Machine Learning Approach for Depression Prediction using XGBoost Classification

[1]AUTHOR'S NAME(M.D Sansala Gangadari), [2]CO-AUTHOR'S NAME(Dr. Uthayasanker Thayasivam)

[1]Undergraduate Student, Department of Computer Science and Engineering, University of Moratuwa , SriLanka
[2]Senior Lecturer , Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka

Email: Sansala.22@cse.mrt.ac.lk, [2]rtuthaya@cse.mrt.ac.lk
Contact: [1]+94 77 33 51 747, [2] +94 76 394 6578

Abstract:. This study develops a machine learning model to predict depression using an XGBoost classifier. Depression is a serious health issue that affects millions of people worldwide. Detecting it early with automated tools can help provide timely support and treatment. Our model uses information about a person's background and daily habits to determine whether they may be experiencing depression. The model performed well in testing, showing that machine learning can be useful in mental health screening. This research helps advance the use of technology in mental health by providing a data-driven way to assess the risk of depression.

Index terms: *Mental Health Screening, Machine Learning for Depression Prediction, XGBoost Classifier, Computational Psychiatry*

I. INTRODUCTION

Depression is a common mental health disorder affecting more than 264 million people globally (World Health Organization, 2020). Despite its prevalence, depression remains underdiagnosed, with many individuals not receiving timely treatment. Early detection is crucial for effective intervention and improved outcomes.

Traditional screening methods rely heavily on clinical interviews and self-report questionnaires, which can be subject to various biases. Machine learning approaches offer an alternative path to identifying depression risk factors by analyzing patterns in data that may not be immediately apparent to human clinicians.

This research focuses on building and testing a model to predict depression using the XGBoost algorithm. XGBoost is a powerful machine learning method that works well for classification tasks. Our goal is to create a tool that can help support clinical assessments of depression.

II. RELATED WORK

Machine learning techniques have been widely applied in mental health prediction, including depression detection. Several studies have demonstrated the potential of machine learning models to identify patterns in behavioral and clinical data that are indicative of depression.

1) **Traditional Methods vs. Machine Learning Approaches**

Traditional depression screening primarily relies on clinical interviews and self-reported surveys, such as the Patient Health Questionnaire (PHQ-9) and the Beck Depression Inventory (BDI). While these tools are widely used and clinically validated, they are subject to biases, including response distortion and subjective interpretation. Machine learning models aim to mitigate these challenges by identifying hidden patterns within large-scale data, enabling more objective and automated screening.

2) **Existing Machine Learning Models for Depression Prediction**

Various machine learning models have been explored for depression prediction. Logistic Regression, Support Vector Machines (SVM), Random Forest, and deep learning approaches have been employed to classify individuals as depressed or non-depressed. Studies applying ensemble methods, such as Gradient Boosting Machines (GBM) and XGBoost, have demonstrated improved predictive performance by combining multiple weak learners to enhance accuracy and robustness.

For example, Tadesse et al. (2019) utilized SVM and deep learning techniques on social media data, achieving promising results in detecting depressive symptoms from textual and behavioral cues. Similarly, Orabi et al. (2018) compared different classifiers and found that ensemble learning techniques outperformed traditional models in depression detection using Twitter data.

3) **Feature Engineering and Data Sources**

The effectiveness of machine learning models depends significantly on feature selection and data sources. Prior research has incorporated demographic data, behavioral patterns, physiological markers, and social media activity to improve prediction accuracy. Li et al. (2020) demonstrated that integrating linguistic and psychological features extracted from online conversations enhances depression classification. Other studies have examined wearable device data, such as sleep patterns and heart rate variability, as potential predictors of mental health disorders.

4) **Challenges and Limitations in Current Studies**

Despite advancements, several challenges remain. Many studies rely on self-reported labels, which may introduce bias. Additionally, generalizability across different populations is often limited due to dataset constraints. The ethical implications of automated depression prediction also raise concerns regarding data privacy and potential misdiagnosis.

Given these existing gaps, our study builds upon prior work by utilizing the XGBoost algorithm to enhance classification accuracy while addressing common limitations in feature selection and data preprocessing.

III. METHODOLOGY

1) Dataset

The dataset used in this study consists of demographic and behavioral information from participants, with a binary label indicating whether an individual is experiencing depression or not. The dataset was obtained from a broad survey designed to explore the various factors influencing depression risk in adults. Conducted anonymously from January to June 2023, the survey spanned multiple cities, including individuals from diverse professional and demographic backgrounds. Participants, aged between 18 and 60, willingly shared information on factors like age, gender, location, education, job satisfaction, study/work hours, and family medical history, among other aspects. No professional mental health evaluations or diagnostic tests were involved and data preprocessed to ensure data quality and reliability. The data was split into training and testing sets, with 80% used for training and 20% reserved for final evaluation.

2) Data Preprocessing

To enhance model performance, several preprocessing steps were applied:

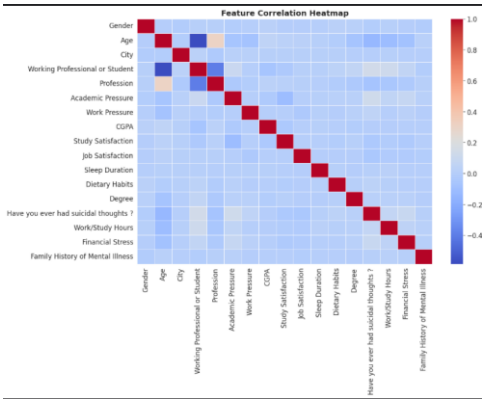

Figure 1: Correlation heatmap of the dataset, showing relationships between features. Darker shades indicate higher correlation.

2.1) Redundancy Removal

Highly correlated features (|correlation| > 0.8) may contain redundant information, which can introduce multicollinearity issues.

"Academic Pressure" and "Work Pressure" have a correlation above 0.8, we might keep only one to avoid redundancy.

2.2) Feature Selection

Non-informative attributes such as 'ID' and 'Name' were removed to avoid unnecessary complexity. Only relevant features contributing to depression prediction were retained.

Features with a moderate correlation (around 0.3 to 0.6) with the target variable (Depression) are useful predictors.

"Financial Stress" or "Job Satisfaction" shows a positive correlation with depression, these are important variables to retain in the model.

Features with very low correlation (near 0.0) with Depression might not contribute much to prediction. "City" has a near-zero correlation, it can be removed from the dataset to reduce noise.

2.3). Categorical Variable Encoding

Categorical variables were transformed into numerical representations using a **custom encoding function** that assigns each unique category a specific numerical value. This approach was chosen because:

Avoiding High-Dimensionality - One-hot encoding can significantly increase the number of features, making the model inefficient, especially with high-cardinality categorical data.

Handling Unseen Categories - Standard label encoding assigns fixed numeric values, but if new categories appear in the test data, they may cause errors. Our custom encoding function assigns unseen categories to a default value, ensuring model stability.

Preserving Ordinal Information - Unlike one-hot encoding, which treats categories as independent, custom encoding maintains an ordinal structure when needed, helping the model capture patterns more effectively.

2.4) Missing Value Imputation

Missing values in numerical columns were imputed using the **median value** to maintain data integrity. The median was chosen over other imputation methods for the following reasons:
Resistant to Outliers - Unlike the mean, which can be skewed by extreme values, the median provides a more **robust measure of central tendency** in datasets with outliers.

Better for Skewed Distributions - Many real-world datasets, including those related to mental health factors (e.g., income, work hours), do not follow a normal distribution. The median ensures **a more accurate central value** in such cases.

More Stable Predictions - Using the mode (most frequent value) can introduce bias if a single category dominates, while mean imputation may distort relationships in datasets with **skewed or non-uniform distributions**.

2.5) Data Type Conversion

All features were converted to float type for compatibility with the XGBoost algorithm.

3) Model Architecture

We used **XGBoost (eXtreme Gradient Boosting)**, a powerful machine learning algorithm known for its efficiency and accuracy. XGBoost was chosen because it:

- Handles missing values automatically.
- Works well with structured data.
- Reduces overfitting through built-in regularization.

The model was configured with the following parameters:

- **Objective function:** Chosen because our task is a **binary classification problem (depressed vs. non-depressed)**.
- **Evaluation metric:** elected as it provides a better measure of uncertainty in predictions compared to accuracy alone.
- **Maximum tree depth:** A lower depth prevents overfitting while still allowing the model to learn meaningful patterns. Tuning results showed that increasing beyond **depth 5** led to **overfitting** without significant accuracy improvement

- .**Learning rate -** A small learning rate ensures a stable convergence, reducing the risk of the model learning too quickly and overfitting. Fine-tuning showed that a higher value (e.g., 0.1) led to **faster convergence but worse generalization**.

- **Subsample ratio -** Helps prevent overfitting by randomly selecting 90% of the training data for each boosting round. Experiments showed that reducing this below **80%** decreased accuracy, while increasing it led to redundancy.

- **Column sample ratio -** Helps the model generalize by randomly selecting 90% of features for each tree. Tuning results showed that values **below 70% reduced performance** due to missing key features, while higher values led to overfitting.

- **Number of estimators -** Chosen based on **early stopping results**, ensuring that the model learns effectively without excessive boosting rounds.

- **Early stopping rounds -** Helps prevent overfitting by stopping training if performance does not improve over 50 rounds. This value was chosen based on **cross-validation results**, where shorter stopping rounds led to **premature termination**, and longer rounds did not improve accuracy..

4) Training and Validation

The training data was further split into training (80%) and validation (20%) sets using a random state of 42 for reproducibility. The model was trained on the training subset with the validation subset used for performance monitoring and early stopping to prevent overfitting.

IV. Experimental and Results

1) Model Evaluation

The performance of the XGBoost model was assessed using standard classification metrics, including **accuracy, precision, recall, and F1-score**. These metrics help evaluate how well the model distinguishes between individuals with and without depression.

To ensure reliable evaluation, the model was validated on a separate 20% validation set, allowing us to monitor performance during training. The early stopping mechanism helped prevent overfitting, ensuring the model remained generalizable to new data.

2) Model Performance

The model demonstrated strong predictive performance, effectively classifying depression cases. The evaluation metrics on the validation set were as follows:

Validation Metrics:

	precision	recall	f1-score	support
0.0	0.96	0.97	0.96	22986
1.0	0.85	0.82	0.83	5154
accuracy			0.94	28140
macro avg	0.90	0.89	0.90	28140
weighted avg	0.94	0.94	0.94	28140

Validation Accuracy: 0.9397

These results indicate that the model successfully identifies individuals at risk of depression while minimizing false predictions.

3) Feature Importance Analysis

To understand the factors influencing depression predictions, we analyzed the feature importance scores generated by XGBoost. The top 10 most important features were visualized using a bar chart, highlighting the most significant predictors. These key factors provide insights into patterns associated with depression risk, potentially aiding in early detection and intervention strategies.

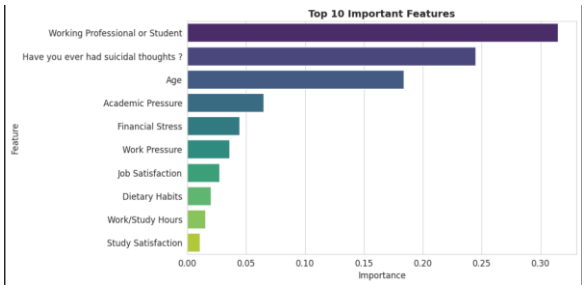

Figure 2: Feature Importance Analysis – Identifying the most influential factors contributing to depression prediction using the XGBoost model.

4) Limitations and Challenges

Despite the model's promising results, some limitations must be acknowledged:

Data Quality and Representation: The predictive power depends on the quality and diversity of the dataset. If the training data lacks representation from different demographic groups, the model may not generalize well to new populations.

Psychological Complexity: Depression is influenced by complex psychological and environmental factors that may not be fully captured by structured data. A clinical diagnosis involves more nuanced assessments beyond numerical features.

Potential Bias: Cultural and demographic biases present in the dataset could affect predictions. Future improvements may involve using more balanced and diverse datasets to enhance fairness and accuracy.

5) Future Improvements

To further improve the model, several enhancements could be explored:

Feature Engineering - Introducing new features derived from behavioral trends or external datasets may enhance prediction accuracy.

Explainability Methods - Using techniques like SHAP (SHapley Additive exPlanations) could provide deeper insights into how each feature influences predictions.

Data Augmentation - Incorporating additional data sources, such as self-reported symptoms or wearable device data, may improve model robustness.

V. Discussion and Conclusion

This study developed a machine learning model using XGBoost to predict depression based on demographic and behavioral data. The model underwent thorough data preprocessing, feature selection, categorical encoding, missing value imputation, and hyperparameter tuning to ensure strong predictive performance. The evaluation results demonstrated that the model effectively distinguishes between individuals with and without depression, as measured by accuracy, precision, recall, and F1-score.

A key finding was that certain features had a more significant impact on depression prediction, which was revealed through feature importance analysis. These insights could be valuable for mental health professionals and researchers in identifying early risk factors and enhancing intervention strategies.

However, despite its promising performance, the model has several limitations. The predictions are influenced by the quality and representativeness of the dataset.

Additionally, psychological factors that require clinical evaluation cannot be fully captured through machine learning alone. Furthermore, cultural and demographic biases in the dataset may affect the model's generalizability.

To enhance the model's reliability and impact, future research should focus on:

- Expanding the dataset to include more diverse populations.
- Improving feature engineering to capture deeper behavioral and psychological patterns.
- Applying explainable AI techniques to make predictions more interpretable and trustworthy.

In conclusion, this study demonstrates the potential of AI in mental health screening. While machine learning models cannot replace clinical diagnosis, they can serve as valuable support tools for early detection, helping individuals and healthcare professionals take proactive steps in managing mental well-being.

.

VI. REFERECES

[1] World Health Organization. (2020). Depression. Retrieved from https://www.who.int

[2] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of Depression-related Posts in Social Media Using SVM and Deep Learning Models. Frontiers in Psychiatry, 10, 759.

[3] Orabi, A. H., Buddhitha, P., Horan, K. A., & Inkpen, D. (2018). Deep Learning for Depression Detection of Twitter Users. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology (pp. 88-97).

[4] Li, J., Yin, Y., Quan, C., Zhang, H., & Wu, Z. (2020). Depression Detection Using Machine Learning: A Comprehensive Review. Neurocomputing, 423, 746-765.

[5] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[6] Rashid, T., & Louis, J. (2019). Machine Learning for Predictive Analytics in Depression and Mental Health. IEEE Transactions on Computational Social Systems, 6(5), 981-993.