HESLB)
6	Living location	The place where the student live while at the University	Qualitative data (Hostel or off campus)
7	Coursework	Marks obtained that include assignments and tests before final examination	Quantitative data (Values from 0 to 50)
8	Final examination	Marks obtained in University Examination	Quantitative data (Values from 0 to 50)
9	Number of instructors	Number of instructors involved in teaching the course in a particular academic year	Quantitative (one or two)
10	Campus location	Location of the campus the students admitted	Qualitative data (town or rural)
11	Remarks	Status of the course as either the student has passed or failed the course	Qualitative data (Pass or Fail)

Source: Researcher (2020)

From Table 4.2 it can be observed that some predictor variables have a direct relation to the output variable Remarks such as Age group, entry category, mathematics grades at o level, CW, campus location, number of instructors and FE. Moreover, some of the predictor variables have shown not to have an influence on the performance of the students in mathematics such as sex, sponsorship and living location. But, the relationship between each other as the predictor variables were calculated and even combined together to find the relationship to the output variable as shown in Table 4.3. Therefore, there were no features that were dropped for training ML algorithms to predict the mathematics performance to degree management students. The calculations of the chi-squares, Z test and ANOVA are in appendices C, D, E, F, G, H, I, J, K, L, M, N, P, and Q. If there is a relationship between variables, 'ok' was applied and 'no' if there was no relationship between variables in Table 4.3. The Significant level used was 0.05 to compare with p values calculated from chi-square, Z test and ANOVA.

Furthermore, Figure 4.1 depicts the scores in feature importance of the predictor variables to the target variable remarks with two classes as pass or fail which were calculated from SelectKBest class in scikit-learn library. It can be observed that the FE at the University is the feature with the highest score of 139 that contributes highly to the students' mathematics performance prediction. The next feature was mathematics ordinary level grades with 118.5 scores that contributed to the mathematics performance prediction at the University. The other feature was the entry category with 24.4 scores that contributed to the University mathematics performance prediction. Also, CW had scores of 20.6 to the target variable remarks in mathematics performance prediction. Likewise, the age of the student had scores of 3.5 contributions to the target variable in mathematics performance prediction while the campus location and number of instructors both had 1.6 scores. The other three predictor variables were gender, sponsorship and living location had scores of 1.1, 0.4 and 0.3 respectively to the contribution in mathematics performance prediction.

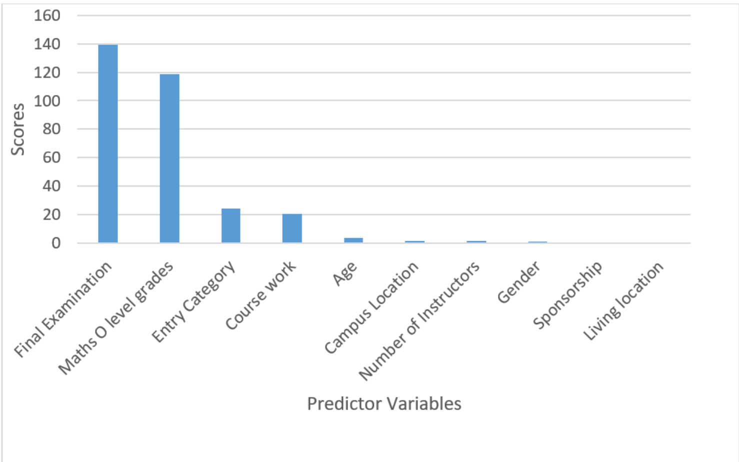

Figure 4.1: Feature Importance to the Target Variable

Source: Researcher (2020)

Table 4.2: Relationship between the Variables in Dataset

Variables	Age group	Sex	Entry category	Maths olevel results	Sponsorship	Living location	CW	FE	Number of instructors	Campus location	Remarks
Age group	ok	no	ok	ok	ok	ok	no	no	ok	ok	ok
Sex	no	ok	no	ok	ok	no	no	no	ok	ok	no
Entry category	ok	no	ok	ok	no	no	no	no	ok	ok	ok
Maths ordinary level results	ok	ok	ok	ok	ok	ok	no	no	ok	ok	ok
Sponsorship	ok	ok	no	ok	ok	ok	no	no	ok	ok	no
Living location	ok	no	no	ok	ok	ok	no	no	no	no	no
CW	no	no	no	no	no	no	ok	ok	no	no	ok
FE	no	no	no	no	no	no	ok	ok	no	no	ok
Number of instructors	ok	ok	ok	ok	ok	no	no	no	ok	ok	ok
Campus location	ok	ok	ok	ok	ok	no	no	no	ok	ok	ok
Remarks	ok	no	ok	ok	no	no	ok	ok	ok	ok	ok

Source: Researcher (2020)

Figure 4.2 describes the DT classifier from the root node up to the branch nodes that classify the students' performance in mathematics as either Pass or Fail. Some of the rules or patterns have been generated from the DT that may be applied to draw conclusion in mathematics performance based on the predictor variables. From rule 1, if the score of the student in FE is above 19.45 and CW is below or equal to 16.5 while during admission mathematics grade in ordinary level was F then the student is likely to fail in mathematics at the University.

Moreover, from rule 2 in Figure 4.2, it can be observed that, if the FE is less than or equal to 19.45 and CW is less than or equal to 24.45 while the mathematics grade in ordinary level during admission was E or F. Then, the student is likely to fail in mathematics at university from which 547 students from branch node majority belongs to the Fail class.

Therefore, from the two samples of observations in patterns from DT classifiers, the admission office at University may consider the minimum grade at the ordinary level in mathematics to be admitted at management degree programmes from D and above so as to reduce the number of failing students in mathematics. This is because from the dataset number of the fail class was 856 students and the majority of students who had failed mathematics in University had also failed at the ordinary level with E or F grade.

Rule 1; if FE > 19.45 and CW<= 16.5 and maths_olevel_results > 5.0 then Remarks = Fails

Rule 2; if FE <= 19.45 and maths_olevel_results >4.5 and CW <= 24.45 then Remarks = Fail

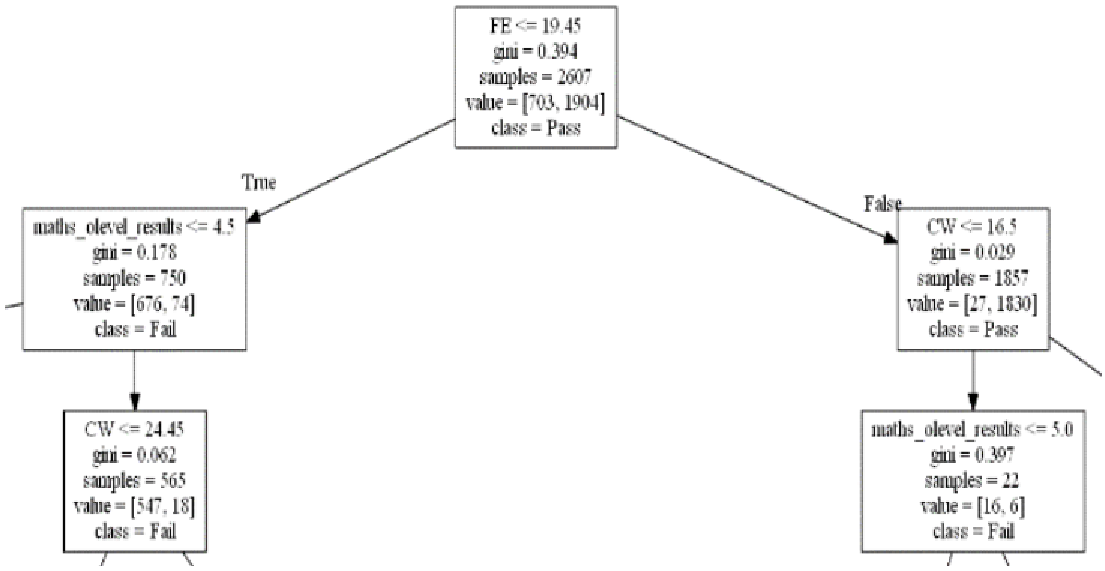

Figure 4.2: Decision Tree Classifier for Pattern and Knowledge Discovery

4.3. Training of ML algorithms for Mathematics Performance Prediction

ML algorithms were trained in Python 3 on the training set after dividing the dataset into a training set (60%), a validation set (20%) and a test set (20%). Therefore, this study trained different ML algorithms based on the binary classification that is widely applied in EDM to predict the performance of the students at primary, secondary, college and HLIs. ML classification algorithms selected based on the literature reviews in EDM for students' performance prediction were DT, K-NN, RF, SVM and MLP. Upon the training of ML algorithms, K-fold cross-validation was applied to validate the models whereby 10-fold cross-validation was used.

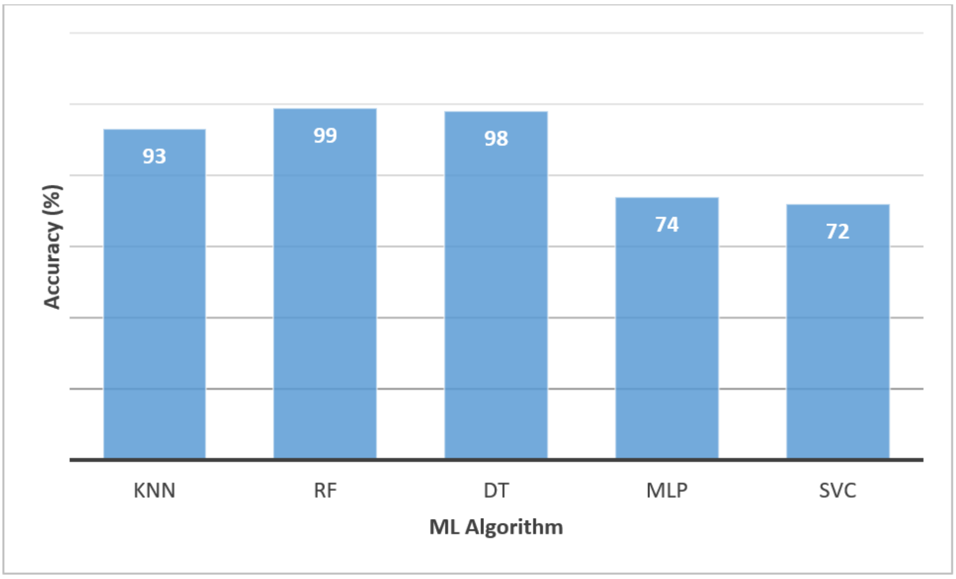

Figure 4.3: Validation Results on Accuracy for Trained ML algorithms

Source: Researcher (2020)

Figure 4.3 depicts the validation results of the trained ML algorithms on accuracy which were K-NN, RF, DT, MLP and SVC. The accuracy was used to determine the best-trained ML algorithms for better generalization to the unseen dataset (testing set) for prediction in mathematics performance. K Nearest Neighbor was one of the best-trained ML algorithms and the validation result was 93% accuracy. Also, the RF ML algorithm was trained and the validation result was 99% accuracy which was found to be the highest among other algorithms. Likewise, the DT ML algorithm had validation results of 98% accuracy as the second algorithm with highest accuracy. Furthermore, MLP had validation results of 74% accuracy and SVC had validation results of 72% accuracy which did not perform better compared to the other three mentioned algorithms.

Therefore, the validation results in this study were compared with that of Zhou (2018) of training ML algorithms with all features due to the dependences of predictor variables to each other (Zhou, 2018). From that study, DT accuracy was 91.5% while in this study was 98%. Moreover, RF from that study had an accuracy of 72.4% while in this study was 99% and for MLP algorithm in this study was 74% while from that study was 88.3%. Likewise, SVC in this study was 72% while 92.6% from that study. Hence, it shows that the trained algorithm can well predict the mathematics performance prediction when trained with all features due to their dependencies.

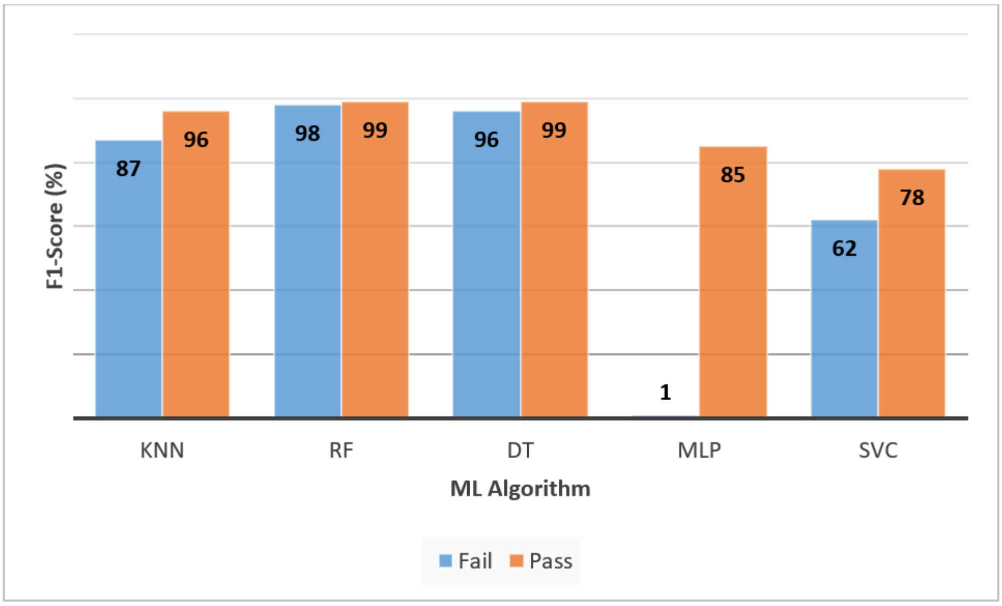

Figure 4.4: Validation Results on F1-Score for Trained ML Algorithms

Source: Researcher (2020)

Similarly, Figure 4.4 depicts the validation results of the trained ML algorithms on F1-score which is the important metric for classification problem that tells how the trained ML algorithm predicts well the pass or fail class. F1-score is also an important metric for the validation of trained ML algorithms as it takes precision and recall simultaneously. One of the studies done by Vihavainen et al. (2013) used the F-measure as the evaluation metric when predicting the students' success in the introductory mathematics course (Vihavainen et al., 2013). Hence, this study applied the F-measure metric so as to validate well the trained ML algorithms. F1-score for fail class and pass class tell how correctly the trained ML algorithm was able to predict the fail class and pass class respectively on validation data set. As it was discussed in the accuracy as the evaluation metric of the best-trained ML algorithms, also K-NN, RF and DT had the best F1-score validation results compared to MLP and SVC. For K-NN, 87% F1-score (Fail class) and 96% F1-score (Pass class) were obtained as well as 98% and 99% F1-score for Fail and Pass class respectively in the RF algorithm. In the DT algorithm, the F1-score results were 96% and 99% for fail and pass class respectively. For MLP algorithm, 1% was the F1-score for prediction of fail class which shows that MLP was not able to classify well while 85% F1-score for pass class. Likewise, SVC had a 62% F1-score for fail class and 78% F1-score for pass class.

Therefore, from the validation results of accuracy and F1-score of trained ML algorithms, it can be observed that considering only the accuracy for the selection of the best-trained ML algorithms is not enough in classification problems. Hence, F1-score should also be considered to have best ML algorithms for prediction on unseen data during evaluation. The purpose of validation results for trained ML algorithms was to avoid overfitting or underfitting when comparing them against evaluation results of the testing set (unseen data) in prediction.

4.4. Evaluation of Trained ML Algorithms

After the training of ML algorithms from the previous section, different evaluation metrics for ML algorithms were applied to determine the performance prediction in mathematics using the testing set. Evaluation metrics such as precision, recall, accuracy and F1-score were calculated from TP, TN, FP FN values of the confusion matrix for each classifier. The testing set was used in the evaluation of the trained ML algorithms to find the values of the evaluation metrics such as correctly classified classes, incorrectly classified classes, precision, recall, F1-Score and accuracy as shown in Table 4.3.

Table 4.3: Evaluation of selected Best Trained ML algorithms in Mathematics Performance Prediction

ML Algorithm	Support	Correctly Classified instances	Incorrectly Classified instances	Precision	Recall	F1-Score	Accuracy
DT	Fail (157) Pass (496)	155 491	2 5	0.98 1.00	0.99 0.99	0.98 0.99	0.99
RF	Fail (157) Pass (496)	156 493	1 3	0.98 1.00	0.99 0.99	0.99 1.00	0.99
KNN	Fail (157) Pass (496)	146 482	11 14	0.91 0.98	0.93 0.97	0.92 0.97	0.96

Source: Researcher (2020)

A comparison of the accuracies as seen in Figure 4.5 depicts the validation and testing results on accuracies of the selected best-trained ML algorithms. The selected trained ML algorithms had almost similar accuracies in validation and testing such as 96% accuracy in testing for the K-NN algorithm while it was 93% in the validation result which is an increase of 3%. Likewise, the accuracy during testing for RF was similar to that during validation that was 99%. Furthermore, the accuracy in the testing set for DT has increased by 1% from 98% in the validation results to 99% in testing results. Therefore, from the comparison of accuracies on validation and testing, it can be observed that no overfitting or underfitting on the best selected trained ML algorithms.

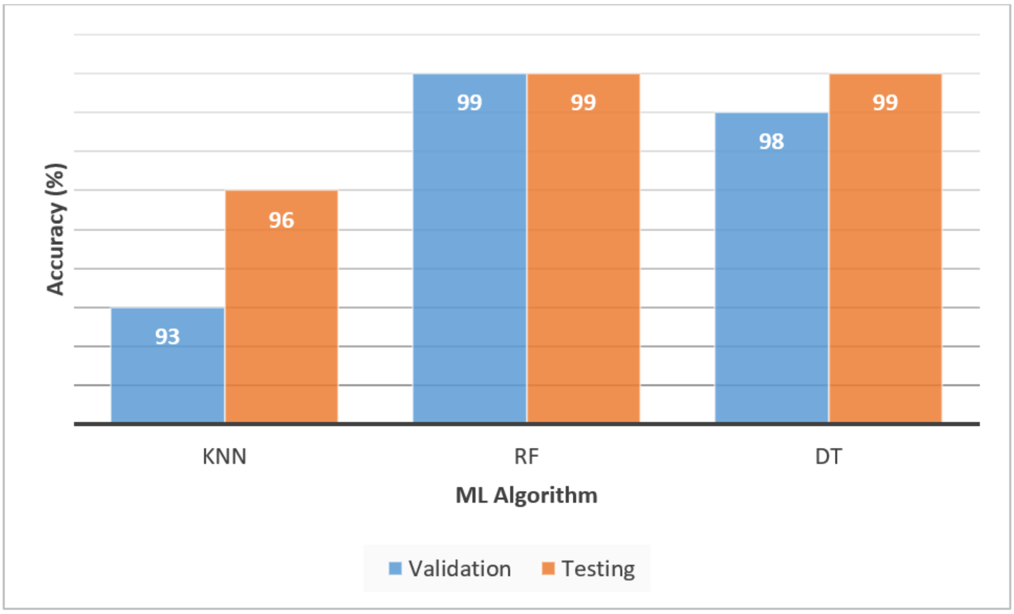

Figure 4.5: Comparison of Validation and Testing Results on Accuracy for Selected Best Trained ML Algorithms

Source: Researcher (2020)

Furthermore, F1-scores in testing were compared with that of the validation during the training of the ML algorithms as shown in Figure 4.6. For the RF algorithm in fail class, F1-score has increased by 1% from 98% in the validation results to 99% in testing results, and also increased by 1% from 99% in the validation results to 100% in testing results. Moreover, the DT algorithm in fail class, F1-score has increased by 11% from 87% in the validation results to 98% in testing results, and also increased by 3% from 96% in the validation results to 99% in testing results. For the K-NN algorithm, F1-score in fail class has decreased by 4% from 96% in the validation results to 92% in testing results, and in the pass class has decreased by 2% from 99% in the validation results to 97% in testing results.

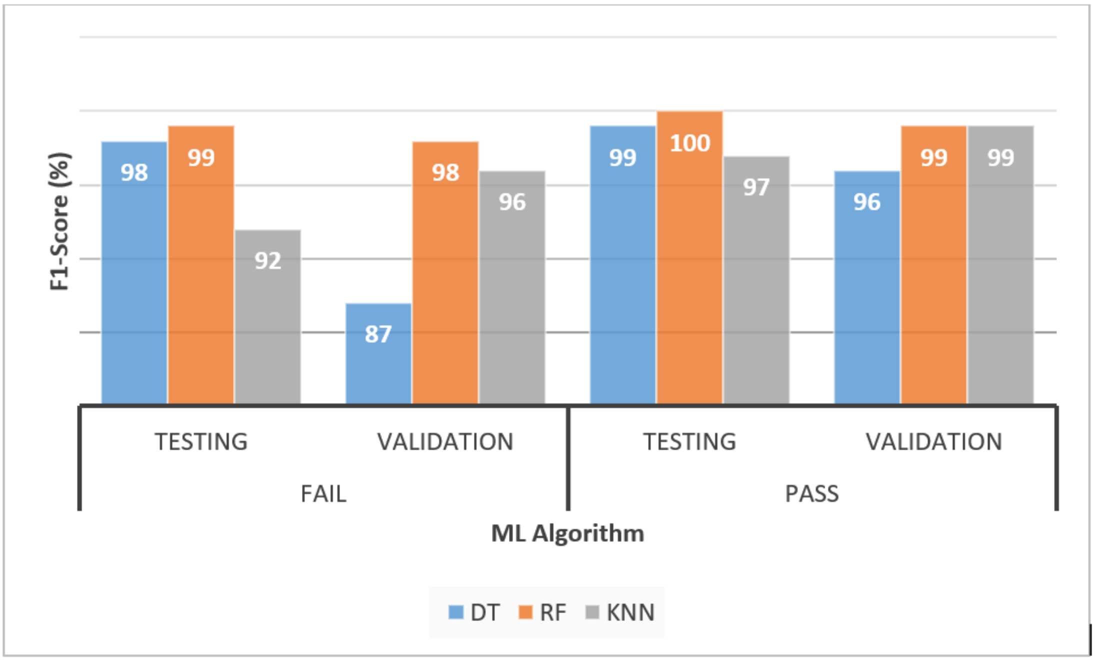

Figure 4.6: Comparison on Validation and Testing Results in F1-scores for Selected Best Trained ML Algorithms

Source: Researcher (2020)

Additionally, it shows that the prediction in mathematics performance using unseen data was done successfully using RF and DT algorithms both with the highest accuracy of 99% followed by K-NN with 96%. Also, the prediction in mathematics performance was successfully on unseen data for a fail class such that the RF algorithm performed best with an F1-score of 99% followed by DT with 98% then K-NN with 92%. Similarly, in pass class for mathematics performance prediction, the RF algorithm had the highest F1-score of 100% followed by DT with 99% and then K-NN with 97%.

The results were compared with the study done by Ünal (2020) who compared the five-level grading that had an accuracy of 71.14% for RF algorithm and binary level grading in mathematics that had an accuracy of 91.39% (Ünal, 2020). Also from that study, DT had an accuracy of 73.42% when five-level grading was applied and then accuracy increased to 89.11% when binary level grading was used. Similarly, the results obtained in this study had high accuracies that involved binary classification as either student will pass or fail in mathematics course as a result accuracies increases as the classification level decreases.

Furthermore, the results were compared to the work of Yadav & Pal (2012) about the performance prediction to engineering students using classification algorithms and obtained the accuracy of 67.8% for the C4.5 algorithm that is suitable for the multi-classification problem that was classifying as either pass, fail or promoted (Yadav & Pal, 2012). In this study, the Classification and Regression Trees (CART) algorithm was applied that is suitable for binary classification problems and found to be among the best in performance prediction for mathematics course to management degree students with 99% accuracy on testing the DT classifier. The difference in the accuracies of the trained ML algorithms might be due to the size of the dataset as in this study 1955 instances were used to train the ML algorithm and 653 were used to test the ML algorithm while Yadav & Pal (2012) used only 90 instances for training the ML algorithm.

The similarity in this study and that of Yadav & Pal (2012) are in the some of the predictor variables in performance of the students such as past performance records like mathematics grades in ordinary level for this study, living location of the students as to whether hostel or off-campus, admission type as either direct or equivalent (form six, diploma or RPL in this study). The number of features as predictive variables in Yadav & Pal (2012) study was 16 from which it supported its ML algorithm accuracy apart from having 90 instances for training the ML algorithm. Therefore, from these two studies on performance prediction, the accuracy of the trained ML algorithm depends on the number of instances to train the ML algorithm and the predictor variables to predict the response variable.

Likewise, the results were compared with that of Kabra & Bichkar (2011) who researched on performance prediction of engineering students using a DT. In that study, the correlation was found between the past performance of the students and the future performance prediction (Kabra & Bichkar, 2011). The accuracy of the trained DT algorithm was 60.46% in such a way that it predicted correctly 209 instances out of 346 instances. It was concluded in that study that past academic performance predict the future performance of the students such that trained DT algorithm was able to obtain a true positive rate (Recall) of 0.907 for a fail class that indicates it can predicts successfully the students who are at risks in engineering during the first year (Kabra & Bichkar, 2011). Likewise in this study, the trained DT algorithm was able to predict successfully the students who are likely to fail in mathematics with a Recall value of 0.99 for fail class. This study similarly to Kabra Bichkar (2011) also considered the ordinary level mathematics performance as the past academic performance to predict the future performance of the students.

Moreover, from Figure 4.7 it can be observed that the RF algorithm was the best as it correctly classified 649 instances on predicting student's performance in mathematics and incorrectly classified only 4 instances out of 653 instances from the testing set. Likewise, the DT ML algorithm classified correctly 646 instances and 7 instances incorrectly and the K-NN algorithm classified 628 instances correctly and 25 instances incorrectly.

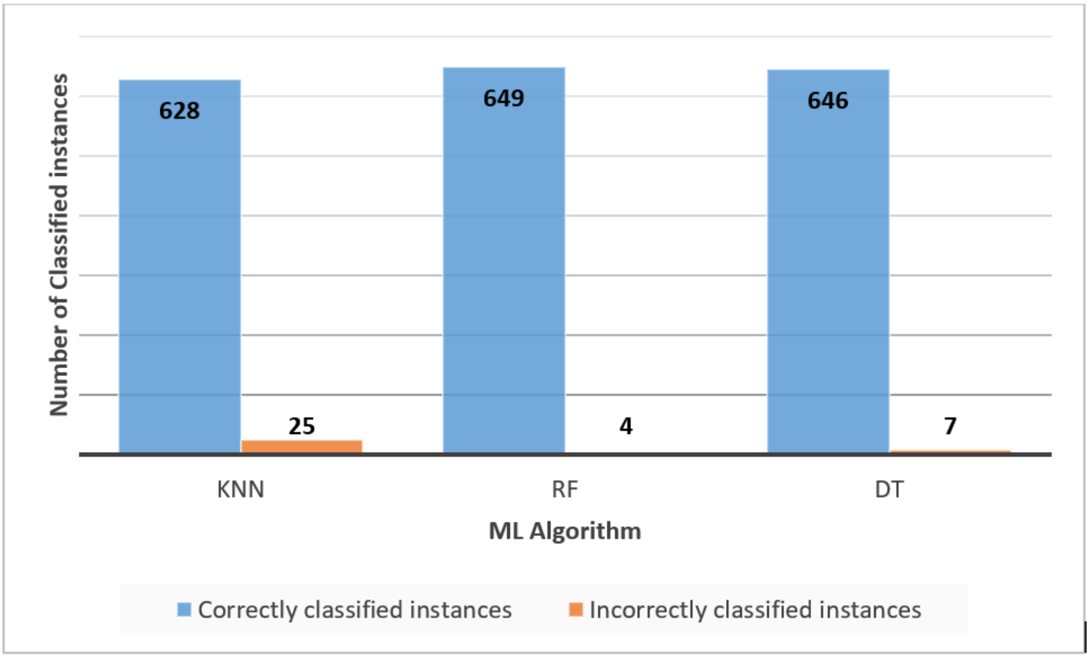

Figure 4.7: Number of Classified Instances With Respect to the Selected Best ML Algorithms

The results of RF in this study were compared with the study of Pandey & Taruna (2014) which involved a RF ML algorithm for students' performance prediction and correctly classified 701 instances while 297 instances were incorrectly classified and the accuracy was 72.04% (Pandey & Taruna, 2014). The predictor variables used in that study were 10 similar to this study but differ in some variables such as aggregate grades in the first and second semesters while in this study considered only the first semester.

Therefore, it can be concluded that the best ML algorithm for mathematics performance prediction in this study was RF with an accuracy of 99% and F1-score of 99% for fail class and 100% for pass class. As a result, the RF predictive model for mathematics performance to degree management students in this study was found to be the best. The trained ML algorithm was saved for future use by the HLIs that offer management degree programmes in mathematics performance as it successfully predicted correctly on the unseen data set. The HLIs that may apply the trained ML algorithm are the ones that offer art-based management programmes. HLIs such as the Institute of social work, Jordan University College, Local Government Training Institute, Moshi Co-operative University and The Mwalimu Nyerere Memorial Academy and others in Tanzania.

4.5. Chapter Summary

The findings and discussion in this chapter were to answer the three research questions based on the specific objectives to attain the main objective of predicting mathematics performance to degree management students using EDM techniques. Subsections involved were identification of predictor variables for performance prediction in mathematics, dataset creation, feature extraction and selection, training of the ML algorithms and the evaluation of the trained ML algorithms. Five selected ML algorithms were trained including MLP, SVC, RF, DT and K-NN and during the validation the best algorithms were RF, DT and K-NN. During the evaluation of the three best ML algorithms, RF ML algorithm was found to be the best in mathematics performance prediction in this study with an accuracy of 99% and F1-scores of 99% and 100% for fail and pass class respectively.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1. Introduction

This chapter of the study is describing the recommendations of this study, future of the work as where other researchers should focus on and the conclusion of this study. The recommendation section was about the opinions of the researcher to various stakeholders regarding the results of this study. Also, future work was about on what can be done in research whereby this study has not covered.

5.2. Conclusion

The study aimed at performance prediction in mathematics for management degree students in the case of MU in Tanzania. The requirements for the performance prediction were to be determined such as the predictor variables to create a dataset to train ML algorithms. The dataset was created after the data collection based on the predictor variables. Then, 60% of the dataset built the training dataset, 20% of the dataset to validate the trained ML algorithms and 20% of the dataset for evaluation of trained ML algorithms that was the prediction of performance in mathematics. The study involved the ML algorithms training in such a way that five ML algorithms were selected based on the literature reviews that are widely applied and showed good performance in EDM for performance prediction. In this study, five selected ML algorithms were trained including MLP, SVC, RF, DT and K-NN and during validation, the best algorithms were RF, DT and K-NN. During evaluation of the three best ML algorithms, RF ML algorithm was found to be the best in mathematics performance prediction in this study with an accuracy of 99% and F1-scores of 99% and 100% for fail and pass class respectively. As a result, RF predictive model for mathematics performance to management degree students was found to be the best in this study. Moreover, DT was able to generate rules that were applied to recommend the minimum grade of D for ordinary level mathematics in admission to degree management students so as to reduce the failure rate. Therefore, the study was successful in achieving all the three specific objectives to reach to the main objective of predicting the mathematics performance to degree management students in HLIs using EDM techniques.

5.3. Recommendations

This section describes the recommendations based on the results of the study in mathematics performance prediction for students pursuing management degree programmes at HLIs. From the findings of this study in identifying the requirements for training ML algorithms, some of the predictor variables had no data because they were not captured. Hence, this study recommended to the government and HLIs that data have to be preserved like education and job of the parents or guardians, interest with the course or programme, attendance of the students in a class for lectures and seminars, and family size of the students. HLIs in Tanzania should be able to capture various data of the students and be stored in digital form for the purpose of being used by the researchers, especially in ML. The data might be used on solving various problems such as drop out of students and failing in subjects or discontinuation from studies.

Moreover, from the results of the first research question on identifying the requirements for training ML algorithms, a DT classifier was applied to draw the pattern and knowledge discovery in mathematics ordinary level grades. Moreover, in feature importance, mathematics ordinary level grades had high scores in contribution to students' performance in mathematics pursuing management programmes at HLIs. Therefore, a minimum grade of D at the ordinary level in mathematics is recommended to TCU and admission offices by this study during admission of first-years to management degree students to the reduce failure rate in mathematics at HLIs.

Furthermore, from the findings of this study, HLIs are recommended to use EDM techniques in solving the students' problems such as failing in courses so as to determine earlier through prediction the students who are likely to fail in the future. The students will be advised earlier by lecturers during the first year of the first semester to avoid supplementary and discontinuation from studies.

5.4. Future Work

In this study, there are some issues that were not addressed that may be implemented by other researchers in the future as follows. The study was based only on the secondary data that were available that resulted in a few numbers of features as predictor variables for students' performance prediction. Therefore, other researchers may consider the primary data so as to have more features by preparing questionnaires to capture data such as parents' job and education, romantic relationship status of the students, time taken to study mathematics privately, and interests with mathematics and family size.

Moreover, the study based on the HLIs to degree students, other researchers may apply the ML algorithms in primary, secondary and colleges for certificate and diploma students as well as considering other programmes such as engineering, computer science, and business programmes at HLIs. Similarly, the study did not consider the online courses in electronic learning such as Moodle to get features like time spent working on course in Moodle, uploaded assignments in Moodle, participation in a group discussion online. Therefore, other researchers may involve the study in online classes.

Also, the study trained five ML algorithms such as K-NN, RF, SVC, DT and MLP as the widely used algorithms in EDM for performance prediction of the students. Therefore, other algorithms such as NB and LR may be applied by other researchers to solve the problem of the students' performance at primary, secondary, college and HLIs. Likewise, the other researchers may consider working on the performance prediction of the students in the final year considering all the courses as either the student will graduate or not. This will assists students to be advised earlier on the ways to follow to be a graduate, as a result, to take measures to avoid discontinuation.

5.5. Chapter Summary

This chapter concluded the work of the study in mathematics performance prediction using the EDM techniques. Also, this chapter covered the recommendations made in this study as to what should be done so as to improve the research works in ML. Moreover, this chapter discussed the future research works that other researchers may do in ML so as to solve various problems in the society such as health problems, educational problems and business problems.

REFERENCES

Abu, A. (2016). Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5), 212–220. https://doi.org/10.14569/ijacsa.2016.070531

Abu, A., Mostafa, S., Emran, A., & Shaalan, K. (2019). *Factors Affecting Students ' Performance in Higher Education : A Systematic Review of Predictive Data Mining Techniques*. *Technology, Knowledge and Learning*. Springer Netherlands. https://doi.org/10.1007/s10758-019-09408-7

Abu Zohair, L. M. (2019). Prediction of Student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education*, *16*(1). https://doi.org/10.1186/s41239-019-0160-3

Agarwal, S. (2013). Data Mining. https://doi.org/10.1109/ICMIRA.2013.45

Allibhai, E. (2018). Hold-out vs. Cross-validation in Machine Learning. Retrieved February 8, 2020, from https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f

Anuradha, C., & Velmurugan, T. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. *Indian Journal of Science and Technology*, *8*(July).

Ashenafi, M. M. (2017). A Comparative Analysis of Selected Studies in Student Performance Prediction, *7*(4), 17–32. https://doi.org/10.5121/ijdkp.2017.7402

Bhavsar, H., & Ganatra, A. (2012). A Comparative Study of Training Algorithms for Supervised Machine Learning. *International Journal of Soft Computing and Engineering*, (January 2012).

Bunkar, K. (2012). Data Mining : Prediction for Performance Improvement of Graduate Students using Classification. *IEEE*, 1–5.

Cortez, P., Silva, A., Trees, D., & Forest, R. (2003). Using Data Mining to Predict Secondary School Student Performance, *2003*(2000).

Creswell, J. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). Thousand Oaks, California: SAGE Publications.

Fussy, D. S. (2018). The status of academic advising in Tanzanian universities. *KEDI Journal of Educational Policy*, *15*(1), 81–98.

Hossin, M., & Sulaiman, M. . (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge*

Management Process (IJDKP), *5*(2), 1–11.

Innocentia, K., Ramaphosa, M., Zuva, T., & Kwuimi, R. (2018). Educational Data Mining to Improve Learner Performance in Gauteng Primary Schools. *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (IcABCD)*, 1–6.

Kabra, R. R., & Bichkar, R. S. (2011). Performance Prediction of Engineering Students using Decision Trees, *36*(11), 8–12.

Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, & Samir Chatterjee. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77.

Kumah, M. S., Region, V., Akpandja, T. K., Region, V., Djondo, B. I., & Region, V. (2016). Factors Contributing To The Poor Performance In Mathematics : A Case Study Among Students ' In Colleges Of Education - Ghana. *ResearchJournali's Journal of Mathematics*, *3*(2), 1–12.

Kumar, M., & Salal, Y. K. (2019). Systematic review of predicting student's performance in academics. *International Journal of Engineering and Advanced Technology*, *8*(3), 54–61.

Kumar, M., Singh, P. A. J., & Handa, D. (2017). Literature Survey on Student ' s Performance Prediction in Education using Data Mining Techniques. *MECS Publisher*, (November), 40–49. https://doi.org/10.5815/ijeme.2017.06.05

Kumar, R. (2011). *Research methodology: A step-by-step guide for beginners*. Los Angeles: SAGE.

Kumar, R. (2012). Classification Algorithms for Data Mining : A Survey. *International Journal of Innovations in Engineering and Technology (IJIET)*, *1*(2), 7–14.

Livieris, I. E., Drakopoulou, K., & Pintelas, P. (2012). Predicting students' performance using artificial neural networks. *In Proceedings of the 8th Pan-Hellenic Conference "Information and Communication Technology in Education.*, 28–30.

Mduma, N., Kalegele, K., & Machuve, D. (2019). Machine learning approach for reducing students dropout rates. *International Journal of Advanced Computer*

Research, *9*(42), 156–169. https://doi.org/10.19101/ijacr.2018.839045

Mellor, D. (2013). Machine Learning For Beginners. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Muhammad, I., & Yan, Z. (2015). Supervised Machine Learning Approches : A Survey. *ICTACT Journal on Soft Computing*, 946–952. https://doi.org/10.21917/ijsc.2015.0133

Mythili, M. ., & Mohamed, S. A. . (2014). An Analysis of students' performance using classification algorithms. *IOSR Journal of Computer Engineering*, *16*(1), 63–69. https://doi.org/10.9790/0661-16136369

Mzumbe University. (2016). Mzumbe University History. Retrieved January 10, 2020, from https://site.mzumbe.ac.tz/

Mzumbe University. (2018). Mzumbe University Examinations and Students' Assessment Criteria Bylaws, (September), 1–48.

Mzumbe University. (2019). *Prospectus 2018/2019*. Mzumbe University-Office of Deputy Vice Chancellor – Academic Affairs.

Nikam, S. S. (2017). A Comparative Study of Classification Techniques in Data Mining Algorithms. *International Journal of Modern Trends in Engineering & Research*, *4*(7), 58–63. https://doi.org/10.21884/ijmter.2017.4211.vxayk

Nisbet, R., Miner, G., & Yale, K. (2018). Theoretical Considerations for Data Mining. In *Handbook of Statistical Analysis and Data Mining Applications* (2nd ed., pp. 21–37). Elsevier Inc. https://doi.org/https://doi.org/10.1016/C2012-0-06451-4

Okwute, A. O., & Musa, C. D. (2018). Investigating Achievement of Continuous Assessment Techniques on Students' Discovery in Mathematics Performance at SSSI Level at GDSS Bwari , Bwari Area Council , Abuja , FCT. *International Journal of Innovative Research in Information Security (IJIRIS)*, *5*(06), 2014–2019. https://doi.org/10.26562/IJIRAE.2018.AUIS10080

Olatunji, S. O., Aghimien, D. O., Oke, A. E., & Olushola, E. (2017). Factors Affecting Performance of Undergraduate Students in Construction Related Disciplines. *Journal of Education and Practice*, *7*(13), 55–62.

Omid, M. (2013). What is the best way to divide a dataset into training and test sets? Retrieved February 12, 2020, from https://www.researchgate.net/post/What is the best way to divide a dataset into training and test sets

Osmanbegovic, E., & Suljic, M. (2012). Data Mining Approach for Predicting Student Performance. *Economic Review: Journal of Economics and Business*, *10*(1), 3–12.

Pandey, M., & Taruna, S. (2014). A Comparative Study of Ensemble Methods for Students' Performance Modeling. *International Journal of Computer Applications*, *103*(8), 26–32. https://doi.org/10.5120/18095-9151

Roy, S., & Garg, A. (2017). Analyzing Performance of Students by Using Data Mining Techniques A Literature Survey, 130–133.

Rustia, R. A., Cruz, M. M. A., P, M. A., & Palaoag, T. D. (2018). Predicting Student ' s Board Examination Performance using Classification Algorithms. *ICSCA*, 1–5.

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Singh, N. T., & Sharma, A. (2016). A Review of Supervised Machine Learning Algorithms. In *3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310–1315). New Delhi.

Sivasakthi, M. (2018). Classification and prediction based data mining algorithms to predict students' introductory programming performance. *Proceedings of the International Conference on Inventive Computing and Informatics, ICICI 2017*, (Icici), 346–350. https://doi.org/10.1109/ICICI.2017.8365371

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for Educational Data Mining: A Review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85–106. https://doi.org/10.3102/1076998616666808

TCU. (2019). *Handbook for Standards & Guidelines for Universiry Education in Tanzania* (3rd ed.).

TCU. (2020). Tanzania Commission for Universities (TCU) Undergraduate Admission Guidebook for 2020 / 2021 Academic Year (For Holders of Form Six

Qualifications).

Ünal, F. (2020). Data Mining for Student Performance Prediction in Education. *Intech*, 11. https://doi.org/10.5772/intechopen.91449

Vihavainen, A., Luukkainen, M., & Kurhila, J. (2013). Using students' programming behavior to predict success in an introductory mathematics course. *Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013*.

Vista College. (2019). The Importance of Higher Education in the 21st Century. Retrieved January 13, 2020, from https://www.vistacollege.edu/blog/resources/higher-education-in-the-21st-century/

Widyahastuti, F., & Tjhin, V. U. (2017). Predicting students performance in final examination using linear regression and multilayer perceptron. *Proceedings - 2017 10th International Conference on Human System Interactions, HSI 2017*, 188–192. https://doi.org/10.1109/HSI.2017.8005026

Wikipedia. (2020). Multilayer Perceptron. Retrieved September 3, 2020, from https://en.wikipedia.org/wiki/Multilayer_perceptron

Yadav, S. K., & Pal, S. (2012). Data Mining : A Prediction for Performance Improvement of Engineering Students using Classification, *2*(2), 51–56.

Zaffar, M., Hashmani, M. A., Savita, K. S., & Rizvi, S. S. H. (2018). A study of feature selection algorithms for predicting students academic performance. *International Journal of Advanced Computer Science and Applications*, *9*(5), 541–549. https://doi.org/10.14569/IJACSA.2018.090569

Zhou, Z. (2018). Student Pass Rates Prediction Using Optimized Support Vector Machine and Decision Tree. *IEEE*, 209–215.

APPENDICES

Appendix A: Univariate analysis for FE

```
In [40]: df.FE.describe()

Out[40]: count    3259.000000
         mean       23.768134
         std        10.020649
         min         0.000000
         25%        18.000000
         50%        23.000000
         75%        30.000000
         max        50.000000
         Name: FE, dtype: float64
```

Appendix B: Univariate analysis for CW

```
In [44]: df.CW.describe()

Out[44]: count    3259.000000
         mean       27.212584
         std         6.070689
         min         0.500000
         25%        23.000000
         50%        27.000000
         75%        31.650000
         max        47.000000
         Name: CW, dtype: float64
```

Appendix C: Z test between Remarks and CW

```
In [82]: df['Remarks'].value_counts()

Out[82]: Pass    2403
         Fail     856
         Name: Remarks, dtype: int64

In [83]: z=(22.522185-28.883404)/np.sqrt(np.square(27.743447)/856+np.square(29.478399)/2403)
         z

Out[83]: -5.665221133379838

In [ ]: p=7.34175E-09
```

Appendix D: Z test between Remarks and FE

```
In [88]: z=(11.828446-28.014565)/np.sqrt(np.square(22.611058)/856+np.square(59.493907)/2403)

In [89]: z

Out[89]: -11.249513159874512

In [ ]: p=1.16437E-29
```

Appendix E: chi square test between maths olevel results and sponsorship

```
In [27]: compute_freq_chi2(df['maths_olevel_results'],df.sponsorship)

Frequency table
===========================
sponsorship           heslb  private
maths_olevel_results
A                         0       10
B                        12       26
B+                        4        5
C                       160      166
D                       593      716
E                        37       50
F                       644      836
===========================
ChiSquare test statistic:  14.41650635393053
p-value:  0.025314271508836986
degree of freedom: 6
```

Appendix F: chi square test between maths olevel results and entry category

```
In [28]: compute_freq_chi2(df['maths_olevel_results'],df.entry_category)

Frequency table
===========================
entry_category        RPL  diploma  form six
maths_olevel_results
A                       0        0        10
B                       0        6        32
B+                      0        1         8
C                       0       41       285
D                       2      321       986
E                       0       40        47
F                       4      310      1166
===========================
ChiSquare test statistic:  57.58133695444822
p-value:  6.205943231931528e-08
degree of freedom: 12
```

Appendix G: chi square test between age group and living location

```
compute_freq_chi2(df['Age Group'],df['living_location'])

Frequency table
============================
living_location  hostel  off_campus
Age Group
0-19                 75         459
20-29               579        2042
30-39                25          68
40-49                 0          11
============================
ChiSquare test statistic:  22.389530681650506
p-value:  5.412230137741571e-05
degree of freedom: 3
```

Appendix H: Chi square between campus location and remarks

```
In [29]: compute_freq_chi2(df.campus_location,df.Remarks)

Frequency table
============================
Remarks          Fail  Pass
campus_location
rural             612  1617
town              244   786
============================
ChiSquare test statistic:  4.9688889330414785
p-value:  0.025807220246515706
degree of freedom: 1
```

Appendix I: chi square test between maths olevel results and sex

```
compute_freq_chi2(df['maths_olevel_results'],df['Sex'])
```

```
Frequency table
============================
Sex                        F    M
maths_olevel_results
A                          4    6
B                         21   17
B+                         2    7
C                        158  168
D                        711  598
E                         51   36
F                        862  618
============================
ChiSquare test statistic:  17.135335537041314
p-value:  0.008798525877424366
degree of freedom: 6
```

Appendix J: chi square test between maths olevel results and remarks

```
compute_freq_chi2(df['maths_olevel_results'],df['Remarks'])
```

```
Frequency table
============================
Remarks               Fail  Pass
maths_olevel_results
A                        0    10
B                        0    38
B+                       0     9
C                        2   324
D                      159  1150
E                       61    26
F                      634   846
============================
ChiSquare test statistic:  562.0579744103247
p-value:  3.549969010659971e-118
degree of freedom: 6
```

Appendix K: chi square test between maths olevel results and age group

```
compute_freq_chi2(df['maths_olevel_results'],df['Age Group'])
```

```
Frequency table
============================
Age Group             0-19  20-29  30-39  40-49
maths_olevel_results
A                        1      9      0      0
B                        7     31      0      0
B+                       2      7      0      0
C                       37    287      2      0
D                      190   1108      9      2
E                        6     52     24      5
F                      291   1127     58      4
============================
ChiSquare test statistic:  335.71003822213686
p-value:  2.0726188076267828e-60
degree of freedom: 18
```

Appendix L: chi square test between remarks and age group

```
compute_freq_chi2(df['Age Group'],df['Remarks'])
```

```
Frequency table
============================
Remarks     Fail  Pass
Age Group
0-19         242   292
20-29        526  2095
30-39         82    11
40-49          6     5
============================
ChiSquare test statistic:  340.63807656868846
p-value:  1.5875870373801625e-73
degree of freedom: 3
```

Appendix M: chi square test between entry category and remarks

```
compute_freq_chi2(df['entry_category'],df['Remarks'])
```

```
Frequency table
============================
Remarks         Fail  Pass
entry_category
RPL                1     5
diploma          298   421
form six         557  1977
============================
ChiSquare test statistic:  109.86291235153169
p-value:  1.3917835527270391e-24
degree of freedom: 2
```

Appendix N: Chi square test between number of instructors and maths olevel grades

```
In [23]: compute_freq_chi2(df.number_of_instructors,df.maths_olevel_results)
```

```
Frequency table
============================
maths_olevel_results   A   B  B+    C    D   E     F
number_of_instructors
one                    8   0   6   68  440  42   466
two                    2  38   3  258  869  45  1014
============================
ChiSquare test statistic:  64.56647960400694
p-value:  5.2890802205245264e-12
degree of freedom: 6
```

Appendix O: Dataset

	Sex	CW	FE	Remarks	maths_olevel_results	living_location	age	entry_category	sponsorship	campus_location	number_of_instructors
0	F	23.0	17.5	Pass	D	off_campus	19	form six	heslb	rural	two
1	F	23.5	17.5	Pass	D	hostel	19	form six	private	rural	two
2	F	25.0	18.5	Pass	D	off_campus	20	form six	private	rural	two
3	M	23.0	20.5	Pass	D	off_campus	20	form six	heslb	rural	two
4	F	24.5	31.5	Pass	D	hostel	20	form six	private	rural	two
...	...	...	...	...	...	...	...	...	...	...	...
3254	F	22.0	26	Pass	D	hostel	24	form six	private	rural	two
3255	M	18.0	27.5	Pass	D	hostel	24	form six	private	rural	two
3256	M	25.0	20	Pass	D	hostel	24	form six	private	rural	two
3257	M	25.5	24	Pass	D	off_campus	24	form six	heslb	rural	two
3258	F	19.0	25.5	Pass	D	off_campus	24	form six	heslb	rural	two

3259 rows × 11 columns

Appendix P: ANOVA between maths olevel results and CW

In [40]:
```
k=7
n=3259
numerator = np.sum([np.square(31.900000-27.212584)/(k-1),
                    np.square(32.090526-27.212584)/(k-1),
                    np.square(31.495556-27.212584)/(k-1),
                    np.square(35.030982-27.212584)/(k-1),
                    np.square(26.718121-27.212584)/(k-1),
                    np.square(23.765632-27.212584)/(k-1),
                    np.square(25.947419-27.212584)/(k-1)])
numerator
```

Out[40]: 23.16066626291767

In [41]:
```
a = np.sum(np.square(df[df['maths_olevel_results']=='A'].CW-31.900000)/(n-k))
b = np.sum(np.square(df[df['maths_olevel_results']=='B'].CW-32.090526)/(n-k))
bb = np.sum(np.square(df[df['maths_olevel_results']=='B+'].CW-31.495556)/(n-k))
c = np.sum(np.square(df[df['maths_olevel_results']=='C'].CW-35.030982)/(n-k))
d = np.sum(np.square(df[df['maths_olevel_results']=='D'].CW-26.718121)/(n-k))
e = np.sum(np.square(df[df['maths_olevel_results']=='E'].CW-23.765632)/(n-k))
f = np.sum(np.square(df[df['maths_olevel_results']=='F'].CW-25.947419)/(n-k))
denominator=a+b+bb+c+d+e+f
denominator
```

Out[41]: 29.252382089487938

In [42]:
```
f=numerator/denominator
f
```

Out[42]: 0.7917531704619921

In [ ]:
```
#link to calculate f statistic
http://onlinestatbook.com/2/calculators/F_dist.html
p=0.5763
```

Appendix Q: chi square between entry category and living location against remarks

```
In [3]: data_crosstab = pd.crosstab([df.entry_category, df.living_location],
                                    df.Remarks, margins = False)
        print(data_crosstab)
```

```
Remarks                         Fail  Pass
entry_category living_location
RPL            hostel              1     0
               off_campus          0     5
diploma        hostel             53   110
               off_campus        245   311
form six       hostel            112   403
               off_campus        445  1574
```

```
In [4]: stat, p, dof, expected = chi2_contingency(data_crosstab)
```

```
In [5]: p
```

```
Out[5]: 7.755733177358027e-25
```

Appendix R: Data Transformation for Ordinary Level Mathematics Grades attribute

Categorical values for Ordinary level Mathematics grades	Corresponding Numerical Values
A	0
B+	1
B	2
C	3
D	4
E	5
F	6

Appendix S: Data Transformation for Gender Attribute

Categorical values for gender	Corresponding numerical values
M	1
F	0

Appendix T: Data Transformation for Sponsorship Attribute

Categorical values for Sponsorship	Corresponding numerical values
Private	1
HESLB	0

Appendix U: Permit to collect Data at Mzumbe University

MZUMBE UNIVERSITY
(CHUO KIKUU MZUMBE)

OFFICE OF THE DEPUTY VICE CHANCELLOR (ACADEMIC)

E-Mail: dvca@mzumbe.ac.tz
Tel: +255 023 2931212
Fax: +255 023 2931213
Cell: +255 0754694029
Website: www.mzumbe.ac.tz

P.O. Box 1
Mzumbe
TANZANIA

Ref. No.: MU/OF/R.2/1/Vol. III/113

Date: 5th August, 2020

Mr. Paul Kavishe Mushi,
P.O. Box 259,
DODOMA - TANZANIA

RE: PERMIT TO DATA COLLECTION AT MZUMBE UNIVERSITY

The caption above refers.

This is to acknowledge receipt of the letter form the VC of the University of Dodoma introducing you as well as requesting the VC of Mzumbe University to grant you permission to meet and talk to the leaders and other relevant stakeholders at Mzumbe University (MU) in connection with your research. We were further, informed that your research topic is ***"performance prediction in Mathematics for Degree Students pursuing Mathematics Programmes Using Education Data Mining Techniques: A Case of Mzumbe University in Tanzania"***.

I am glad to inform you that you have been granted permission to carry out data collection exercise at Mzumbe University as was requested by the Vice Chancellor of the University of Dodoma. You are required to carry this letter with you anytime you are moving around the University for the purpose of data collection and present it to whoever may need to identify you. Up on arrival at MU for the purpose of data collection; please report to the Office of the Deputy Vice Chancellor (Administration and Finance) for logistical arrangements. This data collection permit will be valid from the today (see the date above) to **30th October, 2020**. Should you require an extension of time to complete your assignment, you should do so in writing to the office of the Deputy Vice Chancellor (Academic).

I trust you will find conducting data collection exercise at our university an enjoyable endeavor and enjoy your stay at MU.

Yours Sincerely,

Dr. Frank Theodory (PhD)
FOR DEPUTY VICE CHANCELLOR (ACADEMIC)

MZUMBE UNIVERSITY
P. O. Box 1, MZUMBE
TANZANIA

QUOTATION OF REF. NO IS ESSENTIAL

Appendix V: Request for Research Clearance

THE UNIVERSITY OF DODOMA

OFFICE OF THE DEPUTY VICE CHANCELLOR-ARC

DIRECTORATE OF RESEARCH, PUBLICATIONS AND CONSULTANCY

P.O. Box 259
DODOMA, TANZANIA
TEL: +255-026-2310002 FAX: +255-026-2310012
EMAIL: dvcarc@udom.ac.tz;
Website address:www.udom.ac.tz

Ref. No. MA.84/261/02 25th June, 2020

Vice Chancellor
Mzumbe University

RE: REQUEST FOR RESEARCH CLEARANCE

The purpose of this letter is to introduce to you **Mr. Paul Kavishe Mushi** with **Reg. No. HD/UDOM/00159/T.2018** who is a bonafide student of the University of Dodoma and who is at the moment required to conduct research. Our students undertake research activities as part of their study programmes.

In accordance with government circular letter Ref. No. MPEC/R/10/1 dated 4th July 1980; the Vice-Chancellor of the University is empowered to issue research clearances to staff members and students of the University on behalf of the government and the Tanzania Commission for Science and Technology (COSTECH). I am pleased to inform you that I have granted a research clearance to the student listed above.

I therefore, kindly request you to grant him any help that may help him to achieve his research objectives. Specifically, we request your permission for him to work at your University and to meet and talk with staff from admission office, examination office and ICT and other stakeholders in connection with his research.

The title of his research is "*Performance Prediction in Mathematics for Degree Students Pursuing Management Programmes Using Educational Data Mining Techniques: A Case of Mzumbe University in Tanzania*".

The period of his research is from June to October, 2020 and it will cover planned areas.

Should there be any restrictions, you are kindly requested to advise us accordingly. In case you require further information, please do not hesitate to contact us through the Directorate of Research, Publication and Consultancy. P.O Box 251, Dodoma. Tel. No. + (255) 262310301 Email:research@udom.ac.tz

Yours Sincerely,

VICE CHANCELLOR
THE UNIVERSITY OF DODOMA
P.O. BOX 259
DODOMA TANZANIA

Prof. Faustine K. Bee
VICE CHANCELLOR

Appendix W: Request for Ethical Clearance

THE UNIVERSITY OF DODOMA

OFFICE OF THE DEPUTY VICE CHANCELLOR-ARC

DIRECTORATE OF RESEARCH, PUBLICATIONS AND CONSULTANCY

P.O. Box 259
DODOMA, TANZANIA
TEL: +255-026-2310002
FAX: +255-026-2310012
EMAIL: dvcarc@udom.ac.tz;
Website address:www.udom.ac.tz

Ref. No. MA.84/261/01/40

25th June, 2020

To: Paul Kavishe Mushi
The University of Dodoma

RE: REQUEST FOR ETHICAL CLEARANCE

This is to inform you that the proposal titled "*Performance Prediction in Mathematics for Degree Students Pursuing Management Programmes Using Educational Data Mining Techniques: A Case of Mzumbe University in Tanzania.*" has been granted ethical clearance.

Furthermore, as the Principal Investigator of the study, the following conditions must be fulfilled:

- Progress report is submitted to the University of Dodoma.
- Permission to publish the results is obtained from the University of Dodoma.
- Copies of final publications are made available to the University of Dodoma.
- Sites: Mzumbe University

Approval is valid for a duration provided for under clause five (5) of the Ethical Clearance Form.

Best Regards,

Dr. Ambrose Kessy
Chairperson- Institutional Research Review Committee (IRREC)

C: C: Deputy Vice Chancellor-Academic, Research and Consultancy