ADScAI
Applied Data Science & Artificial Intelligence Symposium

The Impact of Data Cleaning and Model Selection for Depression Prediction: A Comparative Study

Sugith Geesan Munasinghe
dept. Computer Science and Engineering
University of Moratuwa
Sri Lanka.
geesan.22@cse.mrt.ac.lk

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

***Keywords*—Data Cleaning, Model Selection, Classification, Machine Learning**

I. Introduction

Predicting health is a field where machine learning (ML) plays a crucial role [1], offering data-driven insights for early detection and diagnosis. With the increasing availability of internet-based data, there is potential for ML models to leverage digital footprints—such as social media activity, online surveys, and behavioral patterns—to predict mental health conditions like depression. However, while these advancements create new opportunities for mental health monitoring, the effectiveness of ML-based prediction largely depends on the quality of the underlying training data. [2]

Real-world datasets often contain noise, inconsistencies, and mislabeled instances, making data preprocessing a critical step in predictive modeling. Prior research highlights that data cleaning significantly influences model generalization [3], but an important question remains: How to clean data to improve model performance without removing informative data. Furthermore, the choice of machine learning model plays a key role in handling such variations in data quality.

We evaluate multiple classification models—K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and XGBoost—under different data preprocessing conditions to assess how cleaning affects their performance. By analyzing these models, we aim to provide insights into the relationship between data preprocessing, model robustness, and classification accuracy in mental health prediction.The rest of this paper is structured as follows: Section II discusses related work on data preprocessing and model selection. Section III details our methodology, including data sources, preprocessing steps, and model training. Section IV presents experimental results and analysis. Section V concludes the study and outlines future research directions.

II. Literature Review

Recent studies emphasize the importance of data preprocessing techniques in improving machine learning model accuracy, with methods like normalization and outlier removal significantly enhancing performance [3]. Additionally, large-scale data validation systems have been developed to monitor and improve data quality, which ultimately boosts model reliability [3]. Research also indicates that data cleaning has varying effects on model performance, with XGBoost being more reactive to mislabels, because boosting procedures assign higher weights to on those instances that are predicted incorrectly [4]. This suggests that learning from noise may sometimes occur, highlighting the need to carefully evaluate the impact of cleaning on model accuracy. Some comparisons have shown that XGBoost outperforms Random Forest in classification accuracy [4], reinforcing its reliability in handling complex data patterns. Given these insights, we will include both XGBoost and Random Forest in our experiments to better understand their performance.

III. Materials and Methods

We utilized a dataset of adults aged 18–60, containing attributes such as age, gender, location, education level, job satisfaction, study and work hours, and family medical history, among others. Our goal was to evaluate how data quality affects predictive performance in depression risk assessment. By analyzing different preprocessing techniques and model performances, we aimed to determine the best approach for achieving a more accurate and reliable prediction model.

A. Dataset

The dataset consisted of 20 columns, including the target feature, ‘Depression.’ The target variable had more 0s than 1s in the training set, potentially biasing the model toward predicting non-depressed cases. The initial training set contained 140,700 data points, while the test set had 93,800 data points. Both datasets were collected through an anonymous broad survey, ensuring diverse representation. This data served as the foundation for evaluating model performance in predicting depression risk.

https://doi.org/10.31705/ADScAI.2025.10

B. Missing Values and Encoding

Upon examining the train and test datasets, we observed that some fields contained missing values. The following columns had a considerable number of missing values:

- **Profession**: 61,262
- **Academic Pressure**: 187,836
- **Work Pressure**: 46,696
- **CGPA**: 187,836
- **Study Satisfaction**: 187,836
- **Job Satisfaction**: 46,684

We imputed missing values in profession, work pressure, and job satisfaction with placeholders for students, and in CGPA, academic pressure, and study satisfaction for working professionals, as these are likely due to the irrelevance of the field. Other missing values were handled as follows: Financial Stress was imputed with the category average, and diet habits were filled with the mode of all records, following common practices for handling random missing values that are not highly skewed. Degree was set to 'None' as we assumed that missing values in this field indicate the absence of a degree.

For the baseline preprocessing, we applied one-hot encoding to 'Gender' and 'Working Professional or Student,' while label encoding was used for 'City,' 'Profession,' 'Dietary Habits,' and 'Degree.' Additionally, Min-Max normalization was performed to enhance the performance of algorithms like KNN.

C. Model Selection

To evaluate and compare different machine learning models, we applied four algorithms: K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and XGBoost after the preprocessing steps. To enhance model performance, we conducted hyperparameter tuning using cross-validation scoring. Upon analyzing the cross-validation results, XGBoost demonstrated the highest performance among the four models, making it the most promising candidate for accurate prediction of depression risk.

D. Further Cleaning

To enhance model performance, we applied additional data cleaning steps while keeping XGBoost as the selected model. The goal was to improve the cross-validation (CV) score through iterative refinements. First, we addressed potential mislabels in features like Dietary Habits and City, checking the CV score after each modification. Then, we applied target encoding to categorical variables and reassessed the CV score. Finally, we refined the Profession column, continuously evaluating improvements in the model's performance.

IV. Results and Discussion

The cross-validation (CV) scores for each model, showcasing the initial performance of KNN, Decision Tree, Random Forest, and XGBoost before any data cleaning, are shown in Table I.

Model	CV Score
KNN	0.8943
Decision Tree	0.9305
Random Forest	0.9374
XGBoost	0.9377

TABLE I
INITIAL RESULTS FOR MODEL PERFORMANCE

Table II presents the cross-validation (CV) scores at each step of the data cleaning and model tuning process, showing the incremental improvements in model performance after various data preprocessing and hyperparameter tuning steps.

Step	CV Score
Initial (No Cleaning)	0.93770
Cleaning Dietary Habits	0.93786
Cleaning City Mislabels (Method 1)	0.93783
Cleaning City Mislabels (Method 2)	0.93794
Target Encoding	0.93857
Cleaning Profession	0.93831
Hyperparameter Tuning	0.93952

TABLE II
RESULTS AFTER VARIOUS CLEANING AND TUNING STEPS

In the case of city mislabels, two different approaches were tested: in Method 1, unclear mislabels were marked as "Unknown" to avoid introducing potential errors, whereas in Method 2, only obvious mislabels, such as "Tolkata" and "Golkata" being corrected to "Kolkata," were addressed while keeping the rest unchanged. The results indicate that targeted corrections (Method 2) led to a slightly better CV score.

V. Conclusion

Data cleaning plays a crucial role in improving model performance by handling inconsistencies and ensuring better data quality. Our study highlights that different cleaning approaches can have varying impacts on the model's predictive power, emphasizing the need for careful selection of cleaning methods. Additionally, model selection is equally important, as we observed that XGBoost consistently outperformed other models. The results demonstrate that targeted cleaning of categorical variables and using XGBoost can drive consistent performance improvements for depression prediction without compromising data integrity.

References

[1] F. Furizal, A. Ma'arif, and D. Rifaldi, "Application of Machine Learning in Healthcare and Medicine: A Review," Journal of Robotics and Control (JRC), vol. 4, Sep. 2023. doi: 10.18196/jrc.v4i5.19640.

[2] E. Breck, N. Polyzotis, S. Roy, S. Whang, and M. Zinkevich, "Data Validation for Machine Learning," in Proceedings of Machine Learning and Systems (MLSys), 2019. Available: https://proceedings.mlsys.org/paper_files/paper/2019/file/928f1160e52192e3e0017fb63ab65391-Paper.pdf.

[3] K. Maharana, S. Mondal, and B. Nemade, "A Review: Data Pre-Processing and Data Augmentation Techniques," Global Transitions Proceedings, vol. 3, Apr. 2022. doi: 10.1016/j.gltp.2022.04.020.

[4] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks," in 2021 IEEE 37th International Conference on Data Engineering (ICDE), 2021, pp. 13-24. doi: 10.1109/ICDE51399.2021.00009.

[5] Z. Shao, M. Ahmad, and A. Javed, "Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface," Remote Sensing, vol. 2024, p. 665, Feb. 2024. doi: 10.3390/rs16040665.