Effective Machine Learning Based Techniques for Predicting Depression

Author

Lamadi Youssef
Data Science & Engineering Student,
National School of Applied Sciences Al Hociema
youssefflamadi7@gmail.com

Supervision

Aziz Khamjane
Professor of machine learning,
National School of Applied Sciences Al Hociema
akhamjane@uae.ac.ma

***Abstract*—Depression is a critical global issue with profound implications for mental health and overall well-being. Advances in machine learning and the availability of extensive datasets present an opportunity to enhance the early detection of depression. This research leverages a large dataset of 140,000 entries containing diverse features such as age, sleep duration, work/study hours, financial stress, and dietary habits to train and evaluate multiple machine learning models, including CatBoost, XGBoost, Logistic Regression, and Random Forest classifiers. After rigorous evaluation, CatBoost emerged as the most effective model. Subsequently, this model was fine-tuned and validated on a smaller, high-quality dataset of 120 samples, enriched with additional psychosocial factors such as social media usage, loneliness frequency, and marital status. This two-stage approach demonstrates the potential of combining large-scale training with fine-tuning on targeted data to enhance the performance and applicability of depression prediction models. The findings of this study aim to assist clinicians in identifying key factors contributing to depression and pave the way for more personalized and accurate mental health interventions.**

I. Introduction

Depression is one of the most common types of mental illness, significantly impacting individuals' ability to function at work, school, and within family settings. In severe cases, it can even lead to self-harm. Despite its prevalence, only a limited number of studies have prospectively examined a broad range of predictors across multiple domains for new-onset (incidental) depression in adulthood. With the growth of datasets relevant to mental health and the advancement of machine learning, there is now an opportunity to develop intelligent systems capable of recognizing early signs of depression. Symptoms such as negative thinking, reduced concentration, and decreased productivity can be mitigated through early diagnosis, improving the quality of life for both individuals and their families.

The primary objective of this research is to develop an efficient and scalable model to assess whether a person is depressed and to identify effective machine learning techniques for diagnosing depression. According to the World Health Organization (WHO) [1], depression affects approximately 3.8% of the global population, with women more likely than men to experience various forms of depression. Following the COVID-19 outbreak, depression has become an even more severe public health issue, with 322 million people suffering from the condition at any given time. Depression has been linked to a range of chronic illnesses, including diabetes and heart disease, making it the second most significant risk factor for developing such conditions [1]. Additionally, severe depression is a major driver of suicidal tendencies, contributing to over half of the 0.8 million suicides reported worldwide annually, according to statistics.

The growing demand for machine learning algorithms to uncover meaningful patterns from data spans a variety of industries, including healthcare. While these algorithms have been widely employed in psychiatry and other areas of healthcare, their application in mental health remains relatively underutilized. Statistical tests have long been used in psychological profiling and assessments, providing a solid foundation for understanding mental health. However, with the advent of machine learning and its increasing prominence—particularly following the Cambridge Analytica scandal—there has been a shift in focus. Researchers in personality assessment and mental health analysis are progressively adopting machine learning techniques over traditional statistical methods to address prediction validity concerns associated with probabilistic reasoning.

II. Methodology

A. Data Description

This study utilizes two primary datasets: one collected through a survey administered at the National School of Applied Science of Al Hoceima and a second larger dataset sourced from Kaggle to train and test the machine learning models.

Survey Dataset

The survey was conducted between November 13, 2024, and December 10, 2024, at the National School of Applied Science of Al Hoceima. It was designed to gather both socio-demographic and psychological data from participants, with the aim of understanding factors influencing depression. The survey covered various sections, including questions on age, sleep duration, work/study hours, financial stress, social media usage hours, and psychosocial aspects such as body image satisfaction, loneliness, and marital status.

The dataset consists of responses from 120 participants, with 24 identified as depressed and 96 as not depressed. This dataset was used primarily for fine-tuning machine learning models after they were initially trained on the larger Kaggle dataset. Given the relatively small sample size, the survey dataset served as a focused source of information, helping to improve the prediction accuracy for depression specifically within the student population of the National School of

Applied Science Al Hoceima.

Kaggle Dataset

To supplement the survey dataset, a larger dataset was used for initial model training and evaluation. The Kaggle dataset, while containing fewer features compared to the survey dataset, still provided valuable data on aspects such as age, gender, sleep patterns, social media activity, and depression-related indicators. This dataset was instrumental for training and evaluating machine learning models such as CatBoost, XGBoost, Logistic Regression, and Random Forest.

By leveraging the larger Kaggle dataset, the study was able to train models on a more extensive set of entries, capturing general patterns in depression. After identifying CatBoost as the best-performing model, it was further fine-tuned using the survey data to enhance its localization and relevance to the target population at the National School of Applied Science Al Hoceima.

By utilizing both datasets, the research aims to develop a robust model that accurately predicts depression, providing insights relevant to both broader and localized populations.

B. Data Preprocessing

Pre-processing strategies primarily focus on transforming raw data into a comprehensible format. What this implies is that the computer can readily interpret, anticipate, and analyse what is in the data using different machine learning methods.

1) Feature Selections: In this study, the process of feature selection played a crucial role in optimizing the prediction model for depression. Effective feature selection ensures that the model focuses on the most relevant factors, improving both accuracy and computational efficiency while minimizing noise from irrelevant features. Two datasets were utilized in this research: a large dataset with diverse variables, and a survey dataset containing additional psychosocial factors. Each dataset includes a unique set of features that capture essential dimensions of depression, ranging from demographic and lifestyle factors to academic, professional, and psychological attributes. The tables below summarize the variables and their possible values for each dataset, providing a comprehensive overview of the features used in training the prediction model.

TABLE I
VARIABLES FOR PREDICTING DEPRESSION IN A LARGE DATASET.

	Variable	Possible Values
1	Age	18, 19, 21, 22, 23, 24, 25, ...,60
2	Gender	Male, Female
3	Sleep Duration	Less than 5 hours, 5-6 hours, 7-8 hours, More than 8 hours
4	Working Professional or Student	Working Professional, Student
5	Profession	Chef, Teacher, Student, Business Analyst, ...
6	Study Degree	High School, Bachelor's Degree, Master's Degree, PhD
7	Work/Study Hours	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12
8	Financial Stress	1, 2, 3, 4, 5
9	Dietary Habits	Healthy, Moderate, Unhealthy
10	Suicidal Thoughts	Yes, No
11	Family Mental Illness History	Yes, No
12	Depression	0 (Not Depressed), 1 (Depressed)
13	Academic/Work Pressure	1.0, 2.0, 3.0, 4.0, 5.0
14	Study/Job Satisfaction	1.0, 2.0, 3.0, 4.0, 5.0
15	Age_Group	17-19, 20-22, 23-25, 26-28, 29-31, 32-34, 35-37, 38-40, 41-43, 44-46, 47-49, 50-52, 53-55, 56-60

TABLE II
VARIABLES FOR PREDICTING DEPRESSION IN SURVEY DATASET.

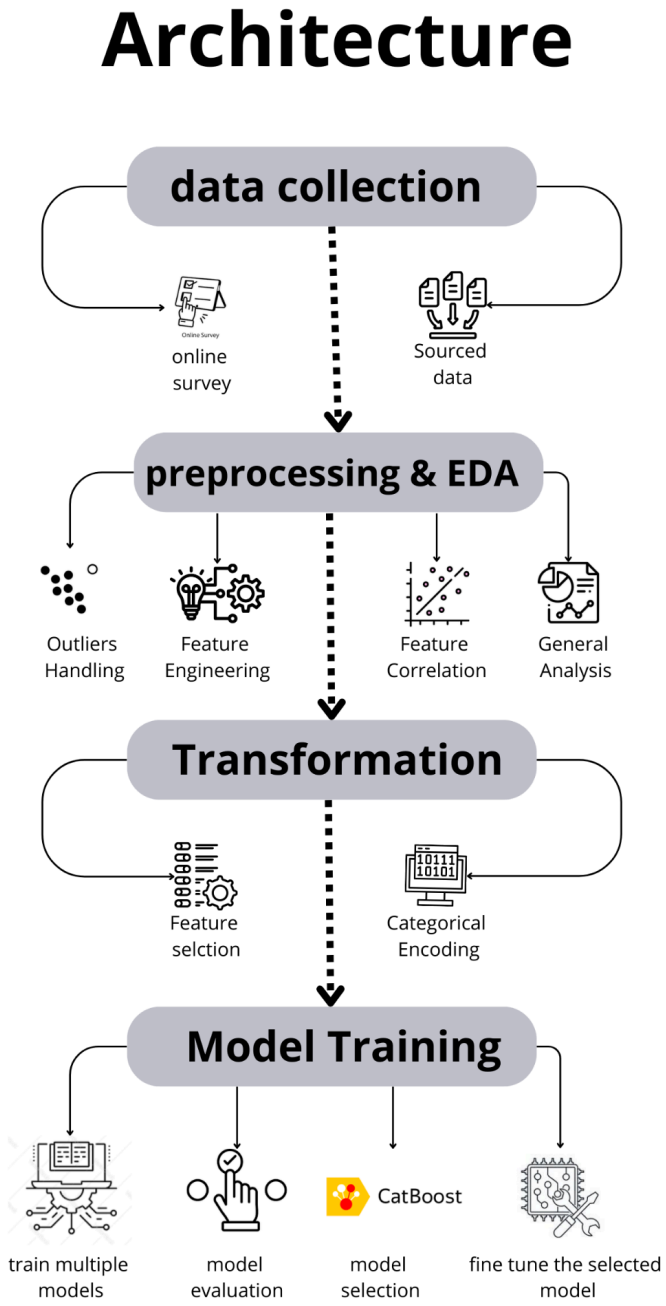
	Variable	Possible Values
1	Age	18, 19, 21, 22, 23, 24, 25, ...,60
2	Gender	Male, Female
3	Sleep Duration	Less than 5 hours, 5-6 hours, 7-8 hours, More than 8 hours
4	Working Professional or Student	Working Professional, Student
5	Profession	Chef, Teacher, Student, Business Analyst, ...
6	Study Degree	High School, Bachelor's Degree, Master's Degree, PhD
7	CGPA	5.0, 6.0, 6.5, 7.0, 7.41, 7.5, 8.0, 8.35, 10.0
8	Work/Study Hours	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12
9	Academic/Work Pressure	1, 2, 3, 4, 5
10	Study/Job Satisfaction	1, 2, 3, 4, 5
11	Financial Stress	1, 2, 3, 4, 5
12	Dietary Habits	Healthy, Moderate, Unhealthy
13	Suicidal Thoughts	Yes, No
14	Family Mental Illness History	Yes, No
15	Marital Status	1.0, 2.0, 3.0, 4.0, 5.0
16	Religious Person	Yes, No
17	Living with Family	Yes, No
18	Living Environment Satisfaction	Yes, No
19	Night/Day Sleep Preference	At night, During the day
20	Physical Activity Frequency	Never, Sometimes, Regularly
	Smoking	Yes, No
21	Alcohol Consumption	Yes, No
22	Social Media Usage Hours	Less than 1 hour, 1-2 hours, 2-4 hours, More than 4 hours
23	Body Image Satisfaction	1, 2, 3, 4, 5
24	Self-Comparison	Yes, No
25	Loneliness Frequency	Never, Sometimes
26	Age_Group	17-19, 20-22, 23-25, 26-28, 29-31, 32-34, 35-37, 38-40, 41-43, 44-46, 47-49, 50-52, 53-55, 56-60
27	Depression	0 (Not Depressed), 1 (Depressed)
28		

2) Data-set Splitting: This study utilised 80% of the data set for training. The remaining 20% of the data-set has been utilised for testing.

C. System Architecture

The architecture of the proposed system is designed to efficiently process the two datasets and apply machine learning algorithms to predict depression. The workflow begins with data collection from the survey and Kaggle datasets, followed by essential preprocessing steps such as data cleaning, feature selection, and handling missing values. Once the data is prepared and transformed, it is fed into different machine learning algorithms for training and evaluation. After selecting the best-performing model, further fine-tuning is done using the survey data to improve the model's accuracy for the target population.

The diagram below illustrates the overall architecture of the system, depicting the key stages from data collection through to model training and prediction.



D. Algorithm selection

In this research, CatBoost, XGBoost, Logistic Regression and Random Forest classifier are used. The specifics of these classifiers are outlined below.

1) CatBoost Classifier: CatCatBoost (Categorical Boosting) is a gradient boosting algorithm specifically designed to handle categorical data efficiently. Unlike traditional gradient boosting methods, CatBoost can process categorical features without the need for extensive preprocessing, such as one-hot encoding. This is achieved through a unique combination of ordered boosting and target-based transformations, which prevent overfitting and ensure optimal model performance.

The key principle of CatBoost is to construct an ensemble of decision trees iteratively. During each iteration *t*, the algorithm builds a new tree $h_t(x)$ that minimizes the loss function *L* by focusing on the residuals (errors) of the previous iterations. Mathematically, the model prediction at iteration *t* is updated as:

$$F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x)$$

where:

- $F_t(x)$ is the prediction of the ensemble at iteration *t*,
- $F_{t-1}(x)$ is the prediction of the ensemble from the previous iteration,
- η is the learning rate,
- $h_t(x)$ is the new decision tree added at iteration *t*.

CatBoost incorporates a feature known as symmetric trees, which helps reduce overfitting and improves computational efficiency by maintaining a consistent tree structure. In symmetric trees, all nodes at the same depth split on the same feature, reducing the search space and ensuring faster training.

CatBoost also applies an ordered boosting technique to maintain unbiased gradient estimation. Instead of relying on all data at once to calculate gradients, it uses only past observations during training. This can be expressed as:

$$\hat{y}_i = g(x_i \mid \mathcal{D}_{past})$$

where:

- $\hat{y}_i$ is the predicted value for observation *i*,
- $g(x_i \mid D_a)$ is the model trained only on past data (i.e., data preceding x_i).

This innovative approach significantly improves the handling of categorical features and enhances the model's robustness. For categorical variables, target-based transformations are performed using conditional distributions, which can be expressed as:

$$TargetEncoding = \frac{\sum_{j \in C} y_j}{|C|}$$

where:

- C represents all data points in the same category as the current observation,
- y_j is the target value for data point *j*,
- |C| is the total number of data points in the category.

The primary goal of CatBoost is to find patterns and relationships in high-dimensional data while handling missing values and categorical variables effectively. Its ability to leverage categorical features directly and its built-in mechanisms for reducing overfitting make it a highly efficient and accurate algorithm for classification tasks.

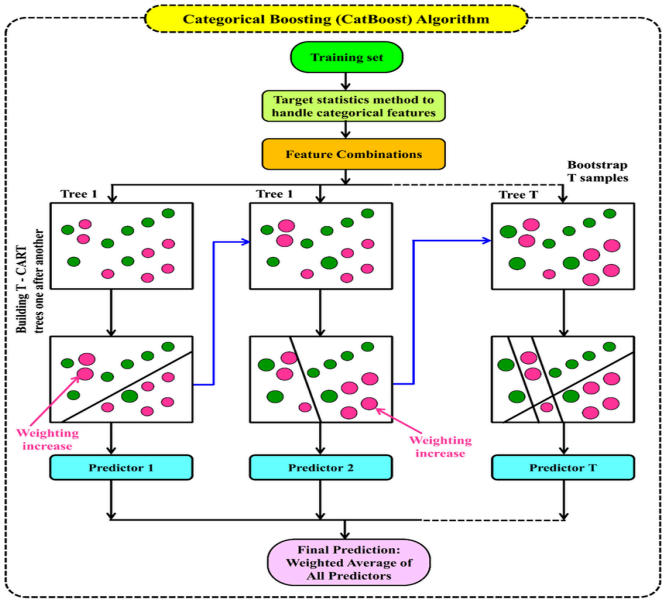


Fig. 1. How does the CatBoost algorithm Work?[2]

2) XGBoost Classifier: XGBoost (eXtreme Gradient Boosting) is an efficient, scalable machine learning algorithm that improves traditional gradient boosting by optimizing speed and accuracy. It builds an ensemble of decision trees, each correcting errors from the previous ones, and uses a regularized objective function to prevent overfitting. Key features include second-order optimization, parallel processing, and the ability to handle missing data. These enhancements make XGBoost particularly powerful for large datasets and high-dimensional tasks, offering both strong performance and computational efficiency.

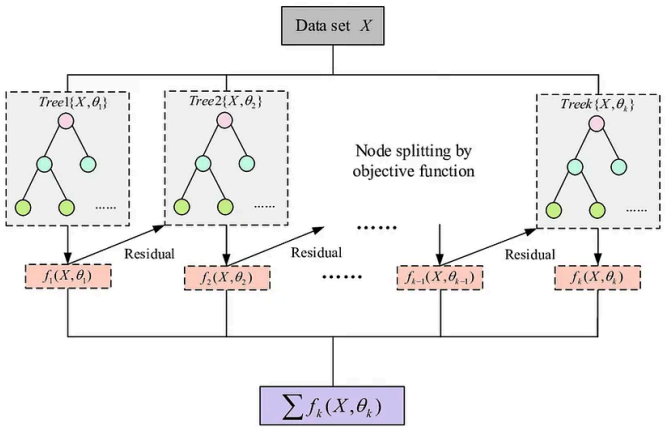


Fig. 1. How does the XGBoost algorithm Work?[3]

3) Logistic Regression: Logistic Regression is a statistical method used for binary classification problems, where the target variable has only two possible outcomes (e.g., depressed or not depressed). Despite its name, it is a linear model that predicts the probability of a data point belonging to

a particular class by using a logistic function, also known as the sigmoid function is:

$$\text{Sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

Here, $z = w\,x + b$, where w represents the weights, x is the feature vector, and b is the bias term. The weights and bias are optimized during training to minimize the difference between the predicted probabilities and the actual target values.

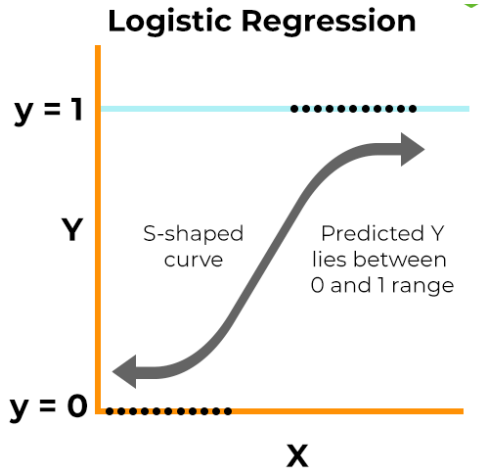


Fig. 2. How does the Logistic Regression algorithm Work? [4]

4) Random Forest Classifier: The Random Forest Classifier is a robust and versatile ensemble learning algorithm used for both classification and regression tasks. It operates by constructing multiple decision trees during training and combining their outputs to make predictions. This approach reduces overfitting and improves the model's generalization performance compared to individual decision trees.

The key principle of Random Forest is to introduce randomness during the training process, which ensures that the individual trees are diverse and less correlated. Two main techniques contribute to this randomness:

Bootstrap Aggregation (Bagging): Random Forest creates multiple subsets of the training data by sampling with replacement. Each subset is used to train an individual decision tree, ensuring that no two trees are identical.

Random Feature Selection: At each split in a tree, the algorithm selects a random subset of features to determine the best split. This prevents the trees from always relying on the same dominant features, encouraging diversity.

The Random Forest Classifier makes predictions by aggregating the outputs of all the trees. For classification tasks, it uses a majority voting system where the class predicted by the most trees is selected as the final output. This ensemble method reduces the risk of overfitting and increases the model's robustness to noise and variability in the data.

Random Forest is non-parametric, meaning it does not assume any specific distribution of the input data. It performs well on datasets with high dimensionality and can handle missing values and categorical variables effectively. Moreover, it provides feature importance metrics, which help in understanding the contribution of each feature to the model's predictions.

Due to its ability to handle complex data and reduce overfitting, Random Forest is widely used in various applications, from medical diagnostics to financial modeling and beyond.

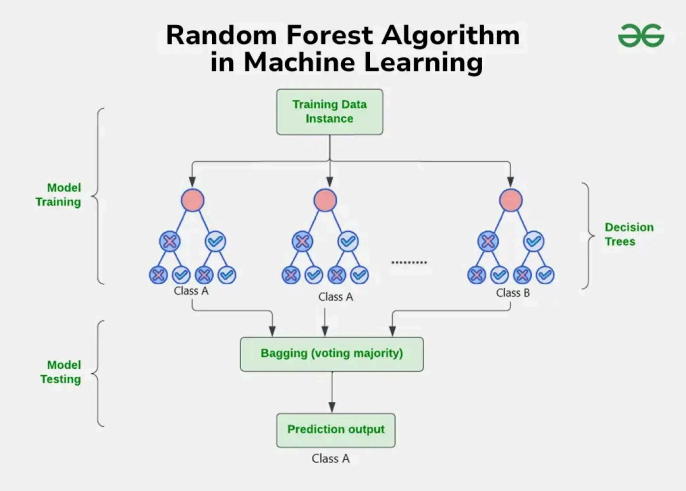


Fig. 4. How does the Random Forest algorithm Work? [5]

E. *Evaluation metrics*

1) Accuracy:

Accuracy measures the proportion of correctly classified instances out of all instances. It is a simple metric but may not be reliable when there is class imbalance.

Mathematical Expression:

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

Where:

- TP = True Positives (correctly predicted positive cases)
- TN = True Negatives (correctly predicted negative cases)
- FP = False Positives (incorrectly predicted positive cases)
- FN = False Negatives (incorrectly predicted negative cases)

2) Precision:

Precision measures the proportion of true positive predictions relative to the total predicted positives. It indicates how many of the predicted positive cases were actually positive.

Mathematical Expression:

$$Precision = \frac{TP}{TP + FP}$$

3) Recall:

Recall measures the proportion of actual positive cases that are correctly identified by the model. It indicates how well the model captures the positive class.

Mathematical Expression:

$$Recall = \frac{TP}{TP + FN}$$

4) F1-Score:

The F1-score is the harmonic mean of precision and recall. It balances the two metrics, providing a single score that accounts for both false positives and false negatives, making it particularly useful when dealing with imbalanced datasets.

Mathematical Expression:

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

5) ROC-AUC (Receiver Operating Characteristic - Area Under Curve):

ROC-AUC evaluates the model's ability to discriminate between positive and negative classes. The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate. The AUC (Area Under the Curve) provides a single value that summarizes the model's performance across all thresholds.

Mathematical Expression: The ROC curve is plotted using:

$$TruePositiveRate(Recall) = \frac{TP}{TP + FN}$$

$$FalsePositiveRate = \frac{FP}{FP + TN}$$

AUC is the area under the ROC curve, where an AUC of 0.5 indicates no discriminative power, and an AUC of 1 indicates perfect classification.

I. Results

After applying various machine learning classifiers, we evaluated their performance using different metrics: accuracy, precision, recall, F1-score, and ROC-AUC. The following tables show the experimental findings for depression prediction, with and without using `RandomizedSearchCV` for hyperparameter optimization.

TABLE I
RESULTS OF THE PROPOSED CLASSIFIERS WITHOUT `RANDOMIZEDSEARCHCV`.

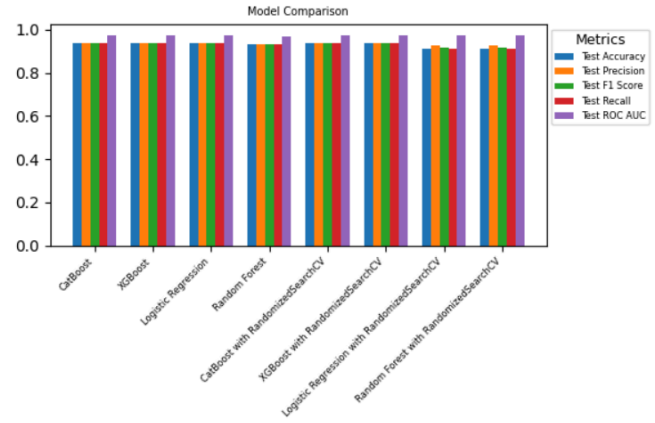
Model	Train Accuracy	Train Precision	Train Recall	Train F1 Score	Train ROC AUC	Test Accuracy	Test Precision	Test Recall	Test F1 Score	Test ROC AUC
CatBoost	0.950	0.950	0.950	0.950	0.984	0.940	0.939	0.940	0.940	0.974
XGBoost	0.948	0.948	0.948	0.948	0.984	0.938	0.937	0.938	0.937	0.973
Logistic Regression	0.937	0.936	0.937	0.937	0.974	0.937	0.936	0.937	0.937	0.973
Random Forest	1.000	1.000	1.000	1.000	1.000	0.935	0.934	0.935	0.934	0.970

TABLE II
RESULTS OF THE PROPOSED CLASSIFIERS WITH `RANDOMIZEDSEARCHCV`.

Model	Train Accuracy	Train Precision	Train Recall	Train F1 Score	Train ROC AUC	Test Accuracy	Test Precision	Test Recall	Test F1 Score	Test ROC AUC
CatBoost with RandomizedSearchCV	0.941	0.941	0.941	0.941	0.977	0.940	0.939	0.940	0.940	0.975
XGBoost with RandomizedSearchCV	0.942	0.941	0.942	0.941	0.977	0.940	0.939	0.940	0.939	0.975
Logistic Regression with RandomizedSearchCV	0.914	0.931	0.914	0.919	0.974	0.912	0.929	0.912	0.916	0.973
Random Forest with RandomizedSearchCV	0.914	0.931	0.914	0.919	0.974	0.912	0.929	0.912	0.916	0.973

II. Result Comparison

Fig. 3. Performance Comparison Graph of Base and Proposed Model



CatBoost:

CatBoost achieved **high test F1-score (0.940)** and **test ROC AUC (0.974)**, indicating excellent performance in distinguishing between depressed and non-depressed classes. Its balanced precision (0.939) and recall (0.940) make it the most suitable model for this imbalanced dataset.

CatBoost's ability to handle categorical features natively and its robust generalization make it the top-performing model.

XGBoost:

XGBoost achieved a **test F1-score of 0.937** and a **ROC AUC of 0.973**, slightly trailing CatBoost. Its performance is comparable and still effective, especially in terms of balancing precision and recall.

This model could be an alternative to CatBoost if interpretability or scalability is preferred.

Logistic Regression:

Logistic Regression performed relatively well, with a **test F1-score of 0.937** and a **ROC AUC of 0.973**. While its simplicity and interpretability are advantageous, it may not capture complex patterns in the data as effectively as ensemble methods.

Random Forest (without tuning):

Random Forest, before tuning, suffered from **overfitting**, evident from perfect training metrics (F1-score = 1.000) but relatively lower test F1-score (0.934) and ROC AUC (0.970). This overfitting indicates poor generalization to unseen data, making it less reliable.

Random Forest (with RandomizedSearchCV):

After hyperparameter tuning, Random Forest showed improved generalization, but its test F1-score dropped to **0.916**. While the tuning mitigated overfitting, the performance remained lower than CatBoost and XGBoost, making it less optimal for this study.

Impact of Hyperparameter Tuning:

For CatBoost and XGBoost, tuning had minimal impact on performance as their default settings were already well-optimized. However, tuning significantly improved Random Forest's generalization by controlling overfitting.

III. Conclusion and Future Work

This research study developed machine learning models aimed at predicting depression and identifying key socio-demographic and psychological attributes that contribute to this condition. By comparing various algorithms, including CatBoost, XGBoost, Logistic Regression, and Random Forest, the study demonstrated that CatBoost and XGBoost achieved the most balanced performance across evaluation metrics such as F1-score, Recall, Precision, and ROC AUC, especially when addressing the class imbalance present in the dataset. These models outperformed others, such as Logistic Regression and Random Forest (without hyperparameter tuning), which faced issues like overfitting in some cases. The study also highlighted the importance of feature-rich datasets in enhancing predictive performance, with the survey dataset offering unique insights compared to the larger Kaggle dataset.

Unlike other studies that primarily focused on accuracy, this research emphasized the importance of balanced metrics, particularly due to the imbalanced class distribution of depression (18.2% depressed vs. 81.8% non-depressed). The results indicate that the CatBoost and XGBoost models are well-suited for depression prediction, providing robust and interpretable outputs that could potentially aid clinicians and researchers in identifying at-risk individuals.

Future work will focus on further enhancing the dataset by integrating clinical bio-markers and other physiological data to improve predictive accuracy and enable more comprehensive analysis. Additionally, expanding the survey dataset to include more diverse populations and incorporating emotional and behavioral features from social media activity will be explored. These enhancements will provide richer inputs for machine learning models and allow for deeper insights into the socio-psychological factors contributing to depression.

Lastly, future research will investigate more advanced ensemble techniques and deep learning approaches, as well as the potential impact of real-time data collection from wearable devices and online interactions. These advancements aim to refine the accuracy, sensitivity, and practical applicability of the models, ultimately contributing to improved mental health diagnosis and treatment strategies, with a long-term goal of reducing depression and anxiety rates while supporting suicide prevention efforts.

References

[1] "Depression," World Health Organization, Sep. 13, 2021. https://www.who.int/news-room/fact-sheets/detail/depression.
[2] https://medium.com/@mohan-gupta/catboost-algorithm-2156129d740d
[3] https://medium.com/@prathameshsonawane/xgboost-how-does-this-work-e1cae7c5b6cb
[4] https://www.simplilearn.com/tutorials/machine-learning-tutorial/logistic-regression-in-python.
[5] https://builtin.com/data-science/random-forest-algorithm
[6] github repository click here