DIABETES DISEASE DIAGNOSIS USING MACHINE LEARNING TECHNIQUES

M. Gnana Pramod
Computer Science and Engineering
National Institute of Technology
Andhra Pradesh, India
421210@student.nitandhra.ac.in

Dr. Selvakumar K
Department of Computer Applications
National Institute of Technology
Trichy, India
kselvakumar@nitt.edu

***ABSTRACT*— Diabetes is a well known term. It is also called as Diabetes Mellitus is a metabolic disorder which is commonly seen in everyone now a days. Diabetes is a disease that result in too much sugar in the blood of living ones. Majorly due to either the pancreas is not producing enough insulin, or the cells of our body not responding properly to the insulin that was produced by the pancreas. If it is not diagnosed or treated properly it may leads to many health complications. Due this this untreated or poorly treated diabetes over 1.5 million people lost their lives. The proposed system utilizes a dataset of clinical data, including patient demographics, medical history, and various diagnostic measurements such as Blood glucose levels, Insulin levels, Body mass index (BMI), Pregnancies, Blood Pressure, Skin Thickness, Diabetes Pedigree Function, Age. And with these features this system predicts either the person is suffering from Diabetes or not which is not properly done by the existing system.**

KEYWORDS

Diabetes Disease Diagnosis, Machine Learning, Clinical Features, Supervised learning algorithms, Feature selection, Dimensionality reduction, Evaluation Metrics.

INTRODUCTION

Diabetes is a chronic metabolic disorder affecting millions of individuals worldwide. Accurate and timely diagnosis of diabetes plays a crucial role in managing the disease and preventing its complications. In recent years, machine learning (ML) techniques have become as powerful tools for disease diagnosis, offering the potential to improve tools the accuracy and efficiency of diagnostic process. It is a prevalent global health issue, with an estimated 463 million adults diagnosed with diabetes worldwide in 2019. Timely and accurate diagnosis of diabetes is crucial for effective management and prevention of complications such as cardiovascular disease, kidney damage, and vison impairment, Traditional methods of diabetes diagnosis involve clinical evaluation, blood tests, and the assessment of various risk factors.

ML algorithms have demonstrated remarkable success in various domains, including healthcare By leveraging large datasets and computational power, ML techniques can uncover complex patterns and relationships withing the data that are beyond human perception. In the context of diabetes diagnosis, ML models can learn from previous patient data to make predictions about the likelihood of an individual having diabetes based on their clinical and demographic information. The supervised learning algorithms are Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVM), enables us to train models on labeled data, where the diabetes status of each patient is known. These algorithms learn from the patterns and relationships within the dataset to make the predictions on new, unseen data.

Feature engineering techniques will be employed to extract valuable information from the dataset, while dimensionality reduction methods, such as principal component analysis (PCA), will help to reduce the feature space and improve computational efficiency. Evaluation metrics such as accuracy, sensitivity, specificity, and AUC-ROC will be used to assess the performance of all ML models and compare them with existing diagnostic methods. In addition to developing accurate ML models, this project also aims to enhance interpretability. By examining feature importance and generating decision rules, we seek to gain insights into the factors influencing diabetes diagnosis. This interpretability aspect can aid healthcare professionals in understanding the underlying mechanisms and contributing factors.

Overall, the development of an ML-based diabetes disease diagnosis system holds immense potential for improving the accuracy. By leveraging the power of ML techniques, we aspire to contribute to better disease management, personalized treatment, and ultimately, improved patient outcomes. Thereby improving patient care and treatment strategies. The symptoms of diabetes are Frequent Urination (polyuria), Excessive Thirst (Polydipsia), Unexplained weigh loss, Fatigue, Weakness, Increased Hunger, Blurred Vision, Slow Healing of Wounds, Frequent Infections and Numbness. The below Fig((i), shows the difference between the blood of a person who is suffering from diabetes and blood of a person who is not suffering from Diabetes.

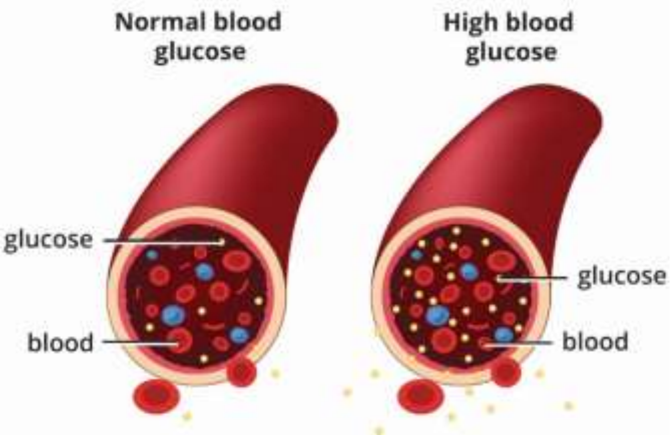

Figure-1: Blood with and without Diabetes .

So here we can see person with diabetes has high glucose where as other one has less glucose.

LITERATURE REVIEW

The article by K Selvakumar, Marimuthu Karuppiah, L Sai Ramesh, SK Hafizul Islam, Mohammed Mehedi Hassan, Giancarlo Fortino, and Kim-Kwang Raymond Choo titled "Intelligent temporal classification and fuzzy rough set-based feature selection algorithm for intrusion detection system in WSNs" was published on May 13, 2019. The framework system's architecture, including the trace dataset, data preprocessing, feature selection, and classification model for training the dataset, was described in this research. It aided me in beginning my work by providing the aforementioned architecture [1]. The study by R Reshma, V Sathiyavathi, T Sindhu, and K Selvakumar titled "IOT based Classification Techniques for soil Content Analysis and Crop Yield Prediction". They suggested a system to forecast the classification of the examined soil datasets in order to identify the crop producing. They utilised a classification strategy to predict the output for this system, which inspired me to use a classification technique to predict diabetes in my own system. On the basis of accuracy value, an appropriate classification method is chosen among Support Vector Machine and Decision Tree approaches. similarly, I applied the above same methods[2].

The paper titled “Analyzing and Comparing Omicron Lineage Variants Protein-Protein Interaction Network Using Centrality Measure” by Marmata Das, K. Selvakumar, P. J.A.. Alphonse was published on 30 march 2023. They examined the Omicron illness, which is related to health, and gathered accurate patient data for this paper. They focused on centrality, a particularly intriguing component of network measurement. Centrality is a commonly used indicator of how important a given node is to the network. They use the central measure to compare omicron lineage variants. As it relates to health, it assisted me in gathering the patient dataset and carrying out the process [3].

The paper titled “Brain FMRI Clustering Using Interaction K-Means Algorithm with PCA” by K. Vijay, K. Selvakumar. In the process of time series data mining based on MRI images of the brain, It has divides into four tasks collection, indexing ,processing and classification. In this paper they succeed with interactive IKM using PCA. They introduced this because their previous approach is not that accurate. Here, K-Mean clustering was performed, and the results were obtained using MRIs. My system benefited from this research because it is related to health [4] .Usually in higher papers the diabetes system is away at a little dataset. However wework with expansive dataset. generally medicinal test requires influence to concentrate on diminishing the therapeutic test. Our system will take a bigger dataset and therapeutic system will overcome problems in calculations for predicting diabetes. [5]. Sajida et al. discusses the role of Adaboost and Bagging methods using J48 decision tree for classifying diabetes with the help of this results we can say a patient diabetic or nondiabetic . The proposed system predicts diabetes with high accuracy with Decision Tree algorithm[6].Dr. Sunita Varma and Mithushi Soni developed a system that they constructed in Python utilising several classification and ensemble approaches. The random forest method produced the greatest results out of all the techniques. With the random forest classification method, they achieved 77% accuracy [7].

The paper titled “Predictive of Diabetes Using Machine Learning Algorithms in Healthcare” by Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali developed a method that made use of six well-known ML algorithms. For both training and testing of the predictive model, 8 attributes were chosen. According to the results SVM and KNN are the most accurate methods for predicting diabetes. The accuracy of both of these algorithms is 77%, which is the highest among the four methods utilised in this work [8]. Using the Random Forest Algorithm, Dr. K. Vijiya Kumar created a model that predicts the occurrence of diabetes. With Random Forest, she likewise obtained the best outcomes [9]. The research by Usama Ahmed, Ghassan proposed a model based on decision level fusion. By using fuzzy logic they used two main techniques in ML. The proposed system got an accuracy i.e. 94.87 which is much bigger than existing systems. From this model we can save my lives of patients. If we diagnose the disease i.e. Diabetes at initial stage then we can control the death ratio [10].

These are all the literatures that I referred and they helped me in building my model to predict diabetes disease. All of them followed ML algorithms .

RELATED THINGS

The below matter explain about complete information about SVM and KSVM as it gives best result generally. The developed technique used it to measure the difference in the patient’s data for Euclidean distance see Eq.1.

$$(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)2} - (1)$$

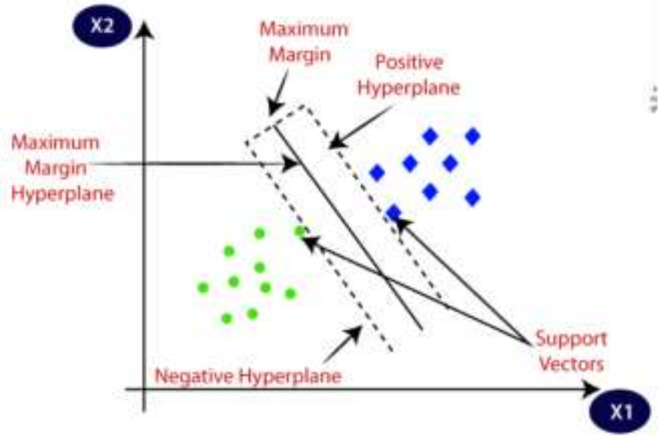

Figure-2: SVM Classifier

The Radial Basis Function (RBF) kernel of SVM is used to as a classifier. By using RBF it is very easy to predict the value of Dependent variable which depends on Euclidean distance from the points to the landmark on RBF curve . The equation of RBF is as shown in Eq-(2).

$$k(x_i , x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right| 2\right) \quad ,\gamma > 0 \quad -(2)$$

Where γ is a kernel parameter.

Hyperplane: In n-dimensional space, there may be several decision borders that can be used to separate the classes, but we must identify the decision boundary that will best enable us to classify the data points. That decision boundary is Hyperplane. we can observe with increase in dimension we get smoother curve.

Support Vectors: The data points or vectors that are the closest to the hyperplane outside and which affect the position of the hyperplane in space are known as Support Vector. Since vectors support the hyperplane they are called as “support vectors.”

METHODOLOGY

1.Data collection: Gather a comprehensive dataset of individuals with and without datasets. Include relevant attributes such as age , gender, body mass index (BMI), blood pressure, glucose levels, insulin levels, pregnancies, skin thickness, Diabetes Pedigree Function.

2. Data Preprocessing: Preprocess the data such that it includes the following steps:

2.1 Importing the data set that contains all the above features . I got my dataset from the website called Kaggle and Libraries that are required like

(i)Numpy- It is used to work with arrays which stores the inputs and outputs of dataset.

(ii)Matplotlib- This library is used to plot charts and graphs of predicted and actual values of all our predictions for our dataset.

(iii)Pandas-This library is used to import dataset and to create the matrix of features and the dependent variable vector.

2.2 Perform data cleaning by handling missing values, outliers, and inconsistences. Fortunately the dataset that is chosen doesn’t have missing if it has we replace them with average of remaining ones.

2.3 Normalize or standardize numerical features to bring them to a similar scale. Here I used Standardization to all the features i.e. Matrix of features to bring to same scale.

2.4 Encode variables that are categorical into numerical representations (e.g., one-hot encoding).

2.5 Split the data set into training sets and testing sets before applying Feature Scale to leverage the maximum value..

3.Feature Scaling:

3.1 Analyze the dataset to identify the most useful features if not consider all and continue to next step with them.

3.2 Utilize techniques like correlation analysis, feature importance, or domain knowledge to select relevant features that are required.

3.3 Remove unwanted features that may hinder model performance.

4. Model Selection:

4.1 Choose an appropriate machine learning algorithm for classification. Commonly used algorithms for disease detection include Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), or Neural Networks. For maximum conditions Kernel SVM is preferrable but not always.

4.2 Consider the dataset size, complexity, interpretability, and performance requirements when selecting the model to get more accuracy and precision.

5.Model Training:

5.1 Train the selected model using the training dataset by following known algorithms. Optimize model hyperparameters through techniques like cross-validation or grid search.

5.2 Evaluate different performance metrics (e.g.,accuracy, precision, recall, F1-score) during training to monitoring model progress so that at final we achieve best model.

6.Model Evaluation:

6.1 Evaluate the trained model using the testing dataset. Assess its performance using various metrics and compare them with the desired objectives. Observe the results in the graphs.

6.2 Perform additional analysis, such as ROC curves or confusion matrices, to understand model behavior and potential trade-offs.

7.Model Optimization

7.1 Fine-tune the model if necessary to improve its performance. Experiment with different techniques, like ensemble learning, regularization, or adjusting decision thresholds .Visualization through the charts and graphs help us to get more optimization.

7.2 Avoid overfitting by validating the model on separate validation sets or using techniques like cross-validation.'

8.Deployment:

8.1 Once satisfied with the model's performance, deploy it for real-world usage because when it achieved to our requirements then it would be useful for our future problems and to avoid them.

8.2 Create a user-friendly interface or integrate the model into an existing system.

9.Continuous Monitoring and Improvement:

9.1 Monitor the model's performance with real-world scenarios. Collect feedback and additional data to refine and enhance the model. Because if it makes any mistakes it would lead to many people's life risk.

9.2 Update the model periodically as new data becomes available or when improvements are identified. These updations make our model more powerful.

9.2 With this we can complete designing our model to predict the diabetes disease.

RESULTS AND DISCUSSION

After completion of training our model and execution of it we should calculate the following things.

I. Accuracy: The accuracy value tells us how efficient our algorithm is for our prediction. Based on its value we finalize our algorithm for our prediction . so more the accuracy more efficient our model.

It can be formulated as:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ Predictions}$$

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN}$$

Where TN= True Negative
TP= True Positive
FP= False Positive
FN= False Negative

II. Confusion Matrix: The table in which we represent Type-1 ,Type-2 errors and actual values is called "Confusion Matrix".

The confusion matrix is very simple to implement, but the terminologies might be difficult for beginners.

n=total Predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

The above table has the following cases:

- **True Negative:** Created Model has given prediction No, and the real or actual value of dataset was also No.
- **True Positive:** Created model has predicted yes, and the actual value of our dataset was also true.
- **False Positive:** Created model has predicted Yes, but the values of our dataset was No, it is also called as **Type-I error**.
- **False Negative:** Created model has predicted No, but the value of our dataset was Yes, it is also called as **Type-II error.**

III. F1 Score: F1 Score is a metric and It is a single score which represents both precision and recall. It is also one of the factors which helps us to get best algorithm for our model.

The formula for calculating the F1 score is given below:

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall}$$

By applying above factors the desired outputs of the project with different algorithms are as follows:

CLASSIFICATION:

It is a Machine Learning technique just like Regression. It is called Classification because it divides or classifies the data

Ex- Saying a Car company is Swift or not
Patient is having Corona or not

Saying a Animal is Cat or Dog

The following are various Classification Algorithms that our model had undergone.

1.LOGISTIC REGRESSION:

Logistic Regression is a Machine Learning algorithm which works on Classification. It calculates probability of an event occurring. It is Extension of Linear Regression. Though it is a Classification we still call it as regression due to his historical association with Linear Regression.

The results or metrics of our model by this algorithm are:

Confusion Matrix = [[89 10]
[24 31]]

Accuracy = 77.92207792207793%

F1 Score = 0.6458333333333333

2.K-NEAREST NEIGHBORS:

K- Nearest Neighbors (KNN) is a Machine Learning Algorithm. It can be used for Regression as well as Classification but mostly for Classification. It's a non-parametric Algorithm. It is also known as Lazy Learner Algorithm. Working of this algorithm is when a new data point is subjected to predict the output this algorithm looks for K Neighbors around it in graph and among that Neighbors our data point is classified into the Neighbors which has majority. Generally default value of k is 5 but it changes accordingly. The value of k will always be positive. Here training data are stored based on neighbors.[11].

So K-NN calculates the distance between the trained and new data points by using the formula

$$(x_{new}, x_i) = \sqrt{\sum_{j=1}^{m}(x_{newj} - x_{ij)}}^{2}$$

Where x_i is trained datapoints

x_{new} is new datapoints

The results or metrics of our model by this algorithm are:

Confusion Matrix = [[90 9]
[22 33]]

Accuracy = 79.87012987012987

F1 Score = 0.6804123711340206

3. SUPPORT VECTOR MACHINE:

Support Vector Machine is one of the most popular Supervised Learning Algorithms, which is also used for both Classification and Regression problems. But most frequently it is used for Classification. The main goal of this algorithm is to create the best line or decision boundary that can separate dimensional space into classes so that we can easily put our new data point in the correct category in future. SVM is very good when we have no idea on data. It works for both structured and unstructured data. But the only drawback is all parameters need to be set correctly to get best results[12].

This best line or decision boundary that is used to separate the space is known as " Hyperplane.". This SVM algorithm chooses the extreme vectors that help in creating hyperplane. These extreme vectors or points are known as "Support Vectors", and hence the algorithm is known as "Support Vector Machine". Hyperplane separating two classes is shown in Figure-2.

The results or metrics of our model by this algorithm are:

Confusion Matrix= [[89 10]
[24 31]]

Accuracy = 77.92207792207793%

F1 Score = 0.6458333333333333

4.KERNEL SUPPORT VECTOR MACHINE:

This algorithm is useful when our data points on the graph are not seperable. It uses Kernel's Trick which is used to convert the input data into a high-dimensional feature space. It uses Kernel Function to get the outputs. There are various types of kernel function linear kernel function, Radial basis function (RBF) etc. The point at tip of the RBF curve which is closer to landmark is consider as one class , and The point farther to the landmark of RBF curve is consider as another class. The Radial Basis Function which is used for our algorithm is shown in eq-(2).

The Gaussian Radical basis of function kernel is

$$k(x, l^{i)} = e^{\frac{-(x-l^{i})^2}{2\sigma^2}}$$

Where k=kernel

X=vector

l^i=landmark

The results or metrics of our model by this algorithm are:

Confusion Matrix = [[92 7]
[23 32]]

Accuracy = 80.51948051948052%

F1 Score = 0.6808510638297872

5.NAIVE BAYES:

Naïve Bayes algorithm is a supervised algorithm which is based on " Bayes Theorem" and it is used for solving classification problems. It includes high-dimensional training dataset. It predicts by using probabilities.It mainly works on probabilities of object. It estimates and handles the missing values by ignoring probability. It is sensitive in preparation of inputs. It causes problems we increase the number of training dataset [12].

Here it uses basic Baye's formula to find the output

$$P({}^{A}/_{B}) = \frac{P({}^{B}/_{A}) * P(A)}{P(B)}$$

Where P(A) and P(B) are probabilities of A and B

$P({}^{A}/_{B})$ and $P({}^{B}/_{A})$ are conditional probabilities

The results or metrics of our model by this algorithm are:

Confusion Matrix = [[85 14]
[21 34]]

Accuracy = 77.27272727272727%

F1 Score = 0.6601941747572815

6.DECISION TREE:

Decision Tree is also a supervised learning technique which can be used for both regression and classification, but most of the case used for classification problem. It is a Tree-Structured Classifier. Here splitting of our data set is done by using principle maximization. We build the tree using CART (classification and Regression Trees) algorithm. It is easy to understand this algorithm. Sometimes it behaves unstable due to tree data structure so it is often inaccurate[05]

The results or metrics of our model by this algorithm are:

Confusion Matrix = [[73 26]
[24 31]]

Accuracy = 67.53246753246753%

F1 Score = 0.5535714285714286

7.RANDOM FOREST:

Random Forest is also a popular machine learning algorith that belongs to supervised algorith. It work on the principle of Ensemble Learning. Combinig small machine learning algorith or using single algorithm many times to make big machine learrning algorithms is called " Ensemble Learning".

It follows the following steps:

1.Pick a random K data points from the training set and build the decision tree associated to that K data points.

2. choose the number N tree of trees want to build and repeat step 1.

3.For a new data point make each one of N tree of trees , predict the category to which the data point belongs, and assign the new data point to the category that wins majority vote.

The results or metrics of our model by this algorithm are:

Confusion Matrix = [[88 11]
[23 32]]

Accuracy = 77.92207792207793%

F1 Score = 0.6530612244897959

CONCLUSION AND FUTURE WORK:

The above results we conclude that Kernel SVM is the best algorithm with highest accuracy i.e. 80.519% and with highest F1 Score 0.68085 to predict diabetes in blood. The following results show the prediction of diabetes done by Kernel SVM for our dataset.

TRAINING SET:

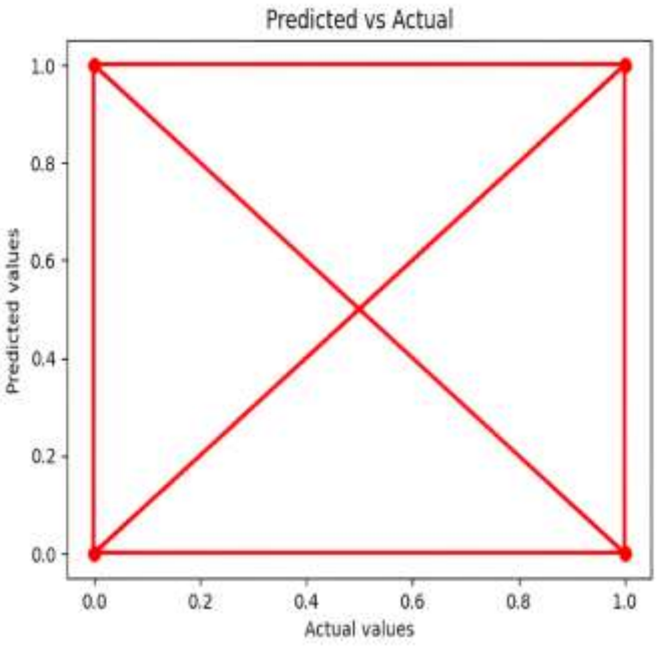

Figure-3:Predicted Vs Actual for Training Set

The above Figure-3 shows results of our model predictions for training set . we can see that some predicted values are wrong when compare to actual values . The four dots above tells

1. If it is 0, model predicted 0
2. If it is 0, model predicted 1
3. If it is 1, model predicted 1
4. If it is 0, model predicted 0

Where 0-Person is not suffering from Diabetes
1-Person is suffering from Diabeteas

TEST SET:

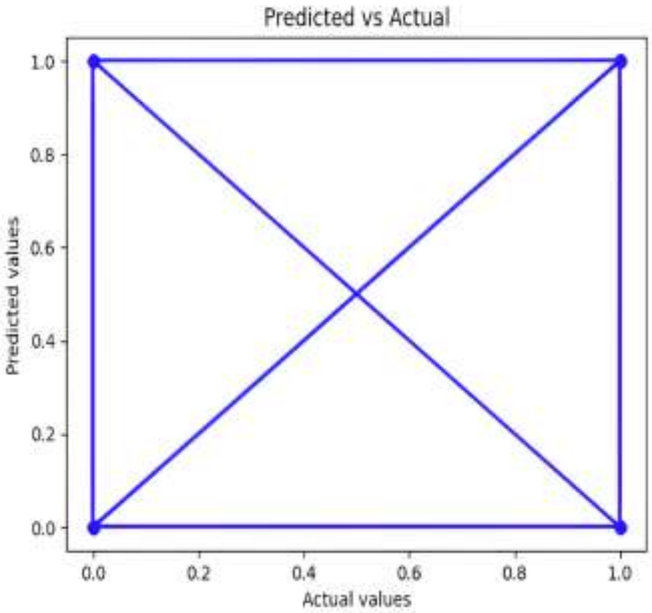

Figure-4: Predicted Vs Actual for Test set.

The above Figure-4 shows results of our model predictions for test set. We can see that very few predicted values are wrong. And the points on the graph are indicated as same as above Training set.

In conclusion, the developed machine learning model for diabetes disease diagnosis achieved a commendable performance with an Accuracy of 80.519%, F1 score of 0.68. The model demonstrated its ability to accurately classify individuals as diabetic or non-diabetic, leveraging important features such as BMI, glucose levels, and insulin levels. The Cumulative Accuracy Profile (CAP) curve analysis further validated the model's predictive power, outperforming a random model across different percentiles of total instances. These findings have significant implications for early detection and intervention in diabetes cases , potentially leading to improved patient outcomes and reduced complication. The study tried to solve diagnosis diabetes disease.

REFERENCES:

[1] K. Selvakumar *et al.*, "Intelligent temporal classification and fuzzy rough set-based feature selection algorithm for intrusion detection system in WSNs," *Inf Sci (N Y)*, vol. 497, pp. 77–90, Sep. 2019, doi: 10.1016/j.ins.2019.05.040.

[2] Institute of Electrical and Electronics Engineers, *Proceedings of the 4th International Conference on loT in Social, Mobile, Analytics and Cloud (ISMAC 2020) : 7-9 October 2020*.

[3] M. Das, K. Selvakumar, and P. J. A. Alphonse, "Analyzing and Comparing Omicron Lineage Variants Protein–Protein Interaction Network Using Centrality Measure," *SN Comput Sci*, vol. 4, no. 3, May 2023, doi: 10.1007/s42979-023-01685-5.

[4] Adhiparasakthi Engineering College, Institute of Electrical and Electronics Engineers. Madras Section., and Institute of Electrical and Electronics Engineers, *IEEE sponsored International Conference on Communication & Signal Processing : ICCSP-2015 : 2nd-4th April 2015*.

[5] Ieee, "Diabetes Disease Prediction Using Data Mining."

[6] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 1578–1585. doi: 10.1016/j.procs.2018.05.122.

[7] M. Soni and S. Varma, "Diabetes Prediction using Machine Learning Techniques." [Online]. Available: www.ijert.org

[8] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare 1."

[9] K. Vijiyakumar, B. Lavanya, I. Nirmala, and S. S. Caroline, *Random Forest Algorithm for the Prediction of Diabetes*.

[10] U. Ahmed *et al.*, "Prediction of Diabetes Empowered With Fused Machine Learning," *IEEE Access*, vol. 10, pp. 8529–8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

[11] M. Alehegn and R. Joshi, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach," *International Research Journal of Engineering and Technology*, 2017, [Online]. Available: www.irjet.net

[12] Surya Engineering College and Institute of Electrical and Electronics Engineers, *Proceedings of the 3rd International Conference on Computing Methodologies and Communication (ICCMC 2019) : 27-29, March 2019*.