ADScAI
Applied Data Science & Artificial Intelligence Symposium

Comparative Analysis of Machine Learning Models for Depression Risk Prediction

Lakshani Dehiwala Liyanage
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
subhagya.22@cse.mrt.ac.lk

Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
rtuthaya@cse.mrt.ac.lk

***Keywords*—Depression Prediction, machine learning, classification model, encoding techniques**

I. Introduction

Depression, affecting 3.8% of the global population and contributing to over 700,000 suicides annually [1], underscores the importance of early detection for effective intervention. This study compares the performance of several machine learning algorithms—Random Forest(RF), XGBClassifier(XGB), LightGBM, and CatBoost—in predicting depression risk based on lifestyle and demographic factors [1]. The models are evaluated using accuracy, precision, recall, F1-score, and advanced techniques such as ROC curves, and calibration plots. The findings aim to provide insights into the potential of ML for early depression risk detection and enhancing intervention strategies.

II. Literature Review

Several studies have explored machine learning for mental health prediction, often relying on clinical assessments and self-reported questionnaires, which may introduce biases. Machine learning models, including XGBoost, GBDT, RF, CatBoost, and LightGBM, have shown effectiveness in predicting depression risk [1]. Additionally, EEG data has been used for depression detection, comparing different preprocessing and classification approaches [2]. However, many studies rely on specialized datasets, such as EEG recordings or social media text, which are not always feasible in real-world healthcare settings. Furthermore, prior research often focuses on a single modeling approach rather than a comparative evaluation of multiple classifiers on structured survey data [3]. Limited attention has also been given to the impact of encoding methods on model performance. Addressing these gaps, this study systematically compares various machine learning models and encoding techniques for depression risk prediction using a publicly available dataset.

III. Materials and Methods

A. Dataset

An anonymized dataset collected between January and June 2023 includes 141,000 adults aged 18–60 from diverse gender and socioeconomic backgrounds [8]. The self-reported dataset includes lifestyle factors like work/study hours and is imbalanced, with 18.1% are classified as at risk for depression.

B. Data Preprocessing

Missing values were imputed based on role: occupation-related variables (e.g., Work Pressure) were median-imputed for professionals and marked as -1 for students, while education-related features (e.g., CGPA) followed the inverse approach. Generic categorical variables were imputed with 'Unknown', and numerical variables used median imputation. High-cardinality categorical features were label-encoded, with CatBoost evaluated using both label encoding and its native categorical handling [5]. Tree-based models used unnormalized numerical features due to their scale invariance.

C. Model Selection and Tuning

The models—RF, XGBoost (XGB), LightGBM, and CatBoost—were tuned using Optuna for hyperparameters. Class weights were applied to the minority class to address class imbalance.

D. Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-score, ROC curves, and calibration plots [6]. For CatBoost, these metrics were compared for both encoding methods to assess their impact on performance.

IV. Results and Discussion

A. Model Performance Comparison

The performance of the CatBoost, XGB, LightGBM, and RF models was evaluated using accuracy, precision, recall, F1-scores and ROC-curves. For precision, recall, and F1-scores, weighted averages were used to address class imbalance. The accuracy scores for each model are summarized in Table 1.

https://doi.org/10.31705/ADScAI.2025.13

TABLE I
MODEL PERFORMANCE COMPARISON WITH ENCODING METHODS

Metrics	Label Encoding				Native Encoding
	RF	XGB	LGBM	CB	CB
Accuracy	0.9374	0.9380	0.9378	0.9385	0.9404
Precision	0.94	0.94	0.94	0.94	0.94
Recall	0.94	0.94	0.94	0.94	0.94
F1-score	0.94	0.94	0.94	0.94	0.94

CatBoost outperformed the other models, achieving the highest accuracy of 94.04%, largely due to its native handling of categorical features. XGB and LightGBM also performed well with accuracies of 93.8% and 93.78%, respectively. RF, while competitive, lagged behind the boosting models in accuracy, likely due to its inability to capture complex data interactions as effectively.

B. ROC and Calibration Plots

Fig. 1 highlights CatBoost's performance (AUC = 0.9766), surpassing other models (avg. AUC = 0.974). It showed slight advantages in the 0.1–0.2 FPR range, with a marginal increase in TPR. CatBoost's improved specificity and calibration, shown in calibration plots, make it ideal for clinical use by reducing false positives and ensuring reliable risk estimates.

C. Feature Important Analysis

As presented in Fig. 2, SHAP analysis revealed that work pressure, academic pressure, job/study satisfaction, and age were the most influential factors in predicting depression risk. Work and academic stress exhibited a strong positive association with depression risk, whereas job/study satisfaction had a protective effect. These findings are consistent with existing mental health research [7].

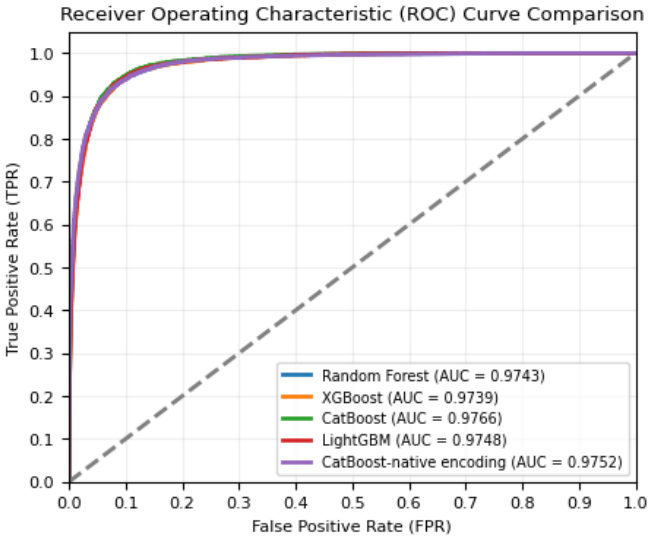

Fig. 1. ROC curves of the evaluated models, with AUC values indicating classification performance.

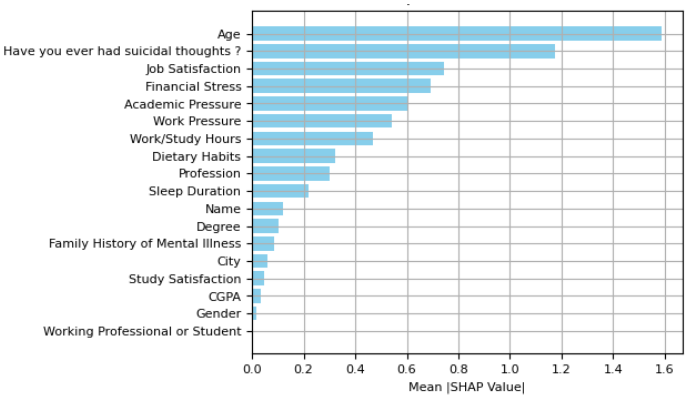

Fig. 2. Feature importance scores for the CatBoost model using native categorical encoding. The top predictors of depression are ranked by their mean SHAP values.

V. CONCLUSION

This study evaluated tree-based models for depression risk prediction, with CatBoost (using native encoding) outperforming others due to its effective handling of categorical features, robustness against overfitting, and efficient processing of complex datasets [5]. Its well-calibrated probabilities improve reliability in predicting depression risk. Feature importance analysis identified key factors influencing depression, providing insights for targeted interventions. Future research could integrate additional data and clinical assessments to enhance predictive accuracy and applicability in mental health evaluation.

REFERENCES

[1] C. Yuan, X. Liu, Q. Xu, Y. Li, Y. Luo, and X. Zhou, "Depression diagnosis and analysis via multimodal multi-order factor fusion," in Proc. Int. Conf. Artif. Neural Netw. (ICANN), Cham, Switzerland, 2024, vol. 15023. Springer.

[2] M. S. Sharif, A. Zorto, A. T. Kareem, and R. Hafidh, "Effective machine learning-based techniques for predicting depression," in Proc. 2022 Int. Conf. Innov. Intell. Informat., Comput., Technol. (3ICT), Sakheer, Bahrain, 2022, pp. 366-371

[3] M. Čukić, V. López, and J. Pavón, "Classification of depression through resting-state electroencephalogram as a novel practice in psychiatry: Review," J. Med. Internet Res., vol. 22, no. 11, p. e19548, Nov. 2020.

[4] M. S. Zulfiker, N. Kabir, A. A. Biswas, T. Nazneen, and M. S. Uddin, "An in-depth analysis of machine learning approaches to predict depression," Current Research in Behavioral Sciences, vol. 2, Nov. 2021.

[5] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 31, 2018.

[6] R. Li, X. Wang, L. Luo, and Y. Yuan, "Identifying the most crucial factors associated with depression based on interpretable machine learning: a case study from CHARLS," Frontiers in Psychology, vol. 15, 2024.

[7] S. S. Kim, M. Gil, and E. J. Min, "Machine learning models for predicting depression in Korean young employees," Front. Public Health, vol. 11, p. 1201054, 2023.

[8] Kaggle, "Predicting Depression Machine Learning Challenge." [Online]. Available: https://www.kaggle.com/competitions/predicting-depression-machine-learning-challenge