Effective Predictive model for diabetes classification using optimized machine learning on imbalanced dataset

G R Ashisha[1,2], Sai Kiran Oruganti[3]

[1] Postdoctoral Researcher, Lincoln University College, Malaysia; [2] Assistant Professor, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India; [3] Associate Professor, Lincoln University College, Malaysia

Email ID: pdf.ashisha@lincoln.edu.my grashisha27@gmail.com

Abstract: Diabetes mellitus is a growing health issue that demands precise and early detection. In recent research, machine learning techniques have shown promising results in assisting medical practitioners with the prediction of the disease. However, the presence of class imbalance and missing values is a common issue with real-world medical datasets, which may impact the prediction performance. In this paper, the authors propose an optimized machine learning framework that uses a class-balanced dataset generated with the Synthetic Minority Over-sampling Technique. The proposed framework is experimented with the Diabetes 130-US Hospitals dataset. Data preprocessing techniques were used to improve the quality of the dataset. Various machine learning algorithms were used, including CatBoost, XGBoost, LightGBM, and stacking. In the experimentation process, the authors observed that the use of ensemble-based machine learning techniques resulted in better classification performance.

Keywords: SMOTE; LightGBM; Machine Learning; Diabetes; Ensemble Technique

---

Introduction

Diabetes mellitus is one of the most common chronic health conditions that affects millions of people worldwide. If not detected and treated in its early stages, it can cause serious health problems, such as heart problems, failure of the kidneys, nerve problems, and vision problems. According to health reports, the number of people suffering from diabetes is on the rise every year. Hence, the detection and prediction of diabetes is an important research area in healthcare analytics.

Machine learning has also come up as a potential solution [1] to deal with medical data sets to identify complex patterns that might not be easily identified using traditional statistical tools. The predictive model using machine learning would help clinicians take accurate decisions based on patient information. However, there are certain issues in the data sets used in healthcare management, such as missing information, noisiness in data, and class imbalance problems in which the number of normal cases is very high in comparison to diabetic cases.

To overcome the challenges, this research is centered on the application of appropriate preprocessing techniques, balancing of the data, and the application of optimized ML algorithms in the improvement

of the accuracy of the diabetes prediction process. The proposed method applies the SMOTE algorithm in balancing the classes and tests the effectiveness of various robust ML algorithms in the classification process.

Related work

There are a number of research studies that have investigated machine learning approaches in diabetes prediction. Researchers have used machine learning algorithms to classify diabetic patients from non-diabetic patients. The machine learning algorithms used in this research are logistic regression, support vector machine, decision tree, and random forest.

Recent studies have proposed more advanced ensemble techniques [2], including XGBoost, LightGBM, and CatBoost, which can be used to analyze medical data. These techniques improve the prediction quality with the help of multiple decision trees and optimized gradient boosting. Deep learning techniques have also been proposed for the analysis of large medical data sets.

One of the biggest challenges in medical prediction problems is still the issue of data imbalance [3]. Oversampling methods such as SMOTE have been used in many applications to improve the performance of the model by creating synthetic minority class samples. Research has suggested that the use of data balancing techniques [4] in combination with ensemble learning techniques can improve the results of prediction in healthcare applications.

Key Contribution

The significant contributions of the present work can be summarized as follows:

- Efficient implementation of a data preprocessing step for the diabetes dataset.
- Use of the SMOTE technique for class imbalance.
- Comparison of the effectiveness of sophisticated machine learning algorithms, namely CatBoost, XGBoost, and LightGBM.
- Implementation of a stacking ensemble model for improved model accuracy.
- Improving the classification results for the diabetes prediction problem using the proposed balancing datasets.

Method, Experiments and Results

The Diabetes 130-US Hospitals dataset was used for the study. First, the attributes that were not necessary, i.e., the patient attributes, were removed to reduce complexity. The attributes, e.g., race, gender, age group, mode of admission, mode of discharge, and medication status, were converted to numerical values using label encoding.

The missing values, which were represented by special symbols, were replaced with 'NaN' values. The median and mode imputation techniques were used for handling missing values in an effective manner. After the preprocessing stage, the dataset had 69,984 patient records with 46 relevant attributes. Since the data is imbalanced, the SMOTE algorithm is used to create synthetic samples of the minority class. Then the data is split into a training set and a testing set based on the proportion of 80:20.

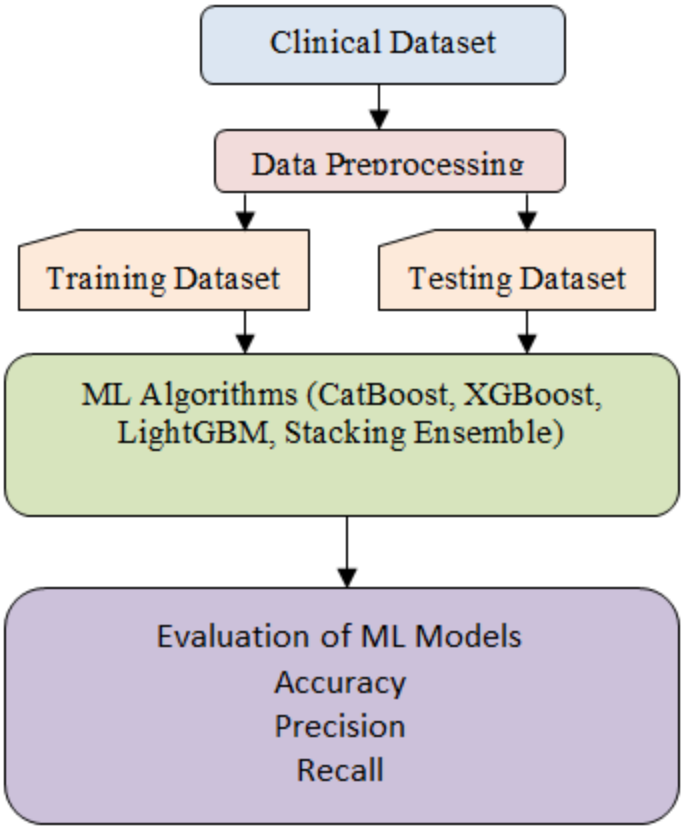

Figure 1. Implementation of proposed method.

Three powerful gradient boosting algorithms were implemented: CatBoost, XGBoost, and LightGBM. In addition, the stacking ensemble method was used. The performance of the models was validated based on the classification performance metrics. Observations from the experiments (Figure 1) showed that the prediction accuracy of the ensemble learning models is high compared to the individual classifiers. The stacking ensemble method recorded the best performance.

Discussions

The results show that the preprocessing and balancing of the data have an important effect on the performance of machine learning models. SMOTE was successful in increasing the count of the minority class, resulting in less biased classification results.

Gradient Boosting algorithms such as XGBoost and LightGBM showed promising results in terms of prediction (Table 1), as they have the capability to handle complex, nonlinear relationships in the data. The stacking model also showed promising results, as it combined the results of multiple classifiers. The results show that the proposed method can help healthcare professionals identify high-risk patients more efficiently.

Table 1. Comparison of all the classifier model

ML Models	Performance metrics							
	Accuracy (%) Before Preprocessing	Accuracy (%) after Preprocessing	Precision (%) Before Preprocessing	Precision (%) after Preprocessing	Recall (%) Before Preprocessing	Recall (%) after Preprocessing	F1-Score (%) before preprocessing	F1-Score (%) after preprocessing
CatBoost	78	81	78	79	77	78	78	79.5
XGBoost	76.5	78	77	75	75	75	74	75
LightGBM	80	81	80	82	79	80	79	81
Stacking Ensemble	80.2	81.3	80	83	78	81	78	81.5

Conclusions

The paper proposed an optimized machine learning framework for diabetes classification using imbalanced medical datasets. The proposed method utilized various preprocessing techniques, SMOTE algorithm, and advanced machine learning algorithms such as CatBoost, XGBoost, LightGBM, and stacking ensemble techniques. The proposed method was evaluated using various experiments, which showed the effectiveness of ensemble learning algorithms in terms of prediction accuracy and reliability. The proposed framework can be utilized as a decision-support tool for the early detection of diabetes and various health-related applications. Future research can be carried out using deep learning techniques and large-scale datasets.

References

1. P. Verma and A. Joshi, "Improving disease prediction accuracy using hybrid machine learning models", IEEE Access, vol. 11, pp. 76512-76524, 2023. https://doi.org/10.1109/ACCESS.2023.3278942
2. M. Patel, R. Mehta and K. Shah, "Predictive analytics for diabetes diagnosis using machine learning", IEEE Access, vol. 11, pp. 67231-67242, 2023. https://doi.org/10.1109/ACCESS.2023.3267812
3. S. Ahmed, M. Khan and A. Khan, "Machine learning approaches for diabetes prediction using healthcare datasets", IEEE Access, vol. 11, pp. 24561-24574, 2023. https://doi.org/10.1109/ACCESS.2023.3245671
4. A. Kumar and S. Singh, "Gradient boosting techniques for medical data classification", IEEE Access, vol. 10, pp. 45671-45682, 2022. https://doi.org/10.1109/ACCESS.2022.3178910