Original Research Article

APPLICATION OF RANDOM FOREST IN MODELING THE PREVALENCE OF DEPRESSION AMONG MURANG'A UNIVERSITY OF TECHNOLOGY STUDENTS

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 associated stigma often leads to misdiagnosis, posing risks such as disability and suicide. The study employed random forest algorithm to model the prevalence of depression among Murang'a University of technology (MUT) students. A sample of 1448 students from the different schools at the university participated in the study by completing questionnaires on sociodemographic and other factors associated with depression. The questionnaires were administered through social media platforms. Participants were selected using proportionate stratified random sampling and simple random sampling to ensure 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. The confusion matrix criteria were used to assess the performance of random forest in modeling depression prevalence among MUT students. Metrics for random forest included, accuracy (0.9868), sensitivity (0.95), specificity (1.00), positive predictive value (1.00), and negative predictive value (0.9824).Implementing targeted interventions founded on identified risk and protective factors and exploring the long-term outcomes of these interventions would contribute to

KeyWords:*Depression; Machine Learning; Random Forest; Prevalence, mental illness*

1. INTRODUCTION

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[1].

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 [2] 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[3] observed that 24.4% of undergraduate students belonging to countries with low and middle-income economies had depressive illnesses.

In Kenya, [4]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%. [5] 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 [6]. 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[7]. 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.

The goal of this research was to apply the random forest algorithm to model the prevalence of depression among Murang'a University of Technology (MUT) students and assess its efficacy in identifying the factors that cause depression.

1.1 Random Forest Algorithm

Random Forest builds multiple decision trees and ensembles them to have a higher accuracy in prediction than each of the individual trees [8]. A decision tree employs tree-structures [9]. 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 [10]

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.

2. LITERATUREREVIEW

[11] employed machine learning methods to analyze the prevalence and predict features of perceived stress among university students in Bangladesh. A sample of 355 students from 28 universities in Bangladesh participated in the research. The researchers used random forest (RF), support vector

machine (SVM), decision tree (DT) and logistic regression (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 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.

[12] evaluated the efficiency of machine learning methods in forecasting the risk factors affiliated with depression and anxiety among Palestinian school children. RF, artificial neural network (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% of 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.

Five machine learning methods namely; DT, NB, RF, SVM and k-nearest neighbor were used in classifying anxiety, depression and stress in [13]. This study employed 348 participants spanning from 20 to 60 years with diverse responsibilities. Data was collected using the 21-item Depression, Anxiety and Stress Scale (DASS-21). The confusion matrix metrics were used to evaluate the performance of these algorithms. Since the classes were imbalanced, accuracy alone was not a sufficient metric in identifying the optimal model. Using the F1 score, the RF algorithm was identified as the best model.

[14] utilized RF, DT, XGBoost (XGB) and Gaussian Naïve Bayes (GNB) in detecting depression among children and adolescents aged 4 and 17 years. A sample of 6310 children was used in the study out of which 5839 were not depressed and 471 were depressed. Using the confusion matrix parameters, RF outperformed the other three models with an accuracy (95%), precision (99%), specificity (100%) and sensitivity (44%).

[15] employed the Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest (RF) models to analyze depression within a sample of 7,224,620 patients under 65 years from the MarketScan commercial claims. For training the models, 75% of the dataset was used while 25% was used for testing the models. Factors linked to depression included mental health-related medications and diagnoses, female gender, chronic diseases in females, and other conditions such as joint disorders. The AUC scores for LASSO and RF were 0.75 and 0.76 respectively. The two models demonstrated effective performance in modeling depression within the MarketScan commercial claims dataset.

3. METHODOLOGY

The study population consisted of 10,127 students from the 7 schools at Murang'a University of Technology. Participants were randomly selected using the proportionate stratified sampling technique 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. Prior to commencing the study, the researcher sought ethical clearance from National Commission for Science, Technology & Innovation (NACOSTI). Informed consent was gathered from the participants by ensuring that the participants were aware of the aim of the research and that 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.

The respondents completed a questionnaire which comprised of sociodemographic characteristics

including gender, age, level of education, year of study, school and sponsorship. The questionnaire also enquired about academic performance, financial situation, social support network, drug and substance abuse and past abuse, trauma and neglect. Depression was measured using the Patient Health Questionnaire (PHQ-9). 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 27 being the highest score[16]. A cut-off point of 10 or greater was used for the screening of depressive symptoms [17].

3.1 Fitting Random Forest Algorithm.

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. A large number of trees was required for greater accuracy and prevent overfitting. The trees were trained using the bagging method which ensured that each decision tree contains slightly different data and was trained slightly differently [18].

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.

RF is a combination of several decision trees and its structure is as shown below:

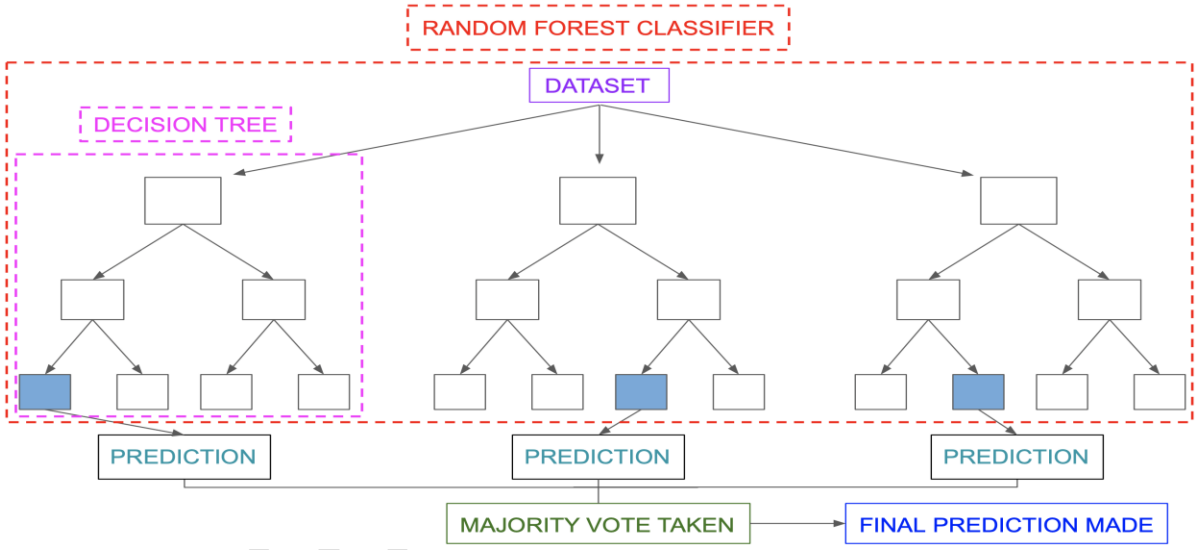

Figure 1: Structure of a random forest

Source: [10]

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 [19]. The larger decline in mean decrease Gini indicates the more essential the explanatory variable is in making accurate predictions.

4. RESULTSANDDISCUSSIONS

Data was gathered from 1448 students (512 female and 936 male). More than half 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 88.66% 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.

Table 1:Relationship between Sociodemographic Variables and Depressive Symptoms

Variable	Category	Total		Depressive Symptoms			
				No		Yes	
		n	%	n	%	n	%
All Students		1448		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 Sponsored	1340	92.54%	996	74.33%	344	25.67%
	Self-Sponsored	108	7.46%	76	70.37%	32	29.63%
Level of Education	Postgraduate	44	3.04%	28	63.64%	16	36.36%
	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%

Key: ***SAES – School of Agriculture and Environmental Sciences***

SBE- School of Business and Economics

SCIT – School of Computing and Information Technology

SEHSS – School of Education, Humanities and Social Sciences

SET – School of Engineering and Technology

SHTM- School of Hospitality and Tourism Management

SPAHS – School of Pure, Applied and Health Sciences

Table 1 above shows the relationship between sociodemographic variables and depressive symptoms among MUT students. 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. In total 34.38% (176) females and 21.37% (200) males had depressive symptoms. Depression levels varied among the 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.

4.1 Random Forest Algorithm

The research employed the R programming language within the RStudio integrated development environment (IDE) to implement and analyze the random forest algorithm. Random forest was fit in R using the randomForest package.

The random forest model was employed to model the predictor variables associated with the dependent variable, depression. The study used 80% of the data for training the model and 20% of the data was used for validating the model. Utilizing more data points during the training process enhances the accuracy of the predictions [20]. Table 2 below shows the results of the training parameters for RF model fit on the training data. Training data was employed to teach the machine learning model in recognizing patterns and establishing relationships between input features and corresponding output labels while testing data was used to assess the performance of a machine learning model by evaluating its prediction on new data.

Table 2:Random Forest 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 hyperparameter mtry controls 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 [21]. This 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.

The parameter ntree refers to the number of decision trees created in the RF model. Increasing the number of trees improves the performance of the RF model up to a certain point. Hence, it is the responsibility of the researcher to determine the number of trees to be used in the RF model [22] in order to provide the best balance between accuracy and efficiency.

The Out-Of-Bag (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 serve as a validation set for the RF algorithm. In the training process, additional trees are incorporated until the OOB error rate reaches a point of stability [23]. A lower OOB error rate indicates better predictive performance of RF on unseen data. The OOB error rate initially decreases, stabilizes and then starts to increase with increasing number of decision trees. The sudden decrease and increase in the OOB error rate when the number of trees is increased beyond the optimal point makes the RF model start to overfit the data making it less effective at generalizing to new data. 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 is making 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.

When using RF for classification tasks, the summary includes a confusion matrix. Including the confusion matrix in the RF model summary provides a concise summary of how well the model is performing on different classes. The confusion matrix gives insights into the model's ability to correctly classify instances and identify any patterns of misclassification. 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.

4.1.1 Variable importance measure for the RF model

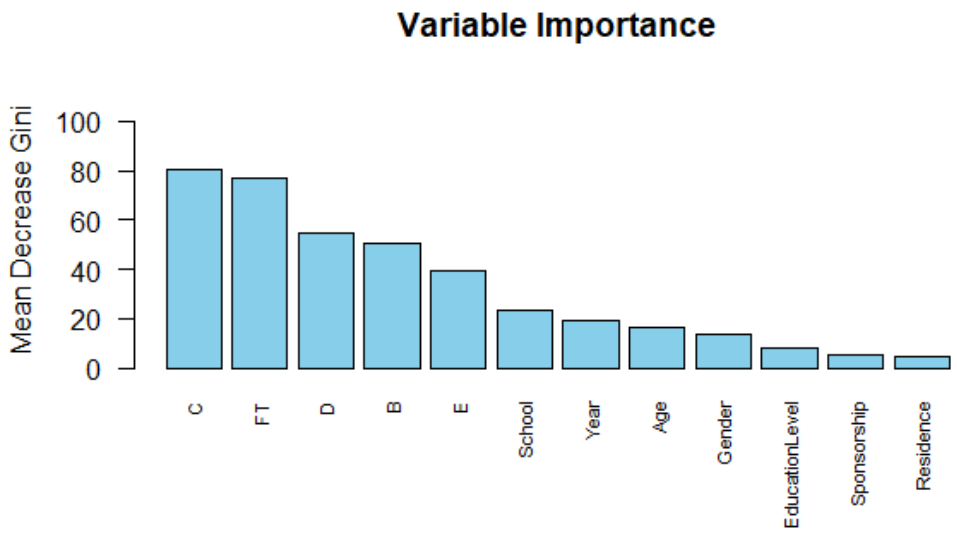

Figure 1: Variable importance for the random forest model

Source: [Author]

From figure 2 above, B represents the predictor variable academic performance, C represents social support network, D represents financial situation, E represents drug and substance abuse and FT represents the predictor variable past abuse, trauma and neglect.

Variable importance measures the contribution of each predictor variable to the predictive accuracy of the model. Figure 2 shows 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 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.

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.

Table 3:Confusion Matrix for the Random Forest Model

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

$$Accuracy = \frac{223 + 76}{223 + 76 + 0 + 4} = 0.9868$$

$$Sensitivity\ (Recall) = \frac{76}{76 + 4} = 0.95$$

$$Specificity = \frac{223}{223 + 0} = 1$$

$$Positive\ predictive\ value = \frac{76}{76+0} = 1$$

$$Negativepredictivevalue = \frac{223}{223+4} = 0.9824$$

RF had an accuracy of 98.68% in predicting depression among MUT students. RF had a sensitivity of 95% in identifying depressed cases among the students. RF had a specificity of 100% in identifying the non-depressed cases. The random forest model correctly identified all (100%) of the real depressed cases (positives). In identifying the real non-depressed cases (negatives) RF correctly identified 98.24% of the real non-depressed cases.

5. CONCLUSION

The findings revealed a depression prevalence of 25.97%, with a higher incidence among female students, those residing off-campus, and those in their third or fourth year 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.

RF showcased remarkable efficiency in classifying depression instances based on a diverse set of features with accuracy of 98.68%, sensitivity of 95%, specificity of 100%, positive predictive value of 100% and a negative predictive value of 100%. The study's thorough analysis, supported by a rich training dataset, contributes valuable insights to the understanding of mental health within the university population.

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.

Ethical Approval:

As per international standards or university standards written ethical approval has been collected and preserved by the author(s).

Consent

As per international standards or university standards, Participants' written consent has been collected and preserved by the author(s).

REFERENCES

[1] World Health Organization (WHO), "Depression."

[2] W. Li, Z. Zhao, D. Chen, Y. Peng, and Z. Lu, "Prevalence and associated factors of depression and anxiety symptoms among college students: a systematic review and meta-analysis," *J Child Psychol Psychiatry*, vol. 63, no. 11, pp. 1222–1230, Nov. 2022, doi: 10.1111/JCPP.13606.

[3] P. Akhtar, L. Ma, A. Waqas, S. Naveed, … Y. L.-J. of A., and undefined 2020, "Prevalence of depression among university students in low and middle income countries (LMICs): a systematic review and meta-analysis," *Elsevier*, Accessed: Jan. 30, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032719332677

[4] A. Mbwayo, M. Kiarie, and J. Ndegwa, "Factors Related to Depression among University Students in Nairobi County, Kenya," 2022, Accessed: Jan. 30, 2024. [Online]. Available: http://repository.daystar.ac.ke/handle/123456789/3976

[5] S. Muriungi, D. N.-S. A. J. of Psychiatry, and undefined 2013, "Effectiveness of psycho-education on depression, hopelessness, suicidality, anxiety and substance use among basic diploma students at Kenya Medical Training," *ajol.infoSK Muriungi, DM NdeteiSouth African Journal of Psychiatry, 2013•ajol.info*, Accessed: Jan. 30, 2024. [Online]. Available: https://www.ajol.info/index.php/sajpsyc/article/view/89883

[6] A. O. Adewuya*et al.*, "Depression amongst Nigerian university students: prevalence and sociodemographic correlates," *SpringerAOAdewuya, BA Ola, OO Aloba, BM Mapayi, OO OginniSocial psychiatry and psychiatric epidemiology, 2006•Springer*, vol. 41, no. 8, pp. 674–678, Aug. 2006, doi: 10.1007/s00127-006-0068-9.

[7] C. Lund, C. Brooke-Sumner, … F. B.-T. L., and undefined 2018, "Social determinants of mental disorders and the Sustainable Development Goals: a systematic review of reviews," *thelancet.comC Lund, C Brooke-Sumner, F Baingana, EC Baron, E Breuer, P Chandra, J HaushoferThe Lancet Psychiatry, 2018•thelancet.com*, Accessed: Jan. 30, 2024. [Online]. Available: https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(18)30060-9/fulltext

[8] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[9] J. Wakiru, "A decision tree-based classification framework for used oil analysis applying random forest

feature selection," *PP 90-Journal of Applied Sciences, Engineering and Technology for Development JASETD*, vol. 3, no. 1, p. 90, 2018, Accessed: Jan. 30, 2024. [Online]. Available: http://41.89.227.156:8080/xmlui/handle/123456789/748

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

[11] R. Rois, M. Ray, A. Rahman, and S. K. Roy, "Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms," *J Health PopulNutr*, vol. 40, no. 1, Dec. 2021, doi: 10.1186/S41043-021-00276-5.

[12] R. Qasrawi, S. P. Vicuna Polo, D. Abu Al-Halawa, S. Hallaq, and Z. Abdeen, "Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis.," *JMIR Form Res*, vol. 6, no. 8, p. e32736, Aug. 2022, doi: 10.2196/32736.

[13] A. Priya, S. Garg, N. T.-P. C. Science, and undefined 2020, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Elsevier*, Accessed: Feb. 06, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920309091

[14] U. M. Haque, E. Kabir, and R. Khanam, "Detection of child depression using machine learning methods," *PLoS One*, vol. 16, no. 12 December 2021, Dec. 2021.

[15] R. Qiu, V. Kodali, M. Homer, … A. H.-… on B. and, and undefined 2019, "Predictive Modeling of Depression with a Large Claim Dataset," *ieeexplore.ieee.orgR Qiu, V Kodali, M Homer, A Heath, Z Wu, Y Jia2019 IEEE International Conference on Bioinformatics and, 2019•ieeexplore.ieee.org*, Accessed: Feb. 06, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8982975/

[16] R. Spitzer, J. Williams, and K. Kroenke, "Test review: patient health questionnaire–9 (PHQ-9)," *Rehabil Couns Bull*, vol. 57, no. 4, pp. 246–248, 2014.

[17] K. Kroenke, R. Spitzer, J. W.-C. D. and Ethnic, and undefined 1999, "Patient health questionnaire-9," *psycnet.apa.orgK Kroenke, RL Spitzer, JBW WilliamsCultural Diversity and Ethnic Minority Psychology, 1999•psycnet.apa.org*, Accessed: Jan. 30, 2024. [Online]. Available: https://psycnet.apa.org/getdoi.cfm?doi=10.1037/t06165-000

[18] H. Sayed, A. William, A. S.- Electronics, and undefined 2023, "Smart Electricity Meter Load Prediction in Dubai Using MLR, ANN, RF, and ARIMA," *mdpi.comHA Sayed, A William, AM SaidElectronics, 2023•mdpi.com*, Accessed: Jan. 30, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/12/2/389

[19] H. Han, X. Guo, H. Y.-2016 7th ieee international conference, and undefined 2016, "Variable selection using mean decrease accuracy and mean decrease gini based on random forest," *ieeexplore.ieee.orgH Han, X Guo, H Yu2016 7th ieee international conference on software engineering and, 2016•ieeexplore.ieee.org*, Accessed: Jan. 30, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7883053/

[20] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation," 2018, Accessed: Jan. 30, 2024. [Online]. Available: https://scholarworks.utep.edu/cs_techrep/1209/

[21] S. R.-J. of I. Medicine and undefined 2017, "Random forest," *meridian.allenpress.comSJ RigattiJournal of Insurance Medicine, 2017•meridian.allenpress.com*, Accessed: Jan. 30, 2024. [Online]. Available: https://meridian.allenpress.com/jim/article-abstract/47/1/31/131479

[22] P. Probst, A. B.-J. of M. L. Research, and undefined 2018, "To tune or not to tune the number of trees in random forest," *jmlr.orgP Probst, AL BoulesteixJournal of Machine Learning Research, 2018•jmlr.org*, vol. 18, pp. 1–18, 2018, Accessed: Jan. 30, 2024. [Online]. Available: https://www.jmlr.org/papers/v18/17-269.html

[23] G. Dudek, "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. Springer International Publishing*, vol. 2, pp. 821–828, 2015.