JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 08 Issue 01
2024

Research Article
https://doi.org/10.56979/801/2024

Predictive Machine Learning Models for Early Diabetes Diagnosis: Enhancing Accuracy and Privacy with Federated Learning

Zill E Huma[1], Nimra Tariq[1*], and Shaharyar Zaidi[2]

[1]Department of Basic Sciences, The Superior University Lahore, Pakistan.
[2]Department of Computer Science, The Superior University Lahore, Pakistan.
Corresponding Author: Nimra Tariq. Email: nimra.tariq@superior.edu.pk

Received: June 23, 2024 **Accepted:** September 29, 2024

Abstract: Millions of people in the world are affected by diabetes, which is a serious chronic illness that have need of early detection for effective management and treatment. Even if they work well, typical techniques for detection are normally very expensive, time-consuming, and invasive. In this regard, machine learning (ML) has become a ground-breaking method for diabetes detection, providing an exact, effective, and non-invasive substitute. We are using the 27,690 instances and nine attributes of the Kaggle diabetes dataset, a combined machine learning model is presented in this paper. We utilized three machine learning algorithms: XG Boost (XGB), Naïve Bayes (NB), and K-Nearest Neighbours (KNN). XGB had the finest accuracy, coming in at 90%. To improve model performance while defending data confidentiality, our methodology includes data collection, pre-processing, training, testing, and parameter adjustment inside a united learning framework. The findings validate machine learning's marvellous potential for enhancing diabetes diagnosis, simplifying early intervention, and lowering medical expenses. Federated learning's integration further keeps patient privacy and data safety, giving it a solid option for extensive clinical use. This work opens the door for more accurate, effective, and accessible healthcare resolutions by highlighting the crucial implication and effectiveness of ML based diabetes prediction.

Key words: Machine Learning; K Nearest Neighbours; XG Boost; Accuracy; Precision; Recall

1. Introduction

The long-term health conditions known as diabetes causes the pancreas to become damaged making the body unable to produce insulin. The major factor sustaining blood glucose levels is insulin. Type 1, type 2 and gestational diabetes are the three different forms of the disease that need to be recognized. The pancreas generates little or no insulin in people is type 1 diabetes. One of the essential treatments for type 1 diabetes is insulin therapy. It typically affects children and young adults (less than 30 years old). Conversely, type 2 diabetes is typically brought on by insulin resistance and is more common in obese and older persons (over 65). Another issue is hyperglycaemia or gestational diabetes which happens during pregnancy [1]. Insulin is a hormone that controls glucose levels in the blood and enables it to enter cells to be used as an energy source. Elevated blood sugar levels result from glucose accumulation in the blood caused by reduced insulin activity. Diabetes can affect a person for a variety of reasons, including high blood pressure, abnormal cholesterol levels, being overweight and physical inactivity. It can also harm the skin, nerves and eyes. If uncontrolled, it can also result in kidney failure and this condition known as diabetic retinopathy [2]. The World Health Organization estimates that 422 million people worldwide have diabetes, and that figure is projected to rise to 693 million by 2045. Diabetes is directly responsible for 1.6 million deaths annually [3]. That is why tools for managing diabetes are necessary to monitor blood sugar,

insulin levels, and meal intake. Activity bands, glucose meters, continuous glucose monitors, and insulin pumps with sensors added are some of these instruments [4].

Worldwide health implications and the importance of early diabetes detection:

Early detection of disease at-risk populations allows health systems to better spend resources, ranking preventative care over costly treatments for more severe problems. By doing this, the total cost of healthcare is decreased while also improving patient outcomes. Further, early diagnosis can encourage better lifestyle choices and treatment compliance, which can lower long-term risks by giving patients more control over their health [5]. While effective, traditional techniques of detecting diabetes include different treatments and lab testing which are very time consuming and expensive. These testing procedures frequently call for several blood draws, fasting, and clinical appointments all of which can be painful and difficult for patients. It's also possible that conventional screening techniques miss early-stage diabetes detection. An individual may for example, have normal fasting glucose levels but unnoticeably endure postprandial (after eating) hyperglycaemia [6]. To lower the quantity of deaths from diabetes the early diabetes treatment and slowing the disease's progression is important and which is depend on early detection [7].

Difficulties and complexities of predicting diabetes through lab testing:

Although diabetes is a complex disease with variable test results, predicting diabetes from lab testing is difficult and time consuming. Individual differences in age, sex, origin, and heredity add to the complexity of early-stage diabetes, which frequently exhibits modest abnormalities [8]. Diabetes risk is greatly influenced by lifestyle variables and coexisting illnesses that are not directly assessed by lab testing. Furthermore, the diagnostic methods and biomarkers available today have their limits. Effective detection necessitates the combination of sophisticated analytics and patient data with lab results. Numerous flaws in lab testing can affect its validity and dependability [9]. The lab test has invasive measurements and some draw backs like:

1. Variability in the Outcomes
2. False Positives/Negatives results
3. Being Overbearing
4. Imperfect scope
5. Highly paid
6. Time consuming
7. Human and Technical Errors
8. Afraid of needles

There is basically two ways for measurement and detection of diabetes. An essential part of predicting and diagnosing diabetes is laboratory testing. Haemoglobin A1C (HbA1c), oral glucose tolerance tests, and fasting blood glucose levels are commonly included in these tests (OGTT). The HbA1c test, in particular, provides a detailed picture of long-term glucose control by measuring the average blood sugar levels over the previous two to three months [10]. Here we have some specific options for collecting data most common is using any website to get useful data or collect data from any hospital which required a lot of time because it has many different measurements of individual's body. On the other hand, data provided from website also has many features [11]. Required features for utilization of machine learning on both type of data is as in figure 2 which is given below:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Age

- Blood glucose level
- Beta cell function
- Urine test
- Systolic test
- Lipoprotein density
- 2-hour serum test

The relevant lifestyle history and health-related indicators are required for the samples such as: age, drinking alcohol, eating habits, blood glucose levels, smoking habits, types of jobs held, gender and the presence of diabetes in the family [12]. A number of tests are used to measure blood glucose levels while measuring diabetes in homes or labs settings using blood extraction. After an overnight fast, the fasting blood glucose test gauges sugar levels to provide a baseline for glucose management. For the purpose of identifying diabetes, tracking its course, and managing it, these blood extraction tests are essential [13].

Figure 1. Blood extraction (left) figure stick glucose monitoring (right)

Therefore, Automatic illness identification can be aided by an effective machine learning approach. The machine learning (ML), a branch of artificial intelligence that has advanced significantly in the field of medicine and prediction of many diseases especially it helps in diabetes. Diagnosing diabetes in medical patients is one of the most difficult and significant responsibilities in medicine [14]. In fact, healthcare systems have been using machine learning techniques for a long time to make decisions based on clinical data. It seems that machine learning is urgently needed in the current environment to facilitate automation with the fewest possible errors, therefore it is eliminating the necessity for human labour. The current approach to detecting diabetes involves laboratory tests like oral glucose tolerance and fasting blood glucose but it takes a lot of time to use this strategy [15]. So, it has been employed in this context by numerous researchers to diagnose diabetes. For the purpose of diabetes prediction, we have suggested the implement of machine learning algorithms like K Nearest Neighbours KNN, Naïve Bayes NB and XG Boost in a Hadoop-based cluster environment. Out of all the methods, the XGB algorithm yields the highest accuracy.

The worth of machine learning-based diabetes prediction:

An approach known as "predictive analysis" uses historical and present data to identify patterns and forecast future occurrences. It combines a range of machine learning algorithms, data mining strategies and statistical methodologies [16]. Predictive analysis can be used to make important judgments and forecasts based on healthcare data. Regression analysis and machine learning can be used in predictive analytics. The goals of predictive analytics are to improve clinical outcomes, optimize resources, improve patient care, and diagnose diseases as accurately as feasible [17]. Millions of individuals worldwide suffer from diabetes, which if left untreated can have serious significances such as nerve damage, cardiovascular disease and diabetic retinopathy. In spite of their efficiency, traditional diagnostic techniques have a number of drawbacks, such as being invasive, time-consuming, expensive, and perhaps producing inaccurate or inconsistent results. This emphasizes the need for more effective and precise techniques for

managing and detecting diabetes early. With machine learning, diabetes prediction can be transformed and many of the drawbacks of conventional diagnostic techniques can be addressed [18]. Large datasets may be quickly analysed by ML models, which can also spot patterns and danger indicators that traditional methods might overlook. The potential of machine learning to predict diabetes to enable early intervention is one of its main benefits. The demand for non-invasive, effective, and scalable diagnostic techniques is rising [19]. By using massive datasets to create prediction models that can identify people at risk of diabetes based on easily accessible characteristics, such as demographic data, lifestyle factors, and fundamental clinical tests, machine learning presents a possible answer [20]. The resulting table 1 summarizes the key characteristics of diabetes prediction using machine learning.

Table 1. Important aspects of diabetes prediction with machine learning

Characteristic	Explanation
Budget-Friendly	By ML early detection and prevention can lower healthcare expenses.
Never-ending Education	As these models are subjected to additional data over time, they can continuously enhance their predicted accuracy.
Assistance for medical professionals	Helps doctors diagnose patients, plan treatments, and keep track of their progress by offering insightful information and decision support.
Early prediction	Large datasets can be analysed by machine learning models to find patterns that point to early diabetic symptoms, allowing for earlier intervention.
Better Accuracy	Machine learning models perform better than conventional techniques, yielding predictions that are more accurate and minimizing false positives and negatives.
The ability to scale	Large-scale deployment of machine learning algorithms enables effective scaling of vast populations.

By taking preventative measures diabetes related complications can be prevented or greatly decreased leading to better patient outcomes and care. Moreover, it reduces the need for costly lab tests and frequent doctor visits, machine learning models can save healthcare expenses. They offer a scalable, effective, and non-invasive approach that can be especially helpful in situations with limited resources. More people at risk of diabetes will receive prompt and suitable care since ML models are reasonable and accessible to a wider community [21]. The development of modified treatment is further supported by the incorporation of machine learning in diabetes prediction. Machine learning algorithms have the potential to improve diabetes management techniques' accuracy and efficiency by modifying risk assessments and actions to the unique profiles of individual patients. Better health results are achieved by addressing each patient's specific demands through the use of a personalised strategy [22].

2. Literature Review

Machine learning models have become extremely effective tools in the recent past for diabetes prediction which providing less invasive and more accurate options than conventional diagnostic techniques. The effectiveness of many machines learning methods, including Decision Trees, Random Forests, Support Vector Machines and some deep learning models like Neural Networks also play important role in predicting diabetes and has been the subject of numerous research [23]. The used algorithms' capacity to evaluate big datasets and spot involved patterns that traditional statistical techniques can miss is a major benefit. For example, the extremely real gradient boosting algorithm XG Boost has established remarkable performance in multiple tests, with accuracy rates reaching up to 90% in predicting the course of diabetes. Although there are some slightly less accurate models like K-Nearest neighbours and Naïve Bayes also. In addition, wearable technology, which tracks physiological indicators like blood sugar levels, physical activity, and sleep patterns in real time, can deliver data to machine

learning models. This ongoing thought offers a more thorough picture of a person's health and permits the early identification of diabetes and its complications. Federated learning is an important expansion in machine learning, especially for the healthcare industry where data security and privacy are serious. A central server is frequently needed for the combination of data in traditional machine learning practises which can provide serious privacy particularly when handling complex health data. Federated learning can relieve these worries by letting ML models to be trained locally on regionalized devices or within distributed institutions. This method improves data privacy while also permitting the use of a variety of datasets, which supports the models' flexibility and generalizability. Federated learning has established major potential in the field of diabetes prediction. Research has shown that FL is capable of professionally compiling model updates from several healthcare providers without revealing raw patient data, protecting patient privacy. This approach has verified especially useful in circumstances where data is dispersed among numerous hospitals or areas with different privacy laws. Because the training data can represent a larger and more varied patient population, researchers have been able to construct diabetes prediction models that are completer and more accurate thanks to the integration of federated learning [24].

3. Methodology

In this study, each participant trained the model locally on their own data, keeping the data private and secure; instead of sharing the raw data, only model updates were sent to a central server, which aggregated these updates to improve the overall model. Federated learning is a devolved machine learning approach where multiple devices or organisations collaboratively train a collective model without replacing their local data. By doing this, a robust and comprehensive model was created while ensuring that sensitive data go on under the control of its original owner, improving privacy and security. There are four modules to our proposed method as shown in Figure 4 which work together to achieve our examination objective. We started by collecting the diabetes dataset and pre-processed it after that. After pre-handling we used to divide the dataset into train and test sets. Then the suggested algorithms are utilized on the research set to give beginning stage diabetes mellitus prediction after this the performance comparison is calculated using evaluation on the tested data set finally. In this section, we'll talk briefly about these stages.

Data collection: The Kaggle website provided the dataset used in this study. The dataset comprises of patients of different age groups also provided information with several factors: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Diabetes Pedigree Function, Age and BMI. Name and gender, on the other hand, will not be included in the classification. Process due to the fact that they are not a part of the calculation. Thus, some variables are utilized for testing and training from the dataset.

Attribute information: The used data is consisting of 27690 instances and 9 attributes from which 8 are used as input variables and last one is as output variable in which diabetes is diagnosis. The output variable consists of 0 and 1 values, yes category is considered as 1 and no category is considered as 0.

Table 2. The following table displays the description of the attributes of the dataset.

Sr. no.	Name of Attribute	Description
1.	Pregnancies	It shows the number of pregnancy occurs.
2.	Glucose	It tells the amount of glucose in blood.
3.	Blood Pressure	It records the pressure of blood.
4.	Skin Thickness	It is record of fat layers of skin.
5.	Insulin	It is measure of the level of insulin.
6.	BMI	It is the ratio of weight and height.

7.	Diabetes Pedigree Function	It counts the genetic possibility of diabetes based on family history.
8.	Age	It is the information about the age of patient.
9.	Diagnosis	It is presences or absence of diabetes (1 if yes and 0 otherwise).

Data cleaning and pre-processing: In machine learning, a step called data pre-processing is utilize where unusual values that could are removed from patterns that are wrong. Invalid records are deleted during this step. It includes selection of features that results in the most important attribute for the application from the dataset with raw input by and large available informational indexes for machine learning application. A portion of the records might contain copies, insignificant as well as irregular data. Therefore, data should be cleaned prior to processing, lessening the size of the dataset, information change for additional machine learning steps. The first 9 variables are regarded as input variables and readmitted is regarded as a class of output variables. The readmitted output variable contains values 0 and 1. Before the dataset is arbitrarily parted into 80 % information for training and the excess 20 % for testing, eliminating outliers, duplicates and remove null data from the dataset is essential.

Data splitting: To evaluate and validate the performance of model data splitting is main step of machine learning. It converts the dataset into two distinct subsets, commonly trading set to train the data and testing to check the final performance of model. The parameters of the model are adjusted and fine-tuned using this subset. It assists in minimizing overfitting and monitoring how well the model performs throughout the training phase. The model's performance is evaluated using this last subset once it has been trained. It offers an objective assessment of how well the model extends to fresh data. By dividing the data like this, we verify the model learns effectively, performs properly, and can be guaranteed to make reliable predictions.

Training: Giving training instances to the machine learning model that has to be skilled is a part of the machine learning algorithm's training process. Labelling the training instances with the output variable's known value is essential. The training instances help the machine learning algorithm to find the patterns for the training dataset and also plotting the graph of the input data variables to the output data variable whose value is to be predicted. The output of this step is a well-trained Machine Learning model that is capable to obtain the best designs from data set. **Testing:** In this step the subset of given dataset is inputted to the trained model for significant the precision of the algorithm. Measuring accuracy, precision, recall, and other performance measures is the major objective of data testing which aims to make sure the model consistently delivers predictions in real-world situations. It is a key step in ensuring that the machine learning model is reliable and effective before it is put into use. In this work three different machine learning classifiers namely K Nearest Neighbours (KNN), Naïve Bayes (NV) and XG Boost (XGB) are measured for tested and evaluated for their performance.

Parameter tuning: A machine learning model's parameter can be changed in order to improve its performance through a process known as parameter tuning. Effective parameter tuning can considerably increase a model's accuracy, reduce overfitting, and enhance its overall predictive potential. This procedure typically employs methods like grid search, random search, or more sophisticated techniques like Bayesian optimization. We can increase the accuracy and effectiveness of the model by fine-tuning them. The quick view of research stages for machine learning models are figure 4:

Algorithms: Algorithms for classification are essential to machine learning. Here are some useful algorithms that are utilized to achieve accuracy are:

K Nearest Neighbours (KNN): The k-nearest neighbours (KNN) algorithm is a supervised machine learning technique that is easy to use and can be applied to both regression and classification tasks. In order to predict a new data point, the algorithm locates the nearest neighbours of the closest data points in the training data set. The neighbour's value is selected from the data set. The closest neighbours are determined

by Euclidean distance, which is mainly defined in terms of the distance between two points P and Q. This distance is defined by the given equation:

$$d(P,Q) = \sum_{i=1}^{n} (P_i - Q_i)^2 \quad (1)$$

Where:

- Pi is i-th component of point P.
- Qi is i-th component of point Q.

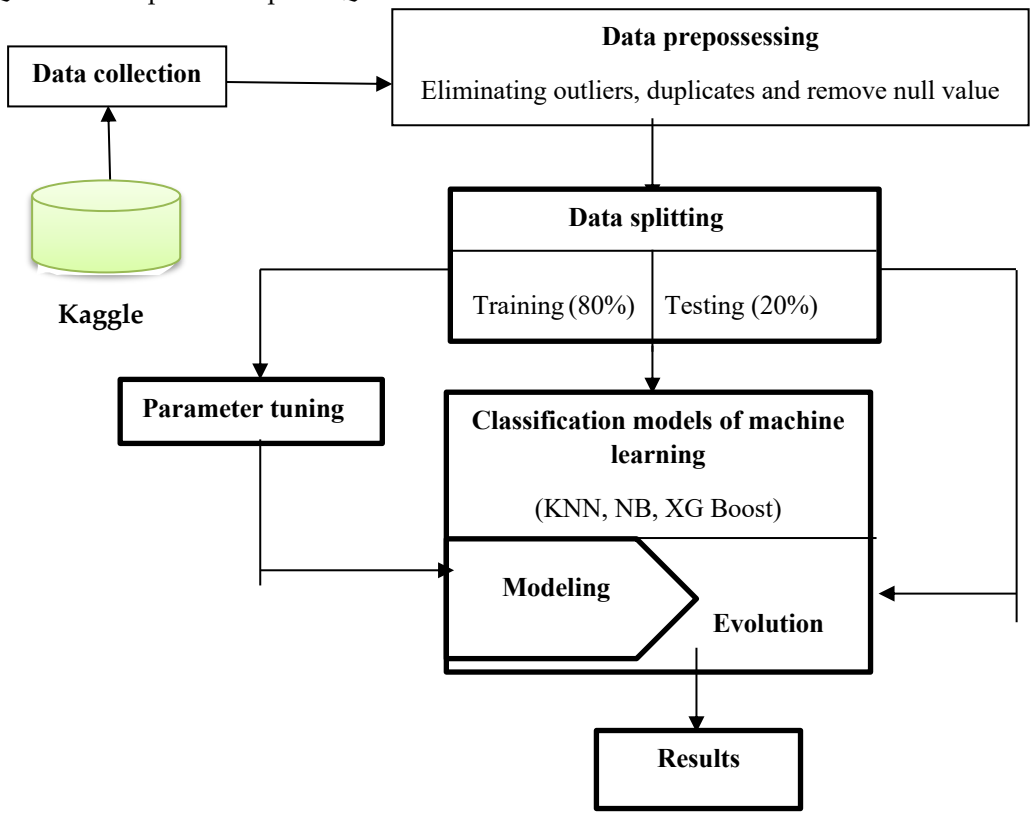

Figure 2. Research stages of machine learning models

Naïve Bayes: An ML algorithm for classification is called NB. Based on the idea that features are conditionally independent once the class label is known, it is based on the Bayes theorem. Because of how straightforward the assumption is, the method is fast and can be used to high-dimensional data. It can manage missing data and is resilient to features that are not relevant. Therefore, its notation is:

$$P(C|X) = \frac{P(C)\ P(C)}{P(X)} \quad (2)$$

Where:

- P (C|X) = target class's next probability.
- P (X|C) = predictor class's probability.
- P(C) = class C's probability being true.
- P(X) = next prior probability.

XG Boost: XG Boost (XGB) is a realistic distributed machine learning platform for scaling tree boosting methods. It is a practical and efficient way to implement the Gradient Boosted Trees technique. In a distributed environment designed for a quick parallel tree structure, the classier is both robust and well-configured. Tens of millions of samples and billions of distributed software samples that are scalable beyond are combined with a single node.

It is represented as:

$$\hat{y}i = \sum_{k=1}^{k} f_x(x_i) \quad (3)$$

Where:

- $\hat{y}i$ is the expected value for the i-th illustration.
- K is the quantity of trees.

- $f_x(x_i)$ is the calculation from the k-th tree to the i-th illustration.

Logistic Regression: A well-liked machine learning technique for binary classification issues is logistic regression. The anticipated values are mapped between 0 and 1 using the logistic function, usually known as the sigmoid function. The following equation is used in logistic regression:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_o+\beta_{11}+\beta_{22}+\cdots+\beta_{nn})}} \quad (4)$$

Where:

- P(y=1|X) is a possibility that given the input attributes X, the output Y will be 1 (positive class).
- β_o The point of intersection.
- $\beta_o + \beta_{11} + \beta_{22} + \cdots + \beta_{nn}$ Are the independent variables' coefficients.

4. Evolution and Results

The purpose of evaluating the prediction outcomes is to determine how well each ML classification technique performs. The confusion matrix will be used to assess and quantify the effectiveness of each machine learning technique in this study.

Accuracy: The measure of precise predictions is accuracy of the models. How accurately the model can work correctly is indicated by the accuracy metric. The accuracy value can be found as:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (5)$$

Different machine learning models show different accuracies among them KNN and XG Boost show more accurate output for the prediction of disease. The accuracy of the used models are as shown in the given graph:

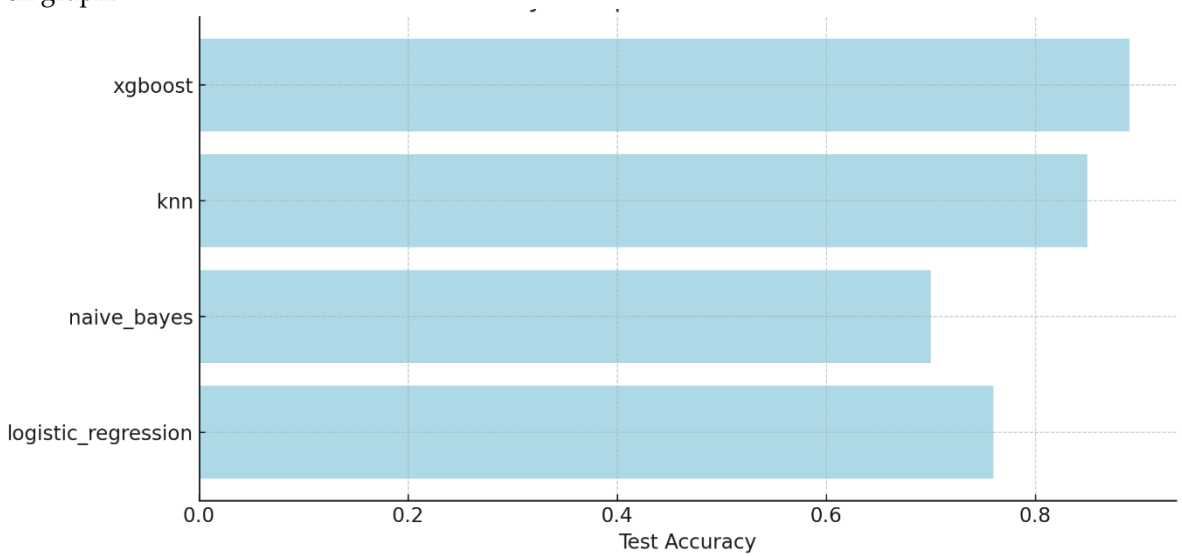

Figure 3. Test accuracy comparison of model

Precision: The parameter which indicates how accurate the outcomes are when only positive predictions are taken into account is known as precision. It only considers how well the algorithm performs while making positive predictions. Therefore, its calculation is performed by:

$$Precision = \frac{TP}{(TP+FP)} \quad (6)$$

Precision is the prediction of all true positive values among all positive values is basically known as precision. Precession of used models is:

Recall: Provides result on the proportion of True Positives that are appropriately classified during the examination. Its calculation will made by:

$$Recall = \frac{TP}{(TP+FN)} \quad (7)$$

Out of all actual positive instance it is the portion of true positive prediction. In simpler terms, recall measures the proportion of actual diabetes cases that the model correctly identifies. A high recall indicates

that the model is good at detecting most people who have diabetes, dropping the risk of missing out on true cases. However, high recall could come at the cost of a higher false positive rate, meaning that more people without diabetes may be wrongly classified as diabetic. Recall of the models are described with the help of given graph:

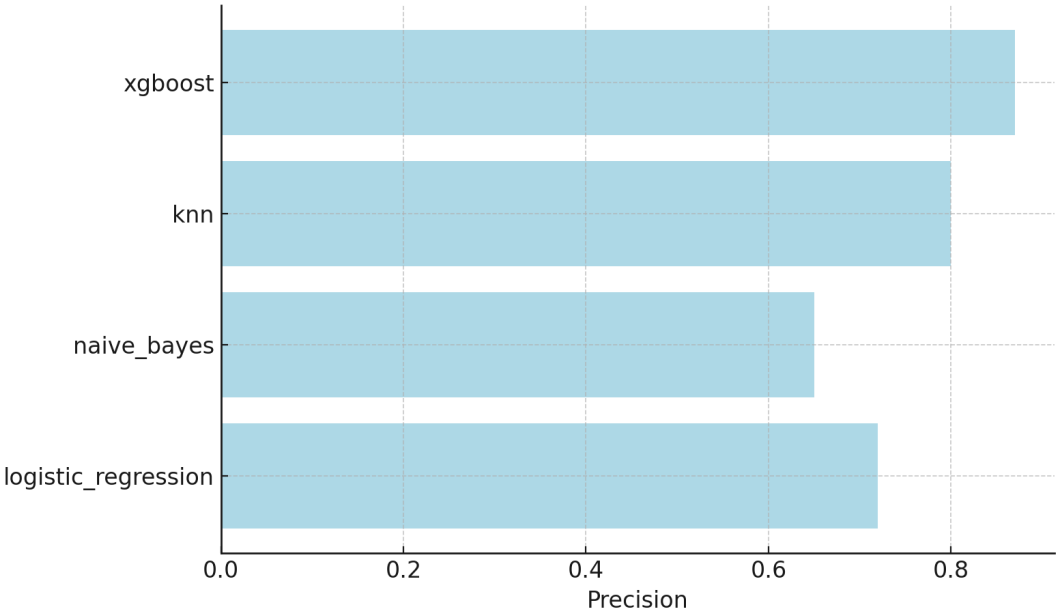

Figure 4. Precision comparison

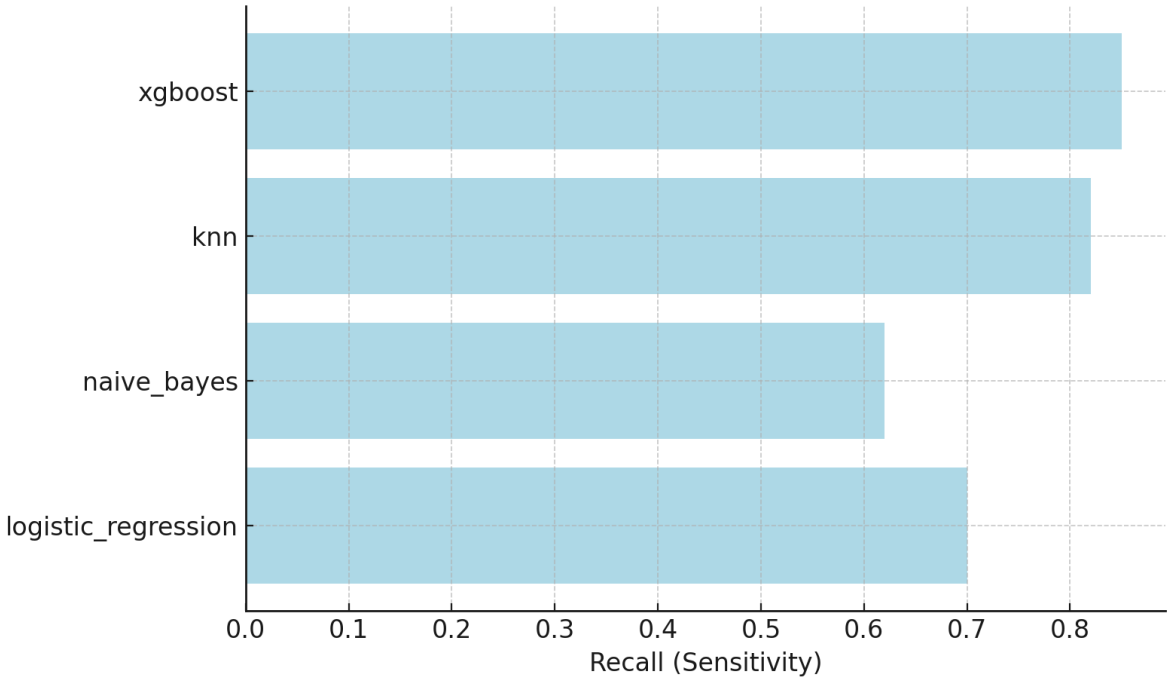

Figure 5. Recall comparison

F1 Score: The balance between precision and recall is F1 score, simply it is the harmonic mean between the two. It is defined as:

$$\text{F1 Score} = \frac{2 \times TP}{(2 \times TP + FN + FP)} \quad (8)$$

A model that accomplishes well in terms of precision and recall that is one that competently diagnoses diabetes while minimizing incorrect classifications has an F1 score that is near to 1.F1 score of KNN, NB, XGB and LG are:

Receiver operating characteristic: The area under the Receiver operating characteristic curve compares the overall ability of the models and distinguish between positive and negative values. It is best way to understand the discriminatory power of ML models. ROC is graphical tool to find the performance of models. It is a calculation metric to measure the total performance of classification methods with beginning values. The ROC curve of the applied models are exposed below:

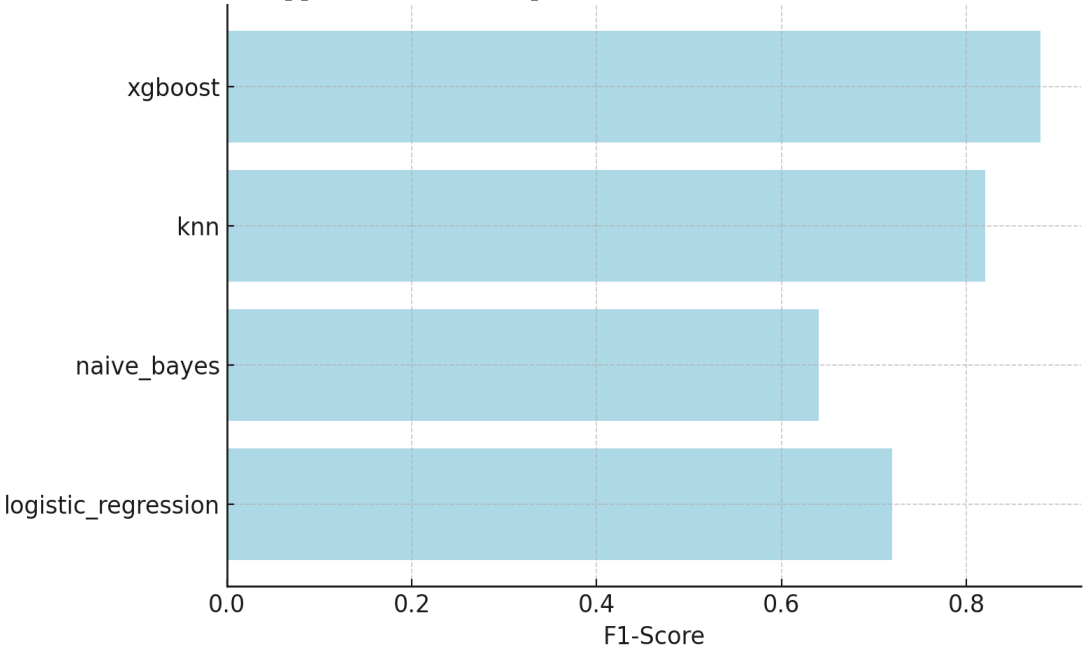

Figure 6. F-1 score comparison

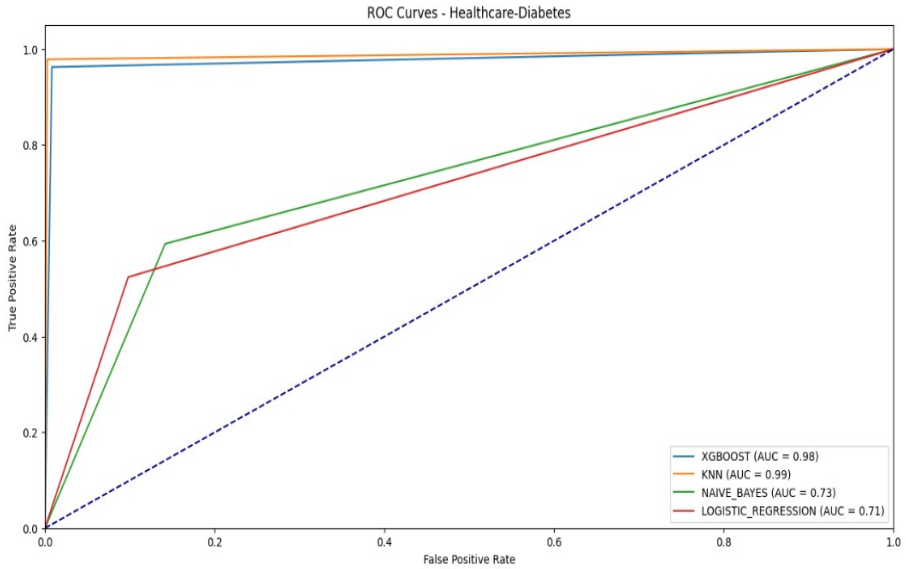

Figure 7. ROC curves of applied model

Area under curve: The area under the ROC curve is known as the AUC. It provides a single figure to express the classifier's total performance. The range of the AUC value is 0 to 1. In healthcare contexts, where the repercussions of false positives and false negatives are significant, a high AUC suggests that the model is good at class distinction.

Confusion matrix: A confusion matrix is a comparative table use to check the effectiveness of machine learning models. It displays how many of the model's predictions were true and false.

The four possible outcomes namely as:

- True positive (TP)
- True negative (TN)
- False positive (FP)
- False negative (FN)

In the meantime, the following metrics will be employed to assess how well each ML technique in this study performs. When true positive predictions should have a higher proportion and false positive predictions a lower percentage, this parameter becomes significant. For example, predicting a patient's test result for diabetes, AIDS or cancer for instance. False positive diagnoses in these situations will ruin patients' lives. Here is confusion matrix result for KNN, NB, XGB and LG.

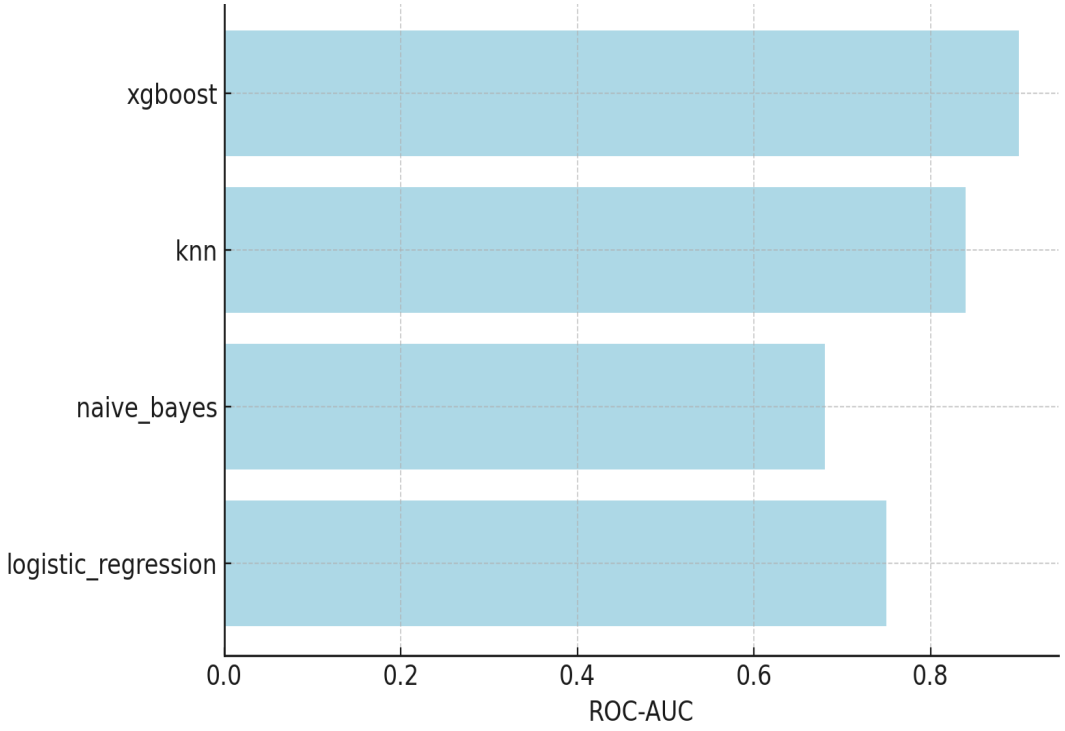

Figure 8. ROC-AUC for used models

K – Nearest neighbours: A confusion matrix for a KNN model is displayed in the image. It delivers an overview of how well the model classified two groups: 0 (not diabetic) and 1 (diabetic).The four quadrants of the matrix are:

- 366 cases of TN occurred when the model accurately predicted class 0 (non-diabetic) data.
- For FP One case in which a non-diabetic case was predicted by the model to be class 1 (diabetes).
- False Negatives FN: four cases in which a diabetes case was classified as class 0 (non-diabetic) by the model.
- True Positives TP: 183 cases in which class 1 (diabetic) predictions made by the model were accurate. The model performs admirably, with very few misclassifications and high accuracy for both classes.

Naïve Bayes: This image displays a confusion matrix for a Naive Bayes model that was used to classify the classes: 0 (non-diabetic) and 1 (diabetic). The matrix contains the following:

- True Negatives TN: 315 instances where the model correctly predicted class 0 (non-diabetic)
- False Positives FP: 52 instances where the model incorrectly predicted class 1 (diabetic) for non-diabetic individuals
- False Negatives FN: 76 instances where the model incorrectly predicted class 0 (non-diabetic) for diabetic individuals

True Positives (TP): 111 instances where the model correctly predicted class 1 (diabetic).

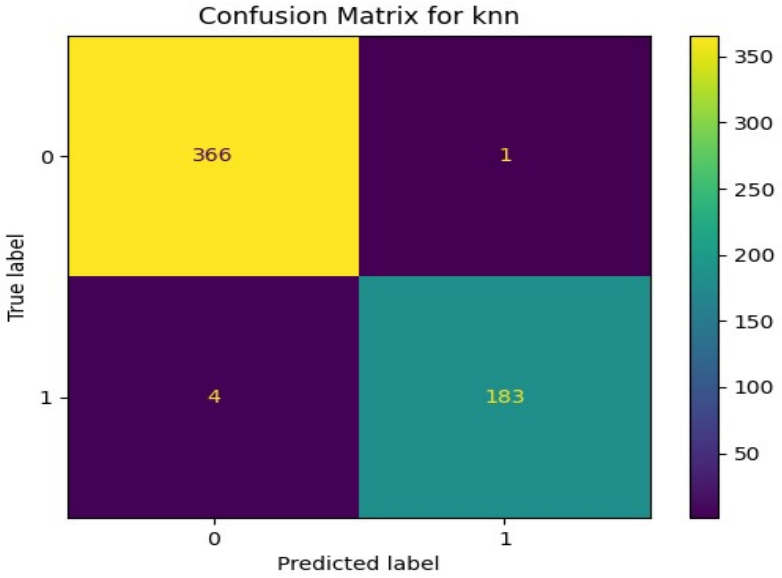

Figure 9. Confusion matrix for KNN

In contrast to the previous KNN confusion matrix, this Naive Bayes model appears to have more misclassifications, with higher false positives and false negatives, indicating lesser accuracy, particularly in detecting true diabetic cases.

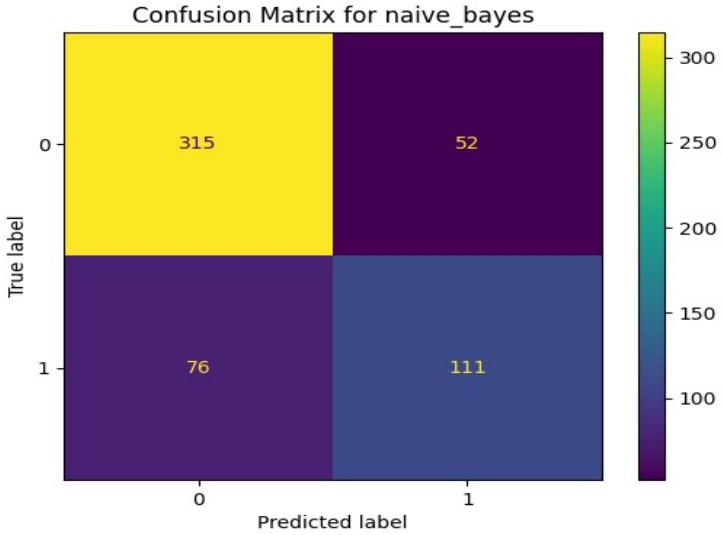

Figure 10. Confusion matrix for naïve-bayes

XG Boost: The picture presents an XG Boost model's confusion matrix, which is used to classify people into two groups: 0 (those without diabetes) and 1 (those with diabetes).The matrix of confusion displays:

- True Negatives TN: 364 cases in which class 0 (non-diabetic) was precisely predicted by the model.
- False Positives FP: three cases in which the model predicted class 1 (diabetic) for people who were not diabetics.
- False Negatives FN: 7 cases where the model predicted class 0 (non-diabetic) for those with diabetes when it should have said class 0 (diabetic).
- True Positives (TP): 180 cases were among them class 1 (diabetic) predictions made by the model were accurate.

With very few improper classifications, the XG Boost model performs commendably. The model is exceptionally accurate in predicting instances that are not diabetic as well as those that are, as seen by its very low false positive and false negative rates. This proves how strong this model is.

Logistic Regression: A logistic regression model is employed in the image to label two classes: 0 (non-diabetic) and 1 (diabetic). The confusion matrix for the model is displayed the following is revealed by the confusion matrix:

- True Negatives TN: 331 cases in which the model predicted class 0 (non-diabetic) variables accurately.
- False Positives FP: 36 cases in which the model predicted class 1 (diabetic) for people who were not diabetics in error.
- False Negatives FN: 89 cases in which the model misclassified diabetes people as class 0 (non-diabetic).
- True Positives TP: 98 cases in which the model predicted class 1 (diabetic) conditions accurately.

The performance of the logistic regression model is middling, and there are a significant amount of false positives and false negatives. Although it predicts non-diabetic patients quite well, its ability to accurately identify diabetes persons is less than that of models such as KNN and XGB.

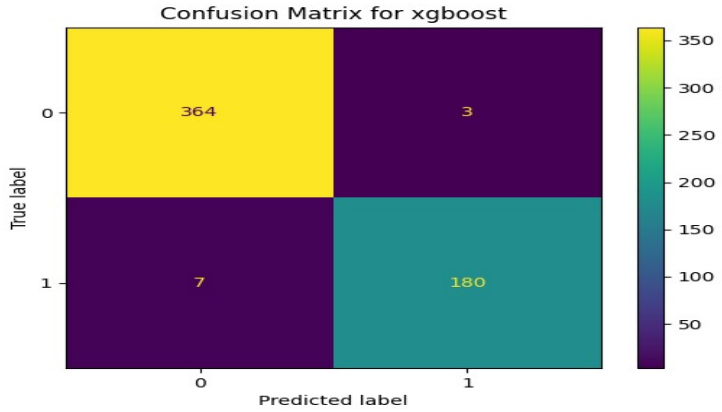

Figure 11. Confusion matrix for XG Boost

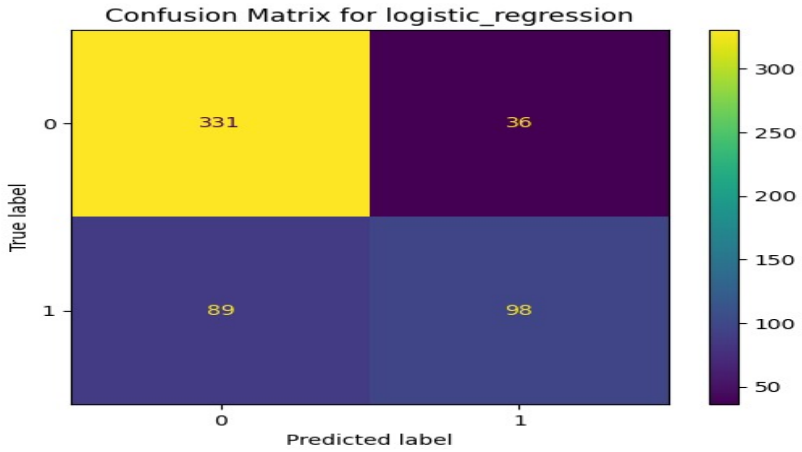

Figure 12. Confusion matrix for logistic regression

Feature correlation matrix: A graph that shows the correlation coefficients between features or variables of the dataset is called a feature correlation matrix. This matrix helps in understanding the correlations between the variables used to detect diabetes. This statistic illustrates the degree of linear relationship between two variables. It falls between -1 and 1:

- A complete positive linear relationship is indicated as +1.
- A complete negative linear relationship is indicated as -1.
- A value of 0 indicates a linear connection.

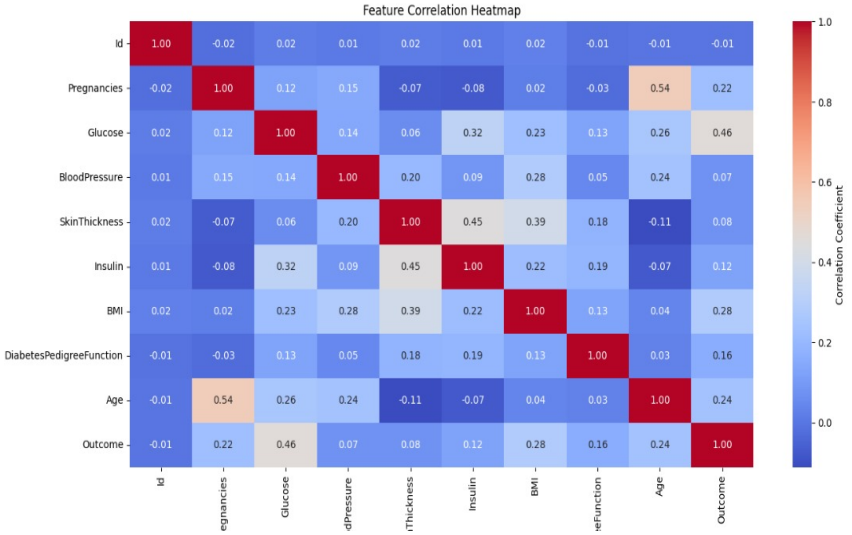

Figure 13. Feature Correlation Matrix of applicable models

	Id	Pregnancy	Glucose	Blood pressure	Skin thickness	Insulin	BMI	Diabetes Pedigree Function	Age	outcomes
Id	1.00	- 0.02	0.02	0.01	0.02	0.01	0.02	- 0.01	-0.01	- 0.01
Pregnancy	-0.02	1.00	0.12	0.15	- 0.07	- 0.08	0.02	- 0.03	0.54	0.22
Glucose	0.02	0.12	1.00	0.14	0.06	0.32	0.23	0.13	0.26	0.46
Blood pressure	0.01	0.15	0.14	1.00	0.20	0.09	0.28	0.05	0.24	0.07
Skin thickness	0.02	- 0.07	0.06	0.20	1.00	0.45	0.39	0.18	-0.11	0.08
Insulin	0.01	- 0.08	0.32	0.09	0.45	1.00	0.22	0.19	-0.07	0.12
BMI	0.02	0.02	0.23	0.28	0.39	0.22	1.00	0.13	0.04	0.28
Diabetes Pedigree Function	- 0.01	- 0.03	0.13	0.05	0.18	0.19	0.13	1.00	0.03	0.16
Age	- 0.01	0.54	0.26	0.24	- 0.11	- 0.07	0.04	0.03	1.00	0.24

Figure 14. This shows how a portion of a feature correlation matrix could seem.

5. Conclusion

Our study concludes with highlighting the considerable advantages of applying machine learning algorithms for the early detection and diagnosis of diabetes. The models we used, K-Nearest Neighbours (KNN), Naïve Bayes (NB), and XG Boost (XGB), demonstrated promising results, with XGB achieving the highest accuracy at 90%. This proposes that machine learning can increase the accuracy of diabetes predictions as well as provide a more real and cost-effective substitute for traditional lab tests. By enabling timely interventions and better disease management, these algorithms can facilitate earlier and more exact

identification of diabetes risk, which can ultimately lead to better patient outcomes and lessen the overall burden of diabetes on healthcare systems worldwide. This is a major progression in medical decision support systems, provided that important benefits in terms of accuracy, efficiency, and resource distribution. Additionally, by lowering the problem on healthcare providers, these models can help healthcare administrations diagnose patients more accurately and efficiently. After our examination we found that XG Boost provide best accuracy, precision recall and F1 score. So, it is best detective model for prediction of diabetes yes or no.

References

1. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(1-2), 1-10.
2. Akhtar, S., Nasir, J. A., Sarwar, A., Nasr, N., Javed, A., Majeed, R., & Billah, B. (2020). Prevalence of diabetes and pre-diabetes in Bangladesh: a systematic review and meta-analysis. BMJ open, 10(9), e036086.
3. Ayon, S. I., & Islam, M. M. (2019). Diabetes prediction: a deep learning approach. International Journal of Information Engineering and Electronic Business, 13(2), 21.
4. Karthi Vijay, D. (2022). The glycaemic control status and tuberculosis treatment outcomes among adult pulmonary tuberculosis patients with diabetes mellitus in Kollam and Pathanamthitta districts of Kerala (Doctoral dissertation, SCTIMST).
5. Khan F. A., Zeb K., AL-Rakhami M., Derhab A., and Bukhari S. A. C., Detection and prediction of diabetes using data mining: a comprehensive review, IEEE Access. (2021) 9, 43711–43735
6. Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., & Juwono, F. H. (2024). Diabetes detection based on machine learning and deep learning approaches. Multimedia Tools and Applications, 83(8), 24153-24185.
7. Kalyankar, G. D., Poojara, S. R., & Dharwadkar, N. V. (2017, February). Predictive analysis of diabetic patient data using machine learning and Hadoop. In 2017 international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC) (pp. 619-624). IEEE.
8. Nithya B, Ilango V. Predictive analytics in health care using machine learning tools and techniques. In2017 International Conference on Intelligent Computing and Control Systems (ICICCS) 2017 Jun 15 (pp. 492-499). IEEE.
9. Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019 Jan 1;165:292-9.
10. Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. Ict Express, 7(4), 432-439.
11. Alehegn, M., Joshi, R., & Mulay, P. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics, 118(9), 871-878.
12. Saru S, Subashree S. Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering. 2019 Apr 2;5(4).
13. Tan KR, Seng JJ, Kwan YH, Chen YJ, Zainudin SB, Loh DH, Liu N, Low LL. Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review. Journal of Diabetes Science and Technology. 2023 Mar;17(2):474-89.
14. Albahra, S. et al. Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data pre-processing and basic supervised concepts. Semin. Diagn. Pathol. 40, 71–87 (2023).
15. El-Bashbishy AE, El-Bakry HM. Pediatric diabetes prediction using deep learning. Scientific Reports. 2024 Feb 20;14(1):4206.
16. Ayon, S. I., & Islam, M. M. (2019). Diabetes prediction: a deep learning approach. International Journal of Information Engineering and Electronic Business, 13(2), 21.
17. Butwall M, Kumar S. A data mining approach for the diagnosis of diabetes mellitus using random forest classifier. International Journal of Computer Applications. 2015 Jan 1;120(8).
18. Alehegn, M., Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology, 4(10), 426-436.
19. Sonar, P., & Jaya Malini, K. (2019, March). Diabetes prediction using different machine learning approaches. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 367-371). IEEE.
20. Xue, J., Min, F., & Ma, F. (2020, November). Research on diabetes prediction method based on machine learning. In Journal of Physics: Conference Series (Vol. 1684, No. 1, p. 012062). IOP Publishing.
21. Paliwal, M., & Saraswat, P. (2022, October). Research on Diabetes Prediction Method Based on Machine Learning. In 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 415-419). IEEE.

22. Su, Y., Huang, C., Zhu, W., Lyu, X., & Ji, F. (2023). Multi-party diabetes mellitus risk prediction based on secure federated learning. Biomedical Signal Processing and Control, 85, 104881.
23. Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. (2023). Reviewing federated machine learning and its use in diseases prediction. Sensors, 23(4), 2112.
24. Islam, H., & Mosa, A. (2021). A federated mining approach on predicting diabetes-related complications: Demonstration using real-world clinical data. In AMIA Annual Symposium Proceedings (Vol. 2021, p. 556). American Medical Informatics Association.