Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 18 No. 14 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijim.v18i14.48669

PAPER

Machine Learning Models to Classify and Predict Depression in College Students

Orlando Iparraguirre-Villanueva[1](✉), Cleoge Paulino-Moreno[2], Andrés Epifanía-Huerta[1], Carmen Torres-Ceclén[3]

[1]Universidad Tecnológica del Perú, Lima, Perú

[2]Universidad católica de Trujillo, Trujillo, Perú

[3]Universidad Católica los Ángeles de Chimbote, Chimbote, Perú

oiparraguirre@ieee.org

ABSTRACT

Depression is an increasingly common mental health condition worldwide and is influenced by various factors such as anxiety, frustration, obesity, medical issues, etc. In severe cases, it can even result in suicide. This study aimed to utilize machine learning (ML) models to categorize and forecast student depression. The research involved analyzing a dataset of 787 college students through a series of steps, including cleansing, model training, and testing using techniques to classify and predict student depression. Three ML models were employed: logistic regression (LR), K-nearest neighbor (KNN), and decision tree (DT). The findings revealed that the LR model achieved the highest accuracy in prediction, with a rate of 77%, 70% recall, and 72% F1 score. Moreover, the study highlighted that two out of five students experience mild depression, around 90% of depressed students do not seek treatment, obese students are 2.5 times more prone to depression, male students are twice as likely to be obese, and male students generally have a higher body mass index (BMI) compared to female students. The study concludes that integrating ML models into the triggers that lead to depression among students.

KEYWORDS

classification, prediction, depression, machine learning (ML), students

1 INTRODUCTION

Depression is an increasingly common mental illness worldwide, affecting an estimated 5.7% of the adult population [1]. Worldwide, approximately 280 million people have depression [2]. Depression is caused by various factors, such as anxiety, frustration, and medical problems [3], [4]. This illness can become a severe health issue, especially when it is recurrent and of moderate to severe intensity, causing suffering to the affected individual and disrupting their daily lives [5], [6]. In the worst-case scenarios, it can lead to suicide [7]. In recent years, suicide has become one of the leading causes of death, with rates on the rise [8]. Over 700,000 people commit suicide every year [9]. Young people in developing countries are particularly

Iparraguirre-Villanueva, O., Paulino-Moreno, C., Epifanía-Huerta, A., Torres-Ceclén, C. (2024). Machine Learning Models to Classify and Predict Depression in College Students. *International Journal of Interactive Mobile Technologies (iJIM)*, 18(14), pp. 148–163. https://doi.org/10.3991/ijim.v18i14.48669

Article submitted 2024-02-20. Revision uploaded 2024-04-18. Final acceptance 2024-04-18.

vulnerable to depression due to the social, economic, educational, and health challenges faced by these nations [10], [11].

The world is constantly changing, and technological development has played an important role in creating new human capabilities [12]. This has led people to work longer hours to keep up with the fast pace of the world, exposing them to physical and mental stress that impacts their health [13]. According to the World Health Organization (WHO) [14], depression rates are increasing globally. For instance, in 2022, Ukraine reported the depression rate at 6.3%, followed by the United States at 5.9%, Australia at 5.9%, Finland at 5.6%, Estonia at 5.9%, Greece at 5.7%, Brazil at 5.8%, Portugal at 5.7%, Belarus at 5.6%, and Peru at 4.8%. As illustrated in Figure 1, countries with the lowest rates of spread are spread across almost all continents, with the Pacific Island region of Oceania having the lowest rates of depression.

Peruvian university students face various personal, academic, economic, health, transportation, and other daily challenges [15]. Another significant factor contributing to depression among university students is the return to face-to-face classes. Throughout the COVID-19 pandemic, many students have experienced severe financial strain on their families [16], making it difficult for them to afford the costs associated with attending college [17]. In Peru, approximately 30% of university students exhibit symptoms of depression, including sadness, irritability, feelings of emptiness, anxiety, guilt, low self-esteem, a lack of hope for the future, thoughts of death or suicide, and a lack of energy [18]. Social media has transformed the way we communicate and has become an essential part of many people's lives [19]. Individuals with depression often have suicidal tendencies and may turn to social networks to express their emotions [20], [21]. However, excessive use of social media among young people is linked to higher rates of anxiety, depression, and sleep disturbances. Modern machine learning (ML) models and techniques are increasingly used for predicting suicide risk. While traditional statistical methods are still utilized, ML models incorporate a wider range of variables for risk assessment, leading to improved accuracy [22]. These prediction systems leverage artificial intelligence (AI) to make highly precise predictions for new scenarios [23]. The dataset for this study comprises eighteen attributes and includes information from 787 undergraduate students at Lahore University in 2021. The study aims to classify and predict student depression using ML techniques, employing cross-validation to determine the most effective classifier among logistic regression (LR), K-nearest neighbor (KNN), and decision tree (DT) models.

2 PREVIOUS STUDIES

The National Institute of Mental Health (NIMH), the Pan American Health Organization (PAHO), researchers, and academics have published work related to mental health. For example, [24] developed a DT-based model to predict the risk of depressive disorder in students. Additionally, the authors in [25] worked on a ML model to find a better classifier and thus achieve improved results in model accuracy. For this purpose, they utilized various models such as K-NN, RF, multilayer perceptron (MLP), support vector machine (SVM), and fuzzy logic, achieving performance levels of 100%, respectively. Moreover, in the paper [26], a ML model was developed to detect and predict depression in students aged 4 to 17. The factors contributing to depression are diverse, ranging from a lack of social support to financial problems and the learning environment. Similarly, the study [27] examined mental health problems among students. They also analyzed ML models to predict mental health, with the SVM model standing out as the most popular and accurate, achieving accuracies between 70% and 96%. Likewise, in the study [28], five ML techniques were analyzed: RF, neural network, DT, SVM, and Naïve Bayes (NB), with SVM and RF yielding the best results in predicting depression. Currently, ML is widely used to predict emotions and psychological characteristics, from design to implementation. In this context, the study [29] designed a prediction model using LR, NB, RF, DT, and K-NN models to identify major disorders such as low self-esteem, Internet addiction, and depression. Technological advancements in recent years have contributed to the health sector with various techniques and tools that have improved results and predictions significantly [29]. Consequently, different ML algorithms have been analyzed to classify and detect depression in students, with the SVM algorithm identified as the most efficient and yielding the best results for detecting depression. Early detection of depression is crucial, as it aids in prevention and evaluation.

In this context, the study [30] utilized ML models with 26 predictor variables to predict suicide risk in Korean adolescents. They employed models such as LR, RF, SVM, ANN, and extreme gradient boosting. The findings revealed that 12.4% of adolescents had a history of suicide due to depression. The extreme gradient boosting model performed the best with 79%, followed by SVM at 78.7%, LR at 77.9%, RF at 77.8%, and ANN at 77.5%. ML models have significantly contributed to the advancement of automated diagnostic methods in various pathologies. For instance, in a study conducted with ML [31] based on brain connectivity, a set of brain imaging data was used for depression detection. The results provided by most ML models lack explicit explanations for individuals, making the predictions difficult to comprehend. In the paper [32], a study was conducted to predict multi-aspect features using deep learning models. The results indicated that the NB model achieved the lowest F1 score, while the multimodal learning model outperformed SVM, NB, and BiGRU. As this model is specifically designed to identify depressed users, the proposed hybrid model based on MLP and CNN (MDHAN) demonstrated the best performance with an 89% F1 score, suggesting that HAN with a multi-aspect strategy achieves significant accuracy in detecting depression. Additionally, in [33], they developed an ML-based model to classify patients based on socio-demographic information, personality traits, and mood, concluding that LR models and the Elastic Net method yielded results with accuracies of 84% and 80%, respectively.

To predict depression using ML models, many challenges must be addressed, such as data quality, character identification, interpretability, ethical and privacy concerns, and individual and temporal variability. When using mental health data to improve accuracy and adaptability, it is necessary to address these challenges to protect data privacy.

3 METHODOLOGY

This section describes the ML models that will be used, as well as the case study that will be developed to address the classification and prediction of depression in college students.

3.1 Logistic regression

The reinforcement learning (LR) model is used for classification and predictive analysis. It is also often used to attempt to correlate the probability of an event occurring [34]. The concept is that the LR model estimates the probability of yielding zero when no event occurs or one when the event occurs based on the explanatory variable's value [35]. In LR, a logit transformation is applied to the probabilities, where the probability of success is divided by the probability of failure [36]. This is refered to as logarithmic probabilities and is represented by equation (1).

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{e^{(\beta_0 + \beta_1 x)} + 1} = \frac{1}{e^{-(\beta_0 + \beta_1 x)} + 1} \tag{1}$$

Equation (1) shows the independent variable with the combinations $\beta_0 + \beta_1 X$, and the dependent variable is the estimated probability $\pi(x)$. The LR model in ML falls under the supervised learning category. In this context, the LR model utilizes the negative log-likelihood as the loss function and employs the gradient descent process to determine the global maximum and obtain the estimates [36].

3.2 K-nearest neighbor

The K-NN algorithm is a nonparametric supervised learning classifier that uses proximity to make classifications or predictions [37]. The algorithm stores the attribute vectors and labels during its training phase for retraining [38]. During classification, K is defined as a user-defined variable, and the unlabeled vector is classified by setting a label among the training attributes deemed most relevant [39]. The Euclidean distance is used for distance metrics for continuous variables, limited to vectors of real values. Equation (2) is utilized for this purpose, while the superposition metric is used for creating variables [40].

$$\mathrm{d}(x, y) = \sqrt{\sum_{i=1}^{n} \left(y_i - x_i\right)^2} \tag{2}$$

The K-NN algorithm is applied in ML for various purposes, primarily classification and prediction. For instance, in data processing, the algorithm is utilized to estimate values; in recommendation engines, it provides automatic recommendations; in finance, it is applied to credit data for risk evaluation and analyzing economic trends; in healthcare, it plays a crucial role in predicting the risk of heart attacks and prostate cancer [41]. Additionally, it is employed in pattern recognition to assist in classifying text and images.

3.3 Decision tree

As a nonparametric regression and classification algorithm, decision trees (DT) are used to predict the attributes of discrete and continuous variables [42]. DT models are constructed similarly to flowcharts, where each node represents an attribute and each branch represents an output. There are three types of nodes in the DT [43]: decision nodes, probability nodes, and end nodes. Decision nodes are usually represented by boxes, probability nodes by circles, and end nodes by triangles, as shown in Figure 1.

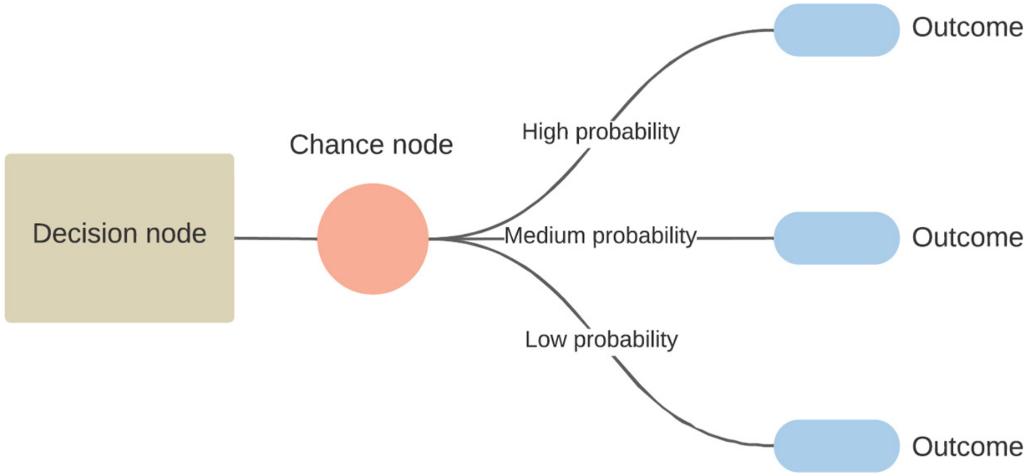

Fig. 1. Decision tree

3.4 Understanding data

The dataset used for processing was obtained from the Kaggle repository and consists of 787 records of university students. A study was conducted using the patient health questionnaire (PHQ) to assess the severity of mental health problems (0–4: not minimal or normal; 5–9: mild; 10–14: moderate; 15–19: moderate–severe; 20–24: severe). The variables analyzed include: php_score: score from the PHQ questionnaire depression severity: an estimate derived from php_score depressiveness: indicating whether the participant has depression diagnosis: indicating if the participant has received a diagnosis from a mental health specialist depression treatment: whether the participants is undergoing treatment for depression suicidal: probability of suicidal tendencies GAD score: score from an anxiety disorder assessment anxiety severity: estimate based on the GAD score anxiety diagnosis; diagnosis by a mental health specialist anxiety treatment: whether the participant is receiving treatment for anxiety Epworth score: score from a test measuring sleepiness: likelihood of daytime sleepiness BMI: body mass index who_bmp: WHO body mass index age: participant's age (between 18 and 31 years) gender: participant's sex: academic year, school year. K-nearest neighbor.

3.5 Data cleaning

This section begins with data cleaning. For instance, the 'dropna()' function is used to remove rows with missing values, while the 'Epworth()', 'loc()', and 'drop()' functions are employed to handle out-of-range data, process incorrect data, manage outliers, and eliminate duplicate data. Subsequently, the tested variables are displayed in Table 1, showcasing the organized variables for which the data type had to be transformed into categories. Also, the number of categories has been reduced, and functions have been defined to sort the categories. Finally, with the 'reset_index()' function, a stratified sample was obtained from the original sample.

Table 1. Data type checking

#	Column	Non-Null Count	Dtype
0	school_year	757 [Not empty]	category
1	age	757 [Not empty]	int64
2	gender	757 [Not empty]	category
3	bmi	757 [Not empty]	float64
4	who_bmi	757 [Not empty]	category
5	phq_score	757 [Not empty]	int6
6	depression_severity	757 [Not empty]	category
7	depressiveness	757 [Not empty]	object
8	suicidal	757 [Not empty]	object
9	depression_diagnosis	757 [Not empty]	object
10	depression_treatment	757 [Not empty]	object
11	gad_score	757 [Not empty]	int64
12	anxiety_severity	757 [Not empty]	category
13	anxiousness	757 [Not empty]	object
14	anxiety_diagnosis	757 [Not empty]	object
Dtypes: category (5), float64 (2), int64 (3), object (8)			

3.6 Exploratory data analysis

In this phase, the tags are first converted into a numerical format and then transformed into a format understandable by the algorithm. A numerical value is assigned to each categorical value. ML algorithms can best determine how tags should be handled. This step is of paramount importance for data processing and is carried out using the sklearn(), pandas(), and NumPy library().

Character extraction is performed utilizing the principal component analysis (PCA) technique, using all the features provided by the dataset to predict, analyze, classify, or group outcomes. The characteristics are closely related to the dimensions. This selection considers a subset of features that are important and eliminates those that do not contribute to classification. Feature extraction enables the creation of valuable information from raw data by combining and transforming core features into new ones until a new dataset can be used by ML models. The distribution of numerical variables is a crucial factor in performing arithmetic operations. Using the ggplot2() function, we scanned to identify the variables and created a chart as shown in Figure 2. In this case, the function identified the variables: age, body mass index (BMI), the score to determine the severity of mental health problems (phq_score), the rate of a generalized anxiety disorder (gad_scor), and the level of daytime drowsiness (Epworth_score).

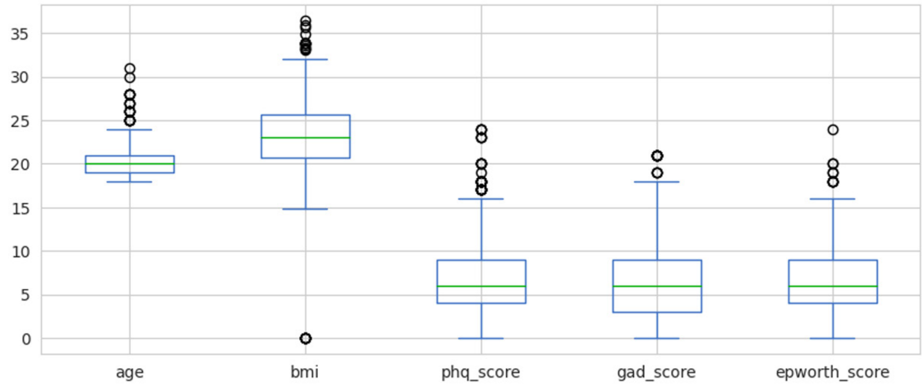

Fig. 2. Distributions of numeric variables

Figure 2 indicates that the median BMI is 25, suggesting that most participants have a normal weight. However, there is a wide range of BMI values, with some participants being overweight and others being lean. Additionally, it shows that the PHQ score is 8, suggesting that the majority of participants do not exhibit symptoms of depression. The median GAD score is 5, indicating that most participants do not have anxiety. Nevertheless, there are some participants with elevated scores that may suggest anxiety. Similarly, the median Epworth score is 8, indicating that most participants do not experience excessive daytime sleepiness.

The principal component analysis (PCA) helps identify the correlation between numerical variables and characteristics. The variables in Figure 3 depict the pairwise correlations within the set of variables. For instance, colors such as red, orange, and yellow signify a positive correlation, whereas colors like blue, green, and brown represent negative correlations. The intensity of the color reflects the strength of the correlation.

It is important to specify that the values in a correlation matrix range from –1 to 1. A value of zero indicates no correlation between the variables. A value of 1 indicates a perfect positive correlation, while –1 indicates a perfect negative correlation. For instance, in Figure 3's matrix, it is evident that there is no significant correlation between BMI, PHQ score, and GAD. There is a weak correlation with the PHQ score of 0.16, suggesting that students with a higher BMI tend to have higher PHQ scores, indicating symptoms of depression. A moderately positive correlation with Epworth of 0.36 implies that students with higher PHQ scores exhibit more symptoms of depression and tend to experience more excessive daytime drowsiness. The matrix also reveals a strong positive correlation with the GAD score, indicating that students with higher PHQ scores have more symptoms of depression and tend to experience more anxiety symptoms.

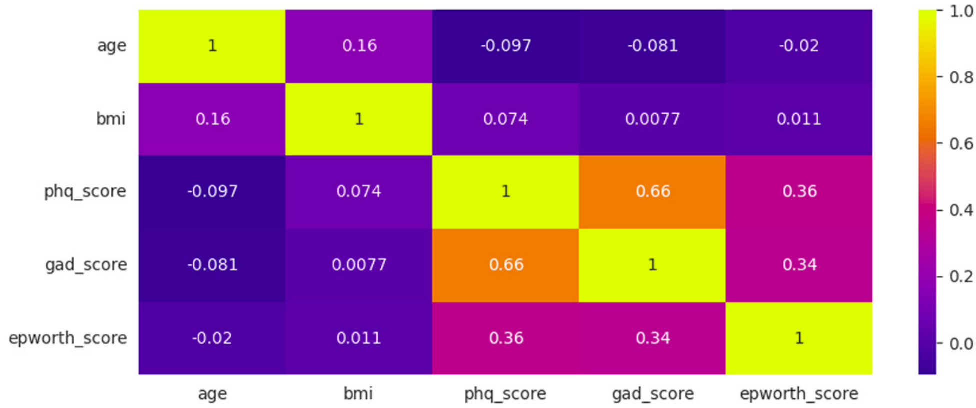

Fig. 3. Correlation matrix between numeric variables

In this work, feature scaling was performed using the standardization technique, which allows the features to be scaled so that the values are centered around the mean with a unit standard deviation. This method is widely used in ML algorithms such as SVM, LR, and neural networks. However, algorithms such as LR, K-NN, and SVM require the features to be normalized. The dataset used has heterogeneous features at each scale due to the different properties they measure. The transformed data is then used for training with functions such as StandarScaler(), which removes the mean and scales each feature to unit variance. However, it can be influenced by outliers. Additionally, there are functions such as scaler.fit_transform(x_train) and scaler.transform(x_test), as shown in Table 2.

Table 2. Function scaling – standardization

Matrix ([		
[–1.25230037,	–0.51501498,	0.01682768],
[0.38909111,	–0.31446922,	0.65269925],
[0.93622161,	–0.2234338,	0.86465644],
…,		
[0.38909111,	–1.11952747,	–1.46687265],
[–0.15803938,	0.27208495,	–0.83100108],
[1.4833521,	–0.8437498,	–0.4070867]
		])

After conducting the EDA, it has allowed us to obtain statistics. For example, the body mass index of a male student is higher than that of a female student; about 26% of students are obese; male students are twice as likely to have obesity and to suffer from severe depression as female students. Female students are twice as likely as male students to experience severe anxiety. Twenty-five percent of female students are more likely to be depressed than male students. 25% of depressed students receiving treatment are suicidal. Students who are depressed have a higher BMI than those who are not depressed. Obsessed students are 2.5 times more likely to be depressed than normal students. Approximately 90% of depressed students do not receive treatment, and finally, 2 out of 5 students have mild depression. Figure 4 shows the proportion of students with severe depression. According to the analysis, female students are more likely to have severe depression; meanwhile, male students are more likely to have moderate to severe depression; likewise, moderate and mild depression. However, female students are more likely to have no depression at all.

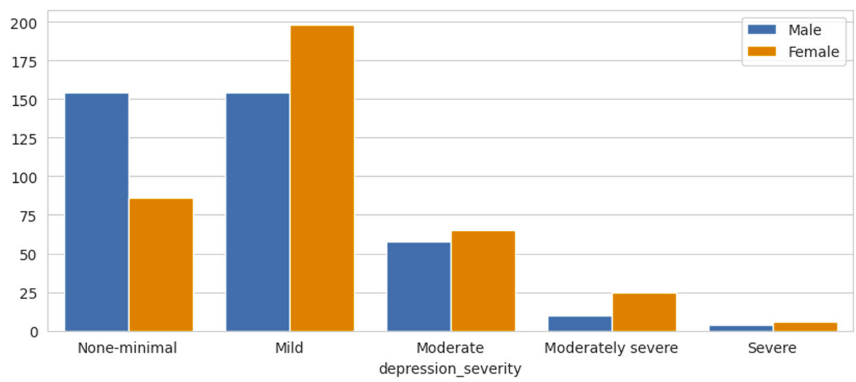

Fig. 4. Proportion of students with severe depression

3.7 Model training and testing

After data processing, in this section, training and testing are performed to obtain the best classifier using the cross-validation technique. Additionally, model fitting is executed using the pipeline, and finally, the model's performance is evaluated. The aim of this study is to determine the most effective classifier among the LR, K-NN, and DT models for predicting depression in students. To achieve this, the cross-validation technique was utilized to assess the model and its performance. This technique involves dividing the dataset into two parts: one for training (80%) and one for testing (20%). Furthermore, the technique involves training the model on the training set, validating it on the test set, and storing the validation results. The following functions were employed for training: logisticregression(), Kneighborsclassifier(), decisiontreeclassifier(), Kfold(), cross_val_score(), and boxplot(), which enabled the acquisition of the results depicted in Figure 5.

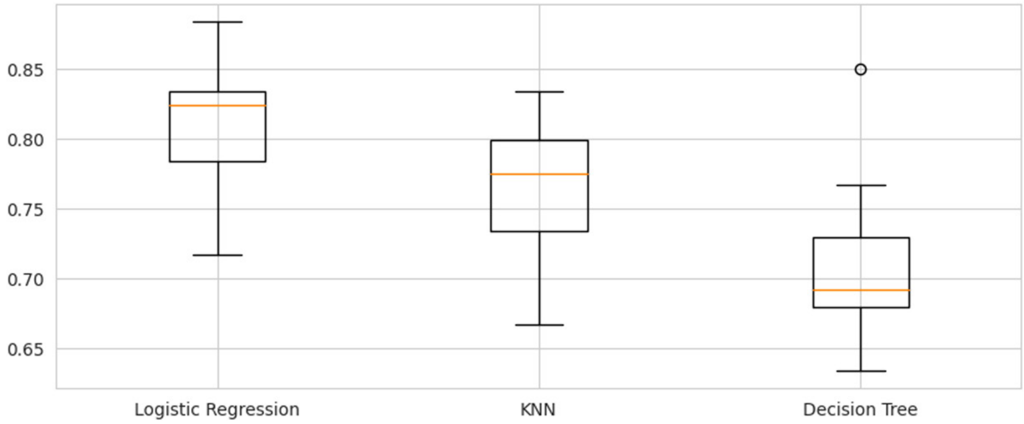

Fig. 5. Cross-validation to obtain the best model

4 RESULTS

This section presents the results of the classification, training, and exploration phases. After processing to rank the best model for prediction, Figure 5 indicates that the LR algorithm achieved the best result with the highest accuracy rate, exceeding 85%. Consequently, the prediction in this study is conducted using this model. Table 3 displays the accuracy index of the three models. Additionally, in Figure 6, the confusion matrix was utilized to assess the classification of the selected ML-based model. The values on the 91st and 26th diagonals represent the correctly estimated values by the model, including true positives and true negatives. Conversely, the values on the second diagonal indicate the model's errors, with 14 false negatives and 19 false positives. The accuracy rate achieved with the LR model is 81%, falling within an acceptable range with a precision of 71%. Hence, it can be concluded that the model is more accurate than precise. Sensitivity, which is the ratio of positive cases identified by the model to the total number of positives, is reported at 48%, indicating the model's ability to detect relevant cases.

Table 3. Data type checking

Accuracy		Accuracy	Recall	F1-Score	Support
LR	0	0.83	0.93	0.88	439
	1	0.71	0.48	0.57	159
	Accuracy			0.81	159
	Average Macro	0.77	0.70	0.72	159
	Weighted	0.80	0.81	0.80	159
KNN	0	0.81	0.90	0.85	439
	1	0.60	0.41	0.49	159
	Accuracy			0.77	598
	Average Macro	0.70	0.65	0.67	598
	Weighted	0.75	0.77	0.75	598
DT	0	0.80	0.79	0.79	439
	1	0.44	0.45	0.44	159
	Accuracy			0.70	598
	Average Macro	0.62	0.62	0.62	598
	Weighted	0.70	0.70	0.70	598

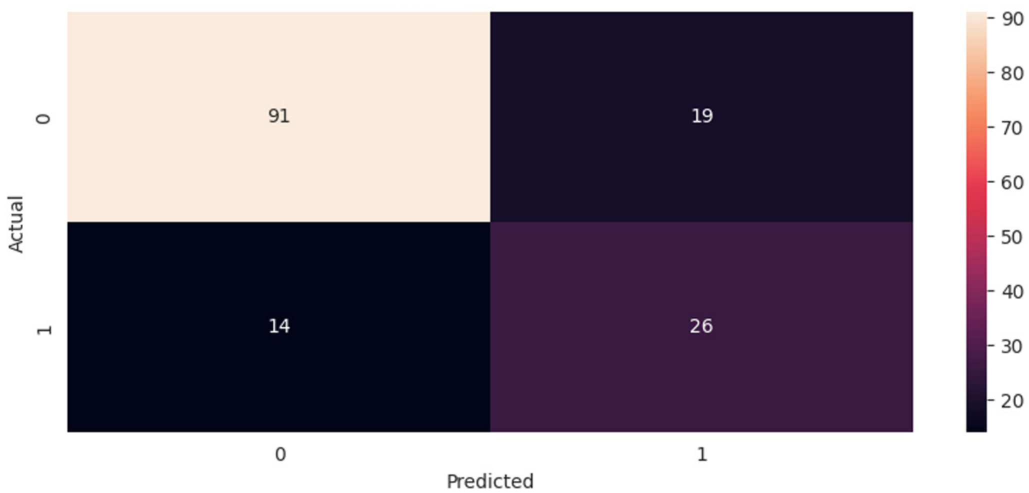

Fig. 6. Matrix of confusion

The ROC curve is used to evaluate the ability of a model to discriminate between two classes. For example, Figure 7 presents the ROC curve of the selected model, showing the relationship between the false positive rate and the true positive rate. The area under the curve (AUC) is 0.74, indicating that the LR model has a good ability to discriminate between positive and negative cases of depression. It is important to note that the ROC curve is not a perfect measure of model accuracy. The performance of the model may vary depending on the size of the dataset used.

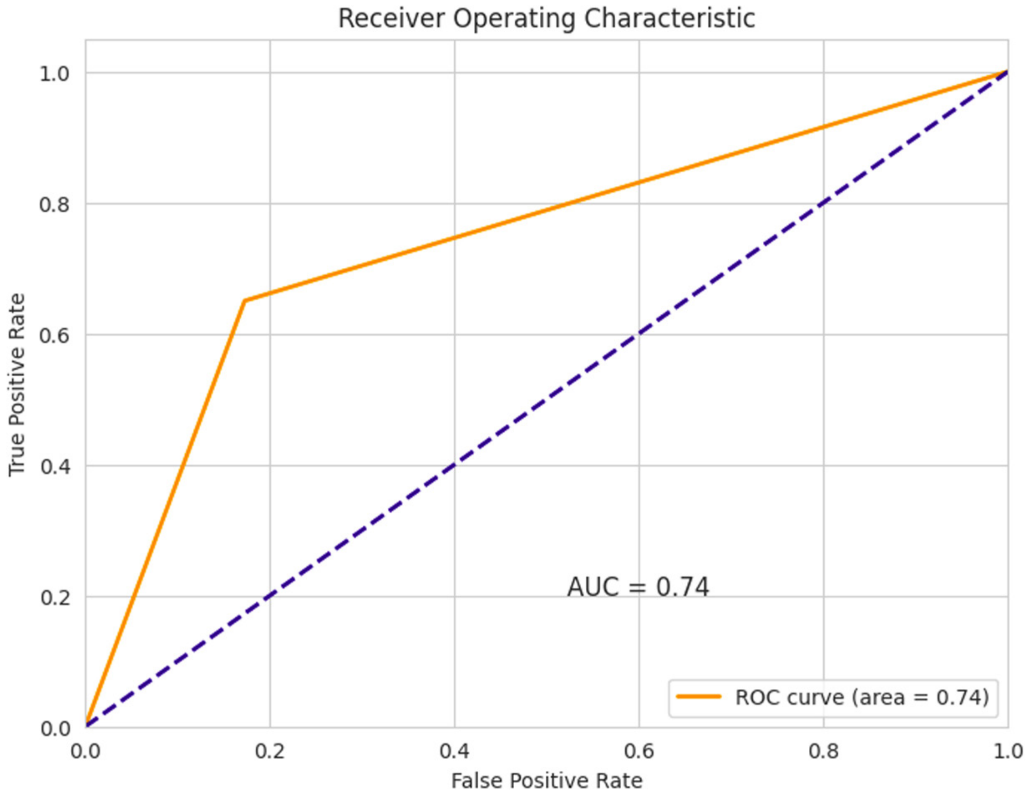

Fig. 7. ROC curve of the model

5 DISCUSSIONS

The performance of ML models in predicting the factors causing depression in students was achieved by applying the cross-matrix with the following metrics: accuracy, precision, recall, and F1 score, for which 18 input variables were used (school year, age, gender, BMI, WHO_BMI, PHQ score, depression severity, depressiveness, suicide, diagnosis of depression, treatment of depression, GAD score, drowsiness, Epworth score, severity of anxiety, anxiety diagnosis, anxiety treatment). Among the models trained in this study: LR, KNN, and DT, the LR model yielded the best training results for predicting depression (LR: 77%; KNN: 70%; DT: 62%). These findings align with the metrics reported in a previous study [28], where ML SVM and RF models were used to predict depression and anxiety, achieving an accuracy of 92.5% for the SVM model and 76.4% for the RF model. Discrepancies in these metrics could be attributed to various factors, primarily the size of the dataset. It is also noteworthy to compare these results with those of another study [26], where DT and RF models of ML were employed to predict depression in children and adolescents aged 4 to 17. The results indicated that the RF model achieved the highest accuracy at 95%. This demonstrates that models can yield varying results based on their training.

As indicated in references [28] and [26], our research has revealed a close relationship between obesity, anxiety, and depression. Moreover, the predictive metrics obtained are comparable to the accuracy rates reported in studies such as [25], which explored mental disorders. The study concluded that the most influential factors for anxiety and depression prediction are obesity, age, BMI, and anxieties, making them the four most significant characteristics for predicting depression.

The contribution of this study lies in classifying the LR model as the most relevant for predicting depression. It has also enabled the ranking of students who are at higher risk of obesity, mild depression, severe anxiety, suicidal thoughts, and the BMI factor directly associated with depression.

Finally, we found that BMI is directly related to depression, suggesting the need for a comprehensive approach to physical and mental health treatment. These findings represent a significant breakthrough in the field of mental health. Consequently, we can develop prevention and early intervention strategies to enhance student well-being.

6 CONCLUSION

The paper evaluated three ML models: LR, K-NN, and DT, based on supervised learning, to predict factors related to depression in undergraduate students. According to the results, the research concludes that the LR model is the most effective and appropriate for predicting student depression, as indicated by the following metrics: accuracy (77%), recall: (70%), and F1 score (72%). This model can be integrated into university information systems to automatically access student depression information based on key variables. Additionally, the case study successfully classified the dataset. For instance, approximately two out of five students have mild depression; around 90% of students with depression do not receive treatment; obese students are 2.5 times more likely to be depressed; male students are twice as likely to become obese; and male students have a higher BMI than female students. This information is crucial for decision-making. University authorities can use it to address issues affecting students' mental well-being. Furthermore, by employing ML techniques such as the LR model, health professionals and decision-makers can implement them in intervention programs to prevent and enhance the health, education, and well-being of students.

Although significant progress has been made in this field, it is essential to acknowledge certain limitations that may influence future research in this area. The study's limitations include sample size, the potential for omitted variables, and the absence of external validation. Moving forward, research could emphasize preventive interventions, longitudinal assessments, test outcomes, and enhancing predictions through statistical models.

7 REFERENCES

[1] WHO, "Depressive disorder (depression)," World Health Organization. Accessed: Oct. 06, 2022. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/depression

[2] Sooma, "Treating depression shouldn't be hard," soomamedical.com. Accessed: Oct. 18, 2022. [Online]. Available: https://soomamedical.com/treating-depression-shouldnt-be-hard/#

[3] F. S. Wicke *et al.*, "The association of depression and all-cause mortality: Explanatory factors and the influence of gender," *J. Affect. Disord.*, vol. 303, pp. 315–322, 2022. https://doi.org/10.1016/j.jad.2022.02.034

[4] E. N. Coman, H. Z. Wu, and S. Assari, "Exploring causes of depression and anxiety health disparities (HD) by examining differences between 1:1 matched individuals," *Brain Sci.*, vol. 8, no. 12, p. 207, 2018. https://doi.org/10.3390/brainsci8120207

[5] N. F. Jamaludin, T. S. M. T. Wook, S. F. M. Noor, and F. Qamar, "Gamification design elements to enhance adolescent motivation in diagnosing depression," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 10, pp. 154–172, 2021. https://doi.org/10.3991/ijim.v15i10.21137

[6] M. A. Taheri and G. Rahimi, "Suicide as approached by Halqeh Mysticism," *Procedia – Social and Behavioral Sciences*, vol. 114, pp. 102–106, 2014. https://doi.org/10.1016/j.sbspro.2013.12.666

[7] R. S. Patel, T. Machado, and W. E. Tankersley, "Eating disorders and suicidal behaviors in adolescents with major depression: Insights from the us hospitals," *Behavioral Sciences*, vol. 11, no. 5, p. 78, 2021. https://doi.org/10.3390/bs11050078

[8] National Institute of Mental Health, "Suicide," nimh.nih.gov, 2024. Accessed: Nov. 01, 2022. https://www.nimh.nih.gov/health/statistics/suicide

[9] M. Chatterjee, P. Kumar, P. Samanta, and D. Sarkar, "Suicide ideation detection from online social media: A multi-modal feature based technique," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100103, 2022. https://doi.org/10.1016/j.jjimei.2022.100103

[10] S. E. Murphy, L. P. Capitão, S. L. C. Giles, P. J. Cowen, A. Stringaris, and C. J. Harmer, "The knowns and unknowns of SSRI treatment in young people with depression and anxiety: Efficacy, predictors, and mechanisms of action," *Lancet Psychiatry*, vol. 8, no. 9, pp. 824–835, 2021. https://doi.org/10.1016/S2215-0366(21)00154-1

[11] J. K. Tan *et al.*, "Prevalence and associated risk factors of undiagnosed depression among people with epilepsy in a multiethnic society," *Epilepsy Res.*, vol. 178, p. 106772, 2021. https://doi.org/10.1016/j.eplepsyres.2021.106772

[12] L. Markauskaite *et al.*, "Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI?" *Computers and Education: Artificial Intelligence*, vol. 3, p. 100056, 2022. https://doi.org/10.1016/j.caeai.2022.100056

[13] C. Son, S. Hegde, A. Smith, X. Wang, and F. Sasangohar, "Effects of COVID-19 on college students' mental health in the United States: Interview survey study," *J. Med. Internet Res.*, vol. 22, no. 9, p. e21279, 2020. https://doi.org/10.2196/21279

[14] "Depression and other common mental disorders," WHO, 2022. https://www.who.int/publications/i/item/depression-global-health-estimates

[15] R. Alejandro-Salinas, A. C. Rojas-Cueva, F. J. Visconti-Lopez, M. L. Osorio-Martinez, and C. J. Toro-Huamanchumo, "Factors associated with post-traumatic stress disorder symptoms in the post-quarantine context of the COVID-19 pandemic in Peruvian medical students," *Heliyon*, vol. 8, no. 5, p. e09446, 2022. https://doi.org/10.1016/j.heliyon.2022.e09446

[16] S. M. Islam and A. Habib, "How impact investing firms are responding to sustain and grow social economy enterprises in light of the COVID-19 pandemic," *Journal of Business Venturing Insights*, vol. 18, p. e00347, 2022. https://doi.org/10.1016/j.jbvi.2022.e00347

[17] R. M. R. Parra, "Depression and the meaning of life in university students in times of pandemic," *International Journal of Educational Psychology*, vol. 9, no. 3, pp. 223–242, 2020. https://doi.org/10.17583/ijep.2020.6784

[18] J. Huarcaya-Victoria *et al.*, "Factors associated with mental health in Peruvian medical students during the COVID-19 pandemic: A multicentre quantitative study," *Rev Colomb Psiquiatr*, vol. 52, no. 3, pp. 236–244, 2023. https://doi.org/10.1016/j.rcp.2021.06.002

[19] S. Kim and K. Lee, "The effectiveness of predicting suicidal ideation through depressive symptoms and social isolation using machine learning techniques," *J. Pers. Med*, vol. 12, no. 4, p. 516, 2022. https://doi.org/10.3390/jpm12040516

[20] M. Garg, "Quantifying the suicidal tendency on social media: A survey," *arXiv preprint arxiv:2110.03663*, 2021. https://doi.org/10.48550/arxiv.2110.03663

[21] A. Stathopoulou *et al*., "Mobile assessment procedures for mental health and literacy skills in education," *International Journal of Interactive Mobile Technologies*, vol. 12, no. 3, pp. 21–37, 2018. https://doi.org/10.3991/ijim.v12i3.8038

[22] Y. Huang *et al*., "Comparison of three machine learning models to predict suicidal ideation and depression among Chinese adolescents: A cross-sectional study," *J. Affect. Disord.*, vol. 319, pp. 221–228, 2022. https://doi.org/10.1016/j.jad.2022.08.123

[23] C. H. Espino-Salinas *et al*., "Two-dimensional convolutional neural network for depression episodes detection in real time using motor activity time series of Depresjon dataset," *Bioengineering*, vol. 9, no. 9, p. 458, 2022. https://doi.org/10.3390/bioengineering9090458

[24] Y. Y. Song and Y. Lu, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, 2015. https://shanghaiarchivesofpsychiatry.org/en/215044.html

[25] T. Iliou *et al*., "ILIOU machine learning preprocessing method for depression type prediction," *Evolving Systems*, vol. 10, no. 1, pp. 29–39, 2017. https://doi.org/10.1007/s12530-017-9205-9

[26] U. M. Haque, E. Kabir, and R. Khanam, "Detection of child depression using machine learning methods," *PLoS One*, vol. 16, no. 12, p. e0261131, 2021. https://doi.org/10.1371/journal.pone.0261131

[27] T. Zhang, Z. Zhong, W. Mao, Z. Zhang, and Z. Li, "A new machine-learning-driven grade-point average prediction approach for college students incorporating psychological evaluations in the Post-COVID-19 Era," *Electronics (Switzerland)*, vol. 13, no. 10, p. 1928, 2024. https://doi.org/10.3390/electronics13101928

[28] 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, 2022. https://doi.org/10.2196/32736

[29] S. Aleem, N. U. Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, "Machine learning algorithms for depression: Diagnosis, insights, and research directions," *Electronics,* vol. 11, no. 7, p. 1111, 2022. https://doi.org/10.3390/electronics11071111

[30] J. S. Jung, S. J. Park, E. Y. Kim, K. S. Na, Y. J. Kim, and K. G. Kim, "Prediction models for high risk of suicide in Korean adolescents using machine learning techniques," *PLoS One*, vol. 14, no. 6, p. e0217639, 2019. https://doi.org/10.1371/journal.pone.0217639

[31] M. Mousavian, J. Chen, Z. Traylor, and S. Greening, "Depression detection from sMRI and rs-fMRI images using machine learning," *Journal of Intelligent Information Systems*, vol. 57, no. 2, pp. 395–418, 2021. https://doi.org/10.1007/s10844-021-00653-w

[32] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, pp. 281–304, 2022. https://doi.org/10.1007/s11280-021-00992-2

[33] P. J. C. Suen, S. Goerigk, L. B. Razza, F. Padberg, I. C. Passos, and A. R. Brunoni, "Classification of unipolar and bipolar depression using machine learning techniques," *Psychiatry Res.*, vol. 295, p. 113624, 2021. https://doi.org/10.1016/j.psychres.2020.113624

[34] L. H. John, J. A. Kors, J. M. Reps, P. B. Ryan, and P. R. Rijnbeek, "Logistic regression models for patient-level prediction based on massive observational data: Do we need all data?" *Int. J. Med. Inform.*, vol. 163, p. 104762, 2022. https://doi.org/10.1016/j.ijmedinf.2022.104762

[35] K. V. Mhetre and A. D. Thube, "Development of road safety models by using linear and logistic regression modeling techniques," in *Recent Trends in Construction Technology and Management*, in Lecture Notes in Civil Engineering, M. S. Ranadive, B. B. Das, Y. A. Mehta, and R. Gupta, Eds., Springer, Singapore, vol. 260, 2023, pp. 1205–1225. https://doi.org/10.1007/978-981-19-2145-2_89

[36] Y. Amemiya, "Instrumental variable estimator for the nonlinear errors-in-variables model," *J. Econom.*, vol. 28, no. 3, pp. 273–289, 1985. https://doi.org/10.1016/0304-4076(85)90001-6
[37] S. Raschka, "STAT 479: Machine Learning Lecture Notes," 2018. http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/
[38] E. Fix and J. L. Hodges, "Discriminatory analysis, nonparametric discrimination: Consistency properties," Technical Report 4, USAF School of Aviation Medicine, Randolph Field, 1951.
[39] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967. https://doi.org/10.1109/TIT.1967.1053964
[40] I. D. Wahyono *et al.*, "Emotion detection based on column comments in material of online learning using artificial intelligence," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 3, pp. 82–91, 2022. https://doi.org/10.3991/ijim.v16i03.28963
[41] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, and J. B. O. Mitchell, "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization," *J. Chem. Inf. Model*, vol. 46, no. 6, pp. 2412–2422, 2006. https://doi.org/10.1021/ci060149f
[42] O. Iparraguirre-Villanueva, V. Guevara-Ponce, F. Sierra-Liñan, S. Beltozar-Clemente, M. Cabanillas-Carbonell, and R. Palma, "Sentiment analysis of tweets using unsupervised learning techniques and the k-means algorithm," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 13, no. 6, 2022. https://doi.org/10.14569/IJACSA.2022.0130669
[43] B. Kamiński, M. Jakubczyk, and P. Szufel, "A framework for sensitivity analysis of decision trees," *Cent. Eur. J. Oper. Res.*, vol. 26, no. 1, pp. 135–159, 2018. https://doi.org/10.1007/s10100-017-0479-6

8 AUTHORS

Orlando Iparraguirre-Villanueva Systems Engineer with a master's degree in Information Technology Management, and a PhD in Systems Engineering from the Universidad Nacional Federico Villarreal in Peru. ITIL® Certified, specialized in Business Continuity Management (SBCM), and SCRUM Certified. He is national and international speaker/panelist having presented in Panama, Colombia, Ecuador, Venezuela, Mexico. Orlando has extensive experience in teaching at both undergraduate and postgraduate levels in various universities. He has served as a thesis advisor and jury member at different universities. With a background in management positions in Information Technology, he is also a research professor with publications in high-impact Scopus and WoS-indexed journals (Q1, Q2, Q3, and Q4). His areas of interest include Open-source software, IoT, Augmented Reality, Machine Learning, AI, CNN, text mining, virtual environments, scientific research methodology, and thesis. You can reach him at (E-mail: oiparraguirre@ieee.org).

Cleoge Paulino-Moreno is a Systems Engineer by profession, holding a Master's degree in Educational Computing and Information Technology. He specializes in Public Management and is certified in Windows Server, Linux, and Windows Server. He has an intermediate level of English proficiency. Cleoge has experience in scientific research and has worked as an Innovation and Technology Support Coordinator. He has published scientific articles in peer-reviewed journals, showcasing his dedication to scientific research. His skills include research abilities, diligent work ethic, effective communication, team leadership, and problem-solving.

Andrés Epifanía-Huerta has a master's degree in systems engineering with a specialization in information and communication technologies. He is currently affiliated with the Faculty of Engineering at the Universidad Tecnológica del Perú (UTP). He lectures on topics related to technological innovation, technologies applied to education, and the management of business technologies (E-mail: C25267@utp.edu.pe).

Carmen Torres-Ceclén is a Computer and Systems Engineer with a Master's degree in Computer and Systems, specializing in Information and Communication Technologies. She also holds a PhD in Education, a diploma in university teaching, and is certified by EXIN in ITIL. Carmen is known for her strong sense of responsibility, creativity, adaptability to change, sociable nature, and proactive approach. She possesses extensive knowledge in the IT field and stays current with the latest information technologies.