Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 27s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Diabetes Detection Using Gradient Boosting Classifier (XGBOOST)

Faiz Ahmed Siddiqui[1], Md Akib Alam[2], Sharik Ahmad[*3]

[1]Department Of Computer Science & Engineering, Sharda School of Computing Science & Engineering, Noida, Uttar Pradesh, India.

Email ID: 2023392533.faiz@pg.sharda.ac.in

[2]Department Of Computer Science & Engineering, Sharda School of Computing Science & Engineering, Noida, Uttar Pradesh, India.

Email ID: 2023493399.md@pg.sharda.ac.in

[3]Department of Computer Science & Engineering, Sharda School of Computing Science & Engineering, Noida, Uttar Pradesh, India.

Email ID: sharik.ahmad@sharda.ac.in

Cite this paper as: Faiz Ahmed Siddiqui, Md Akib Alam, Sharik Ahmad, (2025) Diabetes Detection Using Gradient Boosting Classifier (XGBOOST). *Journal of Neonatal Surgery*, 14 (27s), 94-98.

ABSTRACT

Diabetes results from elevated glucose levels in humans and should not be overlooked if left untreated, as it can lead to significant health issues, including heart complications, kidney disorders, hypertension, and eye damage, as well as impact other organs. Early detection of diabetes can help manage the condition effectively. To accomplish this, we aim to predict diabetes in individuals with high accuracy by utilizing various machine learning techniques. These techniques enhance prediction outcomes by developing models from patient data. In this research, we applied machine learning classification and ensemble methods to a dataset for diabetes prediction. The techniques used include K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting (XGBOOST), LightGradientBoosting (LightGBM) and Random Forest (RF). Each model demonstrated varying levels of accuracy when compared to one another. This project identifies a model with superior accuracy, indicating its effectiveness in predicting diabetes. Our findings reveal that the Gradient Boosting Classifier (XGBOOST) method achieved greater accuracy than the other machine learning techniques.

Keywords: *Diabetes, Machine, Learning, Prediction, Dataset.*

1. INTRODUCTION

Diabetes is a noxious disease that occurs world over. Obesity, high blood glucose levels, and multiple other factors are responsible for diabetes. It influences the hormone insulin, causing an abnormal metabolic process in crabs and enhances the pitch of sugar in the blood. Diabetes occurs because the body does not produce sufficient insulin. People with diabetes reach 422 million worldwide According to the World Health Organization (WHO), mainly in low-income or middle-income countries. This number could reach as 490 billion by 2030. Nonetheless, diabetes is common in many countries including Canada, China, India. The population of India is now over 100 million, and the actual number of diabetic patients in India is 40 million. Diabetes is one of a leading causes of global death worldwide .Disease- like diabetes can be predicted at an early stage which can help to control it and save human life. To this end ,this study examined prediction diabetes using different diabetes-related attributes. We employed the Pima Indian Diabetes Dataset and developed several machine-learning classification and ensemble algorithms to predict diabetes at this end. Machine Learning is a technique used to expressly train computers and machine expressly .Different machine-learning approaches formulating different classification and ensemble models from the accumulated dataset.it can be helpful for diabetes prediction .There are many Machine-learning techniques available to use for prediction ,but it is hard to find the best one ,Thus, for this purpose, we applied popular classification and ensemble methods to datasets for prediction

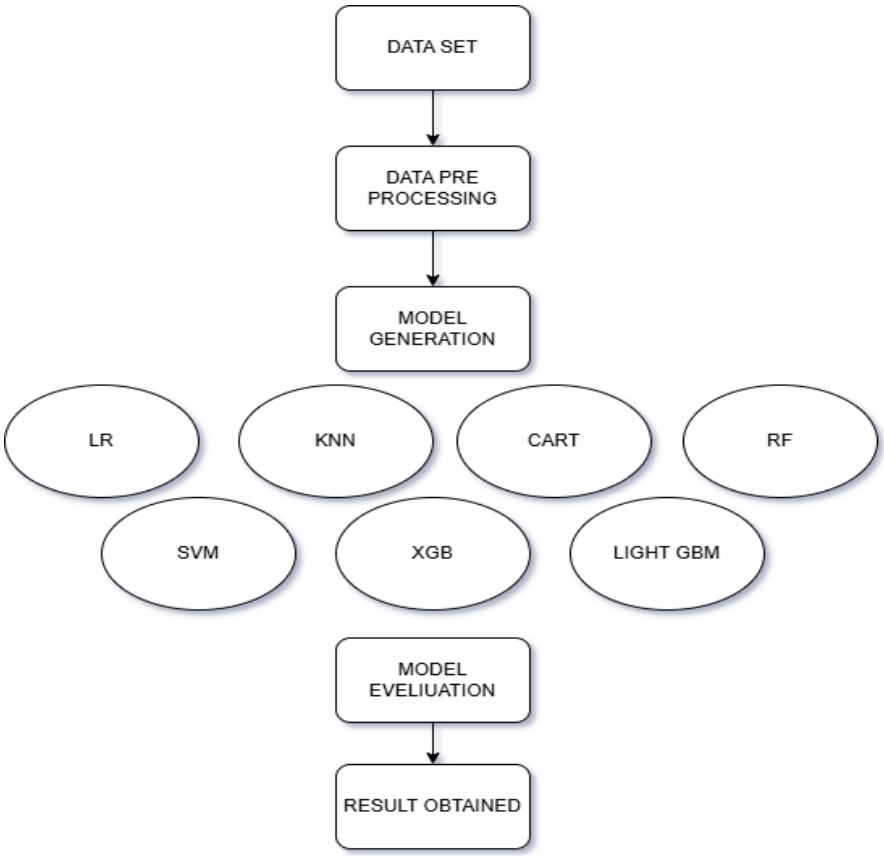
2. LITERATURE REVIEW

"K.VijiyaKumar et al. [11] proposed a system that implemented the Random Forest algorithm for prediction of diabetes with a goal to detect the disease early with high accuracy by using machine learning methods. This model showed better performance in predicting diabetes, showcasing that it could predict the disease effectively, efficiently, and most notably, in

an instant. Nonso Nnamoko et al. [13] studied prediction of diabetes onset by using an ensembler supervised learning method with five popular classifiers that use a meta-classifier to make use of their predictions. The study was compared with comparable studies that implemented the same dataset, revealing that their approach was of greater accuracy in prediction of diabetes onset. Tejas N. Joshi et al. [12] predicted diabetes using three supervised machine learning models: SVM, Logistic Regression, and ANN, suggesting an efficient method of early detection of diabetes. Deeraj Shetty et al. [15] worked on an Intelligent Diabetes Disease Prediction System based on data mining, with an analysis of databases of diabetes patients using algorithms such as Bayesian and K-Nearest Neighbor (KNN) to predict diabetes based on different attributes. Muhammad Azeem Sarwar et al. [10] carried out a study for prediction of diabetes based on six dissimilar studies that implemented different algorithms in health care with a comparison of their performance and accuracy. Comparing various techniques in the study, it was determined which algorithm was most suited for prediction of diabetes. Prediction of diabetes was made a major focus by researchers with an aim of training programs to identify if a patient is diabetic by employing relevant classifiers on the dataset. From past studies, it was mentioned that classification process".

3. METHODOLOGY

System of principles, techniques, and procedures employed in this study.



A. Dataset Discovery

The data were sourced from Kaggle, specifically the Pima Indian Diabetes Dataset. This dataset includes numerous attributes for 768 patients. The ninth attribute serves as the class variable for each data point, indicating outcomes of 0 and 1 for diabetes, which represent negative and positive results, respectively

S No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

Distribution of Diabetic patients:

We developed a model to predict diabetes; however, the dataset was somewhat imbalanced, with approximately 500 instances labeled as 0, indicating no diabetes, and 268 labeled as 1, indicating diabetes.

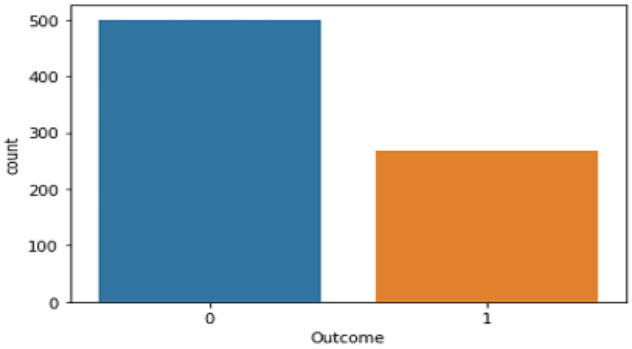


Figure 1: Ratio of Diabetic and Non Diabetic Patient

"**Data Preprocessing** - Preprocessing is an essential step. Most of the health-related datasets have missing values and such impurities that can affect data efficacy. To improve the quality of work and efficacy of the process of mining, data was preprocessed. Preprocessing is significant for getting accurate outcomes and effective predictions with Machine Learning Techniques on a dataset. Preprocessing was done in two steps for Pima Indian diabetes dataset.

1). Removal of missing values - All values with 0 were eliminated. It is not possible to have zero as a value, so these values were removed. By eliminating irrelevant features/instances, we obtain a feature subset, that is, feature subset selection, which decreases complexity in the data and increases speed of processing.

2). Data splitting - Once cleaning was completed, the data were normalized for training and testing purposes. During data splitting, the algorithm is trained on training dataset and the test dataset is reserved. This process of training produces a model based on logic, algorithms, and feature values in the training dataset. Normalization aims to put all attributes in the same range."

Apply Machine Learning- Several classification and ensemble techniques were implemented to predict diabetes. These techniques were implemented on the Pima Indian diabetes dataset. The main aim is to utilize Machine Learning Techniques to test performance, measure accuracy, and find features that play an important role in prediction. The Techniques are summarized below:

Performance Requirements

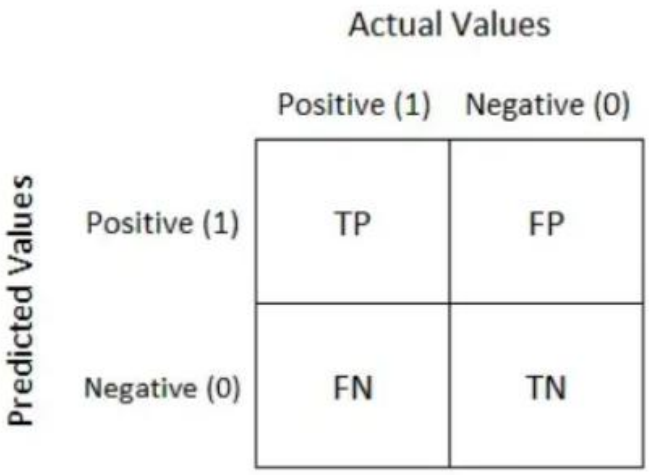
To understand how a model is performing, it's essential to evaluate it properly.

Key evaluation parameters include:

1. Confusion Matrix
2. Classification Report – which includes Precision, Recall, F1 Score, and Accuracy

Confusion Matrix

A confusion matrix is a performance evaluation tool used in classification and regression problems where the outcomes fall into two or more categories. It's a matrix that shows four possible combinations of predicted and actual values



Source :https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Elements of confusion matrix:

True Positive: Positive Prediction

True Negative:Negative Prediction

False Positive : (Type 1 Error)

False Negative :(Type 2 Error)

4. RESULT & DISCUSSION

```
fig = plt.figure(figsize=(15,10))
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
```

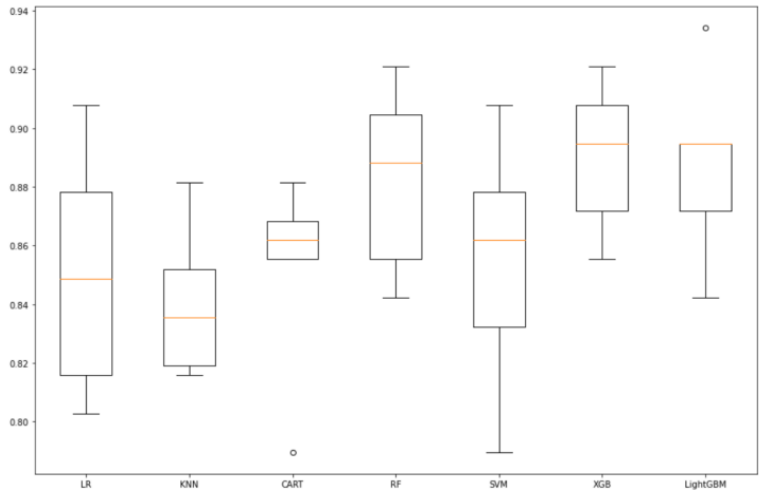
```
LR: 0.848684 (0.036866)
KNN: 0.840789 (0.023866)
CART: 0.857895 (0.024826)
RF: 0.881579 (0.026316)
SVM: 0.853947 (0.036488)
XGB: 0.890789 (0.020427)
LightGBM: 0.885526 (0.024298)
```

In the results, we obtained an accuracy of 89% in the XG Boost compared to the other models. Similar to logistic regression, we obtained an accuracy of 84 percent.

In LightGBM we received 88 percent accuracy

For the Support vector classifier, we achieved an accuracy of 85 percent. In Random Forest, we achieved an accuracy of 88 Percent.

Lowest we received is 84 percent in both KNN and Logistic Regression



In this project, we employed a range of machine learning algorithms to forecast the probability of diabetes. Diabetes is a long-term health condition that arises when the body cannot effectively control blood sugar levels. Without proper management, it can result in severe health issues, such as heart disease, kidney damage, and nerve damage. Detecting diabetes early and predicting it accurately are vital for preventing and managing these complications. We implemented various machine learning methods and trained our model with pertinent datasets to estimate the likelihood of an individual developing diabetes. Our objective was to attain the highest possible accuracy in diabetes prediction, which can aid in early intervention and enhance disease management.

REFERENCES

[1] Al-Zebari, A., & Sengur, A. (2019).” Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection.” 1–4. https://doi.org/10.1109/ubmyk48245.2019.8965542

[2] V. Jithendra, B. Jagadeesh, S. Kusuma, M. Madhusudhan, and R. M. Sai Mohit, “Diabetes Prediction using Machine Learning Techniques,” Journal of Artificial Intelligence and Capsule Networks, vol. 5, no. 2, pp. 190–206, Jun. 2023, doi: 10.36548/jaicn.2023.2.008.

[3] A. Choudhury and D. Gupta, “A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques,” springer singapore, 2018, pp. 67–78. doi: 10.1007/978-981-13-1280-9_6

[4] S. Mishra, P. Chaudhury, B. K. Mishra, and H. K. Tripathy, “An implementation of Feature ranking using Machine learning techniques for Diabetes disease prediction,” Mar. 2016, vol. 2, pp. 1–3. doi: 10.1145/2905055.2905100.

[5] M. J. Uddin et al., “A Comparison of Machine Learning Techniques for the Detection of Type-2 Diabetes Mellitus: Experiences from Bangladesh,” Information, vol. 14, no. 7, p. 376, Jul. 2023, doi: 10.3390/info14070376.

[6] M. Phongying and S. Hiriote, “Diabetes Classification Using Machine Learning Techniques,” Computation, vol. 11, no. 5, p. 96, May 2023, doi: 10.3390/computation11050096.

[7] M. A. Sarwar, M. A. Shah, N. Kamal, and W. Hamid, “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare,” Sep. 2018, pp. 1–6. doi: 10.23919/iconac.2018.8748992.

[8] A. Juneja, V. Kumar, S. Kaur, and S. Juneja, “Predicting Diabetes Mellitus With Machine Learning Techniques Using Multi-Criteria Decision Making,” International Journal of Information Retrieval Research, vol. 11, no. 2, pp. 38–52, Apr. 2021, doi: 10.4018/ijirr.2021040103.

[9] D. Y. Shin, J. K. Hyun, B. Lee, J. W. Park, and W. S. Yoo, “Prediction of Diabetic Sensorimotor Polyneuropathy Using Machine Learning Techniques.,” Journal of Clinical Medicine, vol. 10, no. 19, p. 4576, Oct. 2021, doi: 10.3390/jcm10194576.

[10] A. García-Domínguez et al., “Diabetes Detection Models in Mexican Patients by Combining Machine Learning Algorithms and Feature Selection Techniques for Clinical and Paraclinical Attributes: A Comparative Evaluation.,” Journal of Diabetes Research, vol. 2023, pp. 1–19, Jun. 2023, doi: 10.1155/2023/9713905.