A Comprehensive Study: Mental Health Status Prediction Using Machine Learning

[1]Kailash Pati Mandal, [2]Rana Chakraborty, [3]Sambit S Mondal, [4]Sumanta Karmakar, [5]Anirban Kundu

[1]biltu.cse@gmail.com, [2]rana.jui@gmail.com, [3]sambitsmonda@gmail.com, [4]sumanta.karmakar@gmail.com

[1,2,3,4] Assistant Professor, Asansol Engineering College

[5] Student, Asansol Engineering College

Abstract

Mental health is a growing concern in modern society, especially in high-pressure work environments. With increasing awareness and the availability of large-scale mental health survey data, predictive modeling using machine learning (ML) techniques has emerged as a powerful tool for early diagnosis and support. This research presents a comprehensive pipeline that includes data preprocessing, feature engineering, model building, evaluation, and visualization to predict whether an individual will seek treatment for mental health issues. Multiple models including logistic regression, decision trees, random forests, k-nearest neighbors (KNN), bagging, boosting, stacking, and neural networks are implemented and evaluated. The stacking classifier showed the best performance with an accuracy of 81.7%. The results highlight the importance of factors like work interference, anonymity, and benefits in influencing mental health treatment-seeking behavior.

Keywords: Mental Health, Machine learning, Artificial Intelligence, psychometric

1. Introduction

The mental health crisis is escalating at an alarming rate worldwide, profoundly impacting individuals' personal well-being and professional productivity. According to the World Health Organization (WHO), nearly one in eight people globally live with a mental health disorder, with depression and anxiety being the most prevalent. The workplace, particularly high-pressure industries like the technology sector, has become a significant environment where mental health challenges manifest due to factors such as long working hours, job insecurity, and high expectations. Despite the increasing prevalence of these conditions, many cases remain undiagnosed and untreated due to persistent stigma, lack of awareness, and limited access to mental health resources.

Organizations and policymakers are actively seeking innovative strategies to identify at-risk individuals early and implement timely interventions. Traditional methods, such as self-reported surveys and clinical assessments, are often reactive rather than proactive. However, advancements in artificial intelligence (AI) and machine learning (ML) present an opportunity to revolutionize mental health screening by analyzing behavioral and survey data to detect early warning signs. This study leverages a publicly available mental health survey dataset from tech employees to develop predictive models that assess the likelihood of an individual seeking treatment for mental health issues.

Mental health disorders, including depression, generalized anxiety disorder (GAD), and stress-related conditions such as burnout, have far-reaching consequences. In professional settings, untreated mental health issues lead to decreased productivity, higher absenteeism, and increased turnover rates. The economic burden is substantial, with the WHO estimating that depression and anxiety cost the global economy approximately $1 trillion per year in lost productivity.

Despite these staggering figures, many individuals hesitate to seek help due to fear of discrimination, workplace repercussions, or cultural stigma surrounding mental illness. In high-stress industries like technology, employees often face unique challenges, including isolation (especially in remote work settings), imposter syndrome, and an "always-on"

work culture that exacerbates stress. These factors contribute to underreporting, making it difficult for organizations to implement effective support systems.

Machine learning offers a data-driven approach to identifying individuals at risk of mental health disorders by analyzing patterns in survey responses, behavioral data, and even digital footprints (e.g., social media activity or workplace communication patterns). Unlike traditional diagnostic methods, which rely on self-reporting or clinical evaluations, ML models can process large datasets to detect subtle correlations that may indicate early signs of mental distress.

2. Related Works

The growing interest in mental health prediction using machine learning (ML) has spurred a wide variety of research efforts across different data modalities and algorithms. One prominent line of work has investigated the analysis of textual data, particularly social media content, as a means to detect signs of mental distress, including depression, anxiety, and loneliness [1]. These approaches leverage the language patterns, sentiment, and linguistic features present in user-generated texts to train classifiers that predict mental health states. For example, models have been trained to analyze Reddit and Twitter posts to identify depression-related content with promising accuracy, often utilizing natural language processing (NLP) techniques such as word embeddings and sentiment analysis [2].

In contrast, another stream of studies has utilized structured survey responses and psychometric questionnaires to identify individuals at risk of mental health disorders [3]. These datasets often contain well-labeled, clinically validated features such as demographic information, behavioral indicators, and symptom checklists. Because of their well-structured nature, such data allow researchers to apply traditional ML techniques with greater ease and interpretability.

Among the most commonly used algorithms in these studies are logistic regression and decision trees, primarily due to their simplicity and ease of interpretation [4]. Logistic regression provides a probabilistic framework for binary classification problems and is particularly valuable when feature coefficients are used to interpret the influence of different variables on mental health outcomes. Decision trees, on the other hand, offer a visual representation of decision paths and enable clear understanding of how predictions are made based on feature splits.

More recent studies have shifted toward ensemble learning techniques, such as random forests and gradient boosting machines (GBMs), to enhance prediction accuracy and model robustness [5]. These methods combine multiple weak learners to create a stronger predictive model. Random forests mitigate overfitting by aggregating predictions from several uncorrelated decision trees, while GBMs iteratively minimize error using gradient descent principles. Both techniques have demonstrated superior performance in mental health prediction tasks compared to single classifiers.

Despite these advancements, a key limitation in the existing literature is the lack of comprehensive studies that systematically compare a broad range of classifiers on a single, standardized dataset using uniform preprocessing steps [6]. Variability in data sources, preprocessing methods, and evaluation metrics often makes it difficult to draw generalizable conclusions from cross-study comparisons. This methodological inconsistency restricts the ability to benchmark models effectively and to understand their relative strengths in specific contexts.

The present study aims to address this gap by implementing and evaluating a diverse set of machine learning algorithms—including logistic regression, decision trees, K-nearest neighbors, random forests, bagging, boosting, and stacking—on the same cleaned and encoded dataset. This uniform approach allows for a more reliable assessment of model performance and interpretability in the context of mental health prediction.

3. Proposed Methodology

3.1 Data Acquisition and Exploration

The dataset consists of 1,259 responses to a mental health survey in the tech industry. It includes 27 features such as age, gender, country, benefits, care options, leave policy, and treatment.

3.2 Data Cleaning In the data cleaning phase, columns with excessive missing values such as comments, state, and timestamp were removed to improve data quality. Missing entries in the 'self_employed' and 'work_interfere' columns were imputed with default categories to maintain consistency. The 'Gender' column was standardized into three main

groups: male, female, and trans. Additionally, the 'Age' variable was filtered to remove outliers and binned into defined ranges 0–20, 21–30, 31–65, and 66–100—for better categorical analysis.

3.3 Feature Encoding and Scaling

Categorical variables were encoded using LabelEncoder, and numerical features like 'Age' were normalized using MinMaxScaler.

3.4 Feature Selection

ExtraTreesClassifier was used to determine the most influential features: work_interfere, anonymity, leave, care_options, and benefits.

3.5 Different Models Applied During the Training and Evaluation Process

In this study, a comprehensive model training and evaluation process was conducted using a diverse set of machine learning classifiers to predict mental health conditions. The selection of models ranged from simple linear classifiers to complex ensemble and deep learning techniques, ensuring a robust comparative analysis. Initially, Logistic Regression was employed as a baseline model due to its simplicity and interpretability. It provided valuable insights into feature importance and directional relationships between input variables and the target outcome.

Next, the K-Nearest Neighbors (KNN) algorithm was implemented to evaluate instance-based learning performance. As a non-parametric method, KNN relies heavily on distance metrics, making it sensitive to the scale and distribution of data, which had been carefully preprocessed in earlier stages. The Decision Tree Classifier was also trained to provide an interpretable, rule-based model capable of handling both numerical and categorical features without requiring normalization or scaling.

To enhance predictive accuracy and mitigate overfitting, the Random Forest Classifier, an ensemble method that aggregates predictions from multiple decision trees, was introduced. It outperformed individual tree models by leveraging bagging and feature randomness. Similarly, a Bagging Classifier was employed, which further reinforced the power of bootstrap aggregation in reducing variance. In contrast, AdaBoost, a boosting technique, was used to iteratively focus on misclassified samples, allowing the model to correct its errors and improve generalization.

A Stacking Classifier was also constructed, combining multiple base learners—KNN, Random Forest, Naive Bayes, and Logistic Regression—into a single meta-model. This hybrid approach aimed to capitalize on the individual strengths of each classifier, thereby boosting overall performance through ensemble synergy. Finally, a Deep Neural Network (DNN) built with TensorFlow was trained to assess the capacity of deep learning in handling the mental health prediction task. The DNN architecture included multiple hidden layers and ReLU activations, optimized via backpropagation.

Each model's performance was rigorously evaluated using standard metrics: accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC curve. These evaluation criteria enabled a well-rounded assessment of each algorithm's effectiveness in identifying mental health risks across various classes, ensuring that the final model selection balanced performance with interpretability.

4. Results and Discussion

i) Logistic Regression

Logistic Regression served as the baseline model in this study. It achieved an accuracy of 79.6%, an AUC score of 0.796, and a strong cross-validation AUC of 0.875. Being a linear classifier, logistic regression is straightforward to interpret and effective for binary classification tasks. Its performance reflects that linear decision boundaries can identify some underlying patterns in the data, though it may not capture complex, non-linear relationships effectively. The confusion matrix for the Logistic Regression is shown in Figure 1.

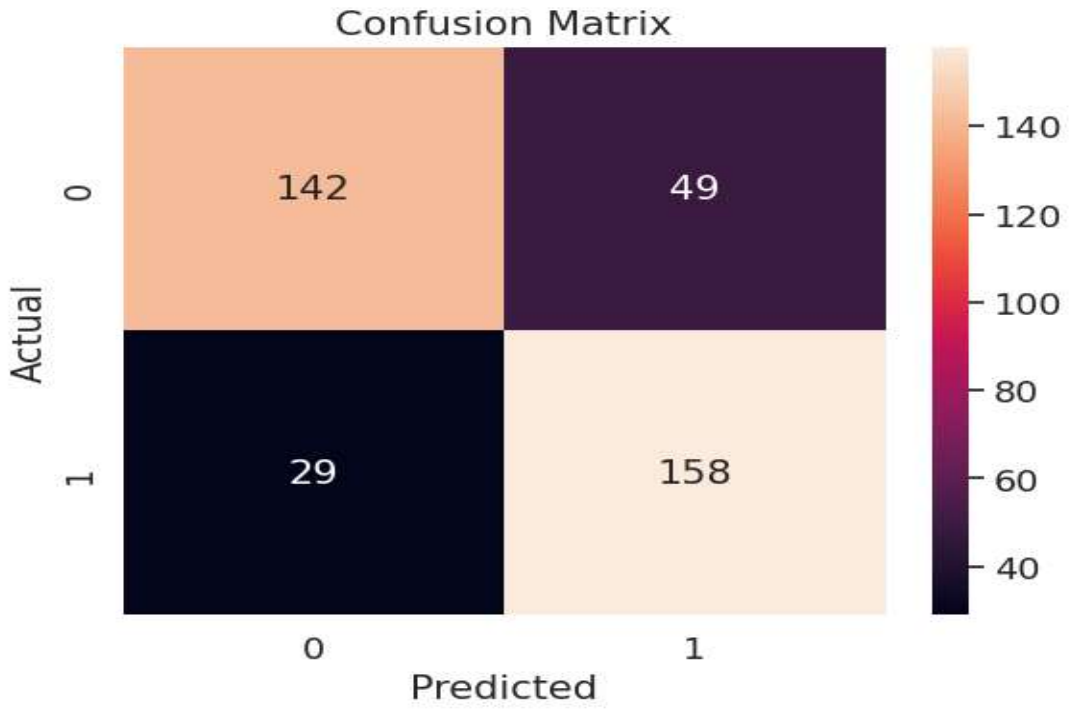

Figure 1. Confusion matrix for Logistic Regression

ii) K-Nearest Neighbors (KNN, k=27)

The K-Nearest Neighbors algorithm slightly outperformed logistic regression, with an accuracy of 80.4%, an AUC score of 0.805, and a CV AUC of 0.878. KNN classifies data points based on the majority class among their nearest neighbors, making it a distance-based, non-parametric method. The improved performance suggests that local pattern recognition is valuable in this dataset, and the high CV AUC reflects consistent generalization across different data splits. The confusion matrix for the KNN is shown in Figure 2.

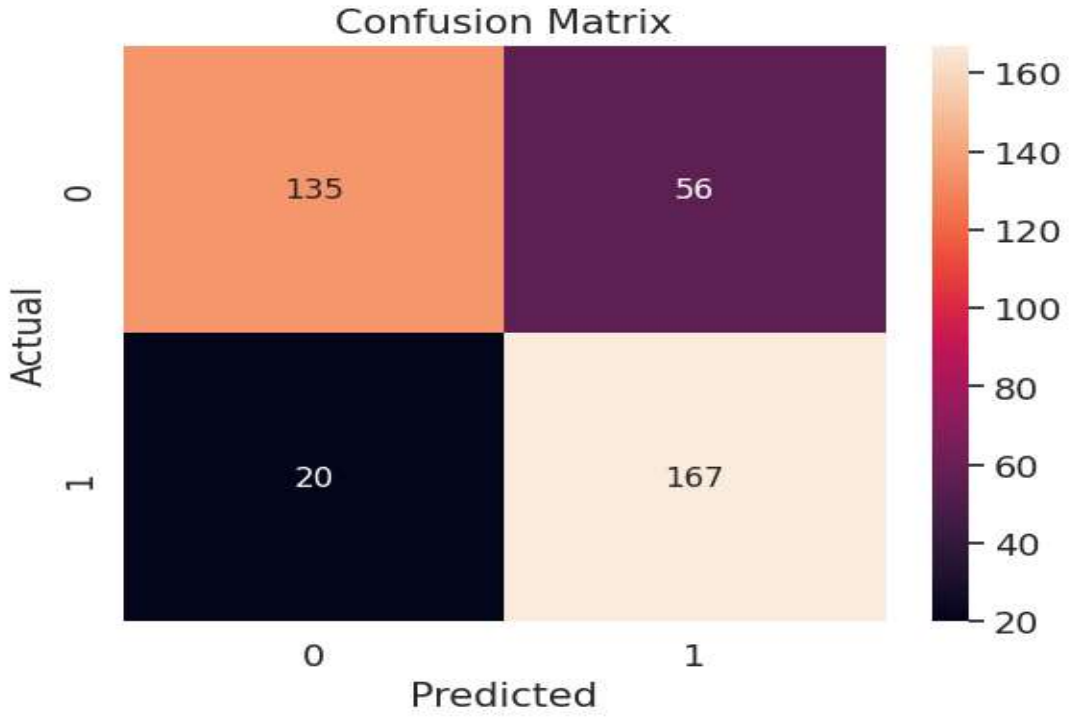

Figure 2. Confusion matrix for K-Nearest Neighbors

iii) Decision Tree

The Decision Tree classifier provided a further improvement, achieving 80.6% accuracy, a 0.808 AUC score, and a CV AUC of 0.882. Decision trees are interpretable and capable of handling mixed data types without preprocessing, making them suitable for real-world applications. However, they are prone to overfitting on training data. Despite this, the strong performance indicates that the model effectively captured relevant patterns in the data. The confusion matrix for the Decision Tree is shown in Figure 3.

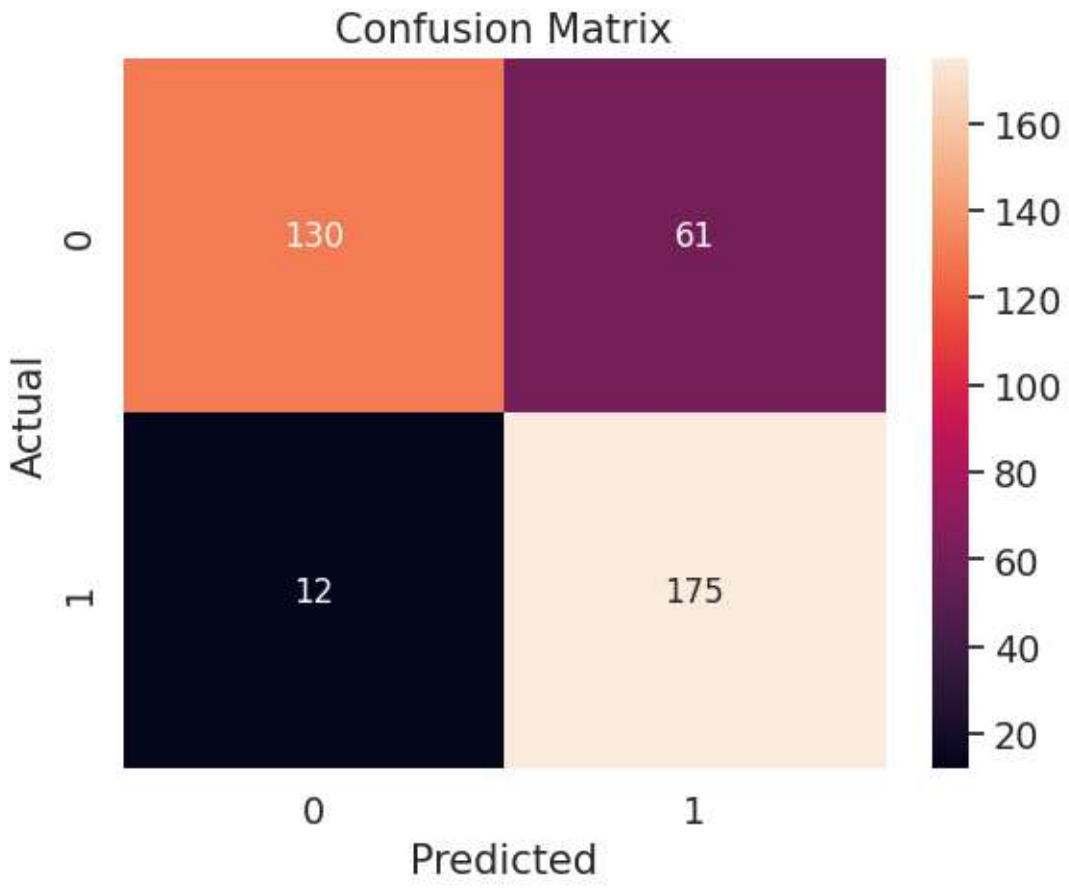

Figure 3. Confusion matrix for Decision Tree

iv) Random Forest

Random Forest performed even better, achieving an accuracy of 81.2%, an AUC score of 0.813, and the highest CV AUC of 0.893 among all models. As an ensemble method built on multiple decision trees, Random Forest reduces overfitting and enhances generalization. The significant gain in CV AUC suggests that this model is highly robust and reliable when applied to different data subsets. The confusion matrix for the Random Forest is shown in Figure 4.

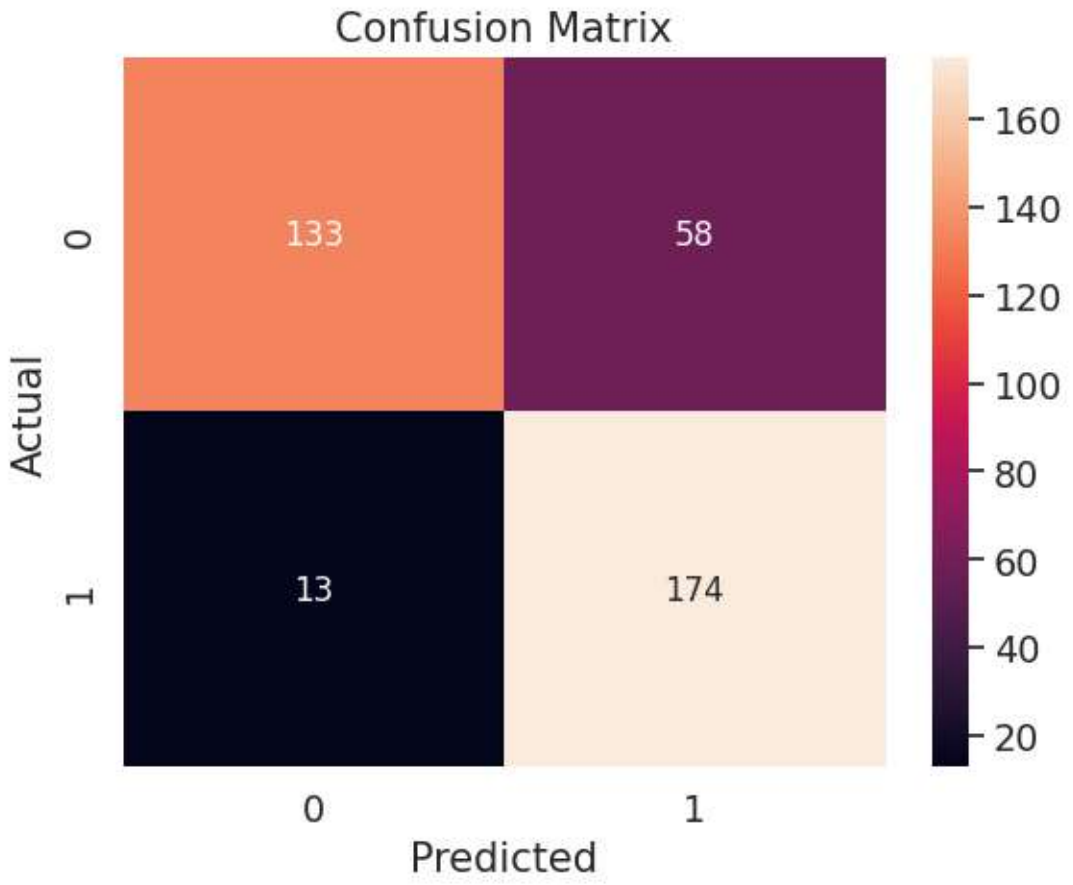

Figure 4. Confusion matrix for Random Forest

v) Bagging Classifier

The Bagging classifier yielded slightly lower results, with 79.1% accuracy, a 0.791 AUC score, and a CV AUC of 0.844. Although bagging is also an ensemble technique like Random Forest, it may not be as effective in capturing complex interactions unless combined with more powerful base learners. Its moderate performance implies reduced variance but limited enhancement in capturing intricate patterns. The confusion matrix for the Random Forest is shown in Figure 5.

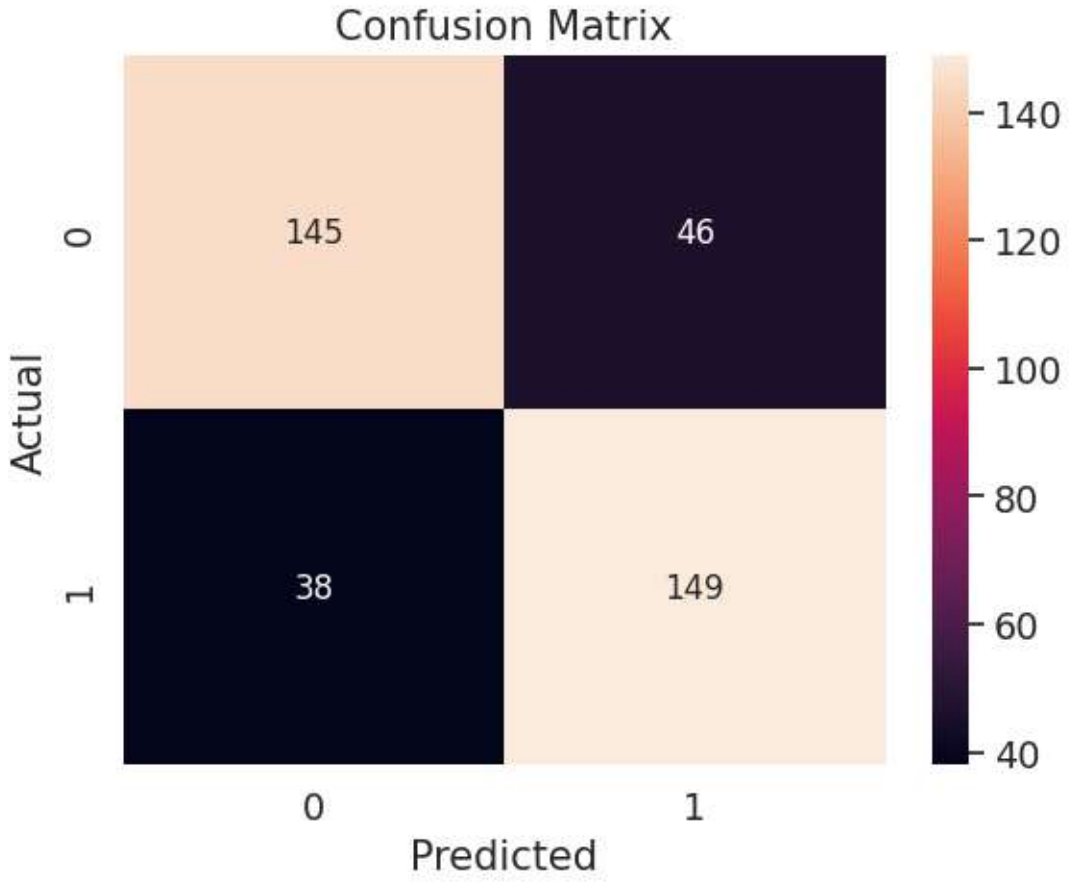

Figure 5. Confusion matrix for Bagging Classifier

vi) Boosting (AdaBoost)

AdaBoost emerged as one of the best-performing models, with an accuracy of 81.7%, an AUC score of 0.818, and a CV AUC of 0.874. Boosting focuses on correcting errors made by previous learners by placing more weight on misclassified samples. This iterative refinement allows AdaBoost to improve its predictive power, making it well-suited for complex classification tasks like mental health prediction. The confusion matrix for the Random Forest is shown in Figure 6.

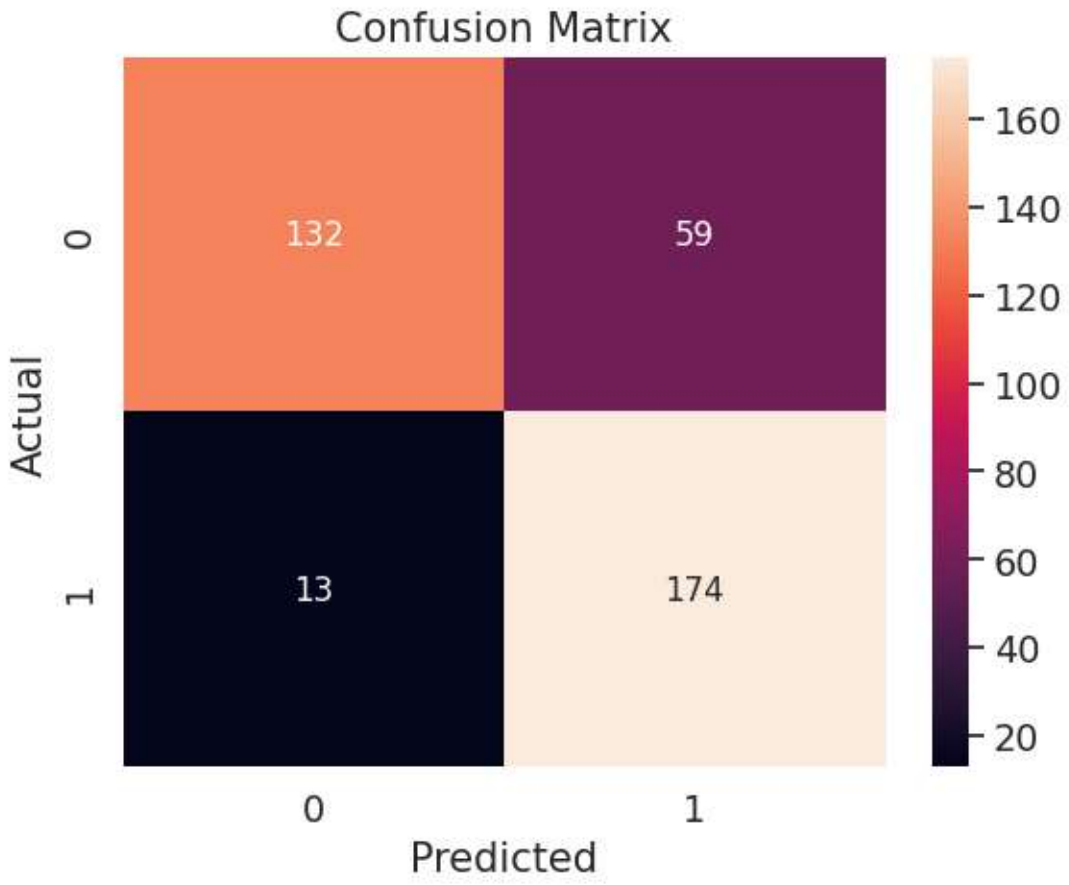

Figure 6. Confusion matrix for Boosting (AdaBoost)

vii) Stacking Classifier

The Stacking classifier also achieved an accuracy of 81.7%, along with an AUC score of 0.817 and a CV AUC of 0.840. While slightly trailing AdaBoost in AUC and CV AUC, stacking benefits from combining multiple base models—KNN, Random Forest, Naive Bayes, and Logistic Regression—under a meta-classifier. This hybrid approach enables the model to leverage the strengths of its components, enhancing its robustness and adaptability. The confusion matrix for the Random Forest is shown in Figure 7.

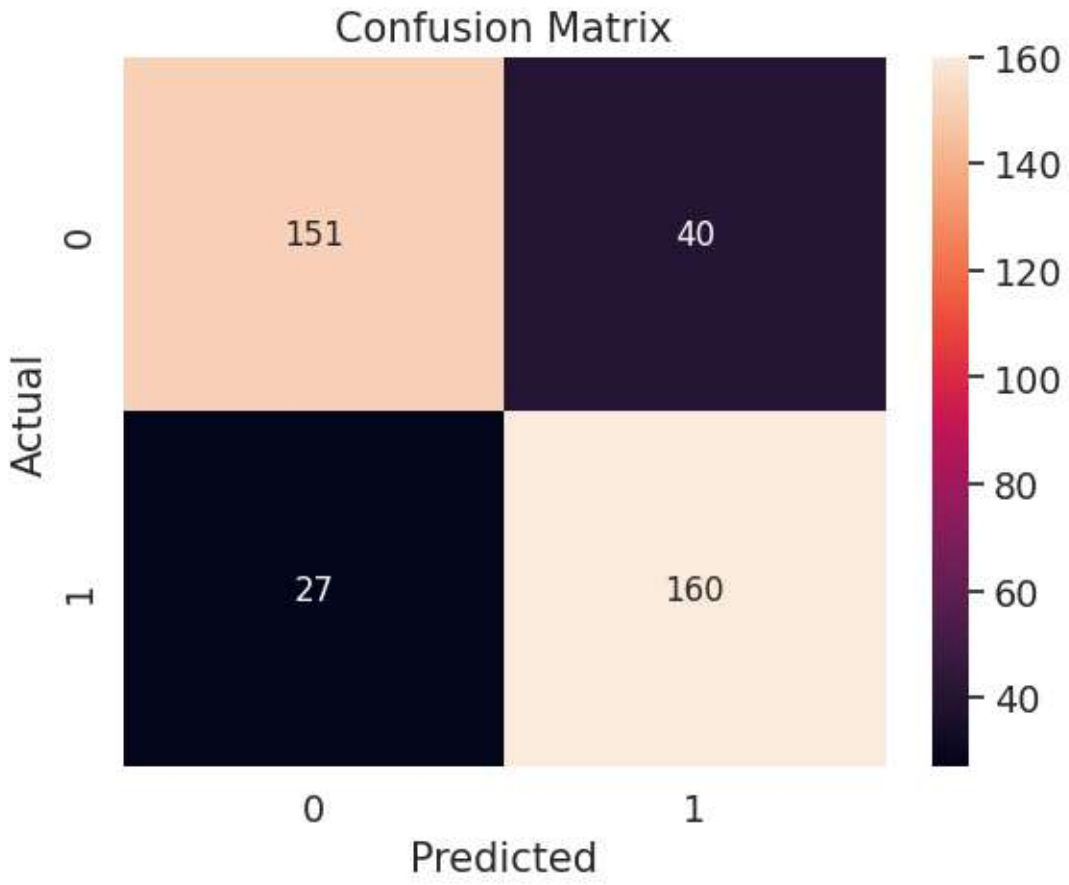

Figure 7. Confusion matrix for Stacking Classifier

The stacking classifier and neural network achieved the highest accuracy of 81.7%. Stacking combined the strengths of multiple classifiers and performed well across metrics. Decision trees and random forests provided strong baselines, while boosting improved generalization. Feature importance analysis confirmed the impact of organizational factors on mental health treatment.

The table in your document summarizes the performance of various machine learning models used for mental health prediction. It compares the models using three key metrics: Accuracy, AUC Score, and Cross-Validation AUC (CV AUC). Comparative analysis has been given in Table 1.

Table 1 Comparative among analysis of different models

Model	**Accuracy**	**AUC Score**	**CV AUC**
Logistic Regression	79.6%	0.796	0.875
KNN (k=27)	80.4%	0.805	0.878
Decision Tree	80.6%	0.808	0.882
Random Forest	81.2%	0.813	0.893
Bagging	79.1%	0.791	0.844
Boosting (AdaBoost)	81.7%	0.818	0.874
Stacking	81.7%	0.817	0.840

5. Future Work

Future work can focus on integrating advanced NLP techniques to analyze textual comments, enabling deeper insights into employee mental health. Developing real-time prediction tools through API deployment will support seamless integration with HR systems. Incorporating longitudinal analysis using time-series data can help track mental health trends over time. Additionally, applying explainability methods like SHAP or LIME will enhance model transparency, fostering greater trust and adoption in real-world applications.

6. Conclusion

This study successfully demonstrated the application of multiple machine learning classifiers to predict mental health treatment-seeking behavior using survey data. The comprehensive pipeline—from data cleaning and feature engineering to model evaluation provides a replicable framework for similar studies. Ensemble models, particularly stacking and random forests, yielded the best results, highlighting the importance of organizational support features. The findings can inform workplace policies aimed at improving mental well-being and early intervention strategies.

References

[1] Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). Quantifying and predicting mental illness severity in online pro-eating disorder communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1171–1184.

[2] Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, 51–60.

[3] Nguyen, T., Tran, T., Luo, W., & Venkatesh, S. (2014). Affective topic modeling for depression detection. *Proceedings of the 5th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 593–600.

[4] Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, 49(9), 1426–1448.

[5] Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. *Health Information Science and Systems*, 6(1), 1–12.

[6] Zhang, Y., Zhang, L., & Zhang, C. (2020). Comparative study of different machine learning techniques for mental health prediction. *BMC Medical Informatics and Decision Making*, 20(Suppl 10), 1–12.