Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Predicting Student Depression Using Machine Learning

Bo Peng*

Waterford Institute, Nanjing University of Information Science & Technology, Nanjing, Jiangsu, China

*Corresponding author: 202283930012@nuist.edu.cn

Abstract. The rising prevalence of depression among students has drawn significant attention. As data mining and artificial intelligence technologies continue to evolve, leveraging behavioral and textual data for early depression prediction offers new opportunities for timely mental health interventions. This study uses 17 depression-related features, including gender, financial stress, academic pressure, and working hours. To evaluate model performance, this study focused on three representative classification algorithms. The first is logistic regression, known for its interpretability; the second is random forest, which leverages ensemble learning; and the third is XGBoost, a powerful gradient boosting framework. To assess model performance, this study considered multiple quantitative indicators. These include the proportion of correct predictions (accuracy), the model's ability to identify true positives (precision and recall), the harmonic means of those two (F1 score), and the area under the ROC curve (AUC), which reflects the overall classification capability. A 50-fold robustness test was also conducted to validate model stability. SHAP plots were utilized to interpret model predictions and to identify the most influential features contributing to depression risk at the end of this paper. The experimental data showed that Logistic Regression had the highest AUC score, which was 0.913, while the AUC scores of Random Forest and XGBoost were 0.906 and 0.903, respectively. Calibration curve analysis verified that Logistic Regression had the best calibration performance. The result of this study supports the feasibility and value of Logistic Regression in predicting student depression.

Keywords: Machine Learning; Students Depression; Model evaluation.

1. Introduction

Depression is a mental health disorder that is prevalent in society and has a significant negative impact on individuals, society, and families. Depression has been identified as the leading cause of disability worldwide. Each year, suicide claims the lives of more than 800,000 individuals, making it the second most common cause of death among people aged 15 to 29 [1]. With the development of the times, the competition among various industries has become more and more intense, which puts higher demands on the student population and also brings more pressure. The risk of depression in the student population is gradually increasing. According to the annual Healthy Minds Study, in recent years, 44% of college students in the United States have reported symptoms of depression, and the number is rising [2]. Due to the particularity of the student group, depression is not easily detected in the early stage, or students who are already suffering from depression choose to conceal it for various reasons, leading to more serious consequences. For example, in terms of physical and mental health, they may lose self-confidence and self-esteem and develop negative emotions such as inferiority and self-blame. In terms of academics, they may show signs of inattention, memory decline, lack of motivation to study and develop an aversion to learning. They may also develop negative emotions such as being withdrawn and introverted, and become unwilling to communicate with others, which in turn leads to tense interpersonal relationships. They may have conflicts and disagreements with family members and friends [3]. Due to the rising number of depressive disorders, traditional statistical methods can no longer accomplish the prediction work, and accurate models and methods are needed for prediction. For the prediction of medical conditions, many studies using machine learning have been reported, previous studies have explored the application of machine learning techniques in predicting psychological conditions such as anxiety, depression, and stress among

college students [4]. Vadher et al. have demonstrated, through their study, that machine learning methods are both feasible and effective when applied to coronary heart disease prediction tasks [5].

In this study, three classification techniques were employed—Logistic Regression (LR), Random Forest (RF), and XGBoost—to develop predictive models using the depression-related factor data within the dataset, and compared the model accuracy, AUC score, ROC curve, and robustness of the three models. The results offer valuable perspectives on how machine learning techniques can be effectively applied to the early identification of depression, contributing to more efficient and data-driven mental health interventions.

2. Datasets

2.1. Dataset contents

This study is based on the publicly available Kaggle dataset "Student Depression Dataset [6]. The data was collected from students at different educational levels (from undergraduate to PhD) from 52 cities in India. The study considers a range of influencing variables such as professionalism, including pressure from academics or work, cumulative grade performance (CGPA), satisfaction with academic and occupational experiences, sleep patterns, nutritional behaviors, educational background, presence of suicidal ideation, time allocation between work and study, financial difficulties, and any familial background related to mental disorders. These factors are all potential causes of depression.

2.2. Data Processing

The data for the current study were cleaned by removing the factor fields that were not relevant to this study (e.g. id). The data were coded so that the target variable Depression was used as a binary label and trained with the positive category of having depression (0 for not having depression and 1 for having depression). The data were subjected to feature normalization operations, and continuous variables (e.g. Age/work hour) were normalized using the Z-score to make the data comparable. For the division of the data set, in order to make the training and evaluation more accurate, the dataset was partitioned into training and testing subsets following an 80:20 ratio. A subsequent check confirmed that the distributions of depressive and non-depressive cases were balanced across both subsets.

3. Modeling

In terms of model selection, relevant studies on disease prediction using machine learning have been conducted around the world. For example, in a previous study, Jatin Gupta and Janmejay Pant et al. used various machine learning models to predict cardiovascular diseases and heart diseases [7,8], Including RF, LR, SVM, NB, and XGBoost. This study aims to predict the presence of depression using factors collected in the dataset that are associated with depressive symptoms. To build classification models, this study applied three learning algorithms: Logistic Regression, Random Forest, and XGBoost. The performance of each model was evaluated from multiple perspectives. These included the proportion of correctly classified instances (accuracy), the ability to identify true positive cases (precision and recall), and a balanced metric that considers both precision and recall (F1 score). In addition, the area under the ROC curve (AUC) was used to assess the model's overall discriminative capability. In addition, ROC curves, SHAP interpretation plots, calibration curves, and robustness line graphs were used to visualize the results.

3.1. Logistics Regression

Logistic Regression is often used in statistics to model the probability of an event occurring in a binary variable, mapping the output of the model to an interval in the range (0,1), with the following prediction function [9].

$$p(x) = \frac{1}{1 + e^{-(w^{\mathrm{T}}x+b)}} \tag{1}$$

In this expression, w refers to the vector of feature weights, x represents the input features, and b denotes the bias component. The function calculates the probability p(x) through the sigmoid transformation.

3.2. Random Forest

Random Forest integrates techniques such as bootstrap aggregation and the random subspace strategy to generate an ensemble of decorrelated decision trees. By aggregating their outputs, it enhances both the model's generalization capability and prediction accuracy. Its role in classification tasks is described as follows [10].

$$\hat{y} = \text{mode}\big(h_1(x),\ h_2(x),\ \ldots,\ h_T(x)\big) \tag{2}$$

In this context, $h_t(x)$ refers to the prediction produced by the tree indexed by t when handling input x. The variable T indicates how many trees are included in the ensemble model, and mode(·) represents the final decision determined through majority voting.

3.3. Extreme Gradient Boosting

Gradient boosting–based ensemble learning methods typically utilize a CART decision tree as the initial weak learner. Through iterative training, additional trees are sequentially built to fit the residual errors, gradually enhancing the overall performance of the model [11].

At every step of the training process, the model aims to reduce the value of the objective function defined below.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(\boldsymbol{x_i})\right) + \Omega(f_t) \tag{3}$$

$\mathcal{L}^{(t)}$ denotes the overall objective function at round t, l(-) is the loss function, $\hat{y}_i$^(t-1) denotes the predicted value of sample i at round t-1, $f_t(x_i)$ denotes the output of the t-th tree for sample i , $\Omega(f_t)$ is the complexity penalty term of the model, Here, T indicates how many leaf nodes are present in the decision tree, w_j refers to the value assigned to the j-th leaf, and γ as well as λ represent the regularization terms used to control model complexity.

4. Experiment Results

4.1. Performance Evaluation and Interpretation

The evaluation summary of the three models is provided in Table 1. Their performance was assessed across several dimensions, such as the correctness of predictions, the ability to identify relevant positive instances, the balance between precision and recall, and the overall discriminative capacity of the model.

Table 1. Comparative evaluation of three models in the context of student depression prediction

Model	Accuracy	Precision	Recall	F1-score	AUC Score
Logistic Regression	0.8384	0.8505	0.8752	0.8627	0.9132
Random Forest	0.8292	0.8416	0.8694	0.8552	0.9061
XGBoost	0.8267	0.8442	0.8601	0.8521	0.9039

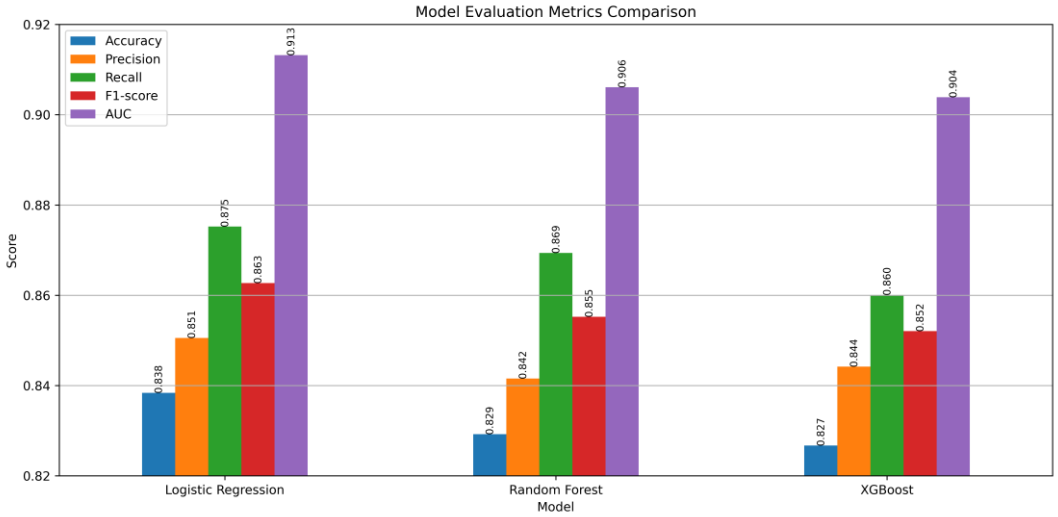

Fig. 1 Comparison of the three models under five performance indicators (Picture credit: Original)

As illustrated in Figure 1, Logistic Regression consistently demonstrates superior performance compared to the other two models. Its advantages are evident across multiple evaluation aspects, such as correctness of classification, ability to detect relevant cases, and overall robustness. These results indicate that the model is particularly reliable when identifying individuals at risk of depression within the positive category.

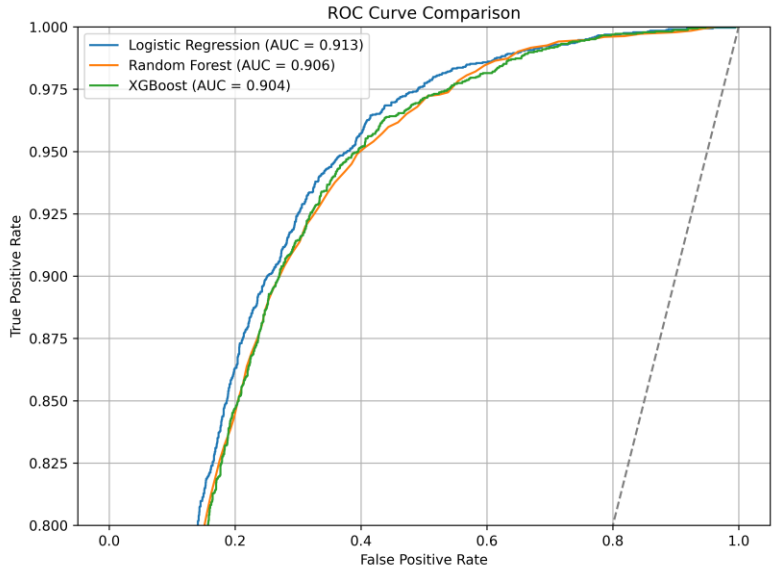

Fig. 2 ROC curves corresponding to the Logistic Regression, Random Forest, and XGBoost models applied in student depression classification (Picture credit: Original)

ROC analysis, with AUC as the main metric, helps determine how effectively a model can tell apart samples belonging to the positive class from those of the negative class, even when the classification threshold varies [12]. According to the ROC curve in Figure 2, it can be seen that all three models outperform the stochastic model, and all of them have strong classification abilities. Logistic Regression consistently achieves the best performance among the three, and it is significantly better than the other two models in the region of low false alarm rate.

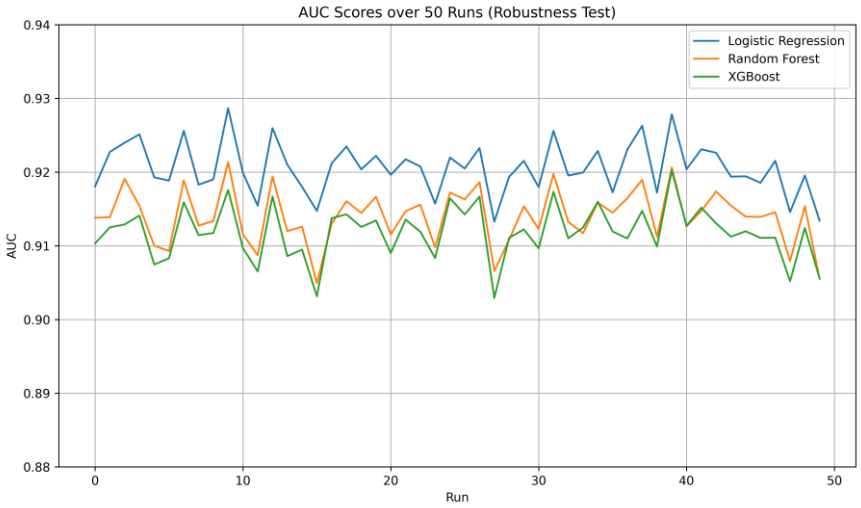

Fig. 3 AUC Scores over 50 Runs (Robustness Test) (Picture credit: Original)

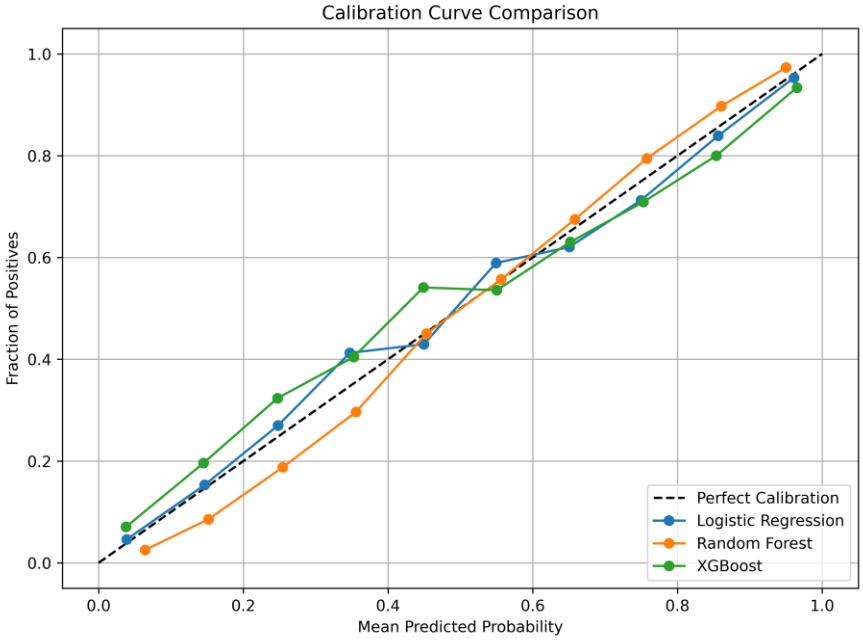

Fig. 4 Calibration curve comparison of Logistic Regression, Random Forest and XGBoost (Picture credit: Original)

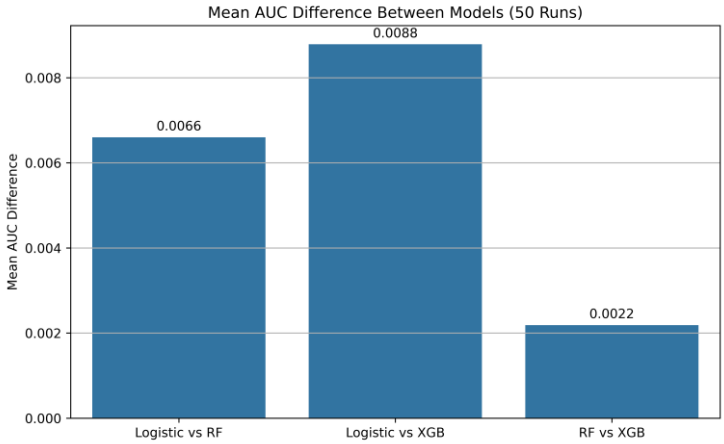

Fig. 5 Mean AUC Difference Between Models (Picture credit: Original)

Reliable interpretations capture stable and inherent relationships within the dataset [13], in order to examine whether the explanation outputs remain consistent across different runs, this study conducted

50 repeated robustness tests, and the results are shown in Figure 3. Combined with Figures 3 and 5, it can be seen that Logistic Regression, Random Forest and XGBoost all have good stability, with AUC values remaining stable between 0.90 and 0.93. Among them, the Logistic Regression model consistently scores higher than the other two models in each round, maintaining between 0.918 and 0.930, showing the best stability and being able to maintain excellent performance under different data divisions. In addition, it is ahead of Random Forest and XGBoost in terms of average AUC.

The Calibration curve in Figure 4 (X-axis is the average probability that the model predicts a positive category, and Y-axis is the proportion of actual positive categories), shows that the Logistics Regression model is very close to the perfect calibration line, and has the best model calibration performance with fluctuations within 0.05 only around the medium probability band (0.4 to 0.6 interval on the X-axis).

From Figures 1 to 5, it can be seen that Logistic Regression has consistent advantages over Random Forest and XGBoost in terms of prediction performance, robustness and calibration ability, and it is suitable as the best model in the prediction of categorized events in this study.

4.2. Feature Importance Analysis

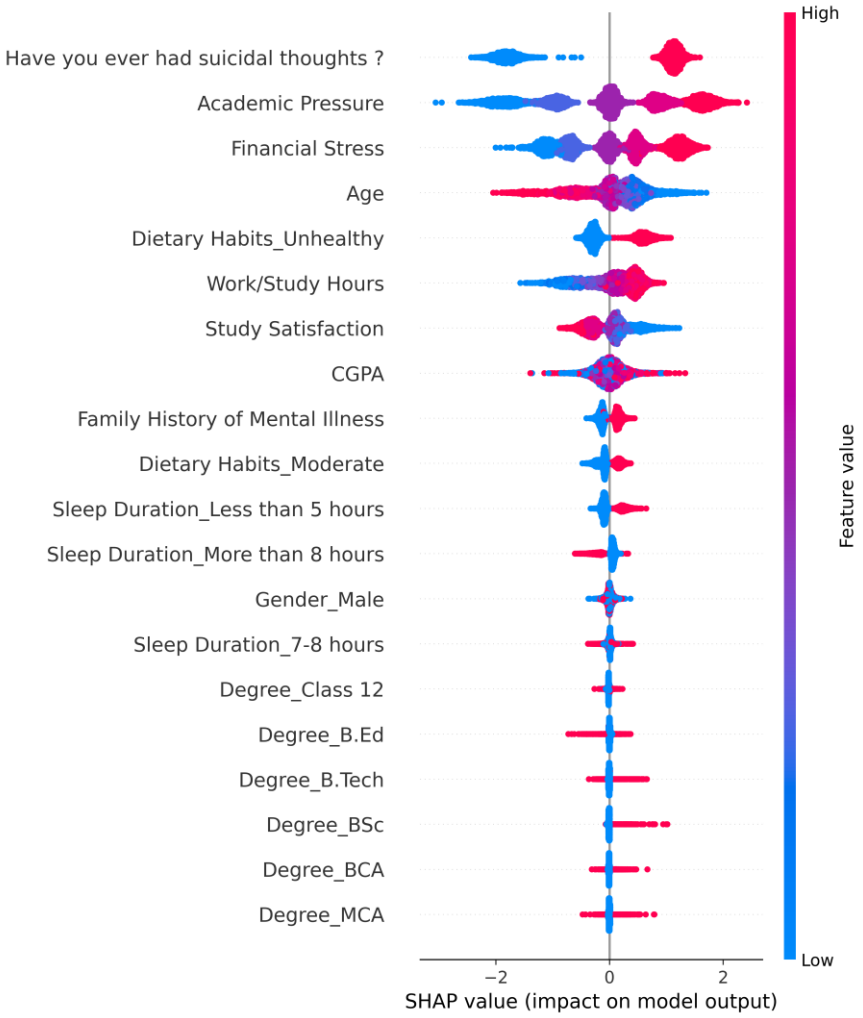

Fig. 6 SHAP Summary Plot (Picture credit: Original)

To explain the importance of feature-related factors in the training models, SHAP plots were utilized to demonstrate how each feature influences the prediction outcome, considering both the strength and direction of the effect[14]. Figure 6 displays the SHAP visualization generated in this study. In the plot, the X-axis indicates the extent to which each feature affects the model's output—the farther a point lies from zero, the stronger its impact. The Y-axis lists the features, arranged vertically

according to their relative importance. The color of each data point reflects the feature's magnitude, where red corresponds to higher values and blue to lower ones. The horizontal placement of each dot illustrates whether the feature contributes positively or negatively to the predicted result—dots on the right suggest an increase in the predicted value, while those on the left suggest a decrease. A larger SHAP represents a higher risk of depression. From Figure 6, it can be concluded that Suicidal thoughts, Academic Pressure and Financial Stress are the three most significant factors affecting depression in the present dataset.

Therefore, schools should conduct regular mental health screenings to identify high-risk students and keep records of them. Set up a 24-hour mental health hotline to intervene in students' suicidal thoughts. Provide mental health education and publicity to guide students to establish a correct mental state.

For students with academic difficulties, schools should optimize the curriculum and assessment system to reduce the academic burden. Establish an effective communication environment between students and teachers. Teachers should be able to monitor students' academic engagement and progress in time and establish a supportive learning atmosphere.

In terms of students' financial pressure, the government and schools should introduce and promote the implementation of relevant financial aid policies to reduce students' financial pressure.

5. Conclusion

In this research, three machine learning models were utilized to perform predictive analysis and model evaluation. The first is Logistic Regression, a widely used linear approach; the second is Random Forest, which builds predictions based on an ensemble of decision trees; and the third is XGBoost, a gradient boosting algorithm known for its efficiency and performance. These models were applied to carry out predictive analysis and model evaluation using 17 features associated with depression found in the dataset, including financial stress, suicidal thoughts, and academic pressure. The performance of the models was assessed based on key evaluation indicators, including accuracy, precision, recall, F1 score and AUC. Among them, Logistic Regression consistently showed the best performance across all evaluation metrics, with an AUC score of 0.913, and maintained optimal results in both the 50-run robustness test and the calibration curve analysis. Random Forest and XGBoost achieved AUC scores of 0.906 and 0.903 respectively, and their performance in certain evaluation metrics was slightly inferior to that of Logistic Regression. Therefore, it can be concluded that Logistic Regression demonstrates practical feasibility and strong performance in predicting student depression.

SHAP plots were used to visually analyze the influence and contribution of each factor in depression prediction. Suicidal thoughts, financial stress, and academic pressure ranked as the top three contributing factors. This result indicates the need for joint efforts from schools and society to reduce the risk of depression among students and help them overcome its negative impact.

This study still has some limitations, mainly in the limited number of machine learning models compared—only three were used. Future research could explore a wider range of models as well as multi-model fusion approaches. In addition, the dataset has geographical limitations (Only the Indian regional dataset was used), which restrict the generalizability of the models in a global context. Future studies should consider using more globally representative datasets.

References

[1] World Health Organization. Depression. 2025-12. Retrieved from https://www.who.int/india/health-topics/depression.

[2] University of Michigan School of Public Health. College Students' Anxiety, Depression Higher Than Ever—But So Are Efforts to Receive Care. 2023-03-09. Retrieved from https://sph.umich.edu/news/2023posts/college-students-anxiety-depression-higher-than-ever-but-so-are-efforts-to-receive-care.html.

[3] Xu L. Preventing Adolescent Depression and Creating a Healthy Growth Environment. Gems of Health, 2025, (6): 87–88.

[4] Malik S. S., Khan A. Anxiety, Depression and Stress Prediction Among College Students Using Machine Learning Algorithms. 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), IEEE, 2023: 1–5.

[5] Vadher H. H., Aryan A., Vamshi K., Rajashekar R., Ashish D., Arora G. D. Unveiling the Potential of Machine Learning: Harnessing Machine Learning for Enhanced Coronary Heart Disease Detection and Intervention. 2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE), IEEE, 2024: 1073–1078.

[6] Shamim A. Student Depression Dataset. 2023. Retrieved from https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset.

[7] Gupta J. The Accuracy of Supervised Machine Learning Algorithms in Predicting Cardiovascular Disease. 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST), IEEE, 2021: 234–239.

[8] Pant J., Singh D., Sharma V., Pant H. K., Bhatt J. A Machine-Learning Approach to Detect Heart Disease. 2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, 2024: 860–863.

[9] Cramer J. S. The Origins of Logistic Regression. Tinbergen Institute Discussion Paper, No. 02-119/4, 2002.

[10] Breiman L. Random Forests. Machine Learning, 2001, 45: 5–32.

[11] Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785–794.

[12] Qin Z.-C. ROC Analysis for Predictions Made by Probabilistic Classifiers. 2005 International Conference on Machine Learning and Cybernetics, IEEE, 2005, 5: 3119–3124.

[13] Hancox-Li L. Robustness in Machine Learning Explanations: Does It Matter?. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020: 640–647.

[14] Marcílio W. E., Eler D. M. From Explanations to Feature Selection: Assessing SHAP Values as Feature Selection Mechanism. 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, 2020: 340–347.