Cluster Innovation Centre
University of Delhi
Dr. Mahima Kaushik
Prof. at Cluster Innovation Centre,
University of Delhi

An In-Depth Analysis of ML Approaches to Predict Depression

Siddharth (152249) Semester IV siddharth899@gmail.com

1. Abstract

Depression affects over 264 million people globally, significantly impacting daily life and well-being. Traditional diagnostic methods are often subjective and time-consuming. This study explores the potential of machine learning (ML) models to enhance depression diagnosis by leveraging socio-demographic and psychosocial data. **Five ML classifiers—K-Nearest Neighbors (KNN), AdaBoost, Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), and Bagging—were evaluated. Advanced feature selection techniques (SelectKBest, mRMR, Boruta) and Synthetic Minority Oversampling Technique (SMOTE) were employed to address class imbalance. Results indicated that AdaBoost combined with SelectKBest achieved the highest accuracy of 90.9%.** The findings suggest that ML approaches, particularly when integrated with effective feature selection and class balancing strategies, can significantly improve depression diagnosis and treatment.

2. Methodology

The study utilized data from 604 participants, including 30 predictor variables and one target variable (depression status). The methodology involved:

1. **Data Splitting**: Dividing the dataset into training (80%) and testing (20%) sets.
2. **Data Encoding**: Converting categorical data into numerical values using label encoding.
3. **Feature Selection**: Applying SelectKBest, mRMR, and Boruta algorithms to extract the most relevant features.
4. **SMOTE**: Addressing class imbalance by generating synthetic samples for the minority class.
5. **Model Training and Evaluation**: Training and evaluating five ML classifiers using accuracy, sensitivity, specificity, precision, F1-score, and ROC-AUC metrics.

.

3. Overview of ML models

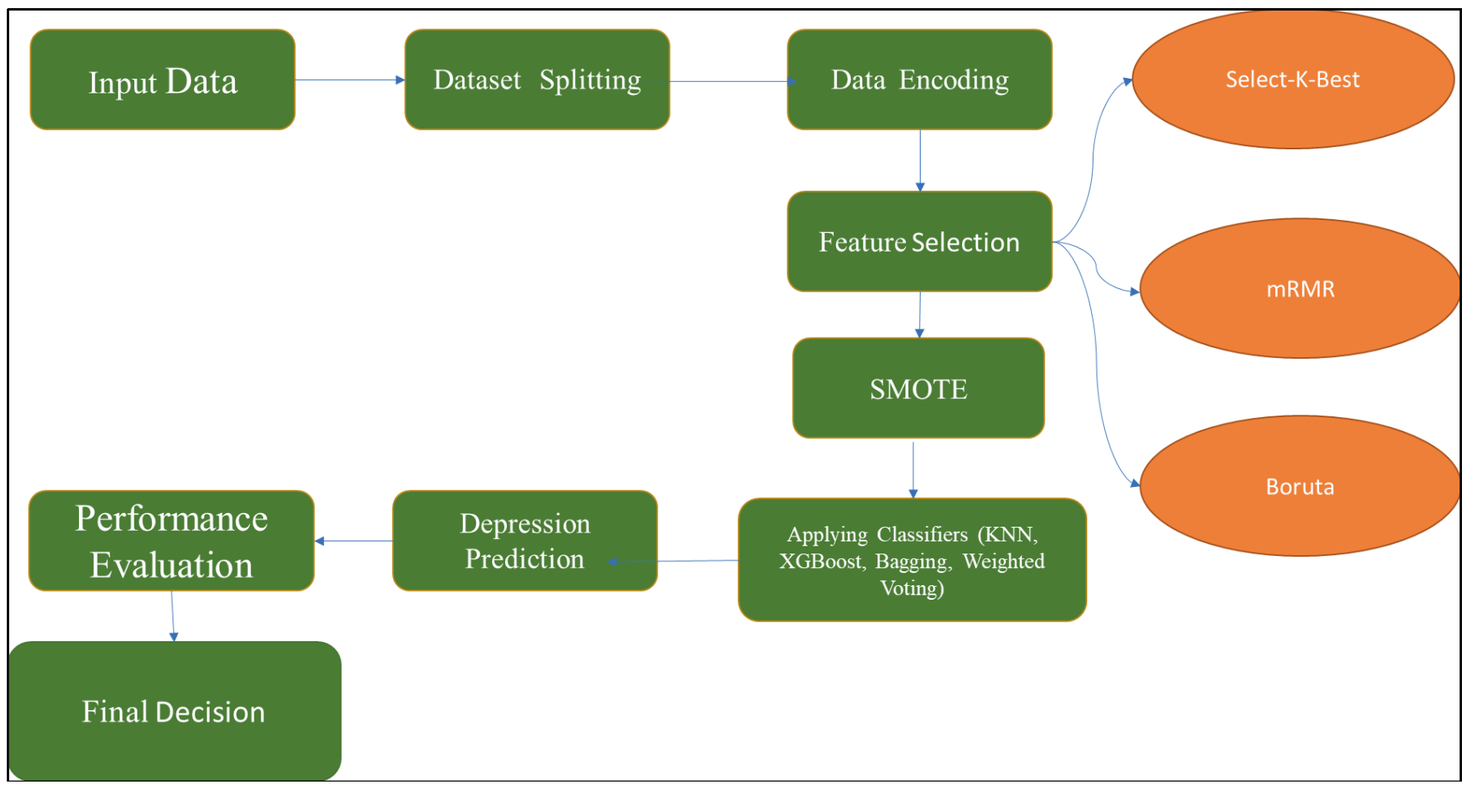
1. **K-Nearest Neighbors (KNN)**: Classifies data points based on the majority class of their nearest neighbors.
2. **AdaBoost**: Combines multiple weak classifiers to create a strong classifier, focusing on misclassified instances.
3. **Gradient Boosting (GB)**: Builds models sequentially, each new model correcting the errors of the previous one.
4. **XGBoost**: An optimized implementation of gradient boosting, including regularization, parallel processing, and tree pruning.
5. **Bagging**: Reduces variance and improves stability by training multiple instances of a base model on different bootstrap samples of the dataset.

4. Pre-Processing and Model Training

- **Pre-processing :**The first step in preparing data involved cleaning and handling missing values, essential in survey-based datasets where inconsistencies and gaps are common. Categorical features, like demographic and psychosocial information, were encoded using techniques like one-hot encoding or label encoding. Continuous variables were normalized to ensure uniform scaling. Additionally, the class imbalance issue was addressed by applying the Synthetic Minority Oversampling Technique (SMOTE). This helped create a balanced training set by generating synthetic samples for the minority class, preventing the models from being biased toward the majority class.

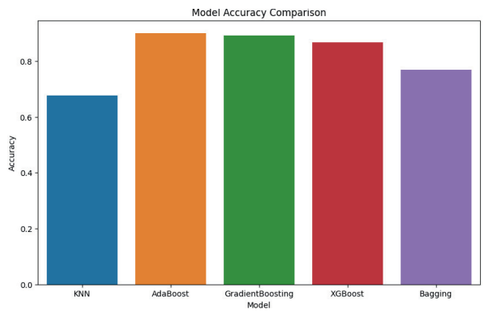
- **Feature Selection:** To identify the most relevant predictors, feature selection methods such as SelectKBest, mRMR, and Boruta were employed. These methods help reduce the feature space, minimize noise, and improve training efficiency by focusing on the most influential features.

- **Model Training**: After preprocessing, the dataset was split into training and testing sets. Each machine learning model was then trained on the training data using cross-validation to ensure generalization. Hyperparameters for each classifier were fine-tuned to optimize predictive performance. The training process involved iterative parameter adjustments and evaluation of model performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

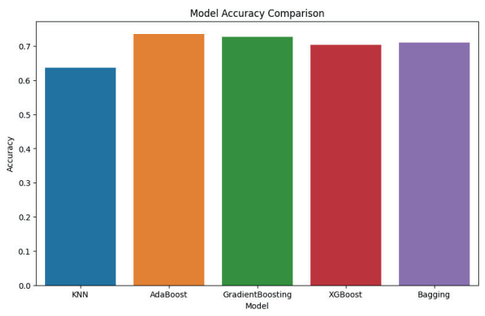


- **Evaluation Metrics:** To compare the models, their predictive performance was measured using metrics that provide insights into different aspects of classification quality. This included sensitivity (recall) to measure the model's ability to identify true positives, specificity for true negatives, and ROC-AUC to provide a comprehensive understanding of model performance.

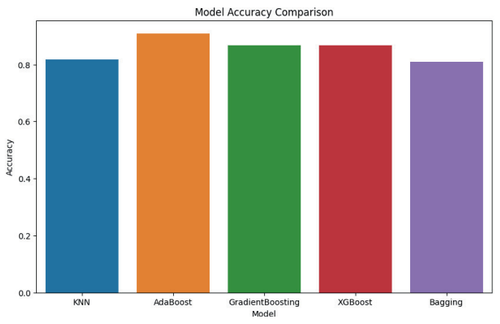
5. Results.



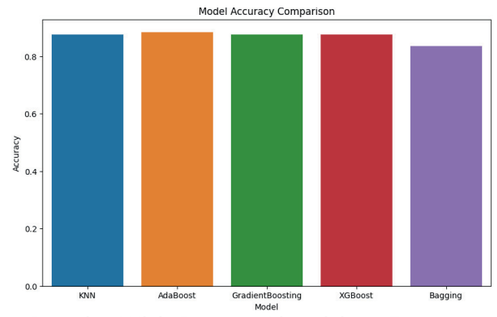
Results Without Feature Selection



Results With MRM features



Results With Select-K-Best features



Results With Boruta Algorithm Features

- **Without Feature Selection**: AdaBoost and Gradient Boosting achieved the highest accuracies of 90.08% and 89.26%, respectively.
- **With SelectKBest**: AdaBoost achieved the highest accuracy of 90.91%, followed by Gradient Boosting and XGBoost (86.78%).
- **With Boruta:** KNN, AdaBoost, and Gradient Boosting all achieved accuracies around 87.60% to 88.43%.
- **ROC Curve Analysis**: AdaBoost consistently showed the highest AUC values across different feature selection methods, indicating superior discriminative power.

6. Conclusion.

This study demonstrated the efficacy of machine learning models in predicting depression. AdaBoost and Gradient Boosting emerged as the most robust models, with feature selection techniques significantly enhancing model performance. Addressing class imbalance with SMOTE further improved predictive capabilities. The integration of advanced ML techniques and feature selection methods provides a promising approach for accurate depression prediction, aiding early detection and intervention.

7. References.

1. Md. Sabab Zulfiker, Nasrin Kabir, Al Amin Biswas, Tahmina Nazneen, Mohammad Shorif Uddin, An in-depth analysis of machine learning approaches to predict depression, Current Research in Behavioral Sciences, Volume 2, 2021, 100044, ISSN 2666-5182, https://doi.org/10.1016/j.crbeha.2021.100044.
2. Choudhury, A.A., Khan, M.R.H., Nahim, N.Z., Tulon, S.R., & Islam, S., Chakrabarty, A. (2019). Predicting depression in Bangladeshi undergraduates using machine learning. In: Proceedings of the 2019 IEEE Region 10 Symposium (TENSYMP). IEEE, pp. 789-794.
3. Cvetković, J. (2017). Breast cancer patients' depression prediction by machine learning approach. Cancer Investigation, 35(8), 569-572.