Diabetes Prediction Using Machine Learning With Smart Web Application

C. Mehavarnan[1], K. Lokaprakash[2], S. Sadath[3], Mr.K. Sathyaseelan[4]
[1, 2, 3, 4] Dept of Information Techonology
[1, 2, 3, 4] Sri Ramakrishna Institute of Technology

Abstract- *Diabetes has become one of humanity's most serious concerns. Diabetes healthcare monitoring services are incredibly important nowadays since remote healthcare monitoring is far more effective than physically visiting hospitals and standing in queues. When a patient has severe chronic diabetes and spends time waiting in queue, anything terrible can happen at any time. There are various traditional procedures to diagnosing diabetes and forecasting preconditions in diabetic patients that are substantially different from computerised methods. Diabetes is caused by a significant increase in the amount of glucose in a blood. Diabetes may cause other issues including heart disease, kidney damage, and blindness. So, in this project, we propose a diabetes disease prediction system that uses machine learning algorithms to forecast diabetes risk, as well as a web framework called Flask to create a user-friendly interface for users to input their data and obtain a risk assessment. To predict diabetes risk, we employed a dataset of 768 patients and trained multiple machine learning algorithms, including logistic regression, decision tree, random forest, support vector machines, K closest neighbour, and XG boost. We evaluated the models using several characteristics, including accuracy, precision, recall, and F1 score. After deciding on the best model, we used Flask to incorporate it into a user-friendly online application. The application's interface was meant to be simple and easy to use, allowing users to enter their data and receive their diabetes risk assessment quickly and simply. Overall, this research shows the power of machine learning and web frameworks in creating a diabetes prediction system that can enhance diabetes prevention and care.*

Keywords- Diabetic prediction, Machine learning, Random forest, Support vector machine, K-nearest neighbor,Prediction.

I. INTRODUCTION

Diabetes is a fairly widespread condition that affects people all around the world. Diabetes is estimated to affect 422 million people globally, according to the World Health Organization(WHO). It might seriously harm a person's health and lead to problems with the heart, kidneys, blood pressure, eyes, and other body organs.

If this condition is caught early on, People could live longer and better lives. Glucose can enter the bloodstream from food thanks to the pancreas production of the hormone insulin in the body. When the pancreas isn't producing enough of that hormone, diabetes results. Comas, failure of the kidneys and the eyes, pathological joint damage, loss of weight, and pathogenic immune responses are all possible complications of diabetes. The subtypes of diabetes are type 1 & type 2, respectively. We'll make advantage of machine learning[ML] methods to create a diabetes diagnostic model for this project, and we'll include it into the Flask web framework. Using Python, Flask is a widely used web framework for developing web applications. With Flask, we can build a web interface where users can input their information, and the model can predict their risk of developing diabetes.

By using machine learning and Flask, we can create an interactive and user-friendly diabetes prediction tool that can help individuals take proactive steps towards managing their health. This project has the potential to make a significant impact on diabetes prevention and management by providing a convenient and accessible tool for people to monitor their health and make informed decisions about their lifestyle choices.

II. OVERVIEW OF MACHINE LEARNING

Machine learning is a subfield of artificial intelligence (AI) that entails the development of statistical models and algorithms that enable computer systems to learn from data and enhance their performance over time without explicitly being big datasets using statistical methods to find patterns or correlations between variables. These patterns are then used to new data to produce predictions or judgements. The taxonomy of machine learning includes three fundamental subcategories: supervised learning, unsupervised learning, and reinforcement learning. Unlabeled data are used to train a model in supervised learning when there is no clearly defined desired outcome.

The process of training a model to make decisions based on comments on the form of rewards or penalties for

each action performed is known as reinforcement learning. Machine learning is used in a broad range of industries, including marketing, finance, healthcare, and social media. It is used for a variety of activities, including automated vehicles, fraud detection, natural language processing, picture and audio recognition, and recommendation systems.

III. RELATED WORKS

Mitushi Soni (2020) presented Diabetes Prediction using ML Techniques using random forest algorithm(RF) in which she was able to predict diabetes of paitents[1]. Her experimental results were to help asst health care to take early prediction and make early decision to cure diabetes and save humans life.

Aishwarya Mujumdar (2019) in her study, We applied different machine learning algorithms to the dataset and performed classification using various methods, among which Logistic Regression yielded the highest accuracy [2]. Here comparison is made betweeen the accuracies of machine learning classifications were evaluated using two distinct datasets.

Quan Zou, Kaiyang , Dehui , Ying Ju and Hua Tang(2018) presented Using machine learning techniques to forecast the onset of Diabetes Mellitus. in which they applied random forest, decision tree and neural network alogorithms on PIMA dataset and Luzhou dataset [3]. Here they obtained predictions on comparison of both dataset and algorithms where random forest showed high accuracy. Due to the data, they couldn't predict the type of diabetes.

Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, (2018) Using Machine Learning to Analyze the Importance of Features for Predicting Diabetes[4].In which they applied logistic regression, Support vector Classifier, Random forest .The random forest yielded the highest accuracy to predict the feature for diabetes prediction.

K.VijiyaKumar, B.Lavanya, I.Nirmala, S, Reworded[5], Using Random Forest Algorithm to Predict Diabetes because of campare to other machine learning techniques the random forest algorithm predict the result for diabetes disease. It predict the result effectively.

Md. Faisal Faruque, Iqbal. Sarker, presented using machine learning technique to predict the diabetes mellitus[6].The machine learning techniques are logistic regression, k-nearest neighbor, Naïve bayes to train and predict the mellitus.The LR is more efficient than other techniques.To predict the diabetes early and quickly.

N. Joshi et .[7]The project "Diabetes Prediction Using Machine Learning Techniques" focuses on before detection of diabetes disease by employing three distinct supervised machine learning techniques: SVM, Logistic Regression, and ANN. The project offers an efficient technique for forecasting the onset of diabetes.

Nonso Nnamoko et al. [8] The study "Predicting Diabetes Onset: An Ensemble Supervised Learning Approach" utilized five commonly used classifiers for the group and combined their outputs with a meta-classifier. The findings were presented and compared to similar studies in the literature that used the similar dataset. The study demonstrated that the proposed approach results in higher accuracy for diabetes onset prediction.

Deeraj Shetty et al. [9] The study proposed an Intelligent Diabetes Prediction System that utilizes data mining techniques to analyze a diabetes patient's database and provide an assessment of the disease. The system employs algorithms such as Bayesian and KNN to analyse several diabetes characteristics to help in illness prediction.

Soumavadeen Manna and Mainak Adhikar are made a diabetes disease prediction using the algorithm of logistic regression and use the PIMA diabetes dataset to made a predictive model. Knowing one's risk of future Diabetes diagnosis, a person can promptly access information on how to control the illness by addressing the primary causative factors[10]. This offers the advantage of effectively controlling Diabetes. It has an accuracy of 78.43%. It may be a very useful model for each and every human being.

Mohamed Rady, Kareem Moussa, and Mahmoud Mostafa used Decision Tree, Random Forest, Nave Bayes, and SVM to Pima Indians Diabetes Data Set [11]. Providing a technique of detection through symptoms that the patient may notice can lead the patient to seek medical treatment more quickly, allowing the patient to be accurately diagnosed and treated.It has a 75.31% accuracy. Machine learning techniques have been implemented in the medical diagnosis system since they have shown to be more accurate in diagnosis, effective in treatments, and cost-effective.

Aiswarya Iyer, Jeyalatha, and Ronak Sumbaly[12], are Diagnosis a Diabetes predcition Using Classification Mining Methods of decision tree and naïve bayes algorithms to predict the diabetes early,it makes to timely treatment the patient easily.

IV. METHODOLOGY

The diabetes dataset is imported along with the required libraries. After pre-processing the data to remove any missing values, the dataset is divided into a Training set and a Test set using an 80/20 ratio.The choice of several machine learning algorithms, including K-Nearest Neighbour, Support Vector classifier(SVC), Decision Tree(DT), Logistic Regression(LR), Random Forest(RF), and XGBoost(XGB), comes in the next phase. Based on the training data, a classifier model is created for each of these algorithms. Each machine learning algorithm's classifier model is tested using the test set, and then the performance of each classifier is compared using a variety of metrics.The highest performing algorithm is chosen in accordance with the findings of this extensive investigation.

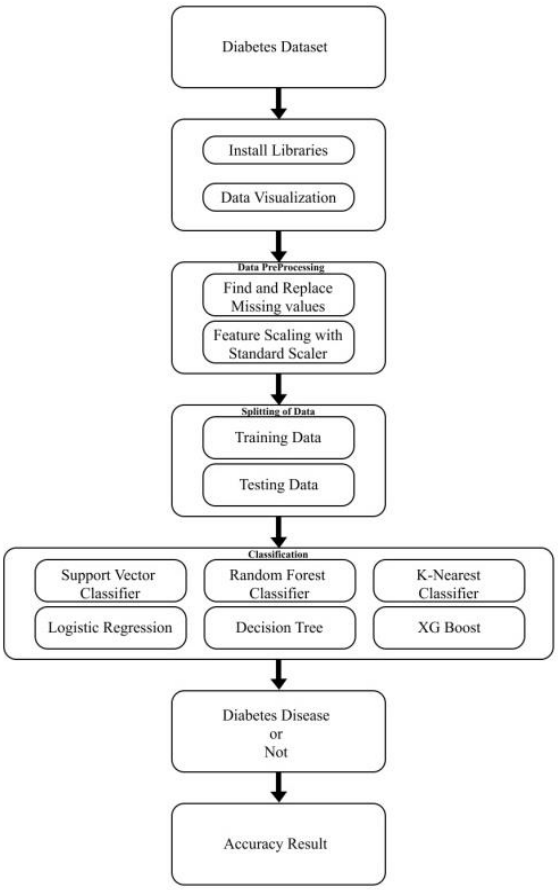

Fig. 1. Flow Diagram

A. Dataset Collection

Obtaining a dataset with pertinent traits and related target variables is important in order to forecast diabetes. The required information was gathered using the Pima Indian Diabetes Dataset, which is accessible in the UCI repository. The dataset comprises 768 patient records, each of which has a number of variables including Age, Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI.The last property of each data point, the class variable, describes whether or not the subject has diabetes. A result with a value of 0 indicates that the person does not have diabetes, whereas a result with a value of 1 indicates that the person does have diabetes. Out of the 768 dataset's data points, 500 are labelled as 0 (negative) and 268 are labelled as 1 (positive).

B. Data Preprocessing

Preprocessing data is an essential step, particularly in healthcare-related data, as it often contains missing values and other contaminants that may affect the quality and effectiveness of the data mining process. Therefore, data preprocessing is performed to enhance the quality and effectiveness of the data obtained after the mining process. Performing data preprocessing is crucial to ensure accurate results and successful predictions using ML methods applied on the dataset. This is especially important for healthcare-related data, which often contains missing values and other impurities. In the case of the Pima Indian diabetes dataset, data preprocessing needs to be performed in two steps to achieve optimal results.

- Missing Values Removal- Instances with a value of zero are removed as they are not possible, resulting in the elimination of irrelevant features/instances and the creation of a feature subset. This process, known as feature subset selection, reduces the dimensionality of the data and improves the processing speed.
- Splitting of data- Prior to the data is cleaned, then normalised before being used to train and test the model. Instead of using the test dataset for training, the algorithm is trained on the training dataset.. Based on the logic, methods, and feature values found in the training data, the model is trained. All of the characteristics are normalised to be on the same scale.

C. Feature Extraction

The process of feature extraction in machine learning is choosing and converting the most pertinent information from raw data to produce a set of features that can be utilised as model input. This process is essential in many applications where the raw data is too complex or too large to be used directly by the model. The accuracy of a model can be improved, the amount of data required for training can be reduced, and the overall process can be sped up by selecting and transforming the most important information using feature extraction. The feature extraction techniques employed are

contingent on the type of data being analyzed and the specific task that needs to be accomplished.

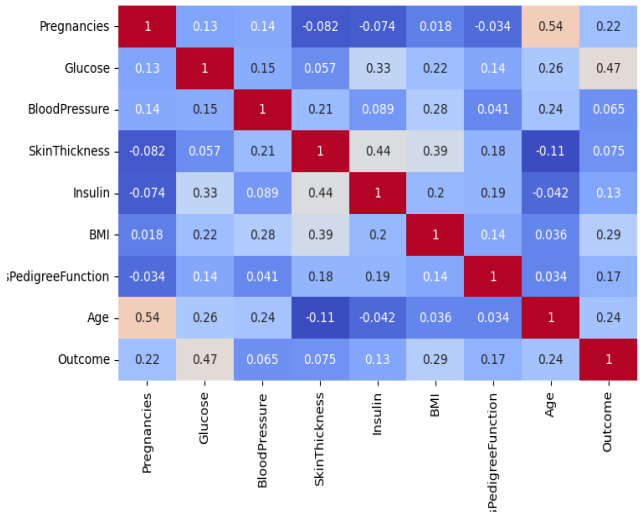

Fig. 2. Correlation matrix

D. Training the Model

Various machine learning techniques that were employed in the prediction model will be covered in this section. When employing ML algorithms, it is critical to divide the data into a training set and a test set. The test set, which constitutes 20% of the data, is used for outcome prediction, while the training set, which constitutes 80% of the data, is used to develop the classification model. We trained the model for training and validation sets to determine the accuracy scores by the algorithm of Random forest, Logistic Regression, K-Nearest Neighbor,Decision Tree, Support Vector Classifier, XG Boost.

1. Random Forest

For applications involving both classification and regression, Random Forest is a kind of ensemble learning that is well-known for its excellent accuracy and capacity to handle enormous datasets.. It was developed by Leo Breiman and works by reducing variance through combining predictions from multiple decision trees. A Random Forest model builds several decision trees during training and outputs the class that represents the mean of the classes or the mean prediction for regression of the individual trees. To achieve the optimal outcome, the algorithm chooses a subset of the best trees and merges them.

2. Logistic Regression

Logistic regression(LR) is a statistical approach that is employed for binary classification tasks. It's a kind of regression analysis where the dependant variable may only take one of two possible values, either 0 or 1. The principal objective of logistic regression is to forecast the probability of a particular sample belonging to a specific class. The LR model is built upon the sigmoid function, which converts any real-numbered value to a value ranging from 0 to 1. This function is used to transform the outcome of the linear regression model to a probability value, and the coefficients of the LR model are learned to minimize the error between the anticipated probability and the actual label.

3. Gradient Boosting

Extreme Gradient Boosting, often known as XGBoost, is a potent machine learning method that is frequently employed for classification, regression, and ranking problems. The goal of XGBoost is to enhance the precision and effectiveness of conventional gradient boosting algorithms. It is built on the gradient boosting framework. Decision trees are combined with XGBoost to create a model that is extremely accurate. The method uses the iterative process of decision tree construction and sample weight adjustments depending on prior mistakes to minimise error or to attain a specific number of trees. The ability of XGBoost to handle missing values and outliers in data is one of its primary advantages. Large datasets and high-dimensional feature spaces are no problem for the method. Additionally, XGBoost offers integrated regularisation strategies to avoid overfitting and enhance the model's generalisation capabilities.

4. Decision Tree

Decision trees are a popular machine learning technique with applications in classification and regression. Recursively dividing the data into subsets depending on the values of the characteristics is how the method operates. The last partitions represent the leaf nodes, which contain the projected class or value, whereas each division represents a decision node. Decision trees (DT) are renowned for their interpretability since they are simple for people to understand and visualise. The advantages of decision trees are their computing efficiency and ability to handle both continuous and categorical inputs.

They can also handle irregularities and data gaps in the dataset. Nevertheless, decision trees are vulnerable to overfitting, which happens when the tree is too complex and begins to include the data noise. Several methods, including pruning and regularisation, have been made to remedy this problem. Pruning is used to remove pointless branches from decision trees, while regularisation modifies the objective

function by including a penalty term to control the complexity of the tree.

5. Support Vector Classifier

A well-liked machine learning(ML) approach for binary classification applications is Support Vector Classifier (SVC). The technique searches a hyperplane that minimises classification error while maximising separating the two classes. A portion of the data points, known as support vectors, that are closest to the hyperplane serve as its definition.

6. K Nearest Neighbor

A preferred machine learning approach known as K-Nearest Neighbour (KNN) can be applied to resolve classification and regression problems. In essence, KNN locates the K data points make up the training dataset that are closest to a given query point, then guesses the class or value of that point based on the labels or values of the K closest neighbours.

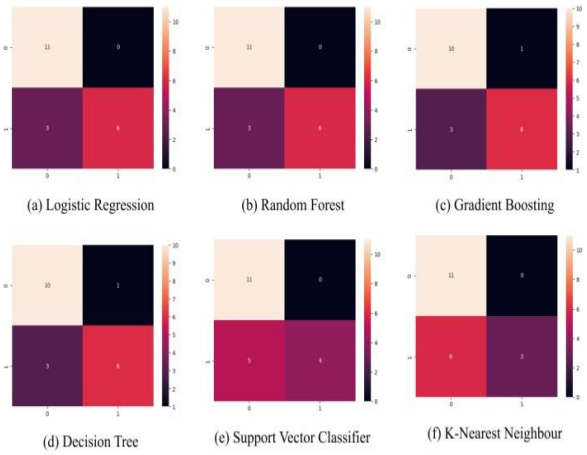

Fig. 2. Confusion matrix

E. Hyper Prameter

Hyperparameters are a type of parameter that can significantly affect a model's performance but are not learnt from data during training. Prior to training the model, the data scientist sets these parameters; they are not learnt during training. Hyperparameter tuning is a vital step in the machine learning process since it is essential for guaranteeing that the model can generalise successfully to new data.

F. Performance Metrics

The evaluation of the suggested method's performance has been conducted using the confusion matrix, as presented in the tables. The confusion matrix has four different outcomes: true positive(TP), true negative(TN), false positive(FP), and false negative(FN), as follows:

- **Accuracy:** It measures the model's total number of accurate predictions and can be measured as a ratio between the no.of correct prediction and total number of test cases of any model as follows:

$$Accuracy = TP + \ / TP + FP + FN + T\text{N}$$

- **Precision:** The proportion of correct positive predictions to total positive predictions is known as precision:

$$\text{Pr}ecison = TP \ / \ TP + T\text{N}$$

- **Recall:** Total positive predictions vs. actual positive values is known as recall:

$$\text{Re}call = TP \ / \ TP + F\text{N}$$

- **F1-score:** F1-score takes precision and recall into account and can be described as follows:

$$F1\text{-}score = 2 * (\ precion * recall \ / \ \ precion + recall\ )$$

Table 1
Logistic Regression Report

	Precision	Recall	F1-score	support
0	0.79	1.00	0.88	11
1	1.00	0.67	0.80	9
accuracy			0.85	20
macro avg	0.89	0.83	0.84	20
weighted avg	0.88	0.85	0.84	20

Table 2
Random Forest Report

	Precision	Recall	F1-score	support
0	0.79	1.00	0.88	11
1	1.00	0.67	0.80	9
accuracy			0.85	20
macro avg	0.89	0.83	0.84	20
weighted avg	0.88	0.85	0.84	20

Table 3
Gradient Boosting Report

	Precision	Recall	F1-score	support
0	0.77	0.91	0.83	11
1	0.86	0.67	0.75	9
accuracy			0.80	20
macro avg	0.81	0.79	0.79	20
weighted avg	0.81	0.80	0.80	20

Table 4
Decision Tree Report

	Precision	Recall	F1-score	support
0	0.77	0.91	0.83	11
1	0.86	0.67	0.75	9
accuracy			0.80	20
macro avg	0.81	0.79	0.79	20
weighted avg	0.81	0.80	0.80	20

Table 5
Support Vector Classifier Report

	Precision	Recall	F1-score	support
0	0.69	1.00	0.81	11
1	1.00	0.44	0.62	9
accuracy			0.75	20
macro avg	0.84	0.83	0.72	20
weighted avg	0.83	0.75	0.73	20

Table 6
K-Nearest Neighbour Report

	Precision	Recall	F1-score	support
0	0.65	1.00	0.79	11
1	1.00	0.33	0.50	9
accuracy			0.70	20
macro avg	0.82	0.67	0.64	20
weighted avg	0.81	0.70	0.66	20

In the result of training model where we classified each method's precision, recall, accuracy, F1 score, etc, We conclude that Logistic Regression(Table 1) and Random forest algorithm(Table 2) has shown better and equal results compared to other four algorithms. K nearest algorithm which is shown in the (Table 6) has the lowest accuracy.

The accuracy score for both logistic regression and Random forest has the same value but we used Random Forest algorithm in our application since it had more accurate result compared to logistic regression. So,We choosen the RF for prediction the diabetes with the web frame work of flask. To build a webpage to predict the diabetes with some features like, BMI, age, insulin, Skin Thickness ,etc… are entered in webpage to diagonsis the result immediately.

The report provides a summary of the model's accuracy, precision, recall, and F1 score for each class in the dataset. Accuracy measures the overall correctness of the model's predictions, while precision measures the proportion of true positives among all positive predictions, and recall measures the proportion of true positives identified by the model among all actual positives.

By calculating their weighted average, the F1 score gauges how well memory and accuracy complement one another.. A classification report can be used to identify which classes the model performs well on and which ones it struggles with. This information can be used to build the model's with performance by adjusting its hyper parameters or using different feature engineering techniques.

V. RESULT

The entire outcomes of the experiment in terms of accuracy, preci- sion, recall, and f1-score are presented above a table. For RF, LR, GB, DT, SVM, and KNN, the accuracy of these models is 85.0%, 85.0%, 80.0%, 80.0%, 75.0%, and 70.0%, respectively. The accuracy comparison table indicates that LR and RF exhibit the highest levels of accuracy, surpassing the other methods.

A. Before Using Hyper parameter Accuracy Score

Table 7

	Model	Tain Score	Test Score
1	Logistic Regression	78.07	85.0
2	Random forest	100.00	85.0
4	K-Nearest	76.74	80.0
3	Support vector	82.75	75.0
0	XGBClassifier	100.00	70.0
5	Decision Tree	100.00	65.0

B. After Using Hyper parameter Accuracy Score

Table 8

	Model	Tain Score	Test Score
1	Logistic Regression	78.07	85.0
2	Random forest	92.51	85.0
0	XGBClassifier	95.32	80.0
5	Decision Tree	82.35	80.0
3	Support vector	77.67	75.0
4	K-Nearest	78.34	70.0

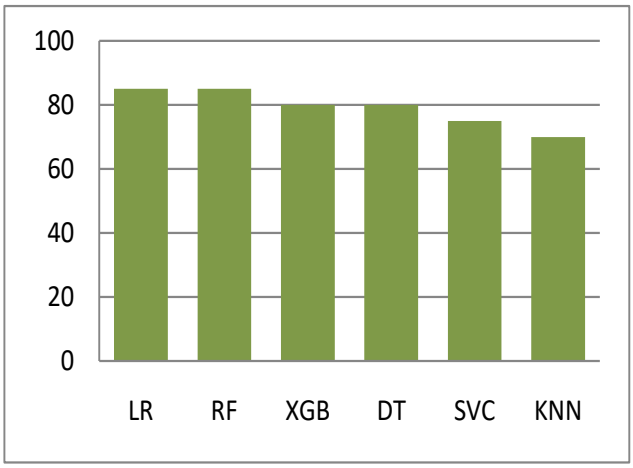

Fig. 3. Accuracy Chart

C. Web Application development using flask

The Flask is a micro web framework that uses Python and enables users to incorporate application functionality in a manner that seems to be a part of the framework itself.

WorkFlow of Application:

- The user sends the necessary information required by the application in a Webpage (Step-1).
- The information is sent to the back-end (Step-2).
- The flask server adopted with the machine learning classification pre- dict the results (Step-3 and Step-4).
- Finally, the predicted result is shown in the webpage (Step-5).

Fig. 4. Predicting Result is Diabetes

Fig. 4. Predicting Result is Normal

VI. CONCLUSION

In our research, Initially, we employed various machine learning[ML] techniques to diagnosis diabetes in individuals and compared their performance. We then conducted experiments to evaluate the efficacy of our proposed approach. Finally, based on the observed results, we developed a smart web application that utilizes the best algorithm to predict the likelihood of diabetes. This application allows individuals to input their clinical data and obtain an accurate diabetes prediction. It is a useful tool for those who wish to monitor their health status or need a routine check-up.

REFERENCES

[1] Mitushi Soni, Dr. Sunita Varma. She was able to forecast patients' diabetes using machine learning techniques that used the random forest method.. Volume 09, Issue 09 (September 2020).

[2] Aishwarya Mujumdar, V. Vaidehi. (2019). Diabetes Prediction Using Machine Learning Algorithm Procedia Computer Science 165:292-299 DOI:10.1016/j.procs.2020.01.047

[3] Quan Zou, Kaiyang Qu and Hua Tang presented Using Machine Learning to Predict Diabetes Mellitus in which they applied random forest, decision tree and neural network alogorithms on PIMA dataset and Luzhou dataset. DOI: 10.3389/fgene.2018.00515.

[4] Debadri Dutta, Debpriyo Paul, Parthajeet, " Utilising Machine Learning to Examine the Value of Features in Diabetes Prediction". IEEE, PP 942-928, 2018.

[5] K.VijiyaKumar, B.Lavanya, Reworded: " The Random Forest classification for Diabetes Prediction ".(ICSCAN), 2019.

[6] Md. Faisal Faruque,Iqbal, " Evaluation of Machine Learning classification for Diabetes Mellitus Prediction ". International Conference on (ECCE), 9 Feb- ruary, 2019.

[7] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, January 2018, pp.-09-13.

[8] Nonso Nnamoko, Abir Hussain,"Using Ensemble Supervised Learning for Diabetes Onset Prediction".IEEE (CEC), 2018.

[9] Deeraj Shetty, S, Nikita Patil,"Utilizing Data Mining for Predicting Diabetes Disease".(ICIIECS), 2017.

[10] Soumavadeen Manna and Mainak Adhikar are made a diabetes prediction using the algorithm of logistic regression and use the PIMA diabetes dataset to made a predictive model using cloud anayticss"2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)

[11] Mohamed Rady, Kareem Moussa, and Mahmoud Mostafa used Decision Tree, Random Forest, Nave Bayes, and SVM to Pima Indians Diabetes Data Set .October 2021,DOI:10.1109/NILES53778.2021.9600091.

[12] Diabetes Diagnosis Using Classification Mining Methods," Data Mining and Knowledge Management Processes of International Journal (IJDKP), Vol. 5, No. 1, January 2015. Aiswarya Iyer, S. Jeyalatha, and Ronak Sumbaly.