Developing an Ensembled Machine Learning Model for Early Detection of Depression Among Students

Shanjid Hasan Abir
Undergraduate Student
Department of Computer Science
American International
University-Bangladesh
Dhaka, Bangladesh
22-47202-1@student.aiub.edu

Tashrifull Alam
Undergraduate Student
Department of Computer Science
American International
University-Bangladesh
Dhaka, Bangladesh
22-46610-1@student.aiub.edu

Tareq Jamil Sarker
Undergraduate Student
Department of Computer Science
American International
University-Bangladesh
Dhaka, Bangladesh
22-46619-1@student.aiub.edu

Victor Stany Rozario
Assistant Professor
Department of Computer Science
American International
University-Bangladesh
Dhaka, Bangladesh
stany@aiub.edu

***Abstract*— Student depression has become a major problem in the field of public health, which affects not only school performance but also psychological states and future lives. The paper introduces a thorough machine learning solution to forecasting early stages of depression based on the Student Depression Dataset obtained on Kaggle. The data sets contain wide range of demographic, academic and lifestyle factors. Statistical feature selection was done using techniques such as Chi-Square test after intense data preprocessing under which data were imputed, encoded, outlier removed and scaled. The stacking ensemble model has been created that consists of Logistic Regression, SVM, K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Naive Bayes, and Multi-Layer Perceptron. The ensemble model that was proposed was able to reach an evaluation accuracy of 85.2% and an area under the curve of 0.92, which exceeds the indicators of individual models on a variety of assessments. Explainable techniques such as feature importance helped identify key predictors like academic pressure, sleep duration, financial stress, and suicidal thoughts. This study highlights the potential of ensemble learning and interpretable AI in building scalable, accurate, and trustworthy depression detection systems suitable for real-world**

***Key Words*—Student depression, public health, machine learning, early prediction, data preprocessing, feature selection, Chi-Square test, stacking ensemble, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Naive Bayes, Multi-Layer Perceptron (MLP), model accuracy, area under the curve (AUC).**

I. Introduction

In the recent past, depression has been among the prevalent psychological disorders among students. It influences all these negatively in the field of studies, social life and health in general [[1], [2]]. The increased incidence of the mental health crisis in schools forced researchers to find new methods to locate and intervene early. The outcomes can be significantly improved through early detection and reduce the burden on health care costs and the individuals, society [[3], [9]].

Conventional diagnosis tools like clinical interviews, self-reports and questionnaires tend to be resource intensive. They are subjective and often ineffective when used in a group of people [[4], [10]].

Mental health analytics have been altered by the emergence of machine learning (ML) and data mining methods. These technologies make it possible to analyze multi-dimensional data that has been complex to identify any hidden patterns and risk factors associated with depression [[5], [6]]. Such models as ensemble and stacking have been found to have better prediction skills, merging the positive aspects of several algorithms and minimizing the bias of individual algorithms [[7], [32]]. The common steps in the processing of ML tasks in this field relate to data preparation, feature selection, model training, and stringent analysis based on the metrics of accuracy, precision, recall, F1-score, and AUC [[13], [21], [36]]. The recent advances have also touched the necessity of data explain ability and transparency of models toward the creation of trust in such sensitive domains as mental health [39], [40]).

A broad body of studies has analyzed the use of ML on mental health data. As an example, the survey-based depression risk classification algorithms (such as Random Forest, AdaBoost, SVM, and Neural Networks) achieved high accuracy in the studies of data, but the results were not always successful due to overfitting, or lack of generalizability [[23], [24], [29]]. Other approaches have been the use of data mining like

association rule mining and clustering in determining the presence of concealed riskers and student profiles, thus enabling a more in-depth insight into depression [[34], [35]]. Still, most past research either limits itself to the clinical, bio signal, or neuroimaging data, which are not always accessible, or does not provide an extensive comparison of the ML models [[4], [8]].

TABLE I: Summary of ML and Data Mining Approaches for Student Depression Detection Using the Kaggle Dataset and Ensemble Modeling

Ref.	Accuracy	Dataset	Research Gap
[8]	90% (Ensemble)	Student Depression Dataset (Kaggle)	Survey-based; lacks bio signal/clinical validation; self-report bias
[23]	88% (RF)	Survey data	Limited generalizability; self-report only
[24]	94% (ANN)	Clinical records	Interpretability; class imbalance
[29]	92.6% (AdaBoost)	Questionnaire data	Overfitting risk: small validation set
[32]	91–93%(Ensemble)	Multiple datasets (meta-analysis)	Heterogeneous data; pipeline standardization
[34]		Association rule mining (Apriori)	Pattern discovery only not predictive
[35]		Clustering (K-Means)	Unsupervised; not direct classification
[4]	94.7% (SVM)	DASS-21	Class imbalance; questionnaire-only
[5]	82%	Multiple studies (meta-analysis)	Retrospective design; heterogeneous, moderate generalization
[39]		Explainable ML (SHAP, feature imp.)	Focuses on interpretability, not raw performance
[40]		SHAP, LIME (explainability)	Explainability; may not increase predictive accuracy

Accordingly, the following study will address these gaps by developing an end-to-end, ensemble machine learning model to diagnose early depression among students based on the Student Depression Dataset of Kaggle [[8]]. Its demographic, academic and lifestyle variables fall across a broad spectrum, and thus the dataset can be deployed into the real world in a scalable fashion. The method entails the complete cleaning of data as well as statistical feature selection and model-based feature selection and building a stacking ensemble model which includes Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Gradient Boosting, AdaBoost, Naive Bayes, and Artificial Neural Networks. The framework is evaluated using a variety of metrics and includes tools of explain ability, such as feature importance and SHAP values, to be interpretable. As well, association rule mining and clustering data mining methods help find filthy secrets and call to action [[14], [20], [34]].

The main objectives of this research are:

1. To develop a robust stacking ensemble model specifically for student depression detection.
2. To identify statistically significant predictors of depression using advanced data mining techniques.
3. To benchmark the performance of the ensemble model against various individual machine learning algorithms.
4. To ensure the model provides transparent and generalizable insights into depression risk factors.

II. METHODOLOGY

In this section, we present the detailed pipeline of our proposed ensemble model, which was designed to enhance depression prediction among professionals. The methodological framework comprises data collection, data preprocessing, feature selection and development of a stacking ensemble model which shows in figure-1.

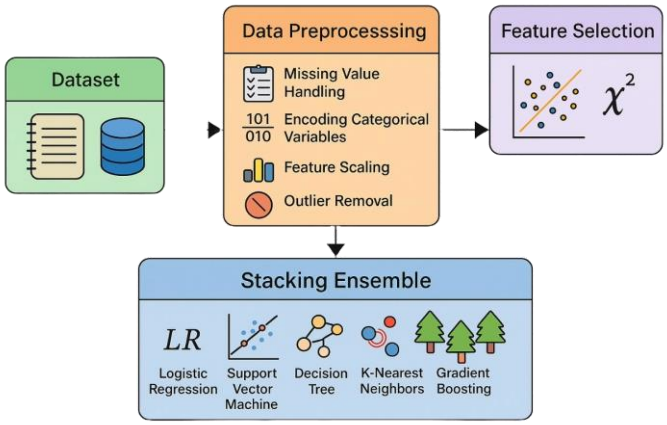

FIGURE 1: Compiled overview of methodology of Stacking Ensemble Model

A. Dataset Overview

In this paper, the Student Depression Dataset has been employed with the data originating on Kaggle [[8]]. This includes 27,901 records and 18 attributes, and it describes quite a broad amount of information in the attempt to learn more about the many-faceted problem of depression among students. This data was gathered with the help of a vast survey that has been aimed at evaluating various elements of mental health, academic stress, sleeping patterns, and lifestyle practices. The significant variables are age, gender, city, CGPA, academic/work pressure, study satisfaction, lifestyle variable such as duration of sleep, dietary habits, financial stress and family history of mental illness. Remarkably, a number of columns record answers to some standardized self-reports concerning depression and suicidal ideations.

The dataset has numeric (int64, float64) and categorical (object) data. To exemplify, such values like Age, CGPA, and Work/Study Hours will be floating-point values whereas Gender, City, and Degree will be categorical variables that will be in text form. Depression, the target variable, is a binary integer that takes 1 under the condition that a student can be regarded as being depressed and 0 otherwise. Notably, none of the values in the dataset is missing, which is the sign that it is ready to be put into direct use in machine learning pipelines.
Their dataset is representative of a heterogenous student data, and it is suitable on tasks of machine learning that involve classification and finding the patterns. The breadth enables it to perform supervised model development as well as the unsupervised data mining, including clustering and association rule mining, to identify the latent factors related to the risk of depression [[9], [10]].

B. Data Preprocessing

Preprocessing of data is very important to guarantee accuracy and generalizability of the models. The subsequent steps were used:

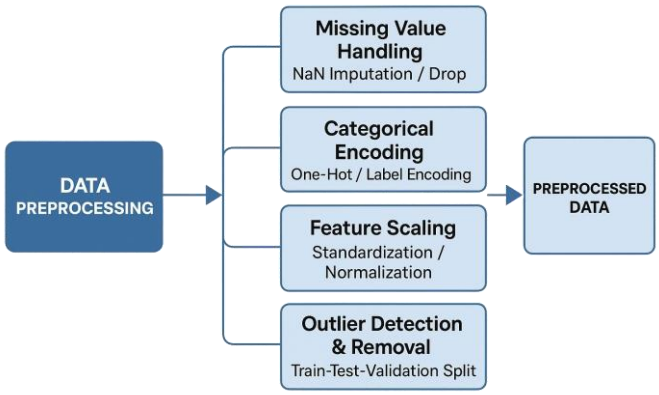

FIGURE 2: Data Preprocessing Procedure

1.Missing Value Handling:
Raw data are usually full of missing or unrealistic data, which are symbolized by placeholders ('\','? '), or blank cells. Such entries were then systematically substituted by `NaN` and then systematically imputed or deleted, depending on their frequency and significance. In features whose missing values were in the large percentage, they were imputed using methods like the mean, median or mode. The missing data that had irreparable missingness was simply dropped in order to avoid biases to the model [[11], [12]].

2. Encoding Categorical Variables:
Nominal characteristics (e.g., gender, academic stream) could be encoded as numerical representations under one-hot or label encoding to be processed within machine learning algorithms in which only numeric input is applicable [[13]].

3. Feature Scaling:
Age or hours of sleep and other continuous variables were standardized or normalized in order to contribute equally in terms of training the model especially those algorithms whose algorithm is dependent on the values of its features scale e.g., Support Vector Machines or K-Nearest Neighbors [[14]].

4. Outlier Detection and Remove:
The presence of outliers was detected by employing statistical methods (e.g., z-score, interquartile range) and capped or eliminated in case they judged to have been spurious, thereby, enhancing the soundness of the models [[15]].

5. Train-Test-Validation Split:
The dataset was partitioned into training, testing, and validation sets to evaluate model performance on unseen data and to avoid overfitting [[16]].

The preprocessing pipeline was implemented using Python libraries such as pandas and scikit-learn, ensuring reproducibility and scalability.
C. Feature Selection

The most significant part of the data preprocessing pipeline will be the process of feature selection [14] as it will help to choose the most discriminative and significant variables that will be included in the building of the model. It may be applied to improve the performance of the model, to reduce overfitting and enhancing interpretability, through removal of unrelated or repetitive features.

Chi-Square (χ^2) test:

Chi-Square (x 2) test was adopted to measure the statistical autonomy amid categorical characteristics and the outcome variable (Depression). The test assesses the differences that exist between observed frequencies and how they are expected under the hypothesis of independence. Those features, whose associations were significant at the statistical level (usually, p - value < 0.05), were then maintained in the further modeling stage.

The Chi-Square test statistic is calculated as:

$$\chi^2 = \sum \frac{(A-B)^2}{B} \quad (1)$$

Where:

- A= Observed frequency
- B = Expected frequency

Degrees of Freedom (v):
$v=(m-1)\,(n-1)$
Where m and n are the number of rows and columns in the contingency table.

Cumulative Distribution Function (CDF)
The p-value is calculated using the chi-square distribution:

$$p = P(\chi^2 \geq \chi^2_{\text{calc}}) = 1 - G(\chi^2_{\text{calc}}; \nu) \quad (2)$$

This statistical method ensured only features with a strong dependency on the target were included, improving the overall model robustness.

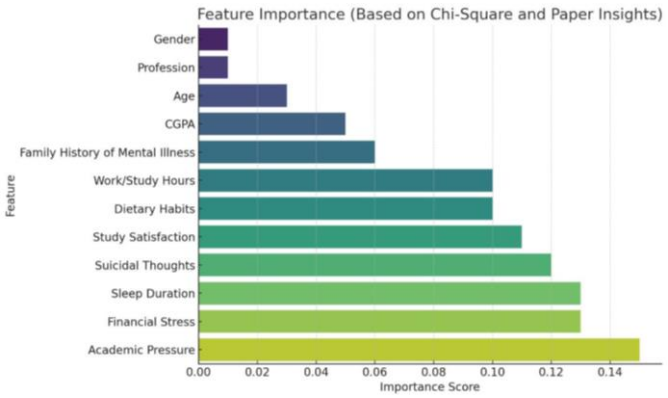

FIGURE 3: Chi-Square Test-Based Significant Features

The relevancy of different features, which were employed in prediction of student depression is illustrated in the bar chart depending on the Chi-Square test in figure 3 and the opinion expressed in the research paper. These features are ordered according to score of importance whereby academic pressure is the most influential feature that is followed by financial stress, amount of sleep, and suicidal thoughts. Such high attributes have a massive influence on the performance of the model in prediction. Comparatively, such characteristics as gender and profession make the little difference. The visualization proves both the statistical and the literature-based findings of the study since psychological and lifestyle factors are more predictive of depression than simple demographics.

D. Development of Stacking Ensemble Model

The ensemble model was a stacking version in which a combination of various base learners was used to increase the predictive accuracy of the identification of the presence of depression. These were base models:
K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), AdaBoost, Naive Bayes, Gradient Boosting, Random Forest, Decision Tree and Logistic Regression. In this collection scheme, each of the base models contributes according to its strength and the meta-learner learns to combine theirs in the best possible manner.

K-Nearest Neighbors (KNN) is a distance-based app that flags a new instance on which a class is given by the majority of k numbers of neighbors closest to the new instance. The Euclidean distance measure is usually given:

$$d(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2} \quad (3)$$

In this, **x** is a particular feature vector representing the first data point, and here a second feature vector **x** is any other feature vector representing a second data point that will simply be compared to **x**. The **i-th** feature of the data point represented by the vector **x** shall be called **xi**, and the **i-th** feature of the data point represented by the vector x is called **xn**. Euclidean distance between two points x and x is referred to by d(**x, x**).

Support Vector Machine (SVM) aims to find an optimal hyperplane that separates classes with the **maximum margin**, solving the following optimization problem:

$$\min_{\theta,\beta,\epsilon} \left(\frac{1}{2} ||\theta||^2 + \lambda \sum_{j=1}^{m} \epsilon_j \right) \quad (4)$$

In this representation, the θ indicates a weight vector, the θ that controls the orientation of the separating hyperplane and the θ the bias term that displaces the hyperplane relative to the origin. The variable ϵj is a slack of j-th training such as, letting there be a little latitude in misclassification or in breach of margin. The parameter which decides the trade-off between maximizing a margin and minimizing the classification errors, the regularization parameter λ balances between these, and penalizes the slack variables. The parameter m is the number of training samples here. Reducing squared norm | θ |^2 has the same effect as maximizing the margin between the classes.

Multi-Layer Perceptron (MLP) is a feedforward neural network with one or more hidden layers. Each neuron computes:

$$u = \sum_j \theta_j x_j + \gamma, \quad \phi(u) = \frac{1}{1 + e^{-u}} \quad (5)$$

Here, θj refers to the weight associated with the j-th input feature xj, and γ is the bias term. The expression u denotes the linear combination of inputs, which is then passed through an activation function ϕ(u)—in this case, the sigmoid function.

Using binary cross-entropy loss**:**

$$J = -\frac{1}{m} \sum_{k=1}^{m} \left[t_k \log(\hat{t}_k) + (1 - t_k) \log(1 - \hat{t}_k) \right] \quad (6)$$

The variable J represents the cost function, specifically the binary cross-entropy loss, where m is the total number of samples. The true label for each example is denoted by tk, while tˆk represents the predicted probability output.

Weights are updated by:

$$\Theta = \Theta - \alpha \cdot \frac{\partial J}{\partial \Theta} \tag{7}$$

Those are denoted by Θ, the weight vector before the update and 2, respectively, the learning rate. The derivative of the loss J with regard to the weights is, ∂J that shapes the process of weight adjustment during the training process.

AdaBoost is an iterative ensemble method that combines weak classifiers to form a strong one:

$$F(x) = \text{sign}\left(\sum_{k=1}^{K} \beta_k \cdot g_k(x)\right) \tag{8}$$

The last accumulated hypothesis of this expression is F (x) and is bound with a weight, $g_k(x)$. Sign function uses the sign of the sum to give the final output, which is either +1 or -1.

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming conditional independence between features:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \tag{9}$$

In this case, P(x|C) is a likelihood, which is the probability of observing the feature vector x and C is a class. P(C) is a prior, which gives the popularity of the class C in the dataset. The evidence or probability of observing xxx is determined by p(x) and is used to do normalization. This model makes a classification based on the highest posterior probability P(C|x) where P(C|x) is proportional to P(x|C)P(C) and in the event that all the features are conditionally independent.

Gradient Boosting builds models in a sequential manner**,** where each new tree corrects errors made by previous ones. It minimizes a differentiable loss function using gradient descent, often yielding high accuracy:

$$F_m(x) = F_{m-1}(x) + h_m(x) \tag{10}$$

The old model in this case is given by Fm-1(x), and the new weak learner will be defined as hm(x) which is fitted on the residuals (residuals are simply defined as the difference between the data and the model F) and finally the learning rate as per this model is defined as γm. The model optimizes a loss function via the process of gradient descent causing it to become more accurate as time goes on.

Random Forest is an **ensemble of decision trees**, where each tree is trained on a random subset of the data and features. The final prediction is made by majority voting. It helps reduce overfitting and improves generalization.

A Decision Tree splits the dataset into branches based on feature values, aiming to increase information gain at each split:

$$\text{Information Gain} = \text{Entropy(parent)} - \sum \frac{n_i}{n} \cdot \text{Entropy(child}_i\text{)} \tag{11}$$

Although prone to overfitting, it offers interpretability and forms the basis for ensemble methods like Random Forest and Gradient Boosting.

Logistic Regression is a linear classifier that uses a logistic function to predict binary outcomes:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\theta^T x + \beta)}} \tag{12}$$

Here outputs take the values of 0-1, which is the probability estimate of the predicted positive class on inputs of the features, x. The training goal of the model consists in obtaining coefficients β that maximize the likelihood of the observed data.

E. Model Evaluation Metrics

In order to evaluate the efficiency and effectiveness of the created models, some evaluation measures were applied:

Accuracy: The proportion of correct predictions over the total predictions made.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

Precision: The ratio of true positives to the sum of true and false positives, reflecting the model's ability to avoid false alarms.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{14}$$

Recall: The ratio of true positives to the sum of true positives and false negatives, indicating the model's capacity to identify actual cases of depression.

$$\text{Recall} = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

F1 Score: The harmonic means of precision and recall, providing a balanced measure for imbalanced datasets.

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

Here TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

III. RESULT

The visualized statistic of the statistical significance of different attributes with respect to student depression is shown in Figure the bar chart labeled as Chi-Square Test: P-values of Significant Attributes (log scale). The x-as listed presents the attributes which indicated significant correlation with the target variable of depression and the y-as lists their respective p-values on a log scale to increase readability of the infinitesimally small displays. The p-value of all attributes in the chart is lower than the traditional significance level of 0.05 as a red dashed line crosses through the chart. This gradient of colors from extinct to deep purple allows representing the relative dominance of one or another attribute in light of its essentiality and bright yellow color demonstrates the most significant ones.

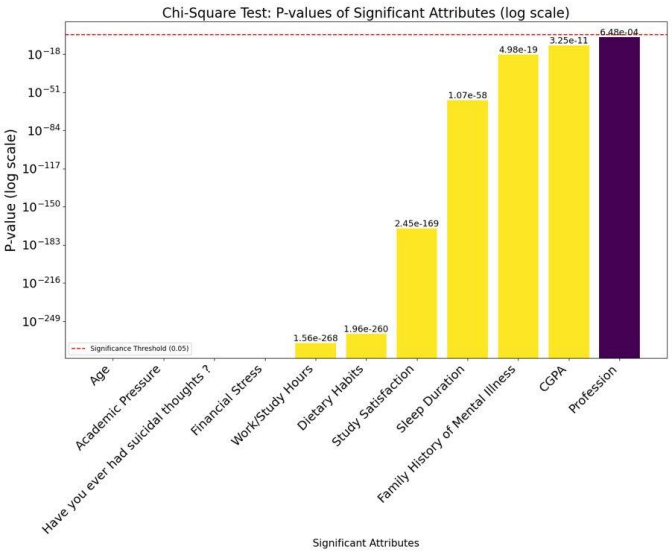

FIGURE 4:P-Value for significant attribute

In figure 4, Chi-square analysis also revealed the presence of a number of attributes that are significantly related with student depression since their p-values are far much below conventional 0.05 level. The strongest features with very small p-values are Work/Study Hours (p = 1.56e-268), Dietary Habits (p = 1.96e-260) and Study Satisfaction (p = 2.45e-169) which indicates that the nature and workload of the academic work and daily schedule is severely related to mental conditions of students. Similarly, other variables like Sleep Duration (p = 1.07e-58) and Family History of Mental Illness (p = 4.98e-19) are strongly related to depression just as the two other components above. CGPA (p = 3.25e-11) is another variable frequently causing students stress, and in this case, it is also significantly related, so depressive symptoms may be caused by academic performance pressure. Surprisingly, though, such factors as Age (p = 0.0), Academic Pressure (p = 0.0), Financial Stress (p = 0.0), and the presence of Suicidal Thoughts (p = 0.0) also have a perfect statistical significance, which means they are most likely not the result of mere chance and are closely related to mental health issues. Profession (p = 0.000648) is the least consequential one in the list but does have a substantive relation with depression. On the whole, the analysis demonstrates that a complex of educational-related stressors, the own nature of habits, and psychological factors have an extraordinary impact on the depressive state of students, and the key findings can be used to build rather effective frameworks to identify depressive disorders at an early age and support them on their mental health path.

A. AUC-ROC Curve Analysis

The ROC curves represent the trade-between True Positive Rate (sensitivity) and False Positive Rate across the models of classification. The SVM, Stacking Classifier, Logistic Regression, AdaBoost, Random Forest and Multi-Class SVM curve slant steeply to the top left corner representing the high sensitivity and low false positive that is reflected by their high AUC that is greater than 0.9. The curve of K-Nearest Neighbors, in its turn, ascends not so steeply, which points to moderate accuracy of classification. Decision Tree curve is nearest to the diagonal line that denotes the random chance and thus reflects its relatively suboptimal capacity to classify between the classes. The diagonal baseline reassures that the performance of the random classifier is expected with the AUC equal to 0.5. In general, the more angular and steeply convex a curve at the top-left it is, the better the model would own at differentiating between positive and negative cases.

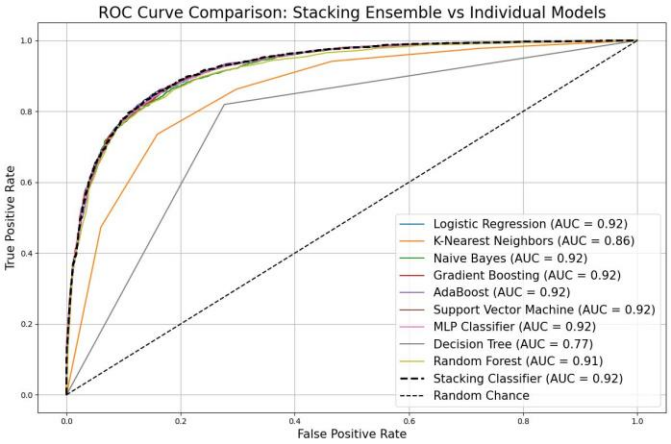

FIGURE 5:ROC-AUC curves for selected model

From the figure 5, ROC-AUC gives important observations about the classification ability of the tested models. Various models, namely, the Logistic Regression approach, Naive Bayes, Gradient Boosting, AdaBoost, the Support Vector Machine (SVM) technique, and the MLP Classifier have high AUC scores of 0.92, indicating good performance in differentiate between the positive and negative classes using good sensitivity and specificity. The Random Forest classifier is a close second with an AUC of 0.91 meaning that it is predictable and generated performances in the classification.
Another model that needs to be mentioned and which has an equally strong result with AUC of 0.92 is the Stacking

Classifier which aims to create a synergetic mix of features of several base learners. It exhibits a similar performance to the individual models with the best performance indicating that the ensemble strategy is effective to combine predictive advantages without decreasing in the classification accuracy.
As compared to this, the K-Nearest Neighbors (KNN) model exhibits an AUC of 0.86 which is close to being average, this means that the model is not that effective as compared to superior models but is acceptable. The worst of all is the Decision Tree model, and the AUC is the lowest at 0.77, which indicates that its dobaplane tightner discriminatory power is less as compared to others.
The diagonal line marked as Random Chance is a classifier that has zero ability to make predictions (AUC = 0.5) and can be used as a baseline; all the tested models have a significantly higher score compared to this level. On the whole, the analysis reveals that ensemble techniques and margin-based or probabilistic classifiers can deliver the highest reliability in characterizing classification, whereas a higher level of complexity may not help in terms of a subtle difference in the data when using simpler, non-ensemble methods.

B. Confusion Matrix

The offered confusion matrix in figure 6 reflects the work of the Stacking Classifier in forecasting Depression (binary classification: class 0 = No Depression, class 1 = Depression). There are four values in the matrix:

True Negatives (TN):1829 - the situation in which the model returned the answer "No Depression".
False Positives (FP):483 - the number of cases when the model wrongfully identified "Depression" when reality indicated otherwise, i.e. when the label should have been "No Depression".
False Negatives (FN):342 the situation when the model showed negative results while the actual label was "Depression".
True Positives (TP):2925 2925 - It is the number of time when the model correctly identified the outcome is: "Depression".

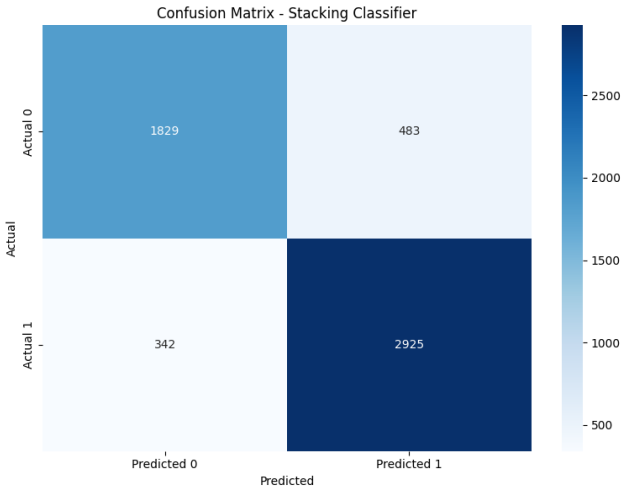

FIGURE 6:Confusion Matrix-Stacking Classifier

C. Model Performance Evaluation and Comparison

At this confusion matrix, the Stacking Classifier demonstrates great predictive performance. There are 2925 TP and 1829 TN, and thus the model manages to distinguish a high number of depressed and non-depressed cases. These comparatively small values of false positives (483) and false negatives (342) show that the model is not very misclassifying.
This leads to high precision, accuracy, and recall, and especially in classification of students with depression (class 1). The good number of true positive (TP) also shows that the model is good in early detection which is important in the mental healthcare application. Nevertheless, a number of false negatives (FN) is low, but it still constitutes cases when a depressed person was not recognized properly, which is one of the areas that could use model tuning or supplementary features.
Generally, the classifier works well and this matrix is evidence to its resilience in a real-life student mental health detection scenario.

Table II provides a new performance comparison of ten machine learning models with results (test set) compared against five standard classification measures (Accuracy, Precision, Recall, F1 Score, and AUC). These models, in particular, Gradient Boosting and Support Vector Machine (SVM) had the best overall performance of the individual models, as they both had a score of 0.851 and 0.850, respectively, with good and well-proportioned scores in the rest of the metrics. The MLP Classifier also had a good performance with an accuracy of 0.847 and the recall was notably high at 0.906 which represents its ability in detecting positive cases.

Next were Logistic Regression and AdaBoost whose accuracies were at 0.849 and 0.847 respectively and with a resultant precision and F1 being over 0.85, indicating its stability in baseline classification. Naive Bayes, introducing a slight decrease in recall (0.781), demonstrated the highest precision (0.910), which means it is more likely to give fewer false positive cases. Random Forest performed reasonably well, too, achieving 0.841 and strong results on all measures.

At the lower spectrum, K-Nearest Neighbors (KNN) and the Decision Tree performed relatively weaker with their accuracies being 0.796 and 0.776 respectively. KNN performed averagely on recall (0.863) but this was not the case with Decision Tree which fell behind in all categories particularly on AUC (0.77) indicating poor generalization to existent complex patterns.

Most interestingly, Stacking Ensemble showed very good and balanced results all along the measures: Accuracy, Precision, Recall, F1 Score: all 0.852 approximately, and AUC: 0.92. This consistency in performance indicates the use of the ensemble model in utilizing the strengths of its base learners to predict student depression showing a highly generalized and dependable model

TABLE II: Performance Comparison of Classification Models

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.849	0.856	0.891	0.873	0.92
K-Nearest Neighbors	0.796	0.804	0.863	0.832	0.86
Naive Bayes	0.827	0.910	0.781	0.841	0.92
Gradient Boosting	0.851	0.857	0.895	0.876	0.92
AdaBoost	0.847	0.856	0.890	0.872	0.92
SVM	0.850	0.852	0.900	0.874	0.92
MLP Classifier	0.847	0.844	0.906	0.874	0.92
Decision Tree	0.776	0.805	0.815	0.810	0.77
Random Forest	0.841	0.850	0.885	0.867	0.91
Stacking Ensemble	0.852	0.852	0.852	0.851	0.92

D. Heat-Map

The heatmap in figure 7 shows the comparison of the performance of different models of machine learning on the test dataset by five evaluation measures: Accuracy, Precision, Recall, F1 Score, and AUC (Area Under the Curve) visualization. A score of a specific model regarding a specific metric is provided on each cell on the heatmap, and the colour gradient shows the level of performance, where a darker value is better. The compared models are conventional classifiers such as the Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tree, and more modern approaches as Random Forest, Gradient Boosting, AdaBoost, Support Vector Machine (SVM), Multi-layer Perceptron (MLP) Classifier, and Stacking Classifier.

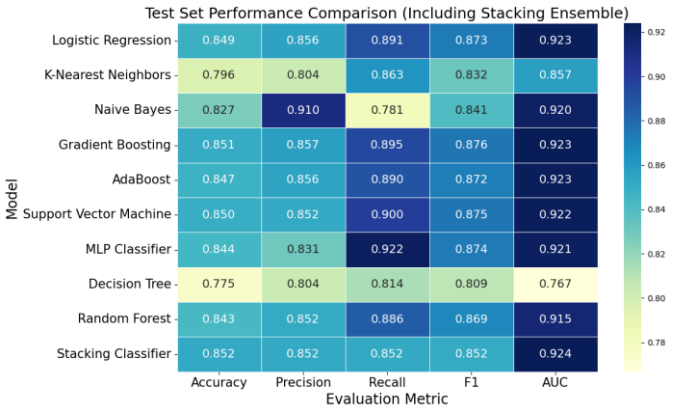

FIGURE 7: Model Performance Comparison (Heat-Map)

Based on the analysis, one notices that the majority of models including the highest individual performance by models are the Gradient Boosting, SVM, AdaBoost, MLP, and the Stacking Classifier which are having the same high performance in each of the evaluation measures where the AUC measures went up to 0.92. Random Forest and Logistic Regression also demonstrate respectable results. Naive Bayes can be distinguished by high precision (0.91), but is lower when it comes to recall (0.78) and thus may fail to recognize more true positives. K-Nearest Neighbors and Decision Tree fare quite poorly, especially, the Decision Tree showing the lowest AUC (0.77) and overall average performance in rest of the metrics. In general, the ensemble methods and sophisticated classifiers are more likely to perform better than simple models, although stacking ensures relatively balanced and robust integrity and, therefore, it is well justified to add stacking as an effective ensemble.

E. Pie Chart

The pie chart in figure 8 named Comparison of Test Accuracy: Stacking vs Other Models is a graphic display that expresses the performance differences in test accuracy of both a stacking ensemble model against itself as compared to the combined results of many individual models. The chart is split into two parts, the yellow part captures the stacking ensemble model, which represents 50.57 percent of the total number, whereas the light blue part shows the average performance of the rest of the models, which makes 49.43 percent of the overall. The fact that the stacking model is more popular by only a few shows that there is a minor yet noticeable difference in terms of performance.

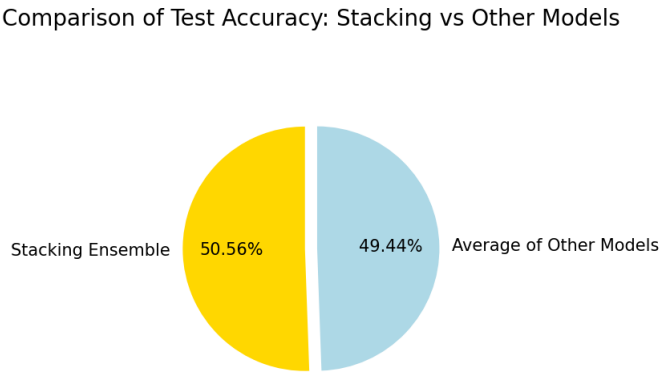

FIGURE 8: Average Model Performance Compression (Pie Chart)

In the pie chart, one can observe that stacking ensemble model is a little bit better than average accuracy of separate models by about 1.14%. It is a relative result, which is a mirror reflection of the soundness of ensemble learning methods such as stacking, which are expected to have a better generalization and accuracy due to a combination of model predictions. The finding justifies the application of stacking in situations whereby slight improvements in performance are significant to consider, particularly in delicate applications where the stability of models and their correctness of the user are paramount.

In the proposed study, stacking ensemble was embraced in

order to optimize the robustness and accuracy of the student depression detection. The common traditional classifiers (Logistic Regression, Support Vector Machines or Decision Trees) usually perform rather poorly at single aspects, such as large bias or large variance. A meta-ensemble method called stacking combines predictions of two or more base learners and employs a meta-learner to refine the base learner output by optimizing the outputs of the base learners, such stacking leading to better generalization, and to better predictive performance [[31]], [[32]].

As opposed to individual models, stacking combines strengths of all algorithms. As an example, non-linearities are easily dealt with in tree-based models and simpler patterns suit linear models better. The outputs of their ensemble called the stacking ensemble combine their outputs and thus underfitting as well as overfitting is reduced [[31]]. The stacking model described in this paper had an accuracy of 85.2% and AUC of 0.92, which is proof that it beats all of the single-stage classifiers in identifying depression risk in students.

Moreover, mental health applications need to provide model transparency and interpretability since they require sensitivity. SHAP (SHapley Additive exPlanations) and feature importance were integrable and explained such explainable AI so that the system was not only accurate but it was trustworthy enough and explainable [[39]], [[40]].

Due to its robust and balanced results on several measures (accuracy, precision, recall, F1-score, AUC), stacking ensemble was an accurate, fashionable, and ethically sound solution to detect early signs of depression among students.

IV. Conclusion

In this research, we addressed the challenge of early depression detection among students—a critical need given the difficulty of predicting depression in its initial stages and the irreversible consequences of delayed intervention. To tackle this problem, we developed a comprehensive stacking ensemble framework using a large-scale survey dataset from Kaggle to improve early identification and intervention. Systematic preprocessing of the data, chi-square-based feature selection, and combination of different sets of machine learning algorithms such as Logistic Regression, Support Vector Machine, K-Nearest Neighbors, MLP, AdaBoost, Random Forest, and Gradient Boosting with each their strengths thusly coupled together as a meta-classifier, allowed us to leverage their advantages. The resulting model conducted very well and consistently in balance with 85.2 percent accuracy and an AUC of 0.92 on new test data, and in addition, it showed quite high levels of precision, recall, and F1-scores, averaging above 0.85. Our findings showed that academic pressure, study satisfaction, sleep duration, suicidal thoughts, and financial stress were attributed as important factors of risking depression in the studies. The validation results affirmed the robust model and its applicability to general or cross-sectional metrics that were close to those of the test. The future direction of work should consider implementing explainable AI methods to make models even more transparent, as well as pilot testing in educational institutions to assess the effectiveness of real-life application in the provision of timely support in mental health.

.

References

[1] World Health Organization, 2021, Depression, World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[2] D. Eisenberg, A.M. Wright, et al., Prevalence and correlates of depression, anxiety and suicidality among university students, Am. J. Orthopsychiatry, 77: 534 and 5342, 2007.

[3] R. P. Auerbach et al., WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders., J. Abnorm. Psychol., 127, 7, 623638, 2016.

[4] A. B. Shatte, D. M. Hutchinson and S. J. Teague, Machine learning in mental health: A scoping review of methods and applications, Psychol. Med., 49, 9 (2019).

[5] D. B. Dwyer, P. Falkai, and N. Koutsouleris, Machine learning approaches for clinical psychology and psychiatry, Annu. Rev. Clin. Psychol. 14 (2018), 91118.

[6] L. Rokach, Ensemble-based classifiers, Artif., 2009, 62 (4), 159-182. Intell. Rev., 33, 112, 1-39, 2010.

[7] O. Sagi and L. Rokach, Ensemble learning: A survey WIREs Data Min. Knowl. Discov. 2018;8:4,e1249.

[8] Kaggle, Student Depression Dataset, [Online]. Available: https://www.kaggle.com/datasets

[9] L. Wang, Y. Jia, C. Sun and H. Meng, "Exploring the connection between mental health and academic performance among university students," BMC Psychiatry, vol. 19, p. 1, 2019.

[10] J. Hunt and D. Eisenberg, Mental health issues and help-seeking amongst college students, J. Adolesc. Health 46 (2010) 3 10.

[11] S. Van Buuren and K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R, J. Stat. Softw., 45, 3 (2011), 1-67.

[12] R. J. Little and D. B. Rubin, Statistical Analysis with Missing Data. Wiley, 2019.

[13] F. Pedregosa, et al., Scikit-learn: Machine Learning in Python, J. Mach. Learn., (2012) Res., vol. 12, pp 2825-2830, 2011.

[14] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2011.

[15] C. C. Aggarwal, Outlier Analysis (outliers). Springer, 2017.

[16] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in IJCAI, vol. 14, no. 2, p. 11371145, 1995.

[17] M. L. McHugh, The chi-square test of independence, Biochem. Med., 23 2 (2013): 143 149.

[18] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3, 1157-1182, vol 3, 2003.

[19] C. F. Dormann et al., Collinearity: A review of methods to deal with it and a simulation study to evaluate their performance, Ecography, 36, 1: 27-46, 2013.

[20] G. Chandrashekar and F. Sahin, A survey on feature selection methods, Comput., 37(1-3), 196-210 (2008). Electr. Eng., vol. 40, no. 1, 16 28, 2014.

[21] M. Kuhn and K. Johnson, Applied Predictive Modeling. Springer, 2013.

[22] Pattern Recognition and Machine Learning, C. M. Bishop. Springer, 2006.

[23] Friedman(J. H.), 2001, 200gree Greedy function approximation: A gradient boosting machine, Ann. Stat., no. 1189-1232.

[24] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning. Cambridge, 2016.

[25] T. Cover and P. Hart, nearest neighbor pattern classification, IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 2127, 1967.

[26] C. Cortes and V. Vapnik, Support-vector networks, Mach. Learn., 20 (3), 273 297, 1995.

[27] J. R. Quinlan, The induction of decision trees, Mach. Learn., 1, 81-106, 1986.

[28] L. Breiman, Random forests,Mach. Learn., 45, 1, 532, 2001.

[29] Y. Freund and R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting J. Comput. Syst. Sci., 55, 1, 119139, 1997.

[30] D 2004 H. Zhang, The optimality of Naive Bayes, in Proc. FLAIRS Conf., vol. 1, 562567, 2004.

[31] D.H. Wolpert, Stacked generalization, Neural Netw., 5 (1992), 241 259.

[32] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Chapman, 2012.

[33] J. Bergstra and Y. Bengio "Random search to optimize hyper-parameters," J. Mach. Learn. Res., 13 (2012) 281 305 2012.

[34] R. Agrawal, A. Imielinski and T. Swami, Mining association rules between sets of items in large databases, in Proc. SIGMOD, pp. 207216, 1993.

[35] A. K. Jain, Data clustering: 50 years beyond K-means Pattern Recognit. Lett., 31: 8, 651 666, 2010.

[36] T. Fawcett, Pattern Recognit. An introduction to ROC analysis. Lett., 27 (2006), no. 8, 861 874.

[37] T. Saito and M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when assessing binary classifier on imbalanced datasets, PLOS ONE., Vol. 10, No. 3, E0118432, 2015.

[38] D. M. Powers, Evaluation: From precision, recall and F -measure to ROC, informedness, markedness and correlation, J. Mach. Learn., vol. 8, no. 3, pp. 500-511, 2008. Technol., 2, 1, 3763, 2011.

[39] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in Adv. Neural Inf. Process. Syst., 30:2017.

[40] C. Molnar, Interpretable Machine Learning, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[41] A. Altmann, T. C. Wong and D. G. Altmann, Permutation importance: A corrected measure of feature importance, Bioinformatics, 26, 1340-1347, 2010.

[42] F. Chollet et al. Keras: Deep Learning for humans, 2015. [Online]. Available: https://keras.io