# Comparative Analysis of Machine Learning Approaches for Diabetes Prediction

Ahmed A. Ibrahim[1]

*1.Misr University for Science and Technology, Egypt*

***Abstract: Diabetes is a common chronic disease with global impact that requires early detection and intervention to mitigate health-related risks. This study focuses on developing predictive models using machine learning methods. Using a dataset that contains various health-related attributes such as age, gender, body mass index, and common medical conditions such as hypertension and heart disease, we aim to create a powerful prediction framework. The main goal is to create a tool for early detection and intervention strategies for diabetes. By harnessing the potential of machine learning technology, this research aims to provide a feasible approach to address the impact of diabetes on individual health outcomes.***

***Keywords: Diabetes, Diagnosis, Machine Learning***

## 1) Introduction

Diabetes remains a common and complex chronic disease that places a heavy burden on health systems and individual well-being worldwide. Its prevalence continues to increase, highlighting the urgent need for effective prognostic tools to enable early detection and intervention [1]. This article explores the field of predictive modeling using machine learning techniques to address this pressing healthcare challenge.

The importance of timely identification in diabetes cannot be overstated. Early detection allows for proactive management, reducing the risk of complications such as heart problems, kidney diseases, etc. [2] and enhancing overall quality of life for affected individuals. Leveraging advancements in machine learning, this research aims to compare and develop a predictive model that harnesses diverse health-related attributes to accurately foresee the onset or progression of diabetes. Attributes such as age, gender, body mass index (BMI), and pre-existing conditions like hypertension and heart disease form the core components of the dataset.

This study aims to conduct a comprehensive analysis of various classification algorithms—Logistic Regression (LR), Naive Bayes (NB), Decision Trees (DT), Random Forest (RF), XGBoost, K-Neighbors (KNN), and Neural Networks (ANN)—in the context of diabetes prediction. The primary objective is to compare the accuracy and precision of these algorithms in classifying diabetes cases.

## 2) Material and Methodology

### 2.1) Material

We used the diabetes prediction dataset [3], which contains 100,000 rows. The distribution of data is 8.82% of patients have diabetes while the rest is healthy, as shown in Figure 1. The dataset contains information about age, gender, body mass index (BMI), and pre-existing conditions like hypertension and heart disease, smoking history, blood glucose levels, HbA1C levels.

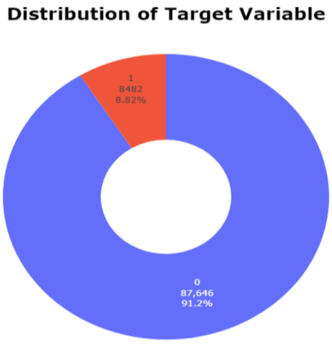


*Figure 1Distribution of Target Variable*

## 2.2) Methodologies:

### *2.2.1)Data cleaning:*

The dataset was analyzed for any duplicates or inconsistencies in data types. Features with less than 5 unique values had their data types changed to categorical, while any inconsistencies or duplicates were removed (3.8%). the smoking_history feature was analyzed and reduced the unique values from 6 to 4.

### *2.2.2)Data Preparation:*

an sklearn Column Transformer was used that would scale the numerical variables using StandardScaler[1] and One Hot Encode the categorical variables

1. $Z = \frac{X - \mu}{\sigma}$ where $\mu$ is the mean and $\sigma$ is the standard deviation.

### *2.2.3)Algorithm Selection:*

a random sample of 5000 instances of each value in target variable was selection and split into training and test data to test the classifiers. Each classifier was given a set of parameters and was tested using 5 fold GridSearch that would try every combination of parameters to find the optimal Algorithm.

### *2.2.4)Model Evaluation:*

Each algorithm in GridSearch would be tested using a few metrics:

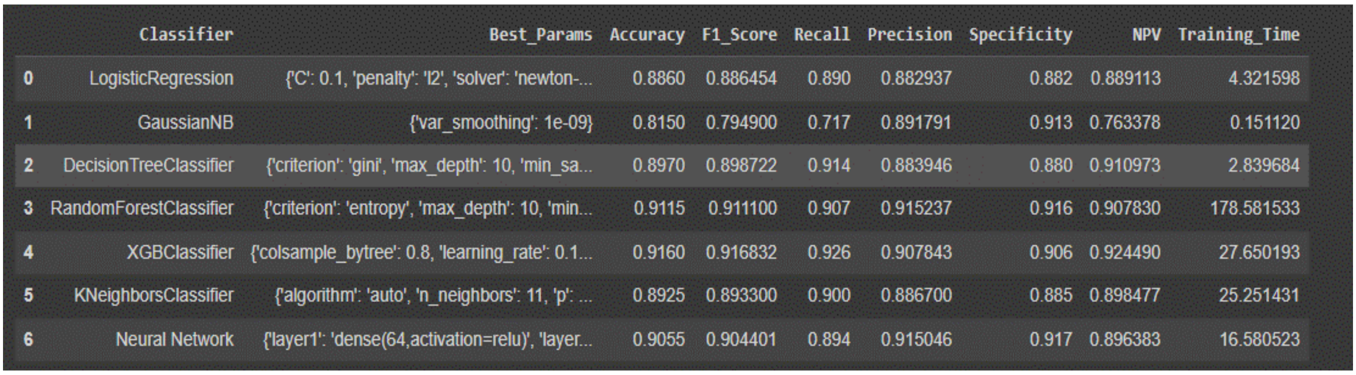
- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- F1 Score = $\frac{2*Precision*Recall}{Precision+Recall}$
- Recall = $\frac{TP}{TP+FN}$
- Precision = $\frac{TP}{TP+FP}$
- Specificity = $\frac{TN}{TN+FP}$
- NPV = $\frac{TN}{TN+FN}$

### *2.2.5) Solving for Imbalanced Data:*

Since the dataset is imbalanced, we compared 2 different ways of solving this problem. Once by up sampling it using Synthetic Minority Over-sampling Technique (SMOTE), and once by down sampling it using RandomUnderSampler.

## 3) RESULTS

The comprehensive evaluation of seven distinct machine learning algorithms for diabetes prediction yielded insightful results regarding their performance. The best performing model was the XGBoost Classifier with an accuracy of 0.916, F1 score of 0.916, Recall of 0.926, Precision of 0.907, Specificity of 0.906, and NPV of 0.924. making the best choice to use. After applying SMOTE and RandomUnderSampler, the XGBoost classifier was tested again using the best parameters from the previous step. Both models were tested on test data that wasn't affected by the data modifications. The model Trained on up sampled data performed significantly better than the one trained on down sampled data. Reaching Accuracy = 0.956, F1 Score = 0.759, Recall = 0.772, Precision = 0.746, Specificity = 0.974, and NPV = 0.977. with an ROC-AUC = 0.87



| | Classifier | Best_Params | Accuracy | F1_Score | Recall | Precision | Specificity | NPV | Training_Time |
|---|---|---|---|---|---|---|---|---|---|
| 0 | LogisticRegression | {'C': 0.1, 'penalty': 'l2', 'solver': 'newton-... | 0.8860 | 0.886454 | 0.890 | 0.882937 | 0.882 | 0.889113 | 4.321598 |
| 1 | GaussianNB | {'var_smoothing': 1e-09} | 0.8150 | 0.794900 | 0.717 | 0.891791 | 0.913 | 0.763378 | 0.151120 |
| 2 | DecisionTreeClassifier | {'criterion': 'gini', 'max_depth': 10, 'min_sa... | 0.8970 | 0.898722 | 0.914 | 0.883946 | 0.880 | 0.910973 | 2.839684 |
| 3 | RandomForestClassifier | {'criterion': 'entropy', 'max_depth': 10, 'min... | 0.9115 | 0.911100 | 0.907 | 0.915237 | 0.916 | 0.907830 | 178.581533 |
| 4 | XGBClassifier | {'colsample_bytree': 0.8, 'learning_rate': 0.1... | 0.9160 | 0.916832 | 0.926 | 0.907843 | 0.906 | 0.924490 | 27.650193 |
| 5 | KNeighborsClassifier | {'algorithm': 'auto', 'n_neighbors': 11, 'p': ... | 0.8925 | 0.893300 | 0.900 | 0.886700 | 0.885 | 0.898477 | 25.251431 |
| 6 | Neural Network | {'layer1': 'dense(64,activation=relu)', 'layer... | 0.9055 | 0.904401 | 0.894 | 0.915046 | 0.917 | 0.896383 | 16.580523 |

*Figure 2 Comparing Different Classifiers*

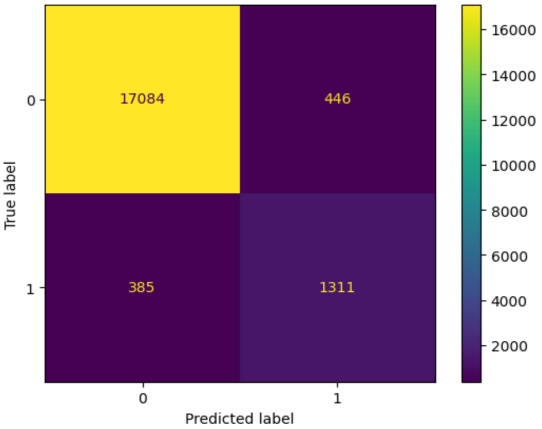


*Figure 4 XGBoost After up sampling*

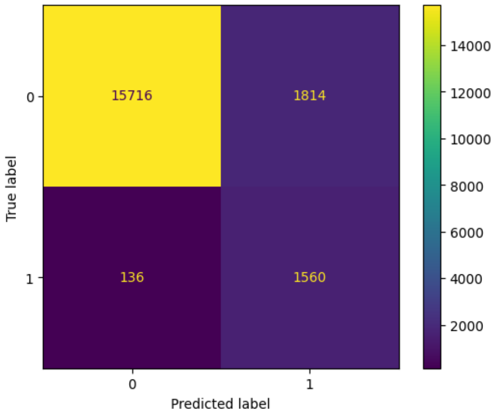


*Figure 3 XGBoost After down sampling*

## 4) Conclusions:

In conclusion, this study conducted a comprehensive analysis of seven machine learning algorithms for predicting diabetes onset, unraveling valuable insights into their comparative performance.

The findings showcased varied strengths and performance nuances across algorithms, emphasizing the importance of tailored model selection for diabetes classification. XGBoost emerged as a frontrunner, exhibiting superior accuracy and discriminatory power, closely followed by Random Forest and Neural Networks. Ensemble methods, particularly XGBoost and Random Forest, demonstrated robust performance across multiple metrics, highlighting their effectiveness in accurate diabetes prediction.

precision, recall, and f1-score evaluations unveiled the nuanced balance between correctly identifying positive cases and capturing a high proportion of actual positives. feature importance analysis elucidated influential attributes crucial for accurate diabetes prediction, aiding in understanding the underlying factors contributing to classification.

## References

[1] *Farajollahi, Boshra & Mehmannavaz, Maysam & Mehrjoo, Hafez & Moghbeli, Fateme & Sayadi, Mohammad. (2021). Diabetes Diagnosis Using Machine Learning. Frontiers in Health Informatics. 10. 65. 10.30699/fhi.v10i1.267.*

[2] Warke M, Kumar V, Tarale S, Galgat P, Chaudhari D. Diabetes diagnosis using machine learning algorithms. International Research Journal of Engineering and Technology. 2019; 6(3): 1470-6.

[3] Mustafa, M. (2023, April 8). *Diabetes prediction dataset*. Kaggle. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset