Machine Learning based Early Predication and Detection of Diabetes Mellitus

Prosanjeet Sarkar
Electronics and Communication Engineering Department
Dr. A. P. J. Abdul Kalam Univeristy
Indore, India
sarkarprosanjeet08@gmail.com

Santosh Pawar
Electronics and Communication Engineering Department
Dr. A. P. J. Abdul Kalam Univeristy
Indore, India
spawarrkdf@gmail.com.

2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI) | 979-8-3503-3091-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAIIHI57871.2023.10489259

***Abstract*— According to a recent survey, diabetes mellitus has become the driving reason for death worldwide. Diabetes mellitus is critical disorders or illness that causes the imbalance of blood glucose levels due to the malfunction of the pancreas, which cannot make insulin hormones in the body; it is a silent and slow killer. It has a high risk for damage, malfunction and failure of human organs like the Kidney, heart, eye and nerve systems. There are several researches for the prediction and detection of Diabetes mellitus. The remarkable public health care infrastructure development for collecting critical and sensitive data. Many intriguing applications of machine learning algorithms are known for their ability to predict and identify diseases early on. The objective of this study is used to develop an effective early detection and prediction of diabetes mellitus using machine learning (ML) algorithms. The ML algorithms were performed on the Pima India Diabetes Dataset (PIDD) to develop the model. The proposed algorithm helps the practitioner to determine the seriousness of diabetes that is useful for the practitioner for the early recommendation of medicine, workout and treatment for the curing of disease. We developed many machine learning models in the experiment, such as Naïve Byase, KNN, Logistic Regression, SVM, decision tree, Random Forest, LightGBM and XGBoost to detect Type-II diabetes mellitus. The improved accuracy table shows the results of various models, and out of that, the XGBoost algorithm gives the highest accuracy of 89.07%.**

Keywords*— Machine learning, Artificial intelligent, Diabetes predication, Classification, Pima India Diabetes Dataset, Accuracy, Model

I. INTRODUCTION

The medical science industry uses Machine learning technology, a subclass of Artificial Intelligence (AI), for early prediction patterns of disease [1]. According to a WHO report, around 600 million people will get infected with diabetes worldwide. Diabetes mellitus is a common chronic disease and a severe challenge in developed and developing countries [2]. Nowadays, people are busy with their lifestyle and, due to this, unable to take care of their health. Most people are addicted to alcohol, smoking and unhealthy food, which causes an imbalance in the human diet [3]. The primary energy source for the human body's work is blood glucose, which comes from the carbohydrates people ingest. The pancreas is a significant organ in the human body that releases insulin, essential for controlling blood sugar levels [4].

In India, people enjoy many festivals by serving sweets and various fast foods and decreasing physical activity, which is one of the severe reasons for diabetes mellitus; rural areas do not have sufficient medical infrastructure to check and verify the diseases. If the human body has an imbalance of sugar levels, it causes Diabetes mellitus. Diabetes is one of the serious issues for Blindness, Kidney disease, heart strokes, nerve disease and lower limb amputation [5]. Diabetes has three types:

Type-1

Type-1 Diabetes results in abnormal glucose levels when the pancreas is cannot create insulin hormones . The actual cause, though, is still being investigated. According to researchers, it is hereditary and most frequently occurs in children [6]. The only solution is to give patients regular health check-ups, healthy diets and insulin doses.

Type -2

Type-2 Diabetes is called non–dependent insuline Diabetes mellitus [7]. Most individuals around the globe have Type-2 Diabetes, which occurs when the body produces less insulin or stops producing it. Researcher studies that Type-2 Diabetes is develop when human body weight increases by more than 20% to that ideal concerning height [8]. The solution is to increase insulin through a regular workout and healthy diet and avoid using Alcohol, Tobacco, etc.

Type 3- Gestational Diabetes

Gestational diabetes is a disorder that some women experience during pregnancy. The placenta resists the body's attempt to absorb insulin during pregnancy, which raises blood sugar levels [9]. This type of diabetes is not recognized before or after pregnancy. It is easily cured with standard medical care and a lifestyle change.

The medical science industry uses various electronic technologies for recording, storing and displaying distinct parts of the patient [10]. This disease-related data is saved for predictive analysis and risk management decisions. The predictive analysis uses machine learning approaches to increase the accuracy of disease diagnosis, improve patient care, make the most significant use of available resources, and enhance clinical results [11]. The paper aims to develop a machine learning classification model using these six algorithms: Naïve Byase (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), XGBOOST (XGB) and Light boost (LGB) that helps in early prediction and detection of diabetes with maximum accuracy.

This paper is structure as follows: Section II- gives a literature review of the early prediction of diabetes by using robust machine learning algorithms. Section III- proposed the methodology used in the development of different

models. Section IV presents the statistical evaluation method for performance measurement of the machine learning model. Section V explain the experiment work and result, section VI explain the conclusion and future work.

II. Related Work

Gaurav Tripathi et al. implemented the model for predicting diabetes disease using four ML classification algorithms: Linear Discriminant Analysis (LDA), K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and RF. The experiment was performed on the Pima India Diabetes database (PIDD) to make the relationship between feature and class. The confusion matrix measures precision, accuracy, F1-score and Recall. The result shows that the maximum accuracy obtained in the Random forest model is 87.66% [12].

Yashi Srivastava et al. used a classification model for the estimation using Microsoft Azure AI service; Microsoft Azure is a powerful deployment platform for machine learning algorithms for easy prediction of diabetes. This Microsoft Azure platform used the Pima India Diabetes Database (PIDD) to examine, achieving a maximum accuracy of 77.8% [13].

Amit Kumar et al. present a paper for classifying diabetes mellitus using a Hybrid model. Various ML algorithms, including Artificial Neural Networks (ANN), SVM, and KNN with hybrid models, were compared in this study. The experiment was performed on the Pima India Diabetes database (PIDD) to check the performance matrix. The hybrid model best performed among all the algorithms with 81.89% accuracy [14].

Umair Muneer Butt et al. have presented a long short-term memory (LSTM) and multilayer perceptron model for classifying diabetic mellitus. They used the Pima India Diabetes Database, collected from the UCI respiratory. The proposed model multilayer perceptron gives a maximum accuracy of 86.08%, and LSTM gives an accuracy of 87.26% [15].

Quan Zou et al. implemented the model using five machine learning classification algorithms, including the DT, RF, ANN, principal component analysis (PCA), and minimum redundancy maximum relevance (mRMR), to forecast the diabetes mellitus disease. The experiment was performed on the Pima India Diabetes Database (PIDD) and analyzed the performance of the machine learning classification algorithm. RF gives the best classification accuracy, 77.21 %, compared to another algorithm [16].

Christobel Y. A et al. present a paper describing a new Class-Wise K-Nearest Neighbor (CKNN) classification. For evaluating the CKNN algorithm and comparing numerous KNN performance metrics, researchers used the Pima India Diabetes Database. Comparing the suggested CKNN algorithm model to the simple KNN, it provides the highest accuracy of 78.16% [17].

Sisodia et al. present a paper on predicting diabetes mellitus disease using several ML algorithms such as NB, SVM and DT. The experiment using the Pima India Diabetes Database (PIDD) and evaluating the confusion matrix to examine accuracy results reveals that the Nave Bayes algorithm outperforms the other three with 76.30% accuracy [18]

III. Proposed Method

In this research work, diabetes classification and early prediction techniques are used in pattern reorganization to classify the dataset output into different classes. The dataset is PIDD to analyze Type-2 diabetes. Now, a more powerful classification algorithm supports day's machine learning technology. These algorithms are also used in the medical field in order to early detection and treatment of the disease. The machine learning models trained using the PIDD dataset to predict future outcomes. To measure the accuracy of the model, use mathematical statistical tools. In this work, our aim to develop a model that shows the early level of diabetes using the machine learning technique, which is used in the medical field by using various significant elements closely related to this disease. Figure 1 shows the flowchart of the proposed model used to build and train the models.

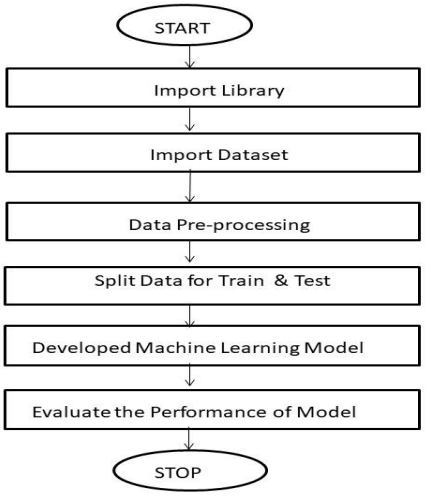

Fig 1. Machine learning flow chart for general model

A. Import Library

In the work, use the Python language and scikit learn/keras library. This library contains the entire classification algorithm that can be used to predicate diabetes.

B. Dataset

We used the University of California's Pima Indian Diabetes Database (PIDD) for this study project. This dataset has essential characteristics that are strongly tied to the illness. There are 500 negative and 268 positive diabetes prediction samples in this dataset, which includes 768 patient records. It contains 1 output class and 8 proper feature columns that can be utilized to determine if a patient has diabetes or not and five random rows of the dataset are shown in the table 1.

Table 1. Some example of PIDD dataset

P	G	B	S	I	BP	PDF	Ag	O
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.	0.351	31	0

					6			
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

C. *Data pre-processing*

During the data collection process, many sample feature values were missed or row duplicates because of fewer health workers and the repetition of jobs. If we use the same dataset, the machine learning algorithm cannot make any proper model with this missing or make a biased model with duplicate values. To solve the above problem, we need to pre-process data before training the model. Data pre-processing process steps is shown in figure 2.

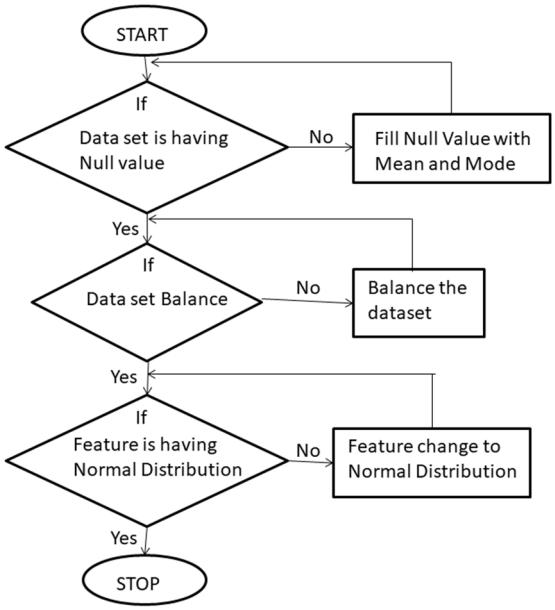

Fig 2. Flowchart for data pre-processing

1) *Check the missing value.*

In this method, we check whether the dataset has null and zero values. If the dataset has null or zero values, we can solve the missing value problem by using two steps. The given dataset has no missing value shown in table2.

a) Deleting the row:- This method can be used only when the dataset contains an extensive record, a few row deletion does not affect the data set records. Our record is small, so there are better approaches than the deletion of the record.

b) Imputation method:- In this procedure, the feature class mean or median is most likely to assign the missing or null value.

Table 2. Features having no missing or zero value

Feature	Number of missing value
Pregnancies (P)	0
Glucose (G)	0
BloodPressure (B)	0
SkinThickness (S)	0
Insulin (I)	0
BMI (BP)	0
DiabetesPedigreeFunction (PDF)	0
Age (Ag)	0

2) *Balance and unbalanced data set*

In the dataset, the repetition row or inequality of the dataset contains an unbalance in the classification area. Suppose the positive class of the data sample contain 90% and the negative class of the data sample 10%. In that case, there is no doubt of bias in the result towards the positive sample side, and the model will give high accuracy but mislead the prediction. To solve the above unbalanced problem, we used Random sampling and oversampling techniques. This work used the oversampling technique to solve the imbalance dataset problem.

3) *Normalization of features*

It is a crucial stage in the pre-processing process. The data sets contain various feature columns, and all are different scales. To make a good machine learning model and to get more accuracy, we need to do all the features on the same scale, which is called normalization. They generally used two techniques called z-score and min-max technique. In this work, we used the z-score technique. Figure 3 shows the random distribution of the dataset features.

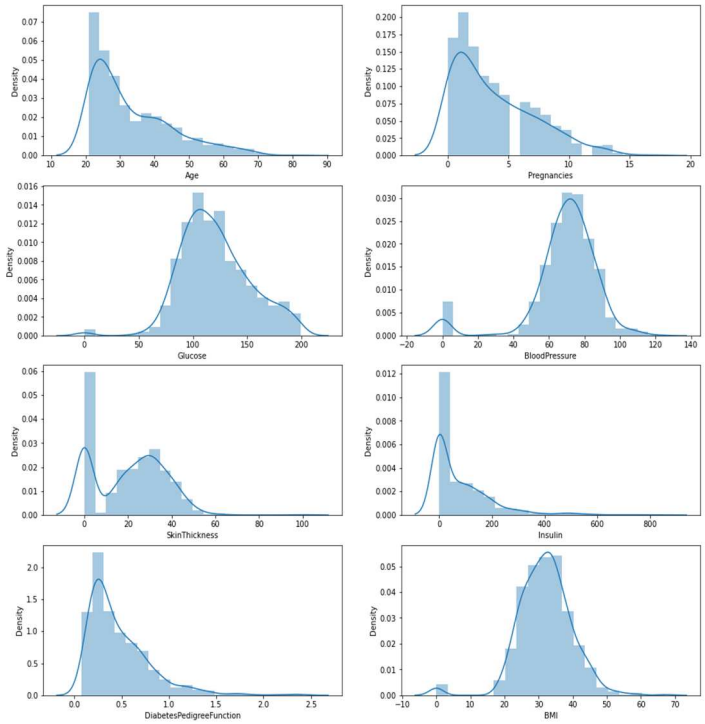

Fig 3. Random distribution of feature data sample

D. *Algorithm used for Modeling*

The following machine learning classification algorithm is used for developing the model described.

1) *Naive Bayes*

In order to solve classification problems, the Native Bayes method is frequently employed in supervised learning. It is the simplest and most effective. It works on the high-dimension training data set. This uses a probabilistic classifier as its guiding principle, formulating forecasts by assessing the probability of a characteristic's presence. It is formulated as equation (1).

$$P(A/B) = \frac{P(\frac{B}{A})P(A)}{P(B)} \quad (1)$$

2) K-Nearest Neighbour (KNN)

KNN is a supervised machine learning technique that may be applied to both regression and classification tasks. This method uses distance and similarity measurements to define the new feature class. The method is sometimes referred to as a lazy learning algorithm due to its practice of saving the training data instead of immediately learning from it. Then, when classifying data, it applies an action to the dataset—the following working steps for the KNN algorithm.

1. In the training phase, load the feature data and class sample of the training sample.
2. Choose the number of K of the neighbours included in the majority class as per Euclidean, Manhattan and Minkowski measuring distance. The distance calculation formula is represented in equations (2), (3) and (4).

- Euclidean distance

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(yi - xi)\text{^}2} \quad (2)$$

- Manhattan distance

$$\mathrm{d(x,y)} = \sum_{i=1}^{m} |yi - xi| \quad (3)$$

- Minkowski distance

$$\mathrm{d(x,y)} = \left(\sum_{i=1}^{m}|yi - xi|\right)^{\frac{1}{P}} \quad (4)$$

3. Place the new data sample in the category for which the neighbour number is farthest away.

3) Logistic Regression

The supervised learning classification problem of logistic regression is straightforward and popular. Logistic regression is used to address classification issues similarly to linear regression, but linear regression is used to determine the outcome of a regression. Based on the available dataset of the independent variable, the central idea uses the probability of an event taking place, such as engaging in voting or abstaining from voting.. The logistic function (sigmoid function) is represented by the following formulas: (5) and (6).

$$\mathrm{f(x)} = \frac{1}{1+e^{-y}} \quad (5)$$

$$y = b_0 + b_1 x \quad (6)$$

4) Decision Tree

The DT algorithm is commonly employed in both classification and regression tasks, however it is particularly preferred for addressing classification difficulties. It visualizes the dataset used to find every decision-making option based on the provided criteria. A decision tree shown in figure 4 consists of two nodes: the decision node and the leaf node. While the decision tree's leaf nodes serve as its output, decision nodes are utilized to create decisions with the potential for several branches.

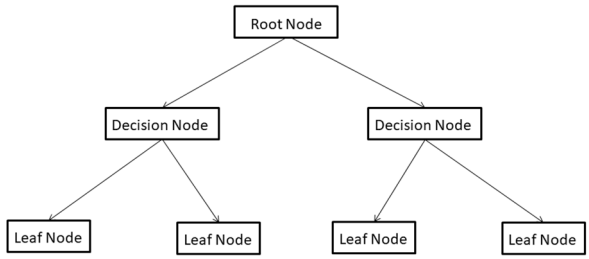

Fig 4. Decision Tree

5) Random Forest

The most effective supervised learning algorithm for classification and regression is random forest. Each decision tree has a significant variance, but when combined in parallel, the volatility is reduced, and the output depends on numerous decision trees rather than just one. RF is an ensemble method that combines different decision trees with bootstrapping, aggregating, and bagging approaches, shown in figure 5. Bagging approaches include randomly selecting rows and features from a dataset to create a sample dataset for each trained random tree.

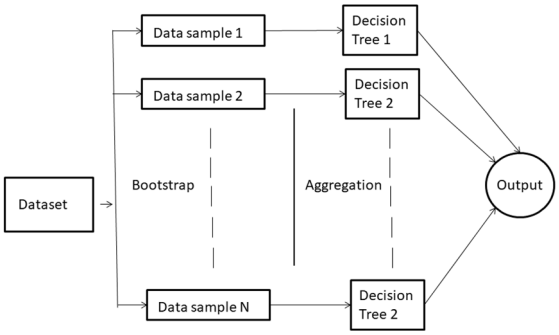

Fig 5. Random Forest tree

6) Support Vector Machine

The most widely used supervised machine learning approach for case regression and classification is called Support Vector Machine (SVM). To simplify the process of assigning a new sample to the relevant feature category, the best decision boundary to partition n-dimensional space into classes is created using the SVM technique. The most challenging task in SVM is selecting the best hyperplane in the dimension space, and the best hyperplane is the one with the most significant distance between data points from the two classes. Support vector refers to the plane closest to the feature sample of the classes. The unidentified sample point is classified according to the hyperplane that corresponds to one of the classes along the hyperplane

7) XGBoost

XGBoost, short for extreme gradient boosting, is a powerful machine learning technique that is used for both regression and classification tasks. The methodology employed is an ensemble approach, namely utilising ensemble bagging and boosting techniques. XGB employs

a gradient boosting approach that consists of three primary components: the loss function, weak learner, and additive module. Data over-fitting is a primary concern for robust algorithms, because gradient boosting, despite being a greedy strategy, has the potential to over-fit a large dataset. Regularization techniques are employed to mitigate over-fitting and enhance the performance of the algorithm.

8) LightGBM

LGBM is a more popular supervised ensemble machine learning algorithm for case regression and classification. LGBM stands for Light gradient boosted machine, which can handle the large size of data. It builds on the gradient-boosting approach by emphasizing cases with higher gradients and adding autonomous feature selection. This may lead to an increase in training efficiency and accuracy.

E. Statistical tools for evaluation of performance

The confusion matrix has shown in figure 6 used to measures the performance of the various classification models.

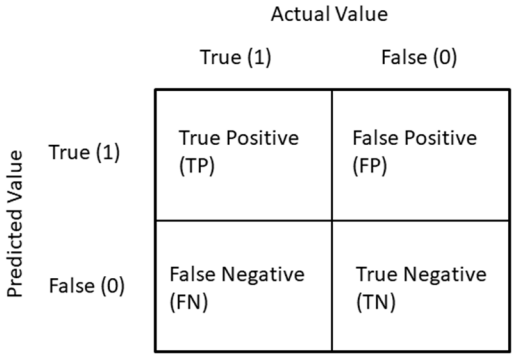

Fig 6. Confusion Matrix

The confusion matrix is calculated following the parameter.

1) Accuracy

Accuracy can be calculated by dividing the summation of TP and TN against the whole population. It measures the weather model correctly and represents the number of positive or negative samples in the dataset. The formula is shown in equation (7).

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (7)$$

2) Sensitivity

Sensitivity for the matrix can be calculated by taking the summation of the main diagonal to the true positive value. It is used to measure any disease in classification technique; it may be represented as equation (8).

$$Sensitivity = \frac{TP}{TP+FN} \quad (8)$$

3) Specificity

It measures the true negative rate and is given in equation (9).

$$Specificity = \frac{TN}{TN+FP} \quad (9)$$

4) Precision

It calculates the ratio of accurately predicted positive results to the total number of predicted positive results and is given in equation (10).

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

5) Error rate

It can be calculated by summation of FP and FN against the whole population. The formula is shown in equation (11).

$$Error\ rate = \frac{FP+FN}{TP+TN+FN+FP} \quad (11)$$

6) F1- Score

It measures the test accuracy, the F1 score obtained by dividing the true positive against the whole positive class value predicated. The range of F1 score is between [0-1]. It tells how precise your classification is and how robust it is. The mathematically is represented in equation (12).

$$F1 - Score = \frac{1}{\frac{1}{Precision}+\frac{1}{recall}} \quad (12)$$

IV. Experimental Work and Result

Eight practical classification algorithms are employed in this study to create a model that aids in the early-stage prediction of diabetes mellitus based on significant characteristics associated with this condition. The robust algorithm is NB, KNN, LR, DT, RF, SVM, XGB and LGB. This experiment used PIDD dataset source from the University of California. In the model analysis, we have to check the description of the data set; we have to find missing values in the features column, imbalance of data, mean, variance, etc. The missing value and class imbalance must be resolved to ensure that the model is as accurate as possible. The over-sampling method resolves the class imbalance issue by substituting the missing value for the features class mean.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	674.0	-5.929975e-17	1.000743	-1.197103	-0.864794	-0.200175	0.796754	2.458301
Glucose	674.0	9.306766e-17	1.000743	-2.571430	-0.691163	-0.161269	0.573744	2.727504
BloodPressure	674.0	-7.997230e-17	1.000743	-4.038964	-0.672784	0.000453	0.673689	4.208178
SkinThickness	674.0	-6.555916e-16	1.000743	-2.084639	-0.609921	-0.437759	0.738594	2.738367
Insulin	674.0	-3.986261e-17	1.000743	-1.757466	-0.494192	-0.494192	0.213166	3.878746
BMI	674.0	-6.254888e-16	1.000743	-2.148163	-0.721719	0.013853	0.610674	2.836554
DiabetesPedigreeFunction	674.0	1.545088e-16	1.000743	-1.296114	-0.734628	-0.320532	0.535734	3.609869
Age	674.0	-3.137945e-16	1.000743	-1.045402	-0.771593	-0.315246	0.597450	3.061728

Fig 7. Shows the description of dataset.

For the diabetes dataset, all the feature columns need to be on the same scale; for that we need to normalise the data and we followed z-score technique. Figure 7 shows the statistical value of the dataset help understand the dataset behaviour. Figure 8 shows the distribution of the diabetes dataset features and found that the columns are asymmetric distributions about the origin. To get a more accurate result, the entire feature column must be converted to the same scale (0-1) using normalization and the same graphically represented as normal distribution.

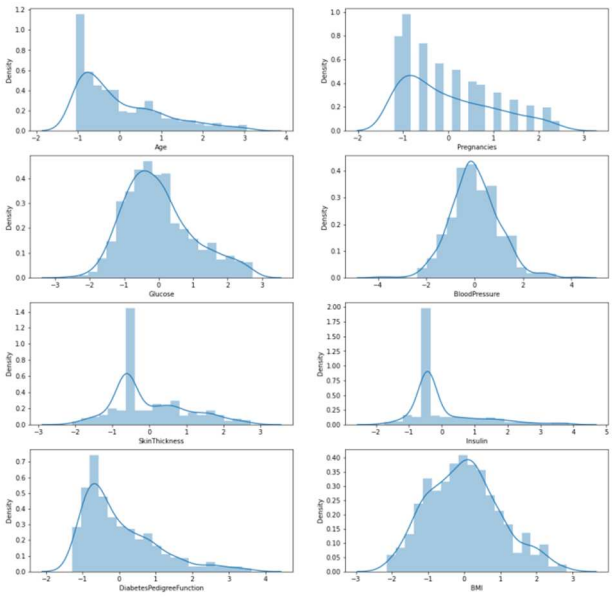

Fig 8. Normalization of Pima India Diabetes Dataset

It has one class label, 768 sample columns, and eight feature columns. '1' for a diabetes patient's positive outcome and '0' for a non-diabetic patient's negative outcome are shown in the output class. Figure 9 displays the binary classification of 768 output class, whereby 268 are diabetes patients and 500 are individuals without diabetes.

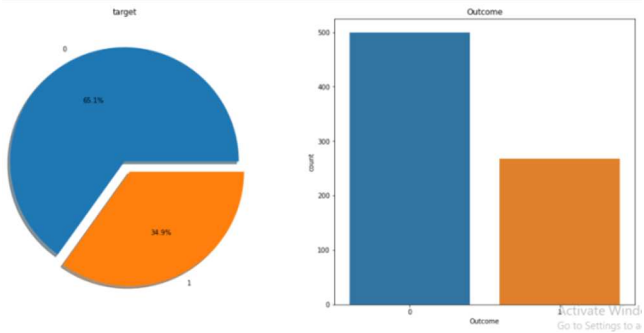

Fig 9. Output Class of PIDD dataset

Table 3. Comparison of the accuracy of the models

Classifier	Accuracy
NB	82.13%
LR	84.86%
KNN	84.07%
DT	80.12%
RF	88.15%
SVM	85.39%
XGB	89.07%
LGBM	88.28%

Table 3 shows the comparison of accuracy of the classification of classical models and found that XGBoost (XGB) outperformed to other classifier model with achieved accuracy of 89.07%, which is the best algorithm to predict diabetes on the Pima India Diabetes Dataset.

V. Conclusions & Future Work

Diabetes is a chronic condition that restricts everyday activities, lowers quality of life, and raises mortality risk in patients. Many researchers' has work on the Pima India diabetes dataset for classification of diabetic and non-diabetic. Medical researchers are proof that only early stage discovery of diabetes can control the dispersion of disease. In this paper, we have modelled various classification algorithms and applied them to the Pima India diabetes dataset to find the best-fit model to get maximum accuracy. The result shows that the XGBoost algorithm gives a maximum accuracy of 89.07%. In the future work, we can extend the accuracy of the model by large dataset and also develop android based application for the high accuracy model for the prediction of diabetes. We can also use machine learning algorithm for easily prediction of other diseases like cancer etc.

References

[1]. Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab, "Significance of machine learning in healthcare:

Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp. 58-73, June 2022.

[2]. Syed Amin Tabish, "Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century?," *International journal of health sciences*, vol. 1, no. 2, July 2007.

[3]. Aslam S, Martin A Holesh JE, *Physiology, Carbohydrates*. Treasure Island, San Francisco: StatPearls Publishing, 2023.

[4]. Julia, and Carol Coupland Hippisley-Cox, "Diabetes treatments and risk of amputation, blindness, severe kidney failure, hyperglycaemia, and hypoglycaemia: open cohort study in primary care," *BMJ (Clinical research ed.)*, vol. 352, March 2016.

[5]. Jane L et al Chiang, "Type 1 Diabetes in Children and Adolescents: A Position Statement by the American Diabetes Association," *Diabetes care*, vol. 41, no. 9, pp. 2026-2044, September 2018.

[6]. W Rodger, "Non-insulin-dependent (type II) diabetes mellitus," *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 145, no. 12, pp. 1571-1581, December 1991.

[7]. Ann Smith Barnes, "The epidemic of obesity and diabetes: trends and treatments," *Texas Heart Institute journal*, vol. 38, no. 2, pp. 142-144, 2011.

[8]. Mahdy H Quintanilla Rodriguez BS, *Gestational Diabetes*. Treasure Island, Florida: StatPearls Publishing, 2023.

[9]. Amit D et al Sonagra, "Normal pregnancy- a state of insulin resistance," *Journal of clinical and diagnostic research : JCDR*, vol. 8, no. 11, November 2014.

[10]. Kharrazi H, Lehmann H, et al Ehrenstein V, *btaining Data From Electronic Health Records*, 3rd ed., Leavy MB, Dreyer NA, Gliklich RE, Ed. United state: Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd Edition, Addendum 2 [Internet], 2019.

[11]. Thomas, and Ravi Kalakota Davenport, "The potential for artificial intelligence in healthcare," *Future healthcare journal* , vol. 6, no. 2, pp. 94-98, June 2019.

[12]. Gaurav & Kumar, Rakesh Tripathi, "Early Prediction of Diabetes Mellitus Using Machine Learning," in *International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, Noida, 2020, pp. 1009-1014.

[13]. P. Khanna and S. Kumar Y. Srivastava, "Estimation of Gestational Diabetes Mellitus using Azure AI Services," in *Amity International Conference on Artificial Intelligence (AICAI)*, Dubai, 2019, pp. 321-326

[14]. Amit Kumar & Agrawal, Pragati Dewangan, "Classififcation of Diabetes Mellitus Using Machine Learning Techniques," *International Journal of Engineering and Applied Science*, vol. 2, no. 5, pp. 145-148, May 2015.

[15]. Umair Muneer et al Butt, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *Journal of healthcare engineering*, vol. 2021, September 2021.

[16]. Quan et al Zou, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Frontiers in genetics*, vol. 9, no. 515, November 2018.

[17]. Christobe Y. Angelinel and Sivaprakasam P., "A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset," *International Journal of Engineering and Advanced Technology*, vol. 2, pp. 396-400, February 2013.

[18]. Deepti and Sisodia, Dilip Singh Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578-1585, 2018.