Improving Diabetes Prediction Accuracy Through Effective Outlier Handling

Nirasha J.A.M
Department of Computer Science and Engineering
University Of Moratuwa
Moratuwa, Sri-Lanka
mayurie.25@cse.mrt.ac.lk

***Abstract*—This study explores the use of machine learning models to predict diabetes using the PIMA Indian Diabetes dataset. The dataset includes health-related features for female patients. Pre-processing steps included handling missing values, outlier removal using the Mahalanobis distance, and class balancing using Adaptive Synthetic Sampling Approach for Imbalanced Learning. Four models—Random Forest, Gradient Boosting, HistGradientBoosting, and XGBoost—were tested. Gradient Boosting performed best, especially after removing outliers, achieving the highest accuracy of 0.9041 and recall of 0.9200. These findings highlight the effectiveness of proper outlier handling in improving prediction accuracy.**

***Index Terms*—Diabetes Prediction, PIMA Indian Dataset, Mahalanobis Distance, Machine Learning**

I. Introduction

Diabetes is a growing health concern worldwide. So early detection is critical to managing its long-term impact. With the rise of healthcare data and machine learning, it is now possible to build predictive models that can support diagnosis and decision-making.

This study focuses on predicting the presence of diabetes using the PIMA Indian Diabetes dataset. The primary goal is to evaluate the performance of various machine learning models and to examine the impact of proper outlier detection and handling on model accuracy.

A key part of this research is data pre-processing, which includes handling missing values, addressing class imbalance, and identifying outliers using statistical techniques. By comparing model performance before and after outlier removal, the study aims to highlight how these pre-processing steps contribute to more accurate and reliable predictions.

II. Methodology

A. Overview of the dataset

The PIMA Indian Diabetes dataset was used for diabetes prediction. This dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases, and all patients included are female. It contains the following predictor variables: number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, Body Mass Index(BMI), diabetes pedigree function, and age. The outcome variable indicates whether a patient has diabetes or not. The dataset consists of 768 entries, of which 268 are diabetic and 500 are non-diabetic. Therefore, the dataset is highly imbalanced.

B. Pre-processing

1) Missing value handling: For the variables Glucose, Blood Pressure, SkinThickness, Insulin, and BMI, there were many values recorded as 0. These values are not realistic, so they were treated as missing values. Mean imputation was used to fill in these missing values. The mean of each variable was calculated separately for diabetic and non-diabetic patients. Then, missing values were filled using the mean value based on the patient's diabetes status.

2) Standerdization: The features were standardized to have a mean of 0 and standard deviation of 1. Numerical columns were scaled using StandardScaler. This helps the model treat all features equally, especially those with different units or ranges.

3) Outlier handling: Outliers are data points significantly different from the rest of the dataset. To detect outliers, the Mahalanobis distance technique was used. Mahalanobis Distance is a statistical tool used to measure the distance between a point and a distribution. It can identify the outliers in multivariate space. Points with a large Mahalanobis distance from the mean were marked as outliers. These data points can be considered outliers and may need to be removed or investigated further.

The figure 1 shows the outliers detected in the multivariate space. If the Mahalanobis distance of a data point was greater than the 95th percentile, it was considered an outlier. A total of 39 such outliers were found. To study the effect of outliers, two models were trained, one with outliers and one after removing outliers. The results were compared to decide which approach gave better performance.

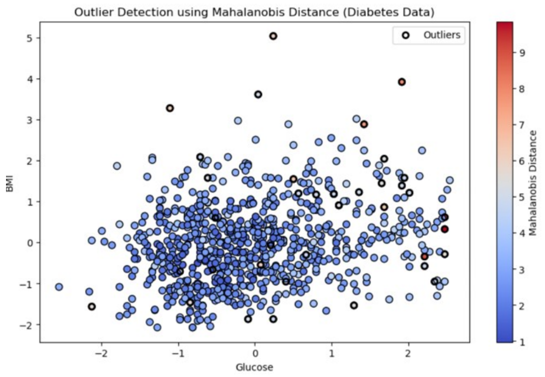

Fig. 1. Outliers detected using Mahalanobis Distance

4) Data Splitting, Class Imbalance Handling: The dataset was first split into diabetic and non-diabetic groups. A stratified split was applied to keep the class balance in both training and testing sets. Each class was divided into 80% for training and 20% for testing. The diabetic and non-diabetic training subsets were then combined to form the final training set, and the corresponding testing subsets were combined to form the final testing set. he training data was shuffled to eliminate any order bias. Since the dataset was imbalanced, with fewer diabetic cases, the Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) technique was used. ADASYN creates synthetic samples for the minority class to help the model learn better. It was applied only to the training data.

C. Model fitting and Evaluation

Several machine learning models were selected based on findings from previous studies. These models include Random Forest, Gradient Boosting, HistGradientBoosting, and XGBoost. Each model was trained and tested to predict whether a person has diabetes. The goal was to compare their performance and choose the best model for this task. The performance of the model was measured using the test set. The evaluation metrics used were Accuracy, Precision, Recall, F1-score, and AUC.

III. Result

A. Model performances with Outliers

TABLE 1
PERFORMANCE COMPARISON BEFORE OUTLIER REMOVAL

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Random Forest	0.8766	0.8431	0.7963	0.8190	0.9540
Gradient Boosting	0.8831	0.8600	0.7963	0.8269	0.9635
HistGradientBoosting	0.8636	0.8511	0.7407	0.7921	0.9485
XGBoost	0.8766	0.8571	0.7778	0.8155	0.9576

The table 1 shows the performance of four models tested with outliers in the dataset. Gradient Boosting gave the best accuracy of 0.8831, best precision of 0.8600, best F1-score of 0.8269 and the highest AUC of 0.9635. This suggests it was the most reliable model in this case. Random Forest and XGBoost also performed well with accuracy scores of 0.8766 and AUC values of 0.9540 and 0.9576. HistGradientBoosting had slightly lower accuracy at 0.8636 and the lowest recall of 0.7407. Overall, Gradient Boosting performed the best when outliers were present.

B. Model Performances after Removing Outliers using the Mahalanobis Distance technique

TABLE 2
PERFORMANCE COMPARISON AFTER OUTLIER REMOVAL

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Random Forest	0.8904	0.8269	0.8600	0.8431	0.9514
Gradient Boosting	0.9041	0.8214	0.9200	0.8679	0.9660
HistGradientBoosting	0.8973	0.8431	0.8600	0.8515	0.9615
XGBoost	0.8973	0.8431	0.8600	0.8515	0.9587

The table 2 shows the performance of the models on the test set after removing outliers using the Mahalanobis Distance method. All models showed improved results compared to when outliers were included. Gradient Boosting achieved the highest accuracy of 0.9041 and the highest recall of 0.9200, indicating strong ability in detecting positive cases. It also gave the best F1-score of 0.8679 and the highest AUC of 0.9660. Random Forest also performed well, with an accuracy of 0.8904 and F1-score of 0.8431. HistGradientBoosting and XGBoost both achieved an accuracy of 0.8973 and F1-score of 0.8515. Their precision and recall scores were also closely matched. Overall, all models benefited from outlier removal, with Gradient Boosting again showing the best overall performance.

C. Feature Importance

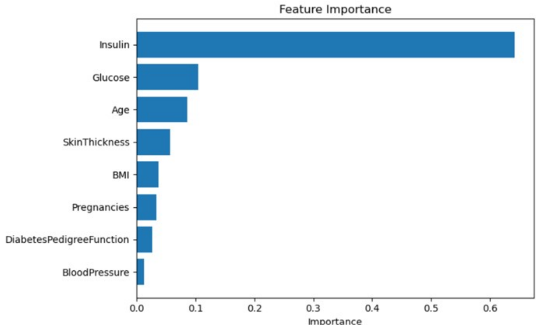

Fig. 2. Feature Importance of Gradiant Boosting Model

The feature importance plot in Figure 2 was obtained using the Gradient Boosting Classifier. It shows that Insulin was the most important feature in predicting diabetes. It had the highest influence on the model. Glucose and Age also showed notable importance. Features like skin thickness and BMI had moderate contribution. Pregnancies and diabetes pedigree function had lower influence. Blood pressure was the least important feature. These results help understand which health indicators the model relied on most.

IV. Conclusion

This study shows that proper data preprocessing plays a key role in building accurate diabetes prediction models. One major finding is that removing outliers made a clear difference. Before outlier removal, the highest accuracy was 0.8831. After removing outliers, it increased to 0.9041 — an improvement of about 2.38Balancing the dataset and cleaning the data helped all models perform better. Among the models tested, Gradient Boosting gave the best results in detecting diabetic cases. The analysis also showed that insulin level, glucose level, and age were the most important features for prediction. These findings can help support early diagnosis of diabetes and guide healthcare decisions. In the future, more advanced models and larger, more diverse datasets could be explored to improve prediction further.

References

[1] M. S. Salih, R. K. Ibrahim, S. R. M. Zeebaree, D. A. Zebari, L. M. Abdulrahman, and N. M. Abdulkareem, ”Diabetic Prediction based on Machine Learning Using PIMA Indian Dataset,” Communications on Applied Nonlinear Analysis, vol. 31, no. 5s, pp. 138, 2024.

[2] A. Mousa, R. Marqas, and others, “A Comparative Study of Diabetes Detection Using The Pima Indian Diabetes Database,” The Journal of The University of Duhok, vol. 26, no. 2, art. no. 24, Oct. 2023.

[3] V. Sharma, “Unlocking the Power of Mahalanobis Distance: Exploring Multivariate Data Analysis with Python,” Medium, Mar. 6, 2023. [Online]. Available: https://medium.com/@the$_{d}aft_{i}ntrovert/mahalanobis-distance-5c11a757b099$