MODELING THE PREVALENCE OF DEPRESSION AMONG MURANG'A UNIVERSITY OF TECHNOLOGY STUDENTS USING MACHINE LEARNING TECHNIQUES

Loise Muthoni Wahome

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Statistics of Murang'a University of Technology

July, 2024

DECLARATION

Student Declaration

I hereby declare that this Thesis is my original work and to the best of my knowledge has not been presented for a degree award in this or any other university. I understand what plagiarism is and I am aware of the University's policy in this regard. I understand that any false claim in respect of this work shall result in disciplinary action in accordance with the University Anti-Plagiarism Policy.

_____ _____

Loise Muthoni Wahome Date

AS400/5688/2021

Supervisors' Approval

The undersigned certify that they have read and hereby recommended for acceptance of Murang'a University of Technology a thesis entitled **"Modeling the Prevalence of Depression among Murang'a University of Technology using Machine Learning Techniques".**

_____ _____

Dr. John Mutuguta, PhD Date

Department of Mathematics and Actuarial Science,

Murang'a University of Technology.

_____ _____

Dr. Euna Nyarige, PhD Date

Department of Mathematics and Statistics,

Machakos University.

DEDICATION

This thesis is dedicated to my family, I am forever indebted for your contribution to my academic and personal growth.

ACKNOWLEDGEMENT

I am immensely grateful to God for His blessings and guidance throughout this journey. My heartfelt appreciation goes to my supervisors, Dr. John Mutuguta and Dr. Euna Nyarige, for their invaluable support and expertise. Dr. Ayubu Anapapa, thank you for your assistance throughout the research. I extend my sincere thanks to the University Management for the sponsorship opportunity and permission to conduct the research at MUT. To the students at MUT who participated in this study, your contributions were indispensable, and I am grateful for your cooperation. I also want to express my heartfelt gratitude to my parents, my beloved husband Ephraim, my family, and my friends for their unending support and belief in my abilities. Your contributions have been instrumental in the completion of this thesis, and I am truly thankful for your presence in my life.

ABSTRACT

Around the world, depression is a prevalent mental illness and it affects the way people think, feel, talk and conduct their daily activities. The stigma associated with depression makes it to be misdiagnosed in the mild or the moderate stages and in its worst state, depression can lead to disability or suicide. Numerous investigations have explored the frequency of depression using different machine learning techniques with results varying across study groups. However, machine learning models have not been utilized to model the prevalence of depression among Kenyan university students. Therefore, this study employed machine learning techniques to model the prevalence of depression among Murang'a University of Technology (MUT) students. The specific objectives were to model the prevalence of depression among MUT students using logistic regression, random forest classifier and support vector machine algorithms, to assess the efficacy of these algorithms in accurately modeling the prevalence of depression among MUT students and determine the best machine learning technique for modeling. A sample of 1448 students from a population of $10{,}127$ students participated in the study by completing questionnaires on sociodemographic and other factors associated with depression. The questionnaires were administered via online platforms. Participants were selected using proportionate stratified random sampling ensuring that a representative sample was chosen from each school. The data gathered was examined using descriptive and inferential statistics. Depression was measured using the Patient Health Questionnaire scale (PHQ-9). Using a cut-off point of 10; 25.97% students had depressive symptoms. This comprised of 19.61% moderate symptoms and 6.35% severe symptoms. Depressive symptoms were significantly more common among male students ($OR = 1.86, p-value = 4.48 \times 10^{-5}, 95\%\ CI: 0.32, 0.92$), students in their third ($OR = 1.84, p-value = 0.00734, 95\%\ CI: 0.17, 1.06$) and fourth ($OR = 2.17, p-value = 0.00407, 95\%\ CI: 0.25, 1.31$) years of study. The variables that were significantly related to depression were, gender ($OR = 0.60, p-value = 0.00363\ , 95\%\ CI: -0.85, -0.17$), social support network ($OR = 0.40, p-value = 4.78 \times 10^{-10}, 95\%\ CI: -1.19, -0.62$), financial situation ($OR = 0.71, p-value = 0.0016\ , 95\%\ CI: -0.55, -0.13$) and past abuse, trauma and neglect ($OR = 2.18, p-value < 2 \times 10^{-16}, 95\%\ CI: 0.60, 0.96$). The confusion matrix criteria were used to select the best machine learning algorithm in modeling depression prevalence among MUT students. The random forest algorithm demonstrated superior performance compared to support vector machine and logistic regression, with accuracy at 98.68%, sensitivity at 95%, specificity at 100%, a positive predictive value of 100%, and a negative predictive value of 98.24%. Following closely was the support vector machine and lastly logistic regression. In conclusion the machine learning models showcased remarkable efficiency in classifying depression instances based on a diverse set of features. The study recommends the utilization of RF algorithm to model the prevalence of depression among university students. Implementing targeted interventions founded on identified risk and protective factors and exploring the long-term outcomes of these interventions would contribute to the evolving field of mental health research within academic settings.

TABLE OF CONTENTS

DECLARATION ..... **ii**
DEDICATION ..... **iii**
ACKNOWLEDGEMENT ..... **iv**
ABSTRACT ..... **v**
LIST OF TABLES ..... **ix**
LIST OF FIGURES ..... **x**
ACRONYMS AND ABBREVIATIONS ..... **xi**
DEFINITION OF TERMS ..... **xii**
CHAPTER ONE: INTRODUTION **1**
1.1 Background of the Study 1
1.1.1 Depression 1
1.1.2 Logistic Regression 3
1.1.3 Random Forest 4
1.1.4 Support Vector Machine 4
1.2 Statement of the Problem 5
1.3 Objectives of the Study 6
1.3.1 General Objective 6
1.3.2 Specific Objectives 6
1.4 Research Questions 7
1.5 Hypotheses 7
1.6 Significance of the Study 8
1.7 Scope and Limitation of the Study 8
1.8 Contributions of the Thesis 9
1.9 Organization of the Thesis 9
CHAPTER TWO: LITERATURE REVIEW **10**
2.1 Introduction 10
2.2 Depression 10
2.3 Logistic Regression (LR) Model 12
2.4 Random Forest (RF) Classifier 14
2.5 Support Vector Machine (SVM) Model 15
2.6 Conceptual Framework 16
2.7 Summary 18

CHAPTER THREE: RESEARCH METHODOLOGY 19
3.1 Introduction 19
3.2 Research Design 19
3.3 Target Population 19
3.4 Sampling Techniques 20
3.5 Sample Size 20
3.6 Data Collection Procedure 21
3.6.1 Pilot Study 21
3.6.2 Validity of the Research Tool 22
3.6.3 Reliability of the Research Tool 22
3.7 Research Procedures 23
3.8 Data Analysis 23
3.8.1 Fitting the Logistic Regression Model 23
3.8.2 Fitting the Random Forest Classifier 26
3.8.3 Fitting the Support Vector Machine 27
3.9 Model Selection 29
3.10 Model Evaluation 30
3.11 Ethical Considerations 31
3.12 Summary 31
CHAPTER FOUR: RESULTS AND DISCUSSION 33
4.1 Introduction 33
4.2 Descriptive Analysis 33
4.2.1 Sociodemographic Characteristics 33
4.2.2 Prevalence of Depressive Symptoms 35
4.2.3 Predictor Variables 38
4.3 Logistic Regression (LR) Model 46
4.4 Random Forest (RF) Model 56
4.5 Support Vector Machine (SVM) Model 59
4.6 Selecting the Suitable Model 60
4.7 Hypotheses Testing 63
4.8 Summary 66
CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS 67
5.1 Conclusion 67
5.2 Recommendations 68

5.2.1 Recommendations for Policy 68
5.2.2 Recommendations for Future Work 68
REFERENCES 69
APPENDICES 75

LIST OF TABLES

Table 3.1: Sample Size for each School 21
Table 3.2: Confusion Matrix Table 29
Table 4.1: Sociodemographic Characteristics 34
Table 4.2: Relationship between Sociodemographic Variables and Depression 36
Table 4.3: Academic Performance 39
Table 4.4: Social Support Network 41
Table 4.5: Financial Situation 42
Table 4.6: Drug and Substance Abuse 44
Table 4.7: Past Abuse, Trauma and Neglect 45
Table 4.8: LR results for Categorical Predictor Variables 48
Table 4.9: LR results for Predictor Variables 52
Table 4.10: RF Algorithm Training Parameters 57
Table 4.11: SVM Algorithm Training Parameters 59
Table 4.12: Confusion Matrix for LR Model 60
Table 4.13: Confusion Matrix for RF Classifier 61
Table 4.14: Confusion Matrix for SVM Algorithm 61
Table 4.15: Metrics of the Confusion Matrix for the three ML Algorithms 62
Table 4.16: Hypothesis test for the LR model 63
Table 4.17: Hypothesis test for the RF model 64
Table 4.18: Hypothesis test for the SVM model 64
Table 4.19: Two-sample Z-test for RF and SVM 65
Table 4.20: Two-sample Z-test for RF and LR 65
Table 4.21: Two-sample Z-test for SVM and LR 66

LIST OF FIGURES

Figure 2.1: Conceptual Framework 17
Figure 3.1: Logistic Sigmoid Curve.....24
Figure 3.2: Structure of a Random Forest.....26
Figure 3.3: SVM Hyperplane.....28
Figure 4.1: Variable Importance for RF Model 58

ACRONYMS AND ABBREVIATIONS

LR	Logistic Regression
MUT	Murang'a University of Technology
PHQ-9	Patient Health Questionnaire
RF	Random Forest
RBF	Radial Basis Function
SVM	Support Vector Machine

DEFINITION OF TERMS

Logistic Regression – It is a supervised machine learning classification technique which is an extension of ordinary linear regression and models dichotomous outcomes; the probability of occurrence or non-occurrence of an event.

Machine Learning – It is an area of artificial intelligence that imitates human intelligence using data and algorithms. Machine learning algorithms utilize various statistical, probabilistic and optimization methods to understand previous experiences and predict significant patterns from large datasets.

Random Forest - It is a supervised machine learning method that builds several decision trees and incorporates them for greater prediction accuracy. It can be used for both classification and regression purposes.

Supervised machine learning – This type of machine learning employs labeled data to train models that categorize information or make predictions about the future.

Support Vector Machine – It is a supervised machine learning method that uses hyperplanes to classify both linear and non-linear data.

CHAPTER ONE
INTRODUCTION

1.1 Background of the Study

1.1.1 Depression

Depression is a prevalent and severe mental illness that impacts millions of individuals worldwide. It can manifest in various ways, including irritability, fatigue, disinterest in activities, feelings of hopelessness, guilt, sleep disturbances, and alterations in eating habits. Depression can range in severity from mild to severe and can significantly impact a person's daily functioning. Sadly, depression can also lead to suicidal ideation and behavior (WHO, 2023).

University students are not immune to depression and the prevalence rates of depression among this population vary across different regions. A recent literature review by (Li *et al.*, 2022) discovered that the global rate of occurrence of depression among university students was 33.6%, with higher rates observed in developing nations at 42.5% and African countries at 40.1%. Medical students also had higher rates of depression of 39.4%, likely due to the demanding nature of their coursework. A study by (Akhtar *et al.*, 2020) observed that 24.4% of undergraduate students belonging to countries with low and middle-income economies had depressive illnesses.

In Kenya, (Mbwayo *et al.*, 2022) analyzed the factors linked to depression among engineering students from University of Nairobi and Technical University of Kenya and found that the prevalence rate was 33%. Muriungi & Ndetei (2013) in assessing the efficiency of psycho-education on depression among students in para-medical fields studying at the Kenya Medical Training College reported the prevalence rates of depression as 20.6% for minimal cases, 12.6% for mild cases, 18.4% for moderate cases, and 48.5% for severe cases.

Factors that are linked to depression among university students are multifaceted and can include a lack of social support, financial stress, poor self-esteem and self-confidence, academic pressure, and substance abuse. For some students, the transition to university life can be challenging, as they navigate new social and academic

environments without the support of familiar networks (Adewuya *et al*., 2006). Additionally, students may face financial difficulties or make comparisons between their financial capacity and that of their peers.

Depression acts as a hinderance to sustainable development and prevents people from realizing their potential as productive members of society (Lund *et al.*, 2018). Knowing the depression prevalence rates among university students is crucial for developing effective measures to promote the well-being and academic success of this population.

People with depressive disorders typically experience a variety of symptoms that have an impact on their feelings, thoughts, and behaviors. WHO (2017) categorized depressive disorders either as major depressive disorder or dysthymia (persistent mild depression). There are two methods of measuring depression: using a standard diagnostic interview which is the clinical diagnosis of depression and using depressive symptoms scales. Patient Health Questionnaire – 9 (PHQ-9) and Beck Depression Inventory-II (BDI-II) are the two scales frequently employed in measuring and categorizing depressive symptoms. BDI-II contains 21 items for measuring and evaluating the extent of depression (Jackson-Koku, 2016). Kroenke *et al.* (2001) developed the PHQ-9 which has 9 items for diagnosing depressive disorders encountered by researchers. Titov *et al.* (2011) discovered that both PHQ-9 and BDI-II have good psychometric properties, but recommended the PHQ-9 due to its fewer attributes and free availability compared to the BDI-II which available on purchase.

The PHQ-9 is a versatile instrument for assessing, detecting, analyzing, and quantifying symptoms of depression (Kroenke *et al.*, 2001). PHQ -9 is self-administered and was initially designed for primary care but it has further been widely used by clinicians and researchers (Adewuya *et al.*, 2006). The PHQ -9 consists of 9 elements and the participants give an overview of how these elements have bothered them over the past two weeks. Each question receives a score between 0 and 3, with 0 signifying not at all, 1 signifying a few days, 2 signifying more than half the days, and 3 signifying virtually every day. A participant's score ranges from 0 to 27, with 27 being the highest (Spitzer *et al*., 2014). Kroenke *et al.* (2001) suggested that a cut-off point of 10 or greater, for the PHQ-9 score is effective for screening of depressive symptoms. Scores from 0 to 4 indicate minimal depression, 5 to 9 indicate mild

depression, 10 to 19 moderate depression, and 20 to 27 severe depression (Kroenke *et al.*, 2001).

Machine learning (ML) is an area of artificial intelligence that involves the use of algorithms to automatically identify patterns and generate insights from data (Mahesh, 2020). ML has found numerous applications across a wide range of domains, from business and finance to healthcare and engineering.

The goal of this research was to apply machine learning techniques to model the prevalence of depression among Murang'a University of Technology (MUT) students and assess the efficacy of logistic regression, random forest and support vector machine algorithms in identifying the factors that cause depression.

1.1.2 Logistic Regression

Logistic regression (LR) is a supervised machine learning technique used for classification purposes (Hosmer *et al.*, 2013). Logistic regression models are classified into three main categories: the binary logistic regression which has two alternative outcomes for the response variable being modelled, multinomial logistic regression which is employed when the response variable includes more than two unordered categories and ordinal logistic regression which is utilized when the response variable involves more than two ordered categories.

For LR to be a valid statistical model, it relies on several assumptions (Stoltzfus, 2011). They include; binary dependent variable, linearity of the independent variables and the log odds of the binary output, no multicollinearity among explanatory variables, independence of observations and a sample size large enough to draw accurate conclusions.

LR is a simple and more efficient statistical model for analyzing data with a binary outcome. It provides a clear understanding of the association between the predictor variables and the output, making it easy to interpret and communicate the results (Stoltzfus, 2011). LR has the capability to identify the most significant variables in a model which is useful when dealing with a large number of predictor variables.

1.1.3 Random Forest

Random Forest (RF) builds multiple decision trees and ensembles them to have a higher accuracy in prediction than each of the individual trees (Breiman, 2001). A decision tree employs tree-structures (Wakiru, 2018). The internal nodes represent the variables of the data, branches indicate the choice made, and leaf nodes represent the outcome. A decision tree just asks a question, and depending on the answer, it is divided into subtrees.

Random Forest is used in solving both classification or regression problems. In regression, the forest picks the mean of all outputs of the trees while in classification random forest chooses the majority of the votes (Mbaabu, 2020).

RF is a non-parametric model that can handle noisy data, missing values and outliers without requiring extensive data preprocessing. It reduces overfitting by combining multiple decision trees which also lead to a higher accuracy. RF also provides information about the relevance of each variable in a model which helps identify the factors contributing to the output.

1.1.4 Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm employed for classification and regression tasks (Mishra *et al*., 2021). SVM was developed by Vapnik and colleagues and it is among the most reliable prediction techniques (Cortes *et al.*, 1995). SVMs are categorized into linear and non-linear SVM. Linear SVM makes the assumption that the data is linearly separable and creates a hyperplane that can split the data into their respective classes. When the data cannot be split linearly, non-linear SVM is utilized, which converts the data into a space with a higher number of dimensions, allowing for the linear separation of data points using kernel functions. The transformed data can then be separated by a hyperplane.

SVMs can handle datasets with large number of explanatory variables using the kernel function that maps high-dimensional data where it can be linearly separable. Since SVMs aim to optimize the distance between the decision boundary and the nearest data points, they can handle noisy data, missing values and outliers without requiring

extensively data preprocessing.

1.2 Statement of the Problem

The World Health Organization reports that depression is a prominent mental health condition, with an estimated 5% of adults affected by the illness (WHO, 2023). People with depressive symptoms are often misdiagnosed due to lack of necessary resources, shortage of skilled medical personnel and the stigma associated with depression.

Depression should be considered just like any other disease and people should feel free to seek help, when need be, without fear of stigma and discrimination. Detection and treatment of depression should be taken into consideration before it leads to extreme cases such as suicide and mental disorder. Awareness should be made on how to identify depression and the risk factors associated with it made known to the general public.

Most young adults in the age of transition to the university get to deal with many changes within their environment which throw them off balance from what they are used to. They have to cope with a new environment where they are fully responsible for what happens around their lives. They get to make decisions for themselves; where to live, what to eat, manage finances among many other decisions that were previously not on them. While some go smoothly through the transition, for some it is an uphill task to adapt to these changes. This may lead to little or no concentration in their studies, leading to underperformance which further adds to stress. These factors greatly contribute to depression among university students.

Machine learning techniques have shown promise in various domains, including healthcare and mental health research. However, there is a lack of comprehensive studies that employ machine learning methods to model the prevalence of depression specifically in the context of university students. Moreover, there exist some limitations in some machine learning models. For example, Decision trees (DT) are susceptible to overfitting if the tree becomes overly complex, and are less effective in capturing complex patterns in the data which reduces their predictive accuracy. K-nearest neighbors (KNN) lack the training phase like other algorithms which creates the inability to capture complex relationships in the data. Neural networks, although

powerful, require substantial amounts of training data to avoid overfitting and training deep networks can be time-consuming. Such models could provide valuable insights into the underlying patterns, risk factors, and predictive factors associated with depression in this population.

Locally, logistic regression was the only machine learning method used to model the prevalence of depression among university students in Kenya. The limitation of using LR only is that it may overlook more complex relationships, interactions among variables and non-linear associations that could be important in understanding the prevalence of depression among university students. This study employed multiple machine learning methods including logistic regression, random forest and support vector machine algorithms to provide a more comprehensive understanding of depression prevalence among university students. RF and SVM handled the non-linear relationship among the variables, outliers and noisy data. Exploring multiple machine learning methods helped identify the key factors contributing to depression, reveal hidden relationships among variables, and enhance the understanding of the complex interactions between risk factors and the occurrence of depression.

1.3 Objectives of the Study

1.3.1 General Objective

The main objective of this study was to model the prevalence of depression among students of Murang'a University of Technology (MUT) using machine learning techniques.

1.3.2 Specific Objectives

i. To model the prevalence of depression among MUT students using logistic regression.

ii. To model the prevalence of depression among MUT students using random forest classifier.

iii. To model the prevalence of depression among MUT students using support vector machine algorithm.

iv. To assess the performance of logistic regression, random forest and support vector machine algorithms in accurately modeling the prevalence of depression among MUT students.

1.4 Research Questions

i. Is logistic regression an effective method in modeling the prevalence of depression among MUT students?

ii. Is the random forest classifier an effective method in modeling the prevalence of depression among MUT students?

iii. Is the support vector machine algorithm an effective method in modeling the prevalence of depression among MUT students?

iv. How do LR, RF and SVM algorithms compare in their ability to accurately model the prevalence of depression among students of MUT?

1.5 Hypotheses

i. **Ho:** Logistic regression is not an effective method in modeling the prevalence of depression among MUT students.

H_1: Logistic regression is an effective method in modeling the prevalence of depression among MUT students.

ii. **Ho:** The random forest classifier is not an effective method in modeling the prevalence of depression among MUT students.

H_1: The random forest classifier is an effective method in modeling the prevalence of depression among MUT students.

iii. **Ho:** The support vector machine algorithm is not an effective method in modeling the prevalence of depression among MUT students.

H_1: The support vector machine algorithm is an effective method in modeling the prevalence of depression among MUT students.

iv. **Ho:** The three machine learning techniques do not differ significantly in their ability to accurately model the prevalence of depression among MUT students.

H_1: The three machine learning techniques differ significantly in their ability to accurately model the prevalence of depression among MUT students, and one method is more suitable than the others.

1.6 Significance of the Study

Depression is a widespread mental health issue among university students which can substantially affect their academic performance, social relationships, and overall well-being. Utilizing machine learning techniques offered a more accurate and effective strategy in predicting and managing depression. The study provided insights into the factors that lead to depression, unveiled hidden relationships between variables and enabled the development of more effective interventions to prevent and treat depression. Moreover, the study raised awareness about the importance of addressing depression and mental health among young people, and encouraged more individuals and organizations to take action in this area. The research added to the current knowledge on depression and mental wellness among university students and the factors that contribute to its occurrence.

1.7 Scope and Limitation of the Study

The study included only the students who were enrolled at MUT excluding the students who had been discontinued and those who had already graduated at the time the research study was conducted. The study obtained data on depression and the factors related to depression such as academic performance, social support network, financial situation, drug and substance abuse and past abuse, trauma and neglect from MUT students. The limitation of this study was the potential for bias in self-reported data since the collected data only relied on the willingness of participants to disclose sensitive information. Ensuring anonymity of participants and confidentiality of the data led sensitive questions being answered with honesty thereby minimizing this limitation. Furthermore, a validated research tool was utilized so that the collected data could be used to draw valid conclusions.

1.8 Contributions of the Thesis

The following are the contributions of this study;

i. The study identified that the RF classifier had the highest performance in accurately modeling the prevalence of depression among MUT students.

ii. The research added to the current knowledge on utilization of machine learning techniques to accurately model the prevalence of depression among university students.

iii. The study recommended implementation of targeted interventions aimed at addressing identified risk factors to enhance the mental well-being of university students.

1.9 Organization of the Thesis

This thesis is organized into five chapters, which are structured as follows;

Chapter one of the thesis introduces the study's background, followed by the statement of the problem, research objectives, questions, hypotheses, significance, justification, scope and limitation and thesis contributions.

Chapter two includes an empirical literature review on depression, LR, RF and SVM models, a summary of literature review findings and a conceptual framework.

Chapter three covers the research methodology which includes the research design, target population, sampling techniques, sample size, data collection procedures, research procedure, data analysis methods, model selection and evaluation and the ethical considerations.

Chapter four presents the results and discussion for the three models, along with the process of model selection, hypotheses testing, and a summary of findings.

Chapter five entails the study’s conclusions and recommendations.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

Numerous investigations have explored depression with the aim of modeling its prevalence among the study population. This chapter addresses the empirical literature review of depression, logistic regression, random forest and support vector machine, the conceptual framework and the last section the summary of the literature review.

2.2 Depression

Haroz *et al.* (2017) peer reviewed academic journals published between August 2012 and May 2015 that contained qualitative literature related to depression to evaluate how depression is felt globally. A total of 16,130 records were found, and of those, 138 had the characteristics the study needed. The 138 records contained information from 77 different nationalities and 170 different target demographics, some of which had multiple samples. The researchers identified the risk factors of depression including drastic life events such as bereavement, financial struggles and childhood neglect, trauma and abuse.

Ibrahim *et al.* (2013) reviewed studies published on the occurrence of depression between January, 1990 and October, 2010 among students in universities. The 24 articles that complied with the guidelines for inclusion and exclusion indicate that almost a third of university students are affected by depression and it is a significant health concern among this population. Additionally, it was found out that 30.6% of students report having depression symptoms on average, which is much higher compared to the 9% prevalence rates reported in general populations. From 16 studies, female students had a higher risk of having depression with a prevalence rate of 29.6% while that of male students was 24.9%. If depression in the university environment is not recognized and managed with more care, susceptibility will grow due to the current economic constraints. Most of the studies were from the Western nations and only 3 studies were from the developing countries which suggests shortage of studies in these regions. Individuals living in these areas are highly susceptible to depression because of financial constraints and insufficient healthcare. Studies conducted in developing

countries are few, and there should be more studies to assess mental health as well as its determinants (Li *et al.*, 2022).

Mbwayo *et al.* (2022) established the factors associated with depression among university students in Nairobi County. The study population consisted of engineering students from University of Nairobi (UoN) and Technical University of Kenya (TUK). A sample of 136 engineering students from UoN formed the experimental group while 137 engineering students from TUK formed the control group. Regression analysis and independent t-tests were employed to examine the relationship between depression and the selected factors among the students. Having dependents, holding a leadership position, and the student's gender were positively correlated with depression. Expenditure, friendships and club membership showed a negative correlation with depression. Using the Beck's Depression Inventory with a cut-off point of 14, the prevalence of depression among engineering students from UoN and TUK was 33%.

Muriungi & Ndetei (2013) examined the efficiency of psycho-education on mental health issues among Kenya Medical Training College (KMTC) students. The mental health issues included, depression, hopelessness, suicidality, anxiety and risk of substance abuse. The study population included first and second-year basic diploma students from the seven campuses of KMTC. The Nairobi campus comprised the experimental group, while the campuses in Nakuru, Port Reitz, Mombasa, Kisumu, Murang'a, and Meru formed the control group. A sample of 1181 students were selected for the experimental group, while 1926 students were assigned to the control group. Data was collected using six self-administered questionnaires; researcher-designed questionnaire for socio-demographic characteristics (SDQ), the Beck's Depression Inventory (BDI), the Beck's Hopelessness Scale (BHS), the Beck's Suicide Ideation Scale (BSIS), the Beck's Anxiety Inventory (BAI), the World Health Organization Alcohol, Smoking and Substance Involvement Screening Test (ASSIST). Baseline characteristics were obtained in the experimental group, followed by 4 two-hour psycho-education sessions covering causes and coping strategies for depression, hopelessness, suicidality, anxiety, and substance abuse. Three months after the initial intervention, reassessment was conducted to evaluate its impact, followed by an additional eight hours of psycho-education delivered in two-hour

sessions. A final reassessment was conducted six months after the initial assessment and the participants were encouraged to contact the investigator for any psychological issues. The control group received no interventions aside from access to healthcare. However, reassessments occurred at 3 and 6 months, with participants given access to a 24-hour helpline for support, including suicidal tendencies. The data was analyzing using Statistical Package for the Social Sciences (SPSS), utilizing both descriptive and inferential statistics. During the first assessment, there were similar prevalence rates of depression between the control and experimental groups: minimal 20.6% vs. 20.8%; mild 12.6% vs. 13.1%; moderate 18.4% vs. 20.2%; and severe 48.5% vs. 45.8%, respectively. At assessment 2, there was no difference in the prevalence of depression among the two groups. During assessment 3, the experimental group showed lower prevalence rates of moderate and severe depression compared to the control group: moderate 5.2% vs. 10.4%; and severe 3.4% vs. 5.0%, respectively. The decrease in depression severity after psycho-education was significant after 6 months. The high levels of depression during the first assessment were due to the challenges the students faced as they transitioned to the college environment. Psycho-education effectively reduced the prevalence of the mental health issues in the experimental group, while the reduction in the control group resulted from the reassessment, prompting the students to seek medical health care.

2.3 Logistic Regression (LR) Model

Othieno *et al.* (2014) assessed the severity of depressive illnesses and described the sociodemographic determinants of depression among Kenyan university students. The research targeted University of Nairobi students and a sample data of 923 students was obtained. In using the Center for Epidemiological Studies - Depression (CES – D) scale to assess depression, the study showed that 41.33% students were more likely to be depressed, with 35.71% exhibiting mild to moderate symptoms and 5.62% experiencing severe depression. LR revealed that students who smoke, drank heavily, and who were older were more likely to suffer from depression. Depression was equally likely to prevail to both male and female students. The study showed that depression is experienced by a significant number of students and is more common among students who come from vulnerable families economically, those residing off campus, the married ones and the first years.

According to a study on depression amongst Nigerian university students, (Adewuya *et al.*, 2006) evaluated the prevalence of depression and investigated its sociodemographic correlates. The researchers examined a sample of 1206 students from south-western Nigeria residing in the halls. Questionnaires were administered to these students and depression was measured using the Mini International Neuropsychiatric Interview (MINI). 8.3% of the students displayed symptoms of depression, 5.6% of these students had mild depression and 2.7% had severe depression. 11.9% of the students who had depressive illness were females and 5.8% were males. Odds ratio (OR) and the 95% confidence interval (CI) were calculated for the factors that were significantly linked to depression: accommodation problems (OR 2.72, 95% CI: 1.79, –4.16), very large family size (OR 2.8, 95% CI: 1.42, –5.73), females (OR 2.21, 95% CI: 1.46, –3.35), excessive smoking (OR 3.67, 95% CI: 2.23, –6.05) and excessive consumption of alcohol (OR 9.44, 95% CI: 3.32, –26.89). Further research was required to create and asses an effective model to forecast the evolution of depression and the measures that can effectively decrease the frequency of depression in this group.

Kuringe *et al.* (2019) assessed the prevalence of anxiety and depression and the factors linked to them among adolescent girls and young women (AGYW) in Tanzania who were not attending school. The study involved the participation of 3014 AGYW between the ages of 15 and 23, who completed an interview and a programmed questionnaire on a tablet. After that, their blood was taken for pre- and post-HIV testing as well as counseling. The researchers utilized the 4-item Patient Health Questionnaire (PHQ-4) as a tool for evaluating depression and anxiety. It comprises of 2 items for screening depression (PHQ-2) and the other 2 for detecting generalized anxiety disorders (GAD-2). To assess how each predictor variable and the response variable related, binary and ordinal logistic regression models were used. Adjusted Odds Ratio (AOR) with 95% CI were calculated for predictor variables that were significantly associated with the response variable. A positive screening for depression was achieved by receiving a score of 3 or more for each of the two primary symptoms, $(PHQ-2 \geq 3)$. HIV infection was detected in 3.6% of study participants. Participants exhibited mild, moderate, and severe cases of anxiety and depression in 32.5% $(30.8\% - 34.2\%)$, 19.7% $(18.3\% - 21.1\%)$, and 6.4% $(5.5\% - 7.2\%)$ of cases, respectively. Having dependents $(AOR =$

$2.51, 95\%\ CI: 1.47, -4.29$) and abuse from sexual partners ($AOR = 1.90, 95\%\ CI: 1.59, -2.27$) were significantly correlated with depressive disorders. Financial security ($AOR = 0.81, 95\%\ CI: 0.70, -0.95$) and moral support ($AOR = 0.82, 95\%\ CI: 0.67, -0.99$) kept the AGYW clear from anxiety and depressive illnesses. Symptoms of anxiety and depression were rampant among the AGYW and a third of them were affected. In order for the AGYW to reach their maximum potential, it was recommended that mental healthcare services should be well developed for testing, diagnosis and management of depressive and anxiety symptoms among this group.

2.4 Random Forest (RF) Classifier

Rois *et al.* (2021) employed machine learning methods to analyze the prevalence and predict features of perceived stress among university students in Bangladeshi. A sample of 355 students from 28 universities in Bangladeshi participated in the research. The researchers used RF, SVM, DT and LR machine learning models to classify and forecast mental stress and their effectiveness was assessed using the confusion matrix. RF had the highest performance with an accuracy of 89.72% and LR had the lowest performance with an accuracy of 74.76% and SVM had an accuracy of 75.7%. In the previous year, stress was recorded by a third of the university students. The significant factors for predicting the levels of stress were pulse rates, blood pressures, sleeping habits, smoking, and education background.

Qasrawi *et al.* (2022) evaluated the efficiency of machine learning methods in forecasting the risk factors affiliated with depression and anxiety among Palestinian school children. RF, ANN, DT, SVM, and Naïve Bayes (NB) algorithms were used for prediction. A sample of 5685 students aged 10-15 years was used in the study. Data on depression was gathered using the 18-item depression self-reported scale (DSRS) while the GAD-7 scale was employed to assess anxiety. The effectiveness of the machine learning models was assessed using the performance metrics of the confusion matrix. 66.67% of the students had moderate depressive symptoms and 7% were severely depressed, while 22% had moderate and severe anxiety symptoms. Males had lower depressive and anxiety levels compared to females. RF had the highest performance with an accuracy of 72.6% and 68.5% in predicting depressive

and anxiety symptoms respectively. Academic achievement, household conflict, economic status, and harassment at school were the factors that were related to depression and anxiety. The study highlighted the significance of utilizing machine learning to respond to challenges related to mental health. Furthermore, early identification and forecasting of factors linked to mental illness could aid in the creation of preventative initiatives that benefit children's physical and mental growth. There was need to incorporate more related factors in order to thoroughly examine the pertinent factors that are linked to anxiety and depression.

Haque *et al.* (2021) employed DT, RF, Gaussian Naïve Bayes (GNB) and XGBoost (XGB) models in detecting depression among children and adolescents. Data from the second Australian Child and Adolescent Survey of Mental Health and Wellbeing 2013–14 was utilized. 70% of the data points were used to train the models while 30% were used for testing the models. The confusion matrix metrics were used to assess the performance of the ML models. RF had the highest performance compared to the other models (accuracy = 95%, precision = 99%, specificity = 100% and sensitivity = 44%). The significant factors crucial for recognizing depression among children and adolescents included feelings of unhappiness, loss of interest, changes in weight, sleep disturbances, restlessness, fatigue, difficulty in concentration and thoughts of suicide.

2.5 Support Vector Machine (SVM) Model

In identifying the people who were mentally affected by the COVID-19 pandemic, Joshi *et al.* (2022) applied machine learning models to depict the impact of the pandemic on depression. The researchers used LR, NB, RF, k-nearest neighbor (k-NN) and SVM. A total of 198 replies were gathered from people in the age group 18-35 years and 70% was used to train the models while 30% was used to test the models. A 7- fold cross validation was used to measure the efficiency of these models. Each feature being studied was given a score between 1 and 5, with 5 indicating most depressed and 1 least depressed. The mean score of each participant was used to classify whether one was depressed or not. A score of more than 3 was used to categorize an individual as depressed while a score of 3 or less indicated that an individual was not depressed. SVM performed better than the other classification algorithms (accuracy = 99.351, precision = 0.991, f1-score = 0.995, recall = 1.0, cross-validation score = 0.988, error = 0.008). Although people have been suffering from

depression for a very long time, it was concluded that COVID-19 worsened people's mental wellness.

Ahuja & Banga (2019) used k-NN, NB, RF and SVM algorithms to identify mental stress among university students. The dataset was obtained from 206 students of Jaypee Institute of Information Technology Noida. The information was divided into two categories; stress before the exams and stress brought about by internet use. Perceived stress scale (PSS) was used to examine a person's stress levels in various settings and the individual's coping skills in those situations. Sensitivity, specificity and accuracy were used to measure the efficiency of the models. SVM had the highest performance with (sensitivity = 75%, specificity = 100% and accuracy = 85.71%), followed by RF with (sensitivity = 100%, specificity = 66.66% and accuracy = 83.33%), NB (sensitivity = 66.66%, specificity = 75% and accuracy = 71.42%) and k-NN (sensitivity = 70%, specificity = 44% and accuracy = 55.55%). The researchers concluded that individuals' mental health could be improved and the population could become mentally sound, by analyzing data and discovering tools like PSS that produce more precise information at a lower cost.

Mutalib (2021) employed DT, ANN, NB, LR and SVM in predicting stress, depression and anxiety among students in institutions of higher learning in Malaysia. A total of 629 students from a learning institute in Kuala Terengganu participated in the study. Data on stress, depression and anxiety was collected using the 21- item Depression Anxiety Stress Scale (DASS-21). The performance of the algorithms was evaluated for each mental health illness using the confusion matrix metrics. In predicting stress, DT had the best performance (accuracy = 84.44%, sensitivity = 54.84%, specificity = 93.27% and precision = 70.85%). SVM outperformed the other models in predicting depression (accuracy = 88.15%, sensitivity = 64.52%, specificity = 95.19% and precision = 80%). In predicting anxiety, ANN had the highest performance (accuracy = 68.89%, sensitivity = 60%, specificity = 75% and precision = 62.26%). Academic environment, financial constraints and deficiency in social support were significant to mental health.

2.6 Conceptual Framework

Figure 2.1 was the conceptual framework which depicted the hypothesized

relationship between the independent variables and the dependent variable. Depression among university students can be caused by various factors such as academic performance, social support network, financial situation, drug and substance abuse and past abuse, trauma and neglect. Poor academic performance may contribute to development of depression. Students experiencing financial difficulties may be more susceptible to experiencing depression. The presence of a strong social support network plays a protective role and help buffer the effects of depression. Drug and substance abuse as a coping mechanism can also contribute to the development or worsening of depressive symptoms. Additionally, a history of abuse, trauma, and neglect can increase the risk of developing depression.

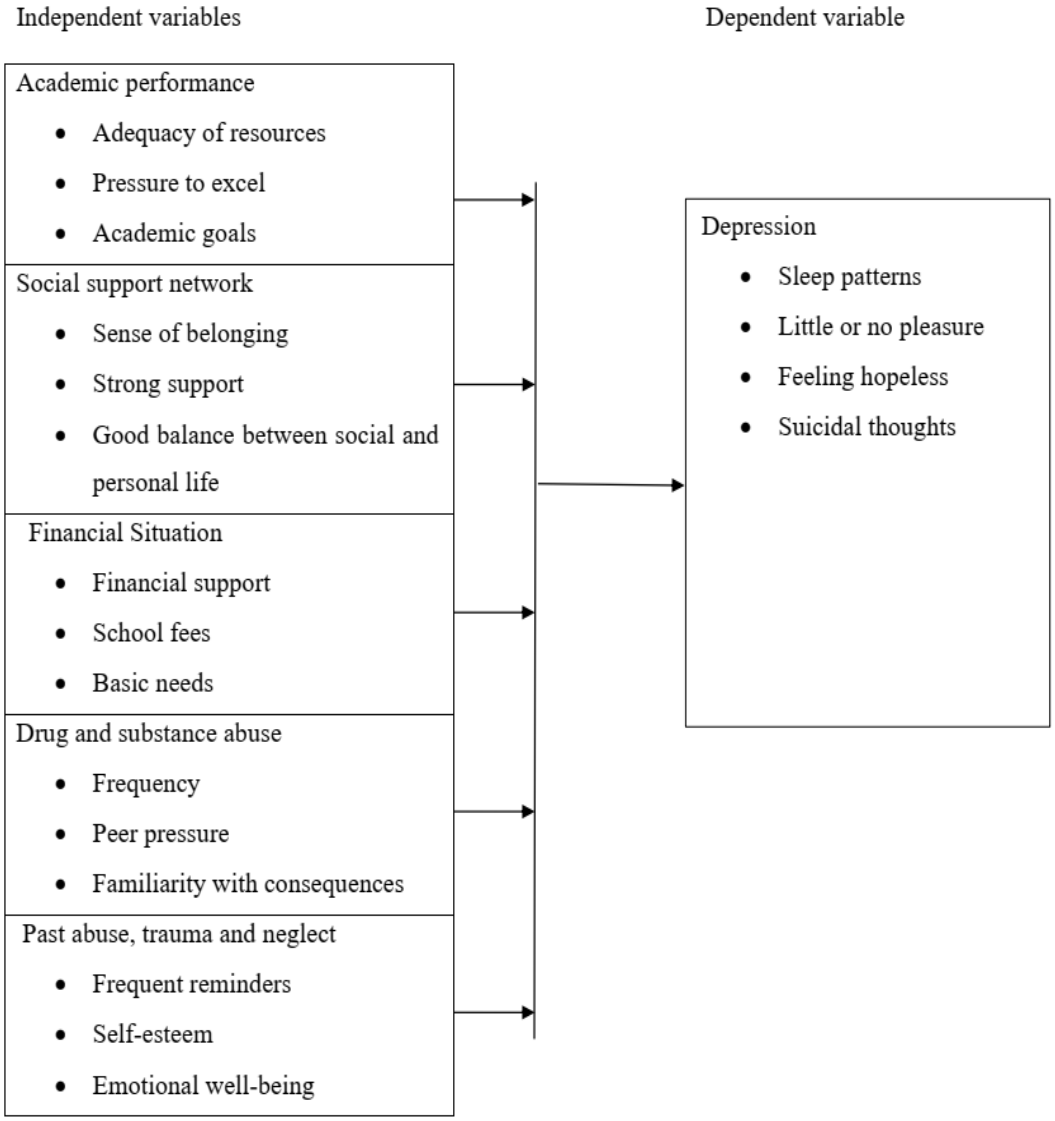


Figure 2.1: Conceptual Framework

2.7 Summary

From the literature review conducted, traditional statistical methods such as regression analysis, independent t-tests, and descriptive and inferential statistics were used to model the prevalence of depression among Kenyan university students. The drawback of these traditional methods is that they rely on assumptions of normality and linearity, which, if violated, can lead to inaccurate results. Moreover, these methods cannot handle large datasets and non-linear relationships among variables. Machine learning methods offer significant advantages over traditional statistical techniques, as they do not rely on assumptions of normality and linearity, can process large datasets, uncover complex relationships, and automatically detect patterns, thereby improving prediction accuracy and providing deeper insights into the factors contributing to depression.

Additionally, further research was necessary to address the occurrence of depression in developing nations since individuals residing in these countries faced a higher susceptibility to depression due to financial challenges and limited access to quality healthcare.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

This chapter entailed the research design, population being studied, sampling technique to be utilized, sample size used, data collection procedure and equipment and data analysis methods. All these components were designed to attain the research study's objectives.

3.2 Research Design

A research design is a well guided plan that aims at ensuring that the research questions are effectively addressed (Lai, 2018). The research design serves as the framework for how data will be collected, measured, and analyzed.

Research design is broadly grouped into quantitative research design which measures variables and tests their relationship and qualitative research design which seeks to gain past information about a phenomenon and generate new ideas. Quantitative research design is further grouped into: experimental, quasi-experimental, descriptive and correlational research. This research employed descriptive research design which was used to accurately and thoroughly depict individuals, situations and identify patterns within a dataset.

3.3 Target Population

Target population is a group of people who satisfy the required set of specifications (Odhiambo *et al.*, 2018). The population for this study comprised of 10, 127 students from the various schools and departments at Murang'a University of Technology (MUT). These schools re; School of Agriculture and Environmental Sciences (SAES), School of Business and Economics (SBE), School of Computing and Information Technology (SCIT), School of Education, Humanities and Social Sciences (SEHSS), School of Education and Technology (SET), School of Hospitality and Tourism Management (SHTM) and School of Pure, Applied and Health Sciences (SPAHS). Understanding this composition was crucial for gaining insights into the prevalence of depression and how it may vary across different groups within the university.

3.4 Sampling Techniques

Sampling is the selection of a sample from the population under study. Since all the 7 schools at MUT participated in the study, proportionate stratified random sampling was utilized to ensure that all the schools were well represented in the same proportion as the population to achieve greater precision. Simple random sampling was applied within each school to ensure that a representative sample from each school was chosen.

The proportionate stratified random sample was obtained using this formula;

$$\frac{\text{Sample Size}}{\text{Population Size}} \times \text{Stratum Size} \tag{3.1}$$

3.5 Sample Size

Sample size refers to the number of individuals included in the study. In this study, the sample size was determined using the Yamanes formula (Yamane, 1967). The Yamanes formula was used by (Rahman *et al.*, 2020) in evaluating the relationship between socio-demographic characteristics and job satisfaction.

$$n = \frac{N}{1 + N(e)^2} \tag{3.2}$$

where n is the sample size, N is the population size and e the margin of error.

$N = 10127$ and $e = 0.025$. Using these values, $n = 1382$

Table 3.1: Sample Size for each School

School	Sample Size	Percent (%)
School of Agriculture and Environmental Sciences (SAES)	35	2.5
School of Business and Economics (SBE)	215	15.5
School of Computing and Information Technology (SCIT)	210	15.2
School of Education, Humanities and Social Sciences (SEHSS)	513	37.1
School of Engineering and Technology (SET)	152	11.0
School of Hospitality and Tourism Management (SHTM)	93	6.7
School of Pure, Applied and Health Sciences (SPAHS)	166	12.0
Total	**1384**	**100**

3.6 Data Collection Procedure

Data collection involves acquiring and assessing data on the relevant variables. This allowed the researcher to evaluate results, test hypothesis and answer the stated research questions. Self-administered questionnaires issued to MUT students via online platforms were the source of data for this study. A questionnaire is a research tool that comprises of questions that gather information from the participants (Chebet & Njuguna, 2020). Closed-ended questions such as Likert scale questions were used to examine the severity of depression among MUT students. Questionnaires were the most efficient means to collect data in the large sample of MUT students quickly. Anonymity of the respondents led to sensitive questions being answered with honesty ensuring accuracy and reliability of the data.

3.6.1 Pilot Study

Arain *et al.* (2010) defined a pilot test as a small study that is intended to test the

methodology and procedures of a larger study. A pilot study is conducted to assess the quality and productivity of the aspects of the research prior to the full-scale study (In, 2017). A sample size of 10% of the final size of the sample is a suitable number for a pilot study (Odhiambo *et al*., 2018). The pilot study involved 139 individuals which was 10% of 1384. The results obtained were used to assess the feasibility of the research tool. Conducting the pilot study was essential for identifying and rectifying ambiguous questions in the research tool, ensuring clarity and enhancing the overall quality and reliability of collected data.

3.6.2 Validity of the Research Tool

Watson (2015) defines validity as the level to which a research tool measures the intended concept accurately. The researcher evaluates whether the inferences and conclusions arrived at based on the questionnaire results are valid (Tsang *et al*., 2017). Validity involves two crucial components, internal validity and external validity. Internal validity relates the degree to which the outcomes of the study are accurate given the method parameters are selected, data is collected, or analysis is carried out while external validity evaluates the applicability of the study results to other study populations (Mohajan, 2017). Face and content validity were used to assess the quality and representativeness of the items for measuring depression. Face validity of the questionnaire was assessed by reviewing the items for clarity and appropriateness. The content validity of the questionnaire was established under the guidance and expertise of my supervisors.

3.6.3 Reliability of the Research Tool

According to (Odhiambo *et al*., 2018) reliability is the extent to which a research tool generates uniform outcomes over a period of time. Internal reliability evaluates how consistently results are obtained across test items measuring the same construct. External reliability refers to how consistent a measurement is when it is repeated by different individuals at different times. The questionnaire's reliability was evaluated using internal consistency reliability, which was gauged using the Cronbach's alpha coefficient. A value of 0.7 or more indicates high reliability of the questionnaire (Mohajan, 2017). A Cronbach's alpha coefficient of 0.89 was achieved indicating a high degree of consistency among the items in the questionnaire.

3.7 Research Procedures

Research procedures are methods used by the researcher to collect the data required for analysis. Since this research utilized primary data, data was gathered by issuing questionnaires to the students. The questionnaire contained closed-ended questions including the validated PHQ-9 and the responses to these questions were analyzed using random forest, logistic regression and support vector machine algorithms to model depression prevalence among MUT students.

3.8 Data Analysis

Data analysis was carried out with regard to the research objectives and hypothesis. This involved fitting a logistic regression model, fitting a random forest classifier, fitting a support vector machine model and choosing the best machine learning technique in modeling the occurrence of depression among MUT students.

Before processing of data, the raw data went through the data preparation process which comprises of data validation and data cleaning to identify and rectify errors (Hameed & Naumann, 2020). After being cleaned, the data was coded for analysis in the R Studio software.

To examine the gathered data, descriptive and inferential statistics were employed. Descriptive statistics gave a summary of the data such as mean and percentage. Inferential statistics was used to draw conclusions to the population from the sample under study.

3.8.1 Fitting the Logistic Regression Model

Fitting a logistic regression model involves estimating the parameters of the model that best describe how explanatory variables relate to the dependent binary variable. This is done using maximum likelihood estimation.

The multiple logistic regression with p parameters is defined as;

$$\pi_i = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)} = \frac{exp(\boldsymbol{X'\beta})}{1 + exp(\boldsymbol{X'\beta})} \qquad (3.3)$$

Where;

π_i is the probability of a binary outcome,

β_0 is the intercept,

$\beta_i, i = 1,2,\ldots,p$ are the regression parameters,

$x_i, i = 1,2,\ldots,p$ are the independent variables,

$\boldsymbol{X}$ is the matrix of independent variables,

$\boldsymbol{\beta}$ is the vector of the regression parameters.

The sigmoid function is used in logistic regression to model the data. It converts the projected values to probabilities. The three most common types of sigmoid functions are: logistic sigmoid function, hyperbolic tangent function and arctangent function. The logistic sigmoid function takes any real values as inputs and gives an output of values between 0 and 1. The hyperbolic tangent function maps any input real values to the range between -1 and 1. The arctangent function is the inverse of the tangent function and transforms any real-valued input to the interval between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$.

The logistic sigmoid curve is as shown below;

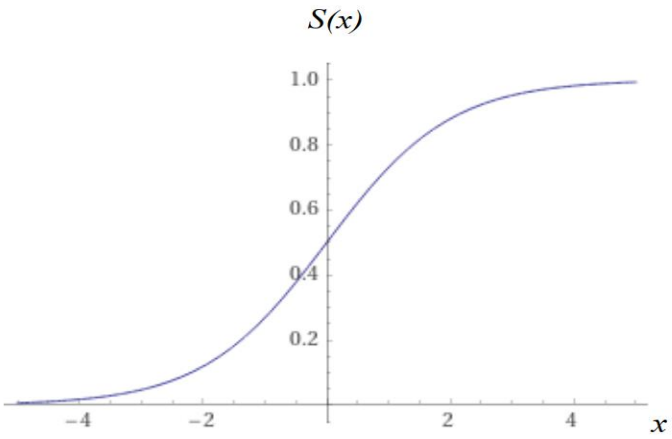


Figure 3.1: Logistic Sigmoid Curve

Source: (Hamdan & Roach, 2022)

where x represents the input variables and $S(x)$ the output variable.

The mathematical formula of the logistic sigmoid function is;

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{e^{x} + 1} \tag{3.4}$$

The response variable Y (depression) took the values 0 and 1.

Odds ratio was used to evaluate the link between the dependent variable and independent variables. Odds ratio is given by;

$$\frac{\pi}{1 - \pi} = exp\,(\beta_0 + \sum_{i=1}^{p} \beta_i x_i) \tag{3.5}$$

Where π is the probability of the event taking place.

3.8.1.1 Parameters Estimation for LR Model

The goal of maximum likelihood estimation is to find the set of parameter values that best explain the link between the predictor variables and the dependent binary outcome, when given a set of data.

Consider the logistic regression function equation (3.3);

The likelihood function is given by;

$$L(\beta; y, X) = \prod_{i=1}^{p} \left[\frac{exp(\boldsymbol{X'\beta})}{1 + exp(\boldsymbol{X'\beta})}\right]^{y_i} \left[1 - \left[\frac{exp(\boldsymbol{X'\beta})}{1 + exp(\boldsymbol{X'\beta})}\right]\right]^{1-y_i} \tag{3.6}$$

The log-likelihood function is given by;

$$\ln\big(L(\beta; y, X)\big) = \sum_{i=1}^{p} \left[y_i ln \left(\frac{exp(\boldsymbol{X'\beta})}{1 + exp(\boldsymbol{X'\beta})} \right) + (1 - y_i) ln \left(1 - \frac{exp(\boldsymbol{X'\beta})}{1 + exp(\boldsymbol{X'\beta})} \right) \right] \tag{3.7}$$

To determine the maximum likelihood estimates, the log-likelihood function is maximized by taking the derivatives with respect to the parameters and setting those derivatives equal to zero.

3.8.2 Fitting the Random Forest Classifier

RF classifier combines multiple decision trees on different subgroups of the input data and takes the majority vote to enhance the accuracy of the data set. The trees were trained using the bagging method which ensured that each decision tree contains slightly different data and was trained slightly differently (Sayed *et al.*, 2023).

The input data was obtained from the following variables; academic performance, social support network, financial situation, drug and substance abuse and past abuse, trauma and neglect. The features to train at each node in the decision tree were random. Each decision tree had an output and random forest took the majority vote from the decision trees to be the outcome.

Figure 3.2 is the structure of the random forest.

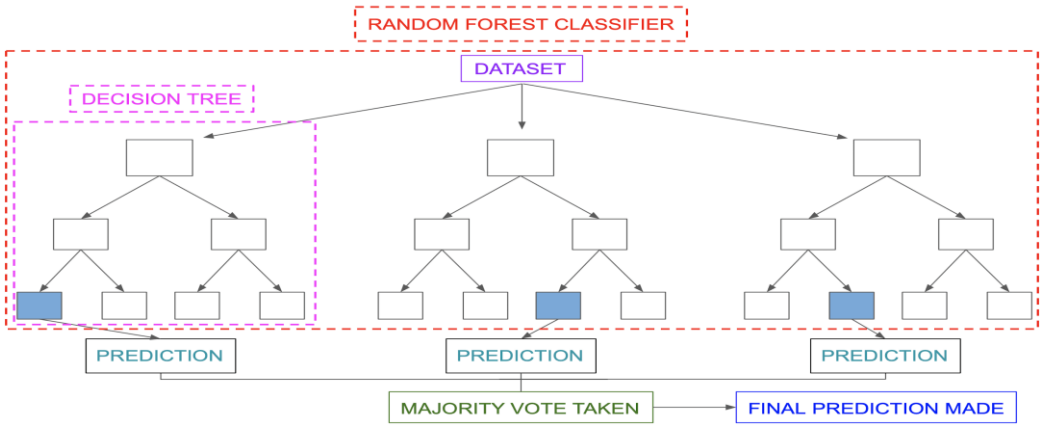


Figure 3.2: Structure of a Random Forest

Source: (Mbaabu, 2020)

The mean decrease Gini method was applied to assess the usefulness of each variable in the model. It represents the average decrease in Gini impurity for a particular variable across all trees in the forest (Han *et al*., 2016). The larger decline in mean decrease Gini indicates the more essential the explanatory variable is in making accurate predictions.

3.8.2.1 Parameters Estimation for RF Classifier

In RF, parameter estimation is crucial for optimizing model performance. This involved adjusting key hyperparameters such as mtry and ntree, while evaluating the model's effectiveness using the Out-Of-Bag (OOB) error rate.

The hyperparameter mtry controlled the number of predictor variables randomly selected for splitting at each node when constructing an individual decision tree. The default value for mtry is the square root of the number of independent variables (Rigatti, 2017).

The parameter ntree refers to the number of decision trees created in the RF model. Increasing the number of trees improved the performance of the RF model up to a certain point. Hence, it was the responsibility of the researcher to determine the number of trees to be used in the RF model (Probst & Boulesteix, 2018) in order to provide the best balance between accuracy and efficiency.

The OOB error rate is a measure of the performance of the RF model on unseen data based on the samples that were not included in the training set for each tree. The OOB samples served as a validation set for the RF algorithm. In the training process, additional trees were incorporated until the OOB error rate reaches a point of stability (Dudek, 2015).

3.8.3 Fitting the Support Vector Machine

SVM algorithm involves the representation of individual data items as points in a space that has n dimensions, where n is a natural number. Classification on the data is then performed by creating a hyperplane that accurately distinguishes the two classes. SVM creates the hyperplane which is determined based on the support vectors. Support vectors are the data points that are critical in determining the position of the

decision boundary (Fletcher, 2009).

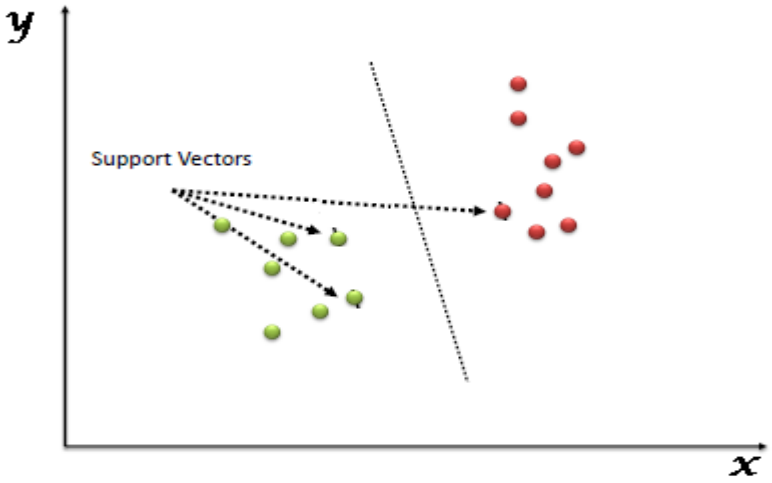


Figure 3.3: SVM Hyperplane

Source: (Rajput & Kaulwar, 2019)

The kernel function was employed in fitting the SVM model as it transforms input features into space with a higher number of dimensions where it may be easier to separate into different classes. The SVM model utilized the radial basis function (RBF) kernel. The RBF kernel which is also known as the Gaussian kernel is especially useful when dealing with non-linear relationships between variables. The feature space consisted of the independent variables; academic performance, social support network, financial situation, drug and substance abuse and past abuse, trauma and neglect.

The formula for the RBF kernel is given by;

$$K(x, y) = exp\ (-\gamma|| x - y ||^2) \quad (3.8)$$

Where;

$K(x, y)$ is the kernel function that measures the similarity between two data points x and y in the feature space,

$|| x - y ||$ is the Euclidean distance between two data points x and y,

γ (gamma) is a parameter that controls the width of the Gaussian distribution.

3.8.3.1 Parameters Estimation for SVM

The tuning of hyperparameters such as gamma and the cost parameter C was essential for balancing model complexity and generalization to new data.

The gamma parameter played a crucial role in shaping the decision boundary of the model. Proper tuning of gamma helped to find the right balance between fitting the training data well and generalizing to new data. The value of gamma is the reciprocal of the number of features in the model (Al-Mejibli *et al*., 2020).

The cost parameter C, is a regularization parameter that controls the balance between achieving a smooth decision boundary and classifying the training points correctly (Han *et al.,* 2012). Smaller values of C resulted in a more flexible model with a wider margin, allowing for increased misclassifications and greater tolerance for outliers. Larger values of C created a more rigid model with a narrower margin prioritizing accurate fitting to the training data which leads to the risk of overfitting.

3.9 Model Selection

A confusion matrix was created for each model to summarize the performance of the models. A confusion matrix is a table utilized to evaluate the efficacy of a machine learning algorithm by comparing the actual and predicted class of a dataset (Visa *et al*., 2011). For binary classification, the confusion matrix is given by;

Table 3.2: Confusion Matrix Table

	N	Estimated Values	
Actual Values		Negative	Positive
	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive (TP) arises when the model correctly estimates a positive case from the dataset. It is sometimes referred to as sensitivity.

True Negative (TN) arises when the model correctly estimates a negative case from the dataset. It is referred to as specificity.

False Positive (FP) arises when the model incorrectly estimates a negative case from the dataset.

False Negative (FN) arises when the model incorrectly estimates a positive case from the dataset.

N represents the total number of observations.

The parameters that were used to estimate the efficiency of each model are recall (sensitivity), specificity, positive predictive value (precision), negative predictive value and accuracy.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.10}$$

$$Positive\ predictive\ value = \frac{TP}{TP + FP} \tag{3.11}$$

$$Negative\ predictive\ value = \frac{TN}{TN + FN} \tag{3.12}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.13}$$

3.10 Model Evaluation

Model evaluation involves assessing the predictive performance of a model to ensure its generalization to new data. During this process, the data is divided into two subsets, the training set and the testing test. Training data is employed to teach a machine learning model in recognizing patterns and establishing relationships between input features and corresponding output labels while testing data is used to assess the

performance of a machine learning model by evaluating its prediction on new data. This study used 80% (0.80 x 1448 = 1145) of the data for training the models and 20% of the data (0.20 x 1448 = 303) was used for validating the models. Utilizing more data points during the training process enhances the accuracy of the predictions (Gholamy *et al*., 2018).

3.11 Ethical Considerations

The researcher abided to various ethical principles during the research process in order to safeguard the welfare of the participants. Informed consent was gathered from the participants by ensuring that the participants were aware of the aim of the research, the benefits of participating and the potential impact it had on their health. The participants were informed that their involvement in the study was optional and they had the right to opt-out of the study whenever they desired.

Anonymity of the participants was upheld by keeping personal information that could lead to their identification private and confidential. This in turn allowed participants to be more honest and open in their responses, which led to more accurate and reliable data.

The researcher ensured that participants were treated with respect throughout the research process. All data gathered was saved and processed in accordance with the relevant regulations and with the proper authorization.

Prior to commencing the study, the researcher sought ethical clearance from National Commission for Science, Technology & Innovation (NACOSTI). This step was essential to ensure that the research conformed to ethical guidelines and safeguarded the rights and well-being of the participants.

3.12 Summary

This chapter outlined the research methodology, including design, population, sampling, sample size, data collection, and analysis. The study employed a descriptive research design to depict the characteristics and patterns of depression among MUT students. The target population included 10,127 students from the seven schools at MUT. Proportionate stratified random sampling was used, with a sample size of 1,382

determined by Yamane's formula.

Data was gathered using anonymous online questionnaires with Likert scale questions. A pilot study with 139 participants confirmed the reliability and validity of the questionnaire, achieving a Cronbach's alpha of 0.89. Data analysis was performed using LR, RF and SVM models in R Studio and the performance of the models was evaluated using the metrics of the confusion matrix. Ethical considerations involved obtaining informed consent, maintaining confidentiality and adhering to NACOSTI guidelines to safeguard the rights and well-being of the participants.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

The research employed the R programming language version 4.2.2 within the RStudio integrated development environment (IDE) to implement and analyze machine learning algorithms, including logistic regression, random forest and support vector machine. RStudio, a dedicated IDE for R, facilitated a seamless workflow by providing a user-friendly interface, efficient script editing tools, and integrated support for package management. The analysis leveraged key R packages such as glm, randomForest and e1071 for LR, RF and SVM respectively. The analysis and results of the three machine learning models are presented in this chapter.

4.2 Descriptive Analysis

This section entails the frequency and percentage distribution of sociodemographic characteristics, association of sociodemographic characteristics and the prevalence of depressive symptoms and the distribution of responses on the predictor variables.

4.2.1 Sociodemographic Characteristics

Table 4.1 provided a detailed overview of the sociodemographic characteristics of the sample population with corresponding frequencies and percentages for each category.

Table 4.1: Sociodemographic Characteristics

Variable	Category	Frequency	Percent (%)
Gender	Female	512	35.35
	Male	936	64.64
Age	18-20	756	52.21
	21-23	552	38.12
	24-26	108	7.46
	27+	32	2.21
Sponsorship	Government	1340	92.54
	Self	108	7.46
Level of Education	Postgraduate	44	3.04
	Undergraduate	1232	85.08
	Diploma	172	11.88
School	SAES	42	2.90
	SBE	219	15.12
	SCIT	222	15.33
	SEHSS	522	36.05
	SET	159	10.98
	SHTM	106	7.32
	SPAHS	178	12.29
Year of Study	First Year	548	37.85
	Second Year	216	14.92
	Third Year	412	28.45
	Fourth Year	272	18.78
	Fifth Year	-	-
Residence	On Campus	152	10.50
	Off Campus	1296	89.50

Data was gathered from 1448 students (512 female and 936 male). More than half of the students 756 (52.21%) were aged between 18 and 20 years and only 32 (2.21%) were 27 years and above. Nearly all the students, 92.54% were government sponsored out of which 78.74% were in their undergraduate level. There was a higher response rate from the first-year students, 37.85% compared to the other years especially fifth

year students who had a 0% response rate. Only 10.5% of the students resided within the university premises.

4.2.2 Prevalence of Depressive Symptoms

Table 4.2 displayed the relationship between sociodemographic variables and depressive symptoms among MUT students.

Table 4.2: Relationship between Sociodemographic Variables and Depression

Variable	Category	Total		Depressive Symptoms			
				No		Yes	
		N	Percent (%)	N	Percent (%)	N	Percent (%)
All Students		1448	100	1072	74.03	376	25.97
Gender	Female	512	35.35	336	65.63	176	34.38
	Male	936	64.64	736	78.63	200	21.37
Age	18-20	756	52.21	580	76.72	176	23.28
	21-23	552	38.12	400	72.46	152	27.54
	24-26	108	7.46	72	66.67	36	33.33
	27+	32	2.21	20	62.50	12	37.50
Sponsorship	Government	1340	92.54	996	74.33	344	25.67
	Self	108	7.46	76	70.37	32	29.63
Level of	Postgraduate	44	3.04	28	63.64	16	36.36
Education	Undergraduate	1232	85.08	916	74.35	316	25.65
	Diploma	172	11.88	128	74.42	44	25.58
School	SAES	42	2.90	39	92.86	3	7.14
	SBE	219	15.12	152	69.41	67	30.59

	SCIT	222	15.33	156	70.27	66	29.73
	SEHSS	522	36.05	370	70.88	152	29.12
	SET	159	10.98	130	81.76	29	18.24
	SHTM	106	7.32	97	91.51	9	8.49
	SPAHS	178	12.29	128	71.91	50	28.09
Year of Study	First Year	548	37.85	444	81.02	104	18.98
	Second Year	216	14.92	156	72.22	60	27.78
	Third Year	412	28.45	288	69.90	124	30.10
	Fourth Year	272	18.78	184	67.65	88	32.35
	Fifth Year	-	-	-	-	-	-
Residence	On Campus	152	10.50	124	81.58	28	18.42
	Off Campus	1296	89.50	948	73.15	348	26.85

Using a cut-off point of 10 from the PHQ-9 scale, 25.97% (376) students had a score of more than 10 thus showing the symptoms of depression. This comprised of 19.61% (284) moderate depressive symptoms and 6.35% (92) severe depressive symptoms. The results align with the previously reported prevalence rates of depression among university students, which were 24.4% (Akhtar *et al.*, 2020) and 30.6% (Ibrahim *et al.*, 2013). However, the prevalence rate was much lower compared to that reported by (Othieno et al., 2014), among University of Nairobi students which was 41.33% and much higher than the prevalence rate of the general population, 9% (Ibrahim *et al.*, 2013).

In total 34.38% (176) females and 21.37% (200) males had depressive symptoms. Depression levels varied among different age groups, 18-20 (176, 23.28%), 21-23 (152, 27.54%), 24-26 (36, 33.33%) and 27+ years (12, 37.5%). From the seven schools at MUT, the prevalence of depression was as follows; SAES (3, 7.14%), SBE (67, 30.59%), SCIT (66, 29.73%), SEHSS (152, 29.12%), SET (29, 18.24%), SHTM (9, 8.49%) and SPAHS (50, 28.09%). From the level of study, a total of 16 (36.36%) postgraduate, 316 (25.65%) undergraduate and 44 (25.58%) diploma students met the diagnosis criteria for depressive symptoms. A total of 104 (18.98%) first year, 60 (27.28%) second year, 124 (30.1%) third year and 88 (32.35%) fourth year students showed the symptoms of depression. Students who resided in the university hostels had comparatively lower depression levels 18.42%, compared to those who resided outside the university, 26.85%. Among the students who resided in the university hostels, 28 (18.42%) students had depressive symptoms and 348 (26.85%) students residing off campus had depressive symptoms.

4.2.3 Predictor Variables

Table 4.3 showed the frequency distribution of the responses on academic performance.

Table 4.3: Academic Performance

	Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
B1	**N**	24	48	416	636	324
	Percent (%)	1.66	3.31	28.73	43.92	22.38
B2	**N**	0	56	180	660	552
	Percent (%)	0.00	3.87	12.43	45.58	38.12
B3	**N**	112	0	532	468	336
	Percent (%)	7.73	0.00	36.74	32.32	23.20
B4	**N**	72	204	332	552	288
	Percent (%)	4.97	14.09	22.93	38.12	19.89
B5	**N**	48	160	320	596	324
	Percent (%)	3.31	11.05	22.10	41.16	22.38
	Total	256	468	1780	2912	1824

B1 to **B5** represented questions participants were asked regarding academic performance. The questions were on a Likert scale of 1-5.

B1 – I am satisfied with my academic performance so far in the university.

1.66% of the students strongly disagreed with this statement, 3.31% agreed that they were satisfied with their academic performance. 28.73% were neutral with regard to their performance academically, 43.92% agreed and 22.38% strongly agreed that they were satisfied with their academic performance.

B2 – I have a clear understanding of my goals and how to achieve them

No student strongly disagreed with this statement while 3.87% disagreed that they had a clear understanding of their goals.12.43% were neutral on knowing and achieving their goals, almost half of the students, 45.58% agreed on having a clear understanding and achievement of the goals and 38.12% strongly agreed with the statement.

B3 - I do not experience the pressure from my family to maintain high academic standards.

Only 7.73% strongly disagreed that they did not experience pressure to perform well academically and no student disagreed with the statement. 36.74% were neutral, 32.32% agreed while 23.20% that they did not receive pressure from their families to maintain high academic standards.

B4 - I feel like I have adequate resources and support from my university to succeed academically.

4.97% strongly disagreed that they had adequate resources and support from the university to excel academically, 14.09% disagreed, 22.93% were neutral in consideration of the statement, 38.12% agreed and 19.89% strongly agreed with this statement.

B5 - I often compare my academic performance to that of my peers, and this comparison contributes to positive feelings of adequacy and self-esteem.

22.38% of the students strongly agreed that comparing their academic performance to that of their peers improved their self-esteem, 41.16% agreed with this statement, 22.10% were neutral on this statement, 11.05% disagreed and 3.31% disagreed with the statement.

Table 4.4 showed the frequency distribution of the participants' responses on the social support network they have.

Table 4.4: Social Support Network

	Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
C1	**N**	0	0	300	656	492
	Percent (%)	0.00	0.00	20.72	45.30	33.98
C2	**N**	88	164	296	540	360
	Percent (%)	6.08	11.33	20.44	37.29	24.86
C3	**N**	40	96	280	636	396
	Percent (%)	2.76	6.63	19.34	43.92	27.35
C4	**N**	44	80	260	584	480
	Percent (%)	3.04	5.52	17.96	40.33	33.15
C5	**N**	48	100	284	616	400
	Percent (%)	3.31	6.91	19.61	42.54	27.62
	Total	220	440	1420	3032	2128

C1 to **C5** represented the questions the students were asked concerning social support network.

C1 - I have a strong support system of friends and family.

33.98% of the students strongly agreed that they had a strong support system of friends and friends, 45.3% agreed while 20.72% were with this statement. Neither of the participants strongly disagreed nor disagreed with this statement.

C2 - I believe I have people I can talk to about my problems.

6.08% of the participants strongly disagreed that they had people they could talk to about their problems, 11.33% disagreed, 20.44% were neutral, 37.29% agreed and 24.86% strongly agreed with this statement.

C3 - I feel like I have a strong sense of belonging in my community.

27.35% strongly agreed that they had a strong sense of belonging in their community, 43.92% agreed, 19.34% were neutral, 6.63% disagreed and 2.76% strongly disagreed with the statement.

C4 - I am confident that I have people in my life who truly understand me and care about me.

33.15% strongly agreed that there were people in their life who truly understood and cared about them, 40.33% agreed, 17.96% were neutral, 5.52% disagreed while 3.04% disagreed with this statement.

C5 - I believe I have a good balance between my social life and my personal life.

3.31% of the participants strongly disagreed that they could balance their social and personal life, 6.91% disagreed, 19.61% were neutral, 42.54% agreed and 27.62% strongly agreed with this statement.

Table 4.5 displayed the distribution of responses from the participants with regard to their financial situation.

Table 4.5: Financial Situation

	Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
D1	**N**	120	336	384	436	172
	Percent (%)	8.29	23.20	26.52	30.11	11.88
D2	**N**	180	368	312	380	208
	Percent (%)	12.43	25.41	21.55	26.24	14.36
D3	**N**	92	300	376	492	188
	Percent (%)	6.35	20.72	25.97	33.98	12.98
D4	**N**	120	304	256	468	300
	Percent (%)	8.29	20.99	17.68	32.32	20.72
D5	**N**	80	232	300	548	288
	Percent (%)	5.52	16.02	20.72	37.85	19.89
	Total	592	1540	1628	2324	1156

D1 to **D5** represented the questions the students were asked on their financial situation.

D1 - I am able to cover all my basic expenses.

11.88% strongly agreed that they were able to cater for all their basic needs, 30.11% agreed, 26.52% were neutral, 23.2% disagreed and 8.29% strongly disagreed.

D2 - I have not had trouble paying my school fees in the past.

12.43% strongly disagreed that they did not have trouble paying their school fees in the past, 25.41% disagreed, 21.55% were neutral, 26.24% agreed and 14.36% strongly agreed with this statement.

D3 - I am satisfied with the financial assistance I receive.

12.98% strongly agreed that they were satisfied with the financial assistance they received, 33.98% agreed, 25.97% were neutral, 20.72% disagreed and 6.35% strongly disagreed with this statement.

D4 - My financial situation does not affect my ability to participate in extracurricular activities or social events.

20.72% strongly agreed that their financial situation did not affect their participation in social events, 32.32% agreed, 17.68% were neutral, 20.99% disagreed and 8.29% strongly disagreed with the statement.

D5 – My financial situation has not affected my relationship with my friends.

19.89% strongly agreed that their financial situation did not affect the relationship with their friends, 37.85% agreed, 20.72% were neutral, 16.02% disagreed and 5.52% strongly disagreed with this statement.

Table 4.6 displayed the distribution of participants' responses on drug and substance abuse.

Table 4.6: Drug and Substance Abuse

	Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
E1	**N**	72	148	152	320	756
	Percent (%)	4.97	10.22	10.50	22.10	52.21
E2	**N**	20	12	44	476	896
	Percent (%)	1.38	0.83	3.04	32.87	61.88
E3	**N**	8	24	80	408	928
	Percent (%)	0.55	1.66	5.52	28.18	64.09
E4	**N**	20	44	112	412	860
	Percent (%)	1.38	3.04	7.73	28.45	59.39
E5	**N**	48	84	128	460	728
	Percent (%)	3.31	5.80	8.84	31.77	50.28
	Total	168	312	516	2076	4168

E1 to **E5** represented the questions that the students were asked regarding drug and substance abuse.

E1 - I have never used drugs or consumed any substances.

52.21% strongly agreed that they had never consumed any drug or substance, 22.1% agreed, 10.5% were neutral, 10.22% disagreed and 4.97% strongly disagreed with this statement.

E2 - I am familiar with the dangers and consequences of drug or substance use.

1.38% strongly disagreed that they were familiar with the consequences of drug and substance abuse, 0.83% disagreed, 3.04% were neutral, 32.87% agreed and 61.88% strongly agreed with this statement.

E3 - I believe that drug or substance use has a negative impact on individuals and society as a whole.

64.09% strongly agreed that drug and substance use had a negative effect on both the individual and the society, 28.18% agreed, 5.52% were neutral, 1.66% disagreed and

0.55% strongly disagreed with this statement.

E4 - Peer pressure influences university students to use drugs and substances.

59.39% strongly agreed that peer pressure influences university students to use drugs and substances, 28.45% agreed, 7.73% were neutral, 3.04% disagreed and 1.38% strongly disagreed with this statement.

E5 - Education and awareness about the dangers of drug and substance abuse can help reduce their use among university students.

50.28% strongly agreed that education and awareness about the dangers of drug and substance abuse can help reduce their use among university students, 31.77% agreed, 8.84% were neutral, 5.8% disagreed and 3.31% strongly disagreed with this statement.

Table 4.7 displayed the distribution of participants' responses on past abuse, trauma and neglect.

Table 4.7: Past Abuse, Trauma and Neglect

	Statement	Not at all	Slightly	Moderately	Considerably	Almost always
F1	**N**	736	372	160	120	60
	Percent (%)	50.83	25.69	11.05	8.29	4.14
F2	**N**	588	388	200	144	128
	Percent (%)	40.61	26.80	13.81	9.94	8.84
F3	**N**	796	256	152	148	96
	Percent (%)	54.97	17.68	10.50	10.22	6.63
F4	**N**	772	308	148	88	132
	Percent (%)	53.31	21.27	10.22	6.08	9.12
F5	**N**	1076	128	152	52	40
	Percent (%)	74.31	8.84	10.50	3.59	2.76
	Total	3968	1452	812	552	456

F1 to **F5** represented questions that students were asked concerning the abuse, trauma and neglect they had experienced previously.

F1 – I have experienced abuse, trauma and neglect in the past.

50.83% reported that they had never experienced abuse, trauma and neglect in the past, 25.69% had experienced abuse, trauma and neglect slightly, 11.05% moderately, 8.29% considerably and 4.14% almost always.

F2 – My past experiences have an effect on my current emotional well-being.

8.84% reported that their past abuse, trauma and neglect almost always had an effect on their emotional well-being, 9.94% had their emotional well-being affected considerably, 13.81% moderately, 26.8% slightly and 40.61% did not have their emotional well-being affected at all.

F3 – I feel that my past abuse, trauma, or neglect has impacted my self-esteem.

6.63% stated that the past abuse, trauma and neglect almost always affected their self-esteem, 10.22% had their self-esteem affected considerably, 10.5% moderately, 17.68% slightly and 54.97% did not have their self-esteem affected at all.

F4 – I frequently experience flashbacks and reminders of my past abuse, trauma or neglect.

53.31% reported that they did not experience any reminders of their past abuse, trauma or neglect, 21.27% experienced it slightly, 10.22% moderately, 6.08% considerably and 9.12% almost always.

F5 - I have sought professional help to cope with the effects of my past abuse, trauma, or neglect.

74.31% stated that they had not sought any professional help to cope with their past experiences at all, 8.84% slightly, 10.5% moderately, 3.59% considerably and 2.76% said they had sought professional help almost always.

4.3 Logistic Regression (LR) Model

The logistic regression model was employed to model the binary dependent variable along with the categorical predictor variables. The sociodemographic characteristics,

gender, age, sponsorship, education level, school, year of study and residence were fit in the LR model. The estimated coefficients, standard (Std) error, z-value, p-value, odds ratio (OR) and the 95% confidence intervals (95% CI) for the predictor variables were presented in Table 4.8.

Table 4.8: LR results for Categorical Predictor Variables

Coefficients	Estimate	Std Error	z-value	p-value	OR	95% CI	
						Lower	Upper
Intercept	-1.21526	0.52586	-2.311	0.02083	0.30	-2.26	-0.19
Gender							
Female (ref)	0	-	-	-	1	-	-
Male	0.61917	0.15171	4.081	4.48×10^{-5}	1.86	0.32	0.92
Age							
18-20 (ref)	0	-	-	-	1	-	-
21-23	0.01703	0.19264	0.088	0.92954	1.02	-0.36	0.39
24-26	0.05690	0.29980	0.190	0.84947	1.06	-0.54	0.64
27+	0.43494	0.50741	0.857	0.39136	1.54	-0.60	1.41
Sponsorship							
Government (ref)	0	-	-	-	1	-	-
Self	0.01141	0.30415	0.038	0.97008	1.01	-0.60	0.60
Level of Education							
Postgraduate	-0.31065	0.47064	-0.660	0.50921	0.73	-1.22	0.63
Undergraduate	0.40711	0.50842	0.801	0.42328	1.50	-0.58	1.42

Diploma (ref)	0	-	-	-	1	-	-
School							
SAES	-1.36427	0.64871	-2.103	0.03546	0.26	-2.85	-0.23
SBE (ref)	0	-	-	-	1	-	-
SCIT	-0.10824	0.23785	-0.455	0.64906	0.90	-0.58	0.36
SEHSS	-0.15183	0.20345	-0.746	0.45550	0.86	-0.55	0.25
SET	-0.47494	0.29703	-1.599	0.10983	0.62	-1.07	0.10
SHTM	-1.91252	0.47841	-3.998	6.4×10^{-5}	0.15	-2.94	-1.04
SPAHS	-0.17084	0.25278	-0.676	0.49915	0.84	-0.67	0.32
Year							
First (ref)	0	-	-	-	1	-	-
Second	0.46342	0.24524	1.890	0.05880	1.59	-0.02	0.94
Third	0.60792	0.22674	2.681	0.00734	1.84	0.17	1.06
Fourth	0.77576	0.27005	2.873	0.00407	2.17	0.25	1.31
Residence							
On Campus (ref)	0	-	-	-	1	-	-
Off Campus	0.18181	0.26314	0.691	0.4896	1.20	-0.32	0.71

Gender was significantly associated with depression ($p - value = 4.48 \times 10^{-5}$) and male students were 1.86 times more likely to be depressed compared to the female students ($OR = 1.86, 95\%\ CI: 0.32, 0.92$). The results are in consistent with the findings of (Olum *et al.*, 2020), indicating that male students are susceptible to depression compared to the female students. Male students often feel isolated and lack the social support networks that female students typically have, which can lead to feelings of loneliness and depression.

The age of the students was categorized as: $18 - 20, 21 - 23, 24 - 26$ and $27 +$ years. The odds of depression for $21 - 23$, $24 - 26$ and $27 +$ years were 1.02, 1.06 and 1.54 times more likely as in $18 - 20$ years ($OR = 1.02, 95\%\ CI: -0.36, 0.39$), ($OR = 1.06, 95\%\ CI: -0.54, 0.64$) and ($OR = 1.54, 95\%\ CI: -0.60, 1.41$) respectively. The older students were more likely to be depressed since they face challenges with their school fees, upkeep, failed units which causes them not to proceed to the next year of study, failed relationships and peer pressure.

The self-sponsored students were 1.01 times more likely to be depressed compared to the Government sponsored students, ($OR = 1.01, 95\%\ CI: -0.60, 0.60$). Self-sponsored students are required to cover their entire school fees, leading to financial constraints. Government-sponsored students receive partial financial assistance from the Government, which covers a portion of their school fees. Despite this support, they are also responsible for managing their personal expenses independently.

Postgraduate students were 0.73 times less likely to be depressed as compared to the diploma students, $OR = 0.73, 95\%\ CI: -1.22, 0.63$. Postgraduate students may have developed higher levels of resilience and adaptability through previous academic experiences, making them more adept at handling stressors associated with their studies. The undergraduate students were 1.5 times more likely to be depressed as compared to the diploma students ($OR = 1.50, 95\%\ CI: -0.58, 1.42$).

Students were classified according to the schools they belonged to, SAES, SBE, SCIT, SEHSS, SET, SHTM and SPAHS. Students from the remaining SAES, SCIT, SEHSS, SET, SHTM and SPAHS were less likely to be depressed with regard to students from SBE. SAES ($OR = 0.26, 95\%\ CI: -2.85, -0.23$), SCIT ($OR = 0.90, 95\%\ CI: -0.58, 0.36$), SEHSS ($OR = 0.86, 95\%\ CI: -0.55, 0.25$), SET ($OR =$

$0.62, 95\%\ CI: -1.07, 0.10)$, SHTM $(OR = 0.15, 95\%\ CI: -2.94, -1.04)$ and SPAHS $(OR = 0.84, 95\%\ CI: -0.67, 0.32)$. The results above on odds ratio may be probably due to the practicability of the courses that are in the schools. SAES and SHTM mainly engage in practical's which help in reduction of stress that may eventually lead to depression.

The year of study had an association with depression as students in their second, third and fourth year of study were more likely to be depressed as compared to those in their first year. Second year $(OR = 1.59, 95\%\ CI: -0.02\ 0.94)$, third year $(OR = 1.84, 95\%\ CI: 0.17\ 1.06)$ and fourth year $(OR = 2.17, 95\%\ CI: 0.25\ 1.31)$. The odds of depression were increasing with the number of years a student had spent at the university. The higher odds are due to accumulated academic stress, the pressure to maintain high academic grades, peer pressure, lack of support networks and constraints with finances.

Students who resided in the University hostels were 0.83 times less likely to be depressed than those who resided in hostels outside the University $(OR = 0.83, 95\%\ CI: -0.71, 0.32)$. The results align with the findings of (Othieno et al., 2014) indicating a significant relationship between depression and the residence of the students. The prevalence was higher among students who were residing off campus. Those who resided off campus were likely to be faced with financial constraints in paying rent and utilities.

The logistic regression model was employed to model the binary dependent variable along with the predictor variables. The estimated coefficients, standard error, z-value, p-value, OR and the 95% CI for the predictor variables were presented in Table 4.9.

Table 4.9: LR results for Predictor Variables

Variable	Estimate	Std. Error	z-value	p-value	OR	95% CI	
						Lower	Upper
Intercept	2.00963	1.12580	1.785	0.07425	7.46	-0.19	4.23
Gender	-0.50992	0.17532	-2.908	0.00363	0.60	-0.85	-0.17
Age	0.07914	0.13176	0.601	0.54809	1.08	-0.18	0.34
Sponsorship	0.23784	0.31633	0.752	0.45212	1.27	-0.39	0.85
Level of education	-0.03302	0.24703	-0.134	0.89366	0.97	-0.51	0.46
School	-0.07019	0.05260	-1.334	0.18207	0.93	-0.17	0.03
Year	0.13206	0.08734	1.512	0.13052	1.14	-0.04	0.30
Residence	0.39320	0.31587	1.245	0.21319	1.48	-0.21	1.04
Academic Performance	-0.02345	0.16644	-0.141	0.88797	0.98	-0.35	0.30
Social Support Network	-0.90434	0.14525	-6.226	4.78×10^{-10}	0.40	-1.19	-0.62
Financial Situation	-0.33912	0.10744	-3.156	0.0016	0.71	-0.55	-0.13
Drug and Substance Abuse	0.01579	0.13326	0.119	0.9057	1.02	-0.24	0.28
Past Abuse, trauma and Neglect	0.77755	0.09109	8.536	$< 2 \times 10^{-16}$	2.18	0.60	0.96

From the independent variables that were fit using the LR model, only gender, social support network, financial situation and past abuse, trauma and neglect had a significant effect on depression since their p-values were less than 0.05.

The intercept was 2.00963 ($OR = 7.46, 95\%\ CI: -0.19, 4.23$), which is the expected value of the dependent variable (depression) when there is no effect of any of the predictor variables.

The coefficient for gender was -0.50992 ($OR = 0.60, 95\%\ CI: -0.85 - 0.17$). The odds of experiencing depression for female students was 0.60 times the odds of the male students. The relationship between gender and depression was statistically significant since the p-value $= 0.00363$.

The coefficient for age was 0.07914 ($OR = 1.08, 95\%\ CI: -0.18, 0.34$). As age increases by 1 unit, the odds of experiencing depression increase by 1.08. The p-value for age was 0.54809 and hence the association between age and depression was not statistically significant and it was removed from the study.

The coefficient for sponsorship was 0.23784 ($OR = 1.27, 95\%\ CI: -0.39, 0.85$). The odds of experiencing depression for self-sponsored students was 1.27 times the odds of the Government sponsored students. The association between sponsorship and depression was not statistically significant since the p-value $= 0.45212$, hence sponsorship was removed from the study.

The coefficient for the level of education was -0.03302 ($OR = 0.97, 95\%\ CI: -0.51, 0.46$). The odds of experiencing depression for undergraduate and postgraduate students was 0.97 times the odds of the diploma students. The relationship between level of education and depression was not statistically significant since p-value was 0.89366 and thus it was removed from the study.

The coefficient for school was -0.07019 ($OR = 0.93, 95\%\ CI: -0.17, 0.03$). The odds of experiencing depression for SAES, SCIT, SEHSS, SET, SHTM and SPAHS was 0.93 times the odds of SBE. The p-value for school was 0.18207 which suggested that the relationship between school and depression was not statistically significant and school was removed from the study.

The coefficient for the year of study was 0.13206 ($OR = 1.14, 95\%\ CI$: $-0.04, 0.30$). For each unit increase in the year of study, the odds of experiencing depression for second, third- and fourth-year students was 1.14 times the odds of the first-year students. The association between the year of study and depression was not statistically significant since the p-value was 0.13052.

The coefficient for residence was 0.39320 ($OR = 1.48, 95\%\ CI$: $-0.21, 1.04$). The odds of experiencing depression for the off-campus students was 1.48 times the odds of the on-campus students. The p-value for residence was 0.21319 which suggested that the relationship between the residence of the student and depression was not statistically significant thus it was removed from the study.

The coefficient for academic performance was -0.02345 ($OR = 0.98, 95\%\ CI$: $-0.35, 0.30$). As academic performance increases by one unit, the odds of experiencing depression decrease by 0.98. The relationship between academic performance and depression was not statistically significant since the p-value $= 0.88797$ was greater than 0.05 and thus academic performance was removed from the study.

The coefficient for social support network was -0.90434 ($OR = 0.40, 95\%\ CI$: $-1.19, -0.62$). When the strength of social support network increases, the odds of experiencing depression decrease by 0.40. The p-value for social support network was 4.78×10^{-10} which suggests that the association between social support network and depression was statistically significant. Mutalib (2021) also discovered that insufficient social support had a significant influence on depression.

The coefficient for financial situation was -0.33912 ($OR = 0.71, 95\%\ CI$: $-0.55, -0.13$). As the financial situation improves by one unit, the odds of experiencing depression decrease by 0.71. The p-value for financial situation was 0.0016 which suggest that the association between financial situation and depression was statistically significant. According to Mutalib (2021), financial constraints were also identified as having a significant impact on depression.

The coefficient for drug and substance abuse was 0.01579 ($OR = 1.02, 95\%\ CI$: $-0.24, 0.28$). Drug and substance abuse increased the odds of experiencing depression by 1.02. The p-value for drug and substance abuse was 0.9057 which indicated that

the association between drug and substance abuse and depression was not statistically significant and it was removed from the study.

The coefficient for past abuse, trauma and neglect was 0.77755 ($OR = 2.18, 95\%\ CI: 0.60, 0.96$). Reminders of past abuse, trauma and neglect increased the odds of experiencing depression by 2.18. The p-value of past abuse, trauma and neglect was $< 2 \times 10^{-16}$ which suggested that a history of past abuse trauma and neglect was statistically significant on depression. The findings align with those of (Flores *et al.*, 2018) that a history of abuse, trauma and neglect has a significant effect on depression.

When the coefficients for the predictor variables were extracted, the LR model was given by;

$$\begin{aligned} ln\left(\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}\right) &= 2.00963 - 0.50992X_1 + 0.07914X_2 + 0.23784X_3 \\ &\quad - 0.03302X_4 - 0.07019X_5 + 0.13206X_6 + 0.3932X_7 \\ &\quad - 0.02345X_8 - 0.90434X_9 - 0.33912X_{10} + 0.01579X_{11} \\ &\quad + 0.77755X_{12} \end{aligned} \tag{4.1}$$

Where;

X_1 is gender

X_2 is age

X_3 is sponsorship

X_4 is level of education

X_5 is school

X_6 is year of study

X_7 is residence

X_8 is academic performance

X_9 is financial situation

X_{10} is social support network

X_{11} is drug and substance abuse

X_{12} past abuse, trauma and neglect

The LR model was then fit using only the explanatory variables that were significant.

The coefficients for this model were extracted in order to write the logistic regression formula.

$$ln\left(\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)}\right) = 2.00963 - 0.50992X_1 - 0.90434X_2 - 0.33912X_3 + 0.77755X_4 \quad (4.2)$$

Where;

X_1 is gender

X_2 is social support network

X_3 is financial situation

X_4 is past abuse, trauma and neglect

4.4 Random Forest (RF) Model

RF classifier was employed to model the predictor variables associated with the dependent variable, depression. Table 4.10 below shows the results of the training parameters for RF model fit on the training data.

Table 4.10: RF Algorithm Training Parameters

Parameter	**ntree**													
	15		**20**		**23**		**24**		**25**		**26**		**27**	
RF Type	Classification													
Mtry	2													
OOB Error Rate	3.68%		2.88%		2.71%		2.71%		2.71%		2.1%		2.62%	
Confusion Matrix	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
0	835	12	846	3	842	7	845	4	844	5	847	2	848	1
1	30	265	30	266	27	269	27	269	26	270	22	274	29	267

The study considered two variables which were randomly selected for splitting to help reduce correlation between trees thus leading to a more diverse and robust ensemble.

This study used 23 trees as the optimal number of trees since the OOB error rate stabilizes between 23 and 25 trees and adding more trees does not lead to a substantial improvement in the predictive performance of the RF model. The RF model made an error of 2.71% in its predictions on the observations that were not included in the construction of each individual tree in the forest.

Class 0 represented the students who were not depressed while class 1 represented the students who were depressed. Using 23 trees, the RF model correctly predicted 842 students as not depressed and 269 students as depressed. A total of 27 students were incorrectly predicted as not having depression while they were actually depressed and 7 students were incorrectly predicted as depressed while they did not have the depressive symptoms.

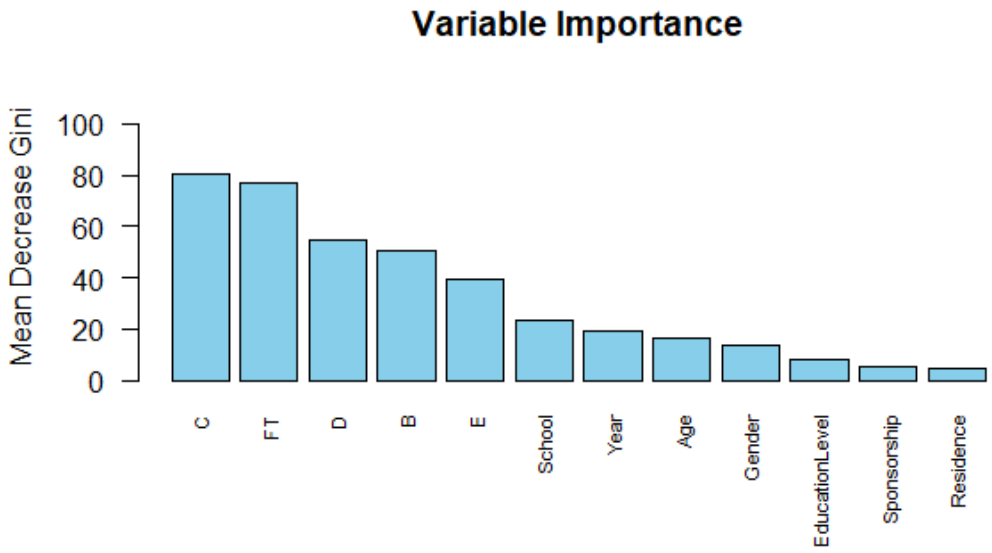


Figure 4.1: Variable Importance for RF Model

From figure 4.1, B represented the predictor variable academic performance, C represented social support network, D represented financial situation, E represented drug and substance abuse and FT represented the predictor variable past abuse, trauma and neglect.

Variable importance measured the contribution of each predictor variable to the predictive accuracy of the model. Figure 4.1 showed the variable importance of the independent variables for the random forest classifier based on their mean decrease Gini. The greater the decline in the mean decrease Gini the more useful the explanatory variable was in making accurate predictions on depression. From the plot above, the variables were arranged in descending order based on their mean decrease Gini. The first three variables that were considered more important in predicting depression in the model were; social support network, past abuse, trauma and neglect and financial situation respectively. The variable which had the least mean decrease Gini was the residence of the student.

4.5 Support Vector Machine (SVM) Model

The SVM algorithm was trained on the training data to model the relationship between the independent variables and the response variable. The RBF kernel was employed in the SVM algorithm to capture the non-linear relationship among the variables. Table 4.11 showed the results for the SVM algorithm using the RBF kernel.

Table 4.11: SVM Algorithm Training Parameters

Parameter	**Cost**				
	0.1	**1**	**10**	**50**	**100**
SVM Type	Classification				
SVM Kernel	RBF				
Gamma	0.0833				
Support Vectors	697	441	393	393	393
Class 0	401	259	259	259	259
Class 1	296	182	134	134	134

There were 12 features in the data and thus the value of gamma was 0.0833.

Further increase of the value of C from C = 10 did not affect the number of support vectors as the model had achieved a level of complexity that effectively captured the training data. Out of the 393 support vectors identified using the training data, 259 (65.9%) students were classified into class 0, indicating an absence of depressive

symptoms, while 134 (34.1%) students were classified into class 1, indicating the presence of depressive symptoms.

4.6 Selecting the Suitable Model

In order to select the most suitable model, a confusion matrix was created for each algorithm. The metrics of the confusion matrix were used to assess the performance of each model. These metrics included; accuracy, sensitivity (recall), positive predictive rate, negative predictive rate and specificity. A confusion matrix table was created for each model using the testing data and the metrics for each algorithm were calculated.

Table 4.12 presented the confusion matrix metrics for the LR model.

Table 4.12: Confusion Matrix for LR Model

	N	Estimated values	
Actual Values		Not Depressed	Depressed
	Not Depressed	186	37
	Depressed	35	45

N = 303

From Table 4.12, the performance metrics for the LR model were as follows.

Accuracy = 0.7624

Sensitivity (Recall) = 0.5625

Specificity = 0.8341

Positive predictive value = 0.5488

Negative predictive value = 0.8416

Table 4.13 showed the metrics of the confusion matrix for the RF classifier.

Table 4.13: Confusion Matrix for RF Classifier

	N	Estimated values	
Actual Values		Not Depressed	Depressed
	Not Depressed	223	0
	Depressed	4	76

N=303

The performance metrics for the RF model from Table 4.13 were;

Accuracy = 0.9868

Sensitivity (Recall) = 0.95

Specificity = 1

Positive predictive value = 1

Negative predictive value = 0.9824

Table 4.14 presented the confusion matrix metrics for the SVM algorithm.

Table 4.14: Confusion Matrix for SVM Algorithm

	N	Estimated Values	
Actual Values		Not Depressed	Depressed
	Not Depressed	219	4
	Depressed	35	45

N = 303

The performance metrics for the SVM model as presented in Table 4.14 were;

Accuracy = 0.8713

Sensitivity (Recall) = 0.5625

Specificity = 0.9821

Positive predictive value = 0.9184

Negative predictive value = 0.8622

Table 4.15 showed the confusion matrix metrics results for the three machine learning models namely logistic regression, random forest and support vector machine.

Table 4.15: Metrics of the Confusion Matrix for the three ML Algorithms

Metric	LR	RF	SVM
Accuracy	76.24%	98.68%	87.13%
Sensitivity (Recall)	56.25%	95%	56.25%
Specificity	83.41%	100%	98.21%
Positive predictive value	54.88%	100%	91.84%
Negative predictive value	84.16%	98.24%	86.22%

The frequency of depression among MUT students' prediction accuracy was found to be high among the three algorithms with random forest having the highest accuracy of 98.68%, followed by SVM with 87.13% and logistic regression with 76.24% using the testing data. Random forest had the highest sensitivity of 95% with the rest having a sensitivity of 56.25% implying that random forest was more accurate in identifying depressed cases among the students. For specificity, random forest had the highest specificity of 100%, followed by SVM with a specificity of 98.21% and logistic regression with 83.41% suggesting that random forest was the best algorithm in identifying the non-depressed cases. The random forest model correctly identified all (100%) of the real depressed cases (positives), the SVM model identified 91.84%, and the logistic regression model identified slightly more than half (54.88%) of the real depressed cases. In identifying the real non-depressed cases (negatives), all three algorithms had high values. Random forest correctly identified 98.24%, SVM identified 86.22%, and logistic regression identified 84.16% of the real non-depressed cases. From table 4.15, it was evident that the random forest classifier was the best

performing machine learning model since it outperformed SVM and logistic regression in classifying the depressed and non-depressed cases. The second-best performing algorithm was SVM followed by logistic regression.

4.7 Hypotheses Testing

The Pearson's Chi-squared test was used to test the hypotheses regarding whether LR, RF and SVM are effective in modeling the prevalence of depression among MUT students. Since the confusion matrix is structured as a 2 x 2 contingency table, the Pearson's Chi-squared test with Yates' continuity correction was employed to slightly adjust the chi-squared statistic thereby providing a more accurate p-value.

Table 4.16 showed the results for the null hypothesis (H_0) that logistic regression is not an effective method in modeling the prevalence of depression among MUT students.

Table 4.16: Hypothesis test for the LR model

Hypothesis	**χ^2 statistic**	**Degrees of freedom**	**p-value**
H_0: Logistic regression is not an effective method in modeling the prevalence of depression among MUT students.	44.926	1	2.05×10^{-11}

The Chi-squared statistic indicates the deviation between the actual and the estimated frequencies. The degrees of freedom were determined by subtracting 1 from the number of categories.

The null hypothesis was rejected since the $p - value = 2.05 \times 10^{-11}$ was significantly below 0.05. This rejection was supported by a chi-squared statistic of 44.926, indicating disparity between observed and expected frequencies. Both the p-value and chi-squared statistic provided evidence that logistic regression is indeed effective in modeling the prevalence depression among MUT students.

Table 4.17 below showed the results for the null hypothesis (H_0) that the random forest classifier is not an effective method in modeling the prevalence of depression among MUT students.

Table 4.17: Hypothesis test for the RF model

Hypothesis	**χ^2 statistic**	**Degrees of freedom**	**p-value**
H_0: The random forest classifier is not an effective method in modeling the prevalence of depression among MUT students.	277.74	1	2.2×10^{-16}

The chi-squared statistic of 277.74, coupled with a small p-value of 2.2×10^{-16} firmly reject the null hypothesis. The outcome highlighted a deviation between the observed and expected frequencies under the null hypothesis, indicating the effectiveness of the random forest classifier in modeling the prevalence of depression among MUT students.

Table 4.18 showed the results for the null hypothesis (H_0) that the support vector machine algorithm is not an effective method in modeling the prevalence of depression among MUT students.

Table 4.18: Hypothesis test for the SVM model

Hypothesis	**χ^2 statistic**	**Degrees of freedom**	**p-value**
H_0: The support vector machine algorithm is not an effective method in modeling the prevalence of depression among MUT students.	124.81	1	2.2×10^{-16}

With a chi-squared statistic of 124.81 and a small p-value of $2.2 \times 10^{-16}$the null hypothesis was emphatically rejected. The disparity between the observed and the

expected frequencies indicated the support vector machine algorithm was an effective method in modeling the prevalence of depression among MUT students.

The study utilized the two-sample Z-test for proportions to examine whether LR, RF and SVM differed significantly in their ability to accurately model the prevalence of depression among MUT students.

Tables 4.19, 4.20, and 4.21 displayed the z-statistic, the tabulated value at the 5% level of significance and the p-value for comparing the accuracy of the models.

Table 4.19: Two-sample Z-test for RF and SVM

Model	Accuracy	Sample size	z-statistic	Tabulated value	p-value
RF	0.9868	303	5.5372	1.65	3.0729×10^{-8}
SVM	0.8713	303			

The z-statistic calculated for comparing the accuracy of RF and SVM models was 5.5372, while the tabulated value at the 5% level of significance was 1.65. Since the z-statistic exceeded the tabulated value, the null hypothesis was rejected. The corresponding p-value was 3.0729×10^{-8}, which was less than 0.05. The low p-value indicated a significant difference between the RF and SVM models in their ability to accurately model the prevalence of depression among MUT students.

Table 4.20: Two-sample Z-test for RF and LR

Model	Accuracy	Sample size	z-statistic	Tabulated value	p-value
RF	0.9868	303	4.2927	1.65	1.7648×10^{-5}
LR	0.7624	303			

The z-statistic obtained from the comparison between the accuracy of the RF and LR models was 4.2927, which exceeds the tabulated value of 1.65 at the 5% significance level. The corresponding p-value was 1.7648×10^{-5}, significantly lower than 0.05. The low p-value confirmed the rejection of the null hypothesis, indicating a significant difference in the ability of the RF and LR algorithms to accurately model the

prevalence of depression among MUT students.

Table 4.21: Two-sample Z-test for SVM and LR

Model	Accuracy	Sample size	z-statistic	Tabulated value	p-value
SVM	0.8713	303	3.6758	1.65	0.0002374
LR	0.7624	303			

The z-statistic obtained for comparing the accuracy of the SVM and LR models was 3.6758, exceeding the tabulated value of 1.65 at the 5% significance level. The corresponding p-value was 0.0002374, which was less than 0.05. This low p-value supported the rejection of the null hypothesis, indicating a significant difference in the ability of the SVM and LR algorithms to accurately model the prevalence of depression among MUT students.

4.8 Summary

This chapter entailed the descriptive and confirmatory statistics of MUT students who participated in the study, the results for each specific objective and hypotheses testing. The results for the first specific objective were achieved by fitting the logistic regression model, the second by fitting the random forest classifier and the third by fitting the support vector machine algorithm. The performance of the three machine learning models was assessed by the metrics of the confusion matrix and the random forest classifier had the best performance among the three methods from the metrics tabulated. The Pearson's Chi-squared test was used to test the first, second, and third hypotheses, while the two-sample Z-test was employed to test the fourth hypothesis.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

In conclusion, this study utilized machine learning techniques, including logistic regression, random forest, and support vector machine to model the prevalence of depression among MUT students. The findings revealed a depression prevalence rate of 25.97%, with a higher incidence among male students, those residing off-campus, and those in the successive years of study. Notably, students with robust social support, academic success, and financial stability exhibited a lower risk of depression, while those engaging in substance abuse or with a history of trauma faced an elevated risk.

Logistic regression achieved an accuracy of 76.24% in distinguishing between depressed and non-depressed cases. It demonstrated a sensitivity of 56.25% for identifying depressed cases and a specificity of 83.41% for non-depressed cases. Additionally, the model displayed a positive predictive value of 54.88% for accurately identifying true depressed cases and a negative predictive value of 84.16% for correctly recognizing true non-depressed cases. Gender, social support network, financial situation and past abuse, trauma and neglect were significantly associated with depression.

The random forest algorithm achieved an accuracy of 98.68% in classifying depressed and non-depressed cases. It demonstrated a sensitivity of 95% for identifying depressed cases, while exhibiting 100% specificity and positive predictive value in accurately recognizing non-depressed and true depressed cases, respectively. Additionally, the algorithm displayed a negative predictive value of 98.24% in identifying true non-depressed cases. Social support network, past abuse, trauma and neglect and financial situation had the highest mean decrease Gini values, respectively, and thus were considered more important in predicting depression.

SVM achieved 87.13% accuracy in differentiating between depressed and non-depressed cases. It exhibited a 56.25% sensitivity for identifying depressed cases and a high specificity of 98.21% for non-depressed cases. Moreover, the model showed a positive predictive value of 91.84% for accurately identifying true depressed cases and

a negative predictive value of 86.22% for correctly recognizing true non-depressed cases.

The machine learning models showcased remarkable efficiency in classifying depression instances based on a diverse set of features, with the random forest algorithm emerging as the most effective. The study's thorough analysis, supported by a rich training dataset, contributes valuable insights to the understanding of mental health within the university population.

5.2 Recommendations

5.2.1 Recommendations for Policy

This study recommended employing the random forest algorithm in modeling the prevalence of depression among university students since it had the ability to identify all the real depressed cases among MUT students and 98.24% of the real non-depressed cases.

Implementing targeted interventions founded on identified risk and protective factors is essential for cultivating the mental health of MUT students. These interventions include personalized counselling, peer mentorships, workshops addressing financial wellness, substance abuse and flexible academic support. Trauma informed counselling is advocated for holistic well-being. These interventions collectively address multifaceted factors influencing mental health, aiming to nurture a resilient and thriving student population at MUT. Additionally, exploring the long-term outcomes of these interventions and their sustained impact on students' mental well-being would contribute to the evolving field of mental health research within academic settings.

5.2.2 Recommendations for Future Work

The dataset should be broadened to incorporate other risk factors causing depression among university students, including extraneous variables such as employment. It should also evaluate whether RF is the best algorithm for modeling the prevalence of depression by comparing it with other machine learning methods such as Naïve Bayes, Artificial Neural Network, and eXtreme Gradient Boosting.

REFERENCES

Adewuya, A. O., Ola, B. A., & Afolabi, O. O. (2006). Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *Journal of affective disorders, 96(1-2), 89-93.*

Adewuya, A. O., Ola, B. A., Aloba, O. O., Mapayi, B. M., & Oginni, O. O. (2006). Depression amongst Nigerian university students. *Social psychiatry and psychiatric epidemiology, 41(8), 674-678.*

Ahuja, R., & Banga, A. (2019). Mental stress detection in university students using machine learning algorithms. *Procedia Computer Science, 152, 349-353.*

Akhtar, P., Ma, L., Waqas, A., Naveed, S., Li, Y., Rahman, A., & Wang, Y. (2020). Prevalence of depression among university students in low- and middle-income countries (LMICs): a systematic review and meta-analysis. *Journal of Affective Disorders, 274, 911-919.*

Al-Mejibli, I. S., Alwan, J. K., & Abd Dhafar, H. (2020). The effect of gamma value on support vector machine performance with different kernels. *International Journal of Electrical and Computer Engineering, 10(5), 5497.*

Arain, M., Campbell, M. J., Cooper, C. L., & Lancaster, G. A. (2010). What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC medical research methodology, 10(1), 1-7.*

Breiman, L. (2001). Random forests. *Machine learning, 45(1), 5-32*

Bukhari, S. R., & Khanam, S. J. (2015). Prevalence of depression in university students belonging to different socioeconomic status. *JPMI: Journal of Postgraduate Medical Institute, 29(3).*

Chebet, D. A. I. S. Y., & Njuguna, R. (2020). Knowledge management practices and service delivery at Oxfam International, Kenya. *International Academic Journal of Human Resource and Business Administration, 3(9), 55-74.*

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20(3), 273-297.*

Dudek, G. (2015). Short-term load forecasting using random forests. In Intelligent Systems' 2014: Proceedings of the 7th IEEE International Conference Intelligent Systems IS'2014, September 24-26, 2014, Warsaw, Poland, Volume 2: Tools, Architectures, Systems, Applications (pp. 821-828). *Springer International Publishing*.

Fletcher, T. (2009). Support vector machines explained. *Tutorial paper, 1-19.*

Flores, R. J., Campo-Arias, A., Stimpson, J. P., Chalela, C. M., & Reyes-Ortiz, C. A. (2018). The association between past sexual abuse and depression in older adults from Colombia. *Journal of geriatric psychiatry and neurology, 31(1), 13-18.*

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation.

Hamdan, M. H., & Roach, D. C. (2022). The sigmoid neural network activation function and its connections to Airy's and the Nield-Kuznetsov functions. *Proof, 2, 108-114.*

Hameed, M., & Naumann, F. (2020). Data preparation: A survey of commercial tools. *ACM SIGMOD Record, 49(3), 18-29.*

Han, H., Guo, X., & Yu, H. (2016, August). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. *In 2016 7th ieee international conference on software engineering and service science (icsess) (pp. 219-224). IEEE.*

Han, S., Qubo, C., & Meng, H. (2012, June). Parameter selection in SVM with RBF kernel function. *In World Automation Congress 2012 (pp. 1-4). IEEE.*

Haque, U. M., Kabir, E., & Khanam, R. (2021). Detection of child depression using machine learning methods. *PLoS one, 16(12), e0261131.*

Haroz, E. E., Ritchey, M., Bass, J. K., Kohrt, B. A., Augustinavicius, J., Michalopoulos, L., ... & Bolton, P. (2017*)*. How is depression experienced around the world? A systematic review of qualitative literature. *Social Science*

& Medicine, 183, 151-162.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). *John Wiley & Sons.*

Ibrahim, A. K., Kelly, S. J., Adams, C. E., & Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. *Journal of psychiatric research, 47(3), 391-400.*

In, J. (2017). Introduction of a pilot study. *Korean journal of anesthesiology, 70(6), 601-605.*

Jackson-Koku, G. (2016). Beck depression inventory. *Occupational Medicine, 66(2), 174-175.*

Joshi, S., Ghosh, A., & Shukla, S. (2022, April). Machine Learning Model for Depression Prediction during COVID-19 Pandemic. *In 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-5). IEEE.*

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The patient health questionnaire (phq-9)–overview. *J. Gen. Intern. Med, 16, 606-616.*

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine, 16(9), 606-613.*

Kuringe, E., Materu, J., Nyato, D., Majani, E., Ngeni, F., Shao, A., ... & Wambura, M. (2019). *Prevalence and correlates of depression and anxiety symptoms among out-of-school adolescent girls and young women in Tanzania: A cross-sectional study*. PLoS One, 14(8), e0221053.

Lai, P. C. (2018). Research methodology for novelty technology. *JISTEM-Journal of Information Systems and Technology Management, 15.*

Li, W., Zhao, Z., Chen, D., Peng, Y., & Lu, Z. (2022). Prevalence and associated factors of depression and anxiety symptoms among college students: a systematic review and meta-analysis. *Journal of Child Psychology and*

Psychiatry, 63(11), 1222-1230.

Lund, C., Brooke-Sumner, C., Baingana, F., Baron, E. C., Breuer, E., Chandra, P., & Saxena, S. (2018). Social determinants of mental disorders and the Sustainable Development Goals: a systematic review of reviews. *The lancet psychiatry, 5(4), 357-369.*

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet], 9, 381-386.*

Mbaabu, O. (2020). Introduction to random forest in machine learning. *Engineering Education (EngEd) Program Section.*

Mbwayo, A., Kiarie, M., & Ndegwa, J. (2022). Factors Related to Depression among University Students in Nairobi County, Kenya. *SSRG International Journal of Humanities and Social Science.*

Mishra, N., Kapil, M., Rakesh, H., Anand, A., Mishra, N., Warke, A., & Panigrahi, P. K. (2021). Quantum machine learning: A review and current status. *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2020, Volume 2, 101-145.*

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University. Economic Series, 17(4), 59-82.*

Muriungi, S. K., & Ndetei, D. M. (2013). Effectiveness of psycho-education on depression, hopelessness, suicidality, anxiety and substance use among basic diploma students at Kenya Medical Training College. *South African Journal of Psychiatry, 19(2), 41-50.*

Mutalib, S. (2021). Mental health prediction models using machine learning in higher education institution. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(5), 1782-1792.*

Odhiambo, M. W., Gachoka, H. G., & Rambo, C. M. (2018). Relationship between age diversity and employee performance of public universities in Western Kenya. *International Journal of Academic Research in Business and Social*

Sciences, 8(11), 223–248.

Olum, R., Nakwagala, F. N., & Odokonyero, R. (2020). Prevalence and factors associated with depression among medical students at Makerere university, Uganda. *Advances in Medical Education and Practice, 853-860.*

Othieno, C. J., Okoth, R. O., Peltzer, K., Pengpid, S., & Malla, L. O. (2014). Depression among university students in Kenya: Prevalence and sociodemographic correlates. *Journal of affective disorders, 165, 120-125.*

Probst, P., & Boulesteix, A. L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research, 18(181), 1-18.*

Qasrawi, R., Vicuna Polo, S., Al-Halawa, D. A., Hallaq, S., & Abdeen, Z. (2022). School children depression and anxiety risk factors assessment and prediction: Machine learning techniques performance analysis. *JMIR formative research (2022) 6(8) e32736*

Rahman, M. M., Chowdhury, M. R. H. K., Islam, M. A., Tohfa, M. U., Kader, M. A. L., Ahmed, A. A. A., & Donepudi, P. K. (2020). Relationship between socio-demographic characteristics and job satisfaction: evidence from private bank employees. *American Journal of Trade and Policy, 7(2), 65-72.*

Rajput, G. G., & Kaulwar, B. H. (2019, March). A comparative study of artificial neural networks and support vector machines for predicting stock prices in National Stock Exchange of India. *In 2019 International Conference on Data Science and Communication (IconDSC) (pp. 1-7). IEEE.*

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine, 47(1), 31-39.*

Rois, R., Ray, M., Rahman, A., & Roy, S. K. (2021). Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms. *Journal of Health, Population and Nutrition, 40(1), 1-12.*

Sayed, H. A., William, A., & Said, A. M. (2023). Smart Electricity Meter Load Prediction in Dubai Using MLR, ANN, RF, and ARIMA. *Electronics, 12(2), 389.*

Spitzer, R. L., Williams, J. B. W., & Kroenke, K. J. R. C. B. (2014). Test review: patient health questionnaire–9 (PHQ-9). *Rehabilitation Counseling Bulletin, 57(4), 246-248.*

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine, 18(10), 1099-1104.*

Titov, N., Dear, B. F., McMillan, D., Anderson, T., Zou, J., & Sunderland, M. (2011). Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cognitive behaviour therapy, 40(2), 126-136.*

Tsang, S., Royse, C. F., & Terkawi, A. S. (2017). Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi journal of anaesthesia, 11(5), 80.*

Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *Maics, 710(1), 120-127.*

Wakiru, J. (2018). A decision tree-based classification framework for used oil analysis applying random forest feature selection. *Journal of Applied Sciences, Engineering and Technology for Development. Volume 2, Issue 1.*

Watson, R. (2015). Quantitative research. *Nursing standard, 29(31).*

World Health Organization (WHO). (2023). Depression. https://www.who.int/news-room/fact-sheets/detail/depression

World Health Organization. (2017). Depression and other common mental disorders: global health estimates (No. WHO/MSD/MER/2017.2). *World Health Organization.*

Yamane, T. (1967). Statistics, an introductory Analysis 2nd Edition: *Horper and Row. New York.*

APPENDICES

Appendix 1: Introductory Letter

Dear Participant,

RE: MODELING THE PREVALENCE OF DEPRESSION AMONG MURANG'A UNIVERSITY OF TECHNOLOGY STUDENTS USING MACHINE LEARNING TECHNIQUES

I am a master's student at Murang'a University of Technology conducting the above research work.

Thank you for dedicating your time to participate in this important research study. Your participation in this study is completely voluntary and confidential. You will not be asked to reveal your name or any personal information that could identify you. The questionnaire contains a series of questions about various factors that may be related to depression and associated factors. Your answers will be maintained with complete anonymity and confidentiality.

I am grateful your time and effort in completing this questionnaire, which is critical to the success of this study.

I would like to express my gratitude for your involvement in this study.

Yours Sincerely,

Loise Wahome

Appendix 2: Questionnaire

The aim of this questionnaire is to collect data on the relevant factors associated with depression to help in modeling depression prevalence among MUT students. Please read each question carefully and respond by filling one blank space in section A. From section B onwards, please choose the number that best represents your response to each statement.

SECTION A: Sociodemographic Characteristics

1. What is your gender?
   Male [ ]
   Female [ ]
2. What is your age?
   a. 18-20 [ ]
   b. 21-23 [ ]
   c. 24-26 [ ]
   d. 27+ [ ]
3. Are you a government-sponsored or self-sponsored student?
   a. Government-sponsored [ ]
   b. Self-sponsored [ ]
4. What is your level of education?
   a. Diploma [ ]
   b. Undergraduate [ ]
   c. Postgraduate [ ]
5. Which school do you belong to at MUT?
   a. School of Computing and Information Technology (SCIT) [ ]
   b. School of Business and Economics (SBE) [ ]
   c. School of Education, Humanities and Social Sciences (SEHSS) [ ]
   d. School of Engineering and Technology (SET) [ ]
   e. School of Hospitality and Tourism Management (SHTM) [ ]
   f. School of Pure, Applied and Health Sciences (SPAHS) [ ]
   g. School of Agriculture and Environmental Sciences (SAES) [ ]
6. What year of study are you in at MUT?
   a. First year [ ]

b. Second year [ ]

c. Third year [ ]

d. Fourth year [ ]

e. Fifth year [ ]

7. Do you live at campus or away from campus?

a. On campus [ ]

b. Off campus [ ]

SECTION B: Academic Performance

The following are statements concerning your academic performance. State your level of agreement using a Likert scale of 1 to 5, where;

$1 = Strongly\ disagree$

$2 = \ Disagree$

$3 = Neutral$

$4 = Agree$

$5 = Strongly\ agree$

Statement	**1**	**2**	**3**	**4**	**5**
I am satisfied with my overall academic performance so far in university					
I have a clear understanding of my academic goals and how to achieve them					
I feel pressure from my family or society to perform well academically in university					
I feel like I have adequate resources and support from my university to succeed academically					
I often compare my academic performance to that of my peers, and this comparison makes me feel worse about myself					

SECTION C: Social Support Network

The following are statements relating to the social support network you have. State your level of agreement using a Likert scale of 1 to 5, where;

$1 = Strongly\ disagree$

$2 = \ Disagree$

$3 = Neutral$

$4 = Agree$

$5 = Strongly\ agree$

Statement	**1**	**2**	**3**	**4**	**5**
I have a strong support system of friends and family					
I believe I have people I can talk to about my problems					
I feel like I have a strong sense of belonging in my community					
I am confident that I have people in my life who truly understand me and care about me					
I believe I have a good balance between my social life and my personal life					

SECTION D: Financial Situation

The following are statements that pertain to your financial situation. Indicate your level of agreement using the Likert scale of 1 to 5, where;

$1 = Not\ at\ all$

$2 = \ Slightly$

$3 = Moderately$

$4 = Considerably$

$5 = Extremely$

Statement	**1**	**2**	**3**	**4**	**5**
I am able to cover all my basic expenses					
I have had trouble paying my school fees in the past					
I am satisfied with the financial assistance I receive					
My financial situation affects my ability to participate in extracurricular activities or					

social events					
My financial constraints have affected my relationship with my partner					

SECTION E: Drug and Substance Abuse

The following are statements concerning drug and substance abuse. State your level of agreement using a Likert scale of 1 to 5, where;

$1 = Strongly\ disagree$

$2 = \ Disagree$

$3 = Neutral$

$4 = Agree$

$5 = Strongly\ agree$

Statement	**1**	**2**	**3**	**4**	**5**
I frequently use drugs or substances					
I am familiar with the dangers and consequences of drug or substance use					
I believe that drug or substance use has a negative impact on individuals and society as a whole					
Peer pressure influences university students to use drugs and substances					
Education and awareness about the dangers of drug and substance abuse can help reduce their use among university students					

SECTION F: Past Abuse, Trauma and Neglect

The following are statements relating to past abuse, trauma or neglect you may have experienced. Indicate your level of agreement using the Likert scale of 1 to 5, where;

$1 = Not\ at\ all$

$2 = \ Slightly$

$3 = Moderately$

$4 = Considerably$

$5 = Almost\ always$

Statement	1	2	3	4	5
I have experienced abuse, trauma and neglect in the past					
My past experiences have an effect on my current emotional well-being					
I feel that my past abuse, trauma, or neglect has impacted my self-esteem					
I frequently experience flashback and reminders of my past abuse, trauma or neglect					
I have sought professional help to cope with the effects of my past abuse, trauma, or neglect					

SECTION G

Over the last 2 weeks, how often have you been bothered by any of the following problems? Use the Likert scale of 0 to 3 to indicate how often these problems have disturbed you. Where;

$0 = Not\ at\ all$

$1 = Several\ days$

$2 = More\ than\ half\ the\ days$

$3 = Nearly\ everday$

Problem	0	1	2	3
Little interest or pleasure in doing things				
Feeling down, depressed, or hopeless				
Trouble falling or staying asleep, or sleeping too much				
Feeling tired or having little energy				
Poor appetite or overeating				
Feeling bad about yourself — or that you are a failure or have let yourself or your family down				
Trouble concentrating on things, such as reading the				

newspaper or watching television				
Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual				
Thoughts that you would be better off dead or of hurting yourself in some way				

Appendix 3: R-codes for Logistic Regression Model

```
getwd()
Data=read.csv("G:\\My Drive\\THESIS Docs\\Thesis\\Depression DataC.csv
attach(Data)
names(Data)

#Find total observations in the data
nrow(Data)
#Summary statistics for variables
summary(Data)
#Find total number of missing values
sum(is.na(Data))
#Removes rows containing missing values
na.omit(Data)
#Data type
class(Data)

library(caret)
library(caTools)
library(MASS)
library(broom)
library(ResourceSelection)

#Split into training and testing set (80% for training)
#Setting the seed for reproducibility
set.seed(123)
split <- sample.split(Data, SplitRatio = 0.8)
training_data <- subset(Data, split == TRUE)
testing_data <- subset(Data, split == FALSE)

#Sums the missing values
```

```
#Sums the missing values
sapply(Data,function(x)sum(is.na(x)))

#Fit a binary logistic Regression Model on the training set for Sociod
model_all=glm(GB~Gender+Age+Sponsorship+EducationLevel+School+Year+Res
summary(model_all)

# Extract the coefficients
coefficientsall <- coef(model_all)

# Calculate odds ratios
odds_ratiosall <- exp(coefficientsall);odds_ratiosall

# Calculate 95% confidence intervals for coefficients
conf_intervalall <- confint(model_all)
conf_intervalall

# Make predictions on the testing data
predicted_probabilities_test <- predict(model_all, newdata = testing_d
```

```
49
50  # Convert probabilities to binary predictions (0 or 1) based on a thre
51  predicted_labels_test <- ifelse(predicted_probabilities_test >= 0.37,
52
53  # Create a confusion matrix for the testing set
54  conf_matrix_test <- table(Actual = testing_data$GB, Predicted = predic
55
56  # Print the confusion matrix for the testing set
57  print(conf_matrix_test)
58
59  # Calculate accuracy for the testing set
60  accuracy_test <- sum(diag(conf_matrix_test)) / sum(conf_matrix_test)
61  accuracy_test
62
63  #Checking the data type
64  class(training_data$Gender)
65  class(training_data$Age)
66  class(training_data$Year)
67  class(training_data$School)
68  class(training_data$EducationLevel)
69  class(training_data$Sponsorship)
70  class(training_data$Residence)
71
72  #Converting the variables to factors
73  training_data$Gender <- factor(training_data$Gender,labels=c("Male","F
74  training_data$Age <- factor(training_data$Age, labels=c("18 - 20","21
75  training_data$Year <- factor(training_data$Year,labels=c("First year",
76  training_data$School <- factor(training_data$School,labels = c("SAES",
77  training_data$EducationLevel <- factor(training_data$EducationLevel,la
```

```
78  training_data$Sponsorship <- factor(training_data$Sponsorship,labels =
79  training_data$Residence <- factor(training_data$Residence,labels = c("
80
81  #Check the Current Levels
82  levels(training_data$Gender)
83  levels(training_data$Age)
84  levels(training_data$Year)
85  levels(training_data$School)
86  levels(training_data$EducationLevel)
87  levels(training_data$Sponsorship)
88  levels(training_data$Residence)
89  set.seed(123)
90  #Reorder the Levels
91  training_data$Gender <- relevel(training_data$Gender, ref = "Female")
92  training_data$Age <- relevel(training_data$Age, ref = "18 - 20")
93  training_data$Year <- relevel(training_data$Year, ref = "First year")
94  training_data$School <- relevel(training_data$School, ref = "SBE")
95  training_data$EducationLevel <- relevel(training_data$EducationLevel,
96  training_data$Sponsorship <- relevel(training_data$Sponsorship, ref =
97  training_data$Residence <- relevel(training_data$Residence, ref = "On
98
99  #Fit a binary logistic Regression Model on the training set for Sociod
100 model_social=glm(GB~Gender+Age+Sponsorship+EducationLevel+School+Year+
101 summary(model_social)
```

```
102
103 # Extract the coefficients
104 coefficientsD <- coef(model_social)
105
106 # Calculate odds ratios
107 odds_ratiosD <- exp(coefficientsD);odds_ratiosD
108
109 # Calculate 95% confidence intervals for coefficients
110 conf_intervalD <- confint(model_social)
111 conf_intervalD
```

Appendix 4: R-codes for Random Forest Classifier

```
getwd()
Data=read.csv("G:\\My Drive\\THESIS Docs\\Thesis\\Depression DataC.csv
attach(Data)
names(Data)

#Find total observations in the data
nrow(Data)
#Summary statistics for variables
summary(Data)
#Find total number of missing values
sum(is.na(Data))
#Removes rows containing missing values
na.omit(Data)
#Data type
class(Data)

library(caret)
library(caTools)
library(MASS)
library(broom)
library(randomForest)

#Split into training and testing set (80% for training)
#Setting the seed for reproducibility
set.seed(123)
split <- sample.split(Data, SplitRatio = 0.8)
training_data <- subset(Data, split == TRUE)
testing_data <- subset(Data, split == FALSE)
```

```
#Sums the missing values
sapply(Data,function(x)sum(is.na(x)))

# Specify the Dependent variable and predictor variables
# Replace with the name of your Dependent variable
target_variable <- "GB"
# Replace with your predictor variable names
predictor_variables <- c("Gender","Age","Sponsorship","EducationLevel"

# Specify the range of values for the number of trees
num_trees_values <- c(10, 15, 18, 20, 23, 24, 25, 26, 27, 28, 29)

# Set a seed for reproducibility
set.seed(123)

##TRAINING DATA
```

```r
##TRAINING DATA

# Initialize a list to store the models
rf_modelstrain <- list()

# Perform cross-validation
for (i in seq_along(num_trees_values)) {
  print(paste("Iteration:", i, "Number of Trees:", num_trees_values[i]
  # Train the random forest model
  trial_rf_modeltrain <- randomForest(
    x =training_data[, predictor_variables],
    y = as.factor(training_data[, target_variable]),
    ntree = num_trees_values[i],
    mtry = 2,   # Adjust as needed
    type = "classification"
  )

  # Store the model in the list
  rf_modelstrain[[i]] <- trial_rf_modeltrain
}
rf_modelstrain[[1]]
rf_modelstrain[[2]]
rf_modelstrain[[3]]
rf_modelstrain[[4]]
rf_modelstrain[[5]]
rf_modelstrain[[6]]
rf_modelstrain[[7]]
```

```r
# Get feature importance
importance_list <- trial_rf_modeltrain$importance

# Print or plot mean_importance to see the variable importance
print(importance_list)
plot(importance_list)

# Plot variable importance
varImpPlot(rf_modelstrain[[5]])

# Assuming rf_modelstrain[[5]] is your random forest model
importance_values <- rf_modelstrain[[5]]$importance

# Check the structure of importance_values
str(importance_values)

# Extract the importance values and variable names
importance_data <- data.frame(Overall = importance_values[, "MeanDecre

# Ensure that importance values are numeric
importance_data$Overall <- as.numeric(importance_data$Overall)

# Sort the data frame by the Overall column in descending order
sorted_importance <- importance_data[order(importance_data$Overall, de

# Print the variable names
print(rownames(sorted_importance))
```

```r
# Check the structure of sorted_importance
str(sorted_importance)

# Convert 'sorted_importance' to a vector
importance_vector <- sorted_importance$Overall

# Create a bar plot
barplot(importance_vector,
        names.arg = rownames(sorted_importance),
        main = "Variable Importance",
        col = "skyblue",
        cex.names = 0.7,
        las = 2,  # Rotate axis labels vertically
        ylim = c(0, max(importance_vector) * 1.4),# Extend y-axis limi
        ylab = "Mean Decrease Gini")


# Make predictions on the testing data
predictionsrf <- predict(trial_rf_modeltrain, newdata = testing_data)

str(testing_data$GB)
str(predictionsrf)
# Convert GB to a numeric or ordered factor if needed
testing_data$GB <- as.numeric(as.factor(testing_data$GB))

# Convert factor to numeric
predictionsrf_numeric <- as.numeric(as.character(predictionsrf))

```

```r
# Create a confusion matrix
conf_matrixrf <- table(Actual = testing_data$GB, Predicted = predictio
conf_matrixrf

# Calculate accuracy
accuracyrf <- sum(diag(conf_matrixrf)) / sum(conf_matrixrf)
accuracyrf
```

Appendix 5: R-codes for Support Vector Machine Algorithm

```r
getwd()
Data=read.csv("G:\\My Drive\\THESIS Docs\\Thesis\\Depression DataC.csv
attach(Data)
names(Data)

#Find total observations in the data
nrow(Data)
#Summary statistics for variables
summary(Data)
#Find total number of missing values
sum(is.na(Data))
#Removes rows containing missing values
na.omit(Data)
#Data type
class(Data)

library(caret)
library(caTools)
library(MASS)
library(broom)
library(e1071)

#Split into training and testing set (80% for training)
#Setting the seed for reproducibility
set.seed(123)
split <- sample.split(Data, SplitRatio = 0.8)
training_data <- subset(Data, split == TRUE)
testing_data <- subset(Data, split == FALSE)
```

```r
# Specify a range of cost values to consider
cost_values <- c(0.1, 1, 10, 50, 100, 500)

# Create a tuning parameter grid
tune_grid <- expand.grid(C = cost_values)

# Specify gamma for the RBF kernel
gamma_value <- (1/12)

# Convert the dependent variable to a factor if it's not already
Data$GB <- as.factor(Data$GB)

##TRAINING DATA
# Initialize a list to store models
model_listtrain <- list()

# Train SVM models with different values of cost and store in the list
for (cost_value in cost_values) {
  svm_modeltrain <- svm(GB ~ ., data = training_data, kernel = "radial
  model_listtrain[[as.character(cost_value)]] <- svm_modeltrain
}
```

```r
# Run a summary for each model in the list
for (cost_value in cost_values) {
  cat("Summary for model with cost =", cost_value, ":\n")
  print(summary(model_listtrain[[as.character(cost_value)]]))
  cat("\n")
}
```

```r
# Make predictions and calculate metrics for each model in the list on
for (cost_value in cost_values) {
  # Extract the model from the list
  svm_model <- model_list[[as.character(cost_value)]]

  # Make predictions on new_data
  svmpredictions <- predict(svm_model, newdata = testing_data)

  # Calculate the confusion matrix
  svmconfusion_matrix <- table(Actual = testing_data$GB, Predicted = s
  cat("Confusion Matrix for model with cost =", cost_value, ":\n")
  print(svmconfusion_matrix)

  # Calculate accuracy
  svmaccuracy <- sum(diag(svmconfusion_matrix)) / sum(svmconfusion_mat
  cat("Accuracy for model with cost =", cost_value, ":", round(accurac
}
```

Appendix 6: MUT Authorization Letter

MURANG'A UNIVERSITY OF TECHNOLOGY

Office of the Director, Postgraduate Studies

E-mail: bps@mut.ac.ke
Website: www.mut.ac.ke

P.O. Box 75-10200
Murang'a, Kenya

Ref: MUT/ARP/PGS/19/2020 /VOL.1 **Date:** 17th July, 2023

Ms. Wahome Muthoni Loise (AS400/5688/2021)
School of Pure, Applied and Health Sciences,
Murang'a University of Technology,
P.O. Box 75-10,200
Murang'a.

Dear Ms. Muthoni,

RE: APPROVAL OF RESEARCH PROPOSAL AND SUPERVISORS.

I am pleased to inform you that the Directorate of Postgraduate Studies on 3rd July, 2023 considered and approved your Masters research proposal entitled ***"Modeling the Prevalence of Depression among Murang'a University of Technology Students Using Machine Learning Techniques"*** and appointed the following as supervisors:

1. Dr. Euna Nyarige - Machakos University
2. Dr. John Wanjau Mutuguta – Murang'a University of Technology

You may now proceed with your data collection subject to obtaining research permit from NACOSTI, if required. You should also begin consulting your supervisors and submit through them quarterly progress reports to the Director Postgraduate Studies through your CoD and School Dean. Progress Reports forms can be accessed in the Student Portal.

It is the policy and regulations of the University that you observe deadlines. The guidelines on Postgraduate supervision can be accessed in the post graduate Handbook.

Your responsibilities as a student will include, among others;

i. Maintain regular consultation with your supervisor(s), at least once a month
ii. Submit quarterly reports on time, through your supervisors, CoD, Dean and to the Director of Postgraduate Studies;
iii. Ensure quality work all through;

iv. Present your research findings at 2-3 seminars/conferences prior to thesis examination.

v. Publish one article from your research findings in a refereed journal prior to thesis examination.

For any further clarification please contact the undersigned.

Yours Sincerely,

DIRECTORATE OF POSTGRADUATE STUDIES
MURANG'A UNIVERSITY OF TECHNOLOGY
17 JUL 2023
P. O. Box 75 - 10200, MURANG'A
Tel: 0771463515, Email: info@mut.ac.ke

Prof. Geoffrey Muchiri, PhD
Director, Postgraduate Studies
GM/PIM

Copy to:

1. Registrar – Academics and Students Affairs
2. Dean, School of Pure, Applied and Health Sciences

Appendix 7: Data Collection Authorization Letter

MURANG'A UNIVERSITY OF TECHNOLOGY

Office of the Registrar
(Administration & Planning)

Cell: +254-0771370824
E-mail: registrar-ap@mut.ac.ke
Website: www.mut.ac.ke

P.O. Box 75 - 10200
Murang'a, Kenya

REF: MUT/EXT. COMM/REG-A&P/25/2022.VOL.1

DATE: 1ST SEPTEMBER, 2023

Loise Muthoni Wahome
P.O. Box 75-10200
MURANG'A

Phone No: 0799 944755

Dear Ms. Wahome,

RE: REQUEST FOR PERMISSION TO COLLECT DATA

Reference is made to your letter dated 15th August, 2023 on the above subject matter.

Your request to collect data on **"Modelling the prevalence of depression among Murang'a University of Technology students using machine learning techniques"** has been granted.

Kindly treat the information you will obtain in strict confidence and thereafter share the findings of your study with us for information purposes

Yours Sincerely,

DR. PEACE B. AGUFANA, PhD.
REGISTRAR, A&P

MUT IS ISO 9001:2015 CERTIFIED

Appendix 8: NACOSTI Authorization Letter

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **720826**

Date of Issue: **09/August/2023**

RESEARCH LICENSE



This is to Certify that Ms.. Loise Muthoni Wahome of Murang'a University of Technology, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Muranga on the topic: Modeling the prevalence of depression among Murang\'a University of Technology students using machine learning techniques for the period ending : 09/August/2024.

License No: **NACOSTI/P/23/28474**

720826

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix 9: Publication

This publication was published from this thesis.

Loise, W. M., Mutuguta, J., & Nyarige, E. (2024). Application of Random Forest in Modeling the Prevalence of Depression among Murang'a University of Technology Students. *Asian Journal of Probability and Statistics, 26(2), 91-100.*