

Diabetes Prediction using Machine Learning Techniques

**V Jithendra¹, R M Sai Mohit², M Madhusudhan³, B Jagadeesh⁴,
Dr. S Kusuma⁵**

^{1,2,3,4}Student, Dept. Of CSE, MITS, Madanapalle, Andhra Pradesh, India

⁵Assistant Professor, Dept. Of CSE, MITS, Madanapalle, Andhra Pradesh, India

Email: ¹19699A0520@mits.ac.in, ²19699A0544@mits.ac.in, ³19699A0524@mits.ac.in, ⁴19699A0517@mits.ac.in,
⁵kusumas@mits.ac.in

Abstract

Now a day due to hectic schedules and sedentary lifestyle people do not follow the proper diet. Poor diet may lead to diabetes, and which could result in various health issues such as heart attacks, strokes, renal failure, nerve damage, etc. When diabetes is accurately detected in its early stage, it can be effectively treated. By using Machine Learning methods, the problem can be easily detected and a solution could be arrived. Early diabetes detection and prediction can be greatly improved with machine learning (ML) approaches. When it is detected in an early stage, it can be resolved quickly. The objective of this research is to provide prediction using various supervised machine learning methods. Seven algorithms are compared with each other to figure out which is the best. The algorithms are Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Gradient Boosting. The evaluation results stated that Logistic Regression is more accurate than other algorithms for the given data set with an accuracy of 82%. After selecting the ML model which is more accurate. A User Interface where users can enter the new data and get results was developed and the results to the user were forwarded through WhatsApp along with some suggestions and precautions.

Keywords: Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Gradient Boosting, Machine Learning (ML), Diabetes.

1. Introduction

Diabetes affects numerous people as a result of its daily increase. Most patients are unaware of their health situation and the challenge before a disease is diagnosed. The main challenge is to make this architecture flexible across several datasets and to resolve or improve the prediction model's correctness. The International Diabetes Federation's seventh edition of the Diabetes Atlas contains the most recent data on DM (IDF). The number of diabetic patients worldwide in 2015 was over 410 million.

To lessen the impact of diabetes and manage the condition, one must focus on a person who is at high risk. The World Health Organization (WHO) specifies additional risk factors for diabetes mellitus as follows:

- Age group above 45 who follows the sedentary lifestyle.
- BMI is more than 24 kg/m².
- Blood sugar levels during fasting are continuously above the normal range or blood glucose is elevated above normal levels (IFG).
- Pregnant women whose age is above 30.
- Having high blood pressure.
- Having high blood levels of triglycerides.
- Having a family history with diabetes.

Diabetes is a condition when there is insufficient insulin in the blood, which leads to deficiencies. Frequent urine and a thirsty are symptoms of elevated blood sugar. If not treated well, it can cause serious problems or even death. Cardiovascular disease, an athlete's foot, and blurry vision are also symptoms of severe difficulty. High blood sugar levels are a sign of prior diabetes. Prior diabetes does not significantly outweigh traditional value. The exocrine gland's failure to produce enough hypoglycaemic agents and its improper response to those agents is what causes diabetes.

To carry out the experiment, a dataset of a patient's medical records is acquired, and various classification methods are then applied to it. The efficiency and accuracy of the employed algorithms are investigated and compared. The algorithm that is most effective for prediction will be selected among the different algorithms that are used. This study uses machine learning approaches to help medical professionals identify and treat diabetes early.

2. Literature Survey

For a prediction they performed 5 different supervised learning algorithms on the dataset that was collected [1]. Additionally, the second author used a heat map to determine the relationship among the features. They used T-Test to identify the features after analyzing how all the features relate to each other. They conclude that Logistic Regression is the most accurate learning technique after testing with various classification models. They are achieving an accuracy of more than 79% with this method.

Following thorough testing, they chose to use the logistic regression technique. It was executed using the fundamental RELU formula, and its behavior was examined [2]. The other study regarding diabetes prediction highlighted that the current value is used to predict the unknown value. The expected value only has two possible outcomes: 0 (No) or 1 (Yes). They used various machine learning methods to predict this diabetes using existing values.

Glucose levels, blood pressure, insulin, body mass index (BMI), age is considered when developing the model for diabetes prediction. Following that, a set of real-world data is used to evaluate the method. They achieved accuracy of 72% for the decision tree and 76.5% for the random forest [3]. Furthermore, the researchers used the support Vector Machine (SVM) and Random Forest (RF) classifiers to predict diabetes. The data set is first pre-processed after being obtained from a clinic.

Some duplicate and incomplete data are removed during the pre-processing of the dataset. The features are chosen from the data set after that pre-processing. The dataset is then prepared for dimensionality reduction. splitted The data set is

spilt into two portions,i.e., training and testing. Later trained with SVM classifier and a Random Forest model, respectively. Two models are then evaluated. the evaluation results showed 81.4% and 83% accuracy respectively [4].

The dataset is spilt as 70% for training and 30% testing . To predict diabetes, the seven best machine learning algorithms are being deployed. For this dataset, random forest achieves maximum accuracy. For the test data set, it provides an accuracy of 79.9%. In consideration of this, they came to the conclusion that, Random Forest algorithm is the most effective [5]. They primarily take into account three data sets to predict diabetes, using the same dataset as the other researchers for prediction. They are thinking about the datasets for heart disease (HD), liver disease, and diabetes disease.

The proposed model has 97% prediction accuracy. To determine the highest accuracy, they have used LR formula, this article used GA to forecast by combining the effects of a sizable number of independent variables.

The proposed classifier's classification result demonstrates that using a variety of variables to predict problem development is more accurate than using a single important factor or a few sets important variables [6]. Researchers have developed a useful approach to predict diabetes with the help of distributed machine learning and big data platforms like Spark. Through the distributed machine learning (ML) built on Apache Spark, their work intends to develop models that can identify diabetes in this circumstance.

and Comparison analysis of five machine learning algorithms is done using three performance metrics, the accuracy, recall, and precision [7].

3. Proposed Work

The proposed methods implement the seven-machine learning (ML) classification models and uses the algorithm which performs well in the Graphical User Interface among the seven algorithms after performing the comparative analysis for the given dataset. After finding the highest accuracy algorithm that is used for testing new data. The user enters the new data in the Graphical User Interface, and it will predict whether the person is diabetic or non-diabetic.

3.1 Methodology

As shown in figure 1 the process will be carried out.

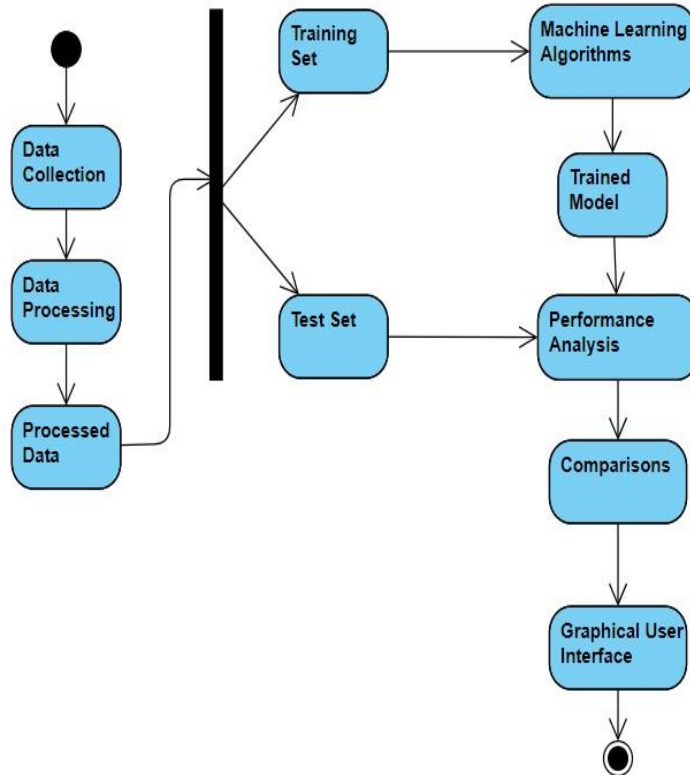


Figure 1. Methodology Representation

3.2 Data Collection

A dataset in machine learning is a group of data points used to develop a model. The accuracy and efficiency of the final model depend greatly on the dataset's quality. It is the process of collecting an existing data. the diabetes dataset from the Kaggle website that contains 768 patient entries like Glucose, Insulin, Skin Thickness, BMI, an Age, DiabetesPedigreeFunction and Outcome was used in the work to perform the analysis.

3.3 Data Pre-processing and Data Analysis

3.3.1 Data Pre-Processing

It is the most important and key aspect in this project. Missing values and other unwanted values will might reduce the effectiveness of the results. Data pre-processing is performed to increase the accuracy and effectiveness of the results. The two step Pre-processing followed in the work is as follows .

1) Missing Values Removal: Remove any value that have null or 0. There can never be a value of zero. This instance is therefore no longer valid. In order to reduce the dimensionality of the data and work more quickly, feature subsets are created by eliminating pointless features and occurrences.

2) Splitting Of Dataset: Splitting of data help us to know how the model will perform for the new data. The dataset is spilt into two portions each 80% and 20%. 80% is for training and 20% is for testing respectively. This process will be done after the above step.

3.3.2 Feature Selection

Performance of a Machine learning models can be improved with the help of a feature selection process. Heat map helps us to know how each and every feature are correlated and helps us to find the most important features that should be included in the model building. In this process four features from the original set of features using heat map for the model building is chosen they are Glucose, BMI, Insulin and Diabetes Pedigree Function. Figure 2 is used for feature selection.

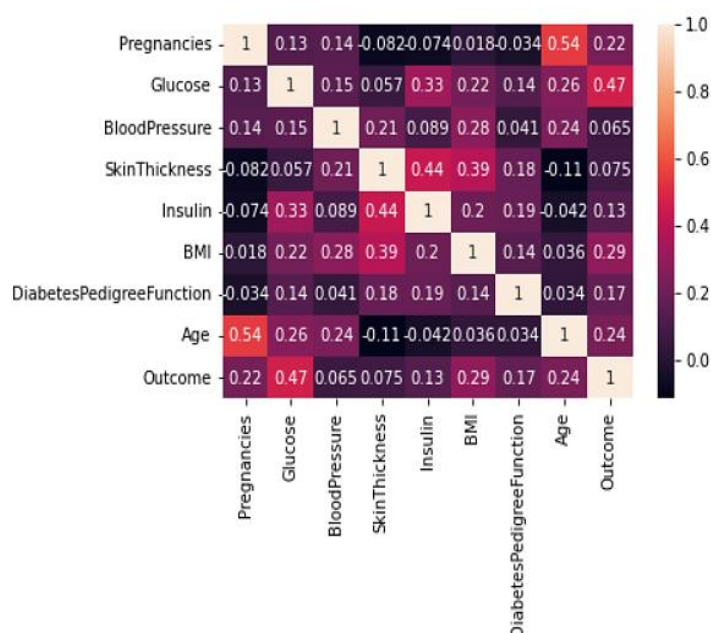


Figure 2. Heat Map

3.4 Model Construction and Prediction

3.4.1 Simulation Tool/ Framework

The whole project is done in the “Jupyter Notebook”. It is an open-source web application which provides an interactive, convenient environment for development and documentation of a python project.

3.4.2 Classification Algorithms

a) Logistic Regression

It is a statistical method used to analyse data and make predictions about the likelihood of an event occurring. When the dependent variable is categorical or binary, this kind of regression analysis is employed. Figuring out how the variable are related to each other is the aim of logistic regression. It uses a logistic function to transform the input values into a probability value between 0 and 1. The logistic regression model calculates the independent variable coefficients that increase the probability of the observed data. The chance that the dependent variable will have a particular value is determined using these coefficients.

b) K- Nearest Neighbor

It is a non-parametric algorithm. It operates by determining the K-nearest points, based on a selected distance metric, to a particular data point. Regression uses the K-nearest neighbor to calculate the predicted number. It is simple to implement and works well in low-dimensional spaces. However, it can be computationally expensive in high-dimensional spaces and requires enough labelled data. The choice of K affects the accuracy of the model, with small values leading to more flexible models and larger values leading to more rigid models. KNN can also suffer from the curse of dimensionality, where the data becomes increasingly sparse, and the distance metric becomes less meaningful in high-dimensional spaces.

c) Support Vector Machine

It is employed for regression and classification analysis. Finding the hyperplane that divides data into various groups is how it functions. The hyperplane is selected for increasing of margin. By utilising kernel functions, this algorithm is able to manage data that can be separated in both linear and non-linear ways. The use of regularisation parameters allows them to effectively operate in high-dimensional environments and avoid overfitting.

d) Naive Bayes

A probabilistic method used in the machine learning for classification jobs is called naive Bayes. The Bayes theorem states that the likelihood of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis, determine. The "naive" premise behind Naive Bayes is that, given class label, the features are conditionally independent. This makes the algorithm more computationally effective and simplifies the likelihood estimate.

e) Decision Tree

In this Internal node represent tests on a feature. Each branch represents the test's result, and each leaf node represents a class name or a numerical value in this tree-like structure. Because they are simple to manage both categorical and numerical data.

f) Random Forest

It is constructed with the help of multiple decision trees and takes one final prediction. Each decision tree is grown using a random subset of features and data samples, making them less prone to overfitting. The final prediction of the Random Forest is the majority vote or an average prediction of all the individual trees. This is most frequently used in classification and regression tasks. This algorithm can handle both categorical and continuous data, and also handles missing values. It can identify feature importance, making it useful for feature selection and data visualization. Random Forest is more efficient and can handle huge datasets.

g) Gradient Boosting

Gradient Boosting that involves in creating an ensemble of weak models that are trained sequentially to correct the errors of the previous models. Each subsequent model tries to minimize the residual errors of the previous models using gradient descent optimization. The prediction is based on a weighted sum of the predictions of all the weak models in the ensemble. The learning rate that controls the contribution of each individual model to the final prediction, is an important hyperparameter in this algorithm. It is known for its ability to handle complex datasets.

3.5 Performance Metrics

We usually use the different machine learning algorithms, and we require certain tools to assess how well they did their work and how they perform. There are many performance metrics that are used in machine learning. In this study, we collect useful data regarding algorithm performance and conduct a comparative analysis using several widely used measures for various tasks such as Accuracy, precision, recall, and f1-score.

Accuracy

It is arguably the most common and basic choice for evaluating an algorithm's effectiveness in classification problems. It can be defined as the percentage of data items that were properly identified out of all observations. Although accuracy has a wide range of applications, it is not necessarily the ideal

performance metric, particularly if the target variable classes in the sample are uneven. Mathematically it can be represented as,

$$A = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Precision

It is a metric for how accurate a binary classification algorithm is. It is defined as the ratio of the model's correct positive predictions to its total positive predictions. Mathematically it can be represented as:

$$P = TP / (TP + FP) \quad (2)$$

Recall

In machine learning, recall is a performance measure that is used to assess performance. The percentage of actual positive instances in the data set that are true positives (i.e., instances that were properly identified as positive). Mathematically it can be represented as:

$$R = TP / (TP + FN) \quad (3)$$

F1-score

It is also known as fscore or f-measure. It will determine how well an algorithm is performing based on recall and accuracy. It has the following mathematical representation:

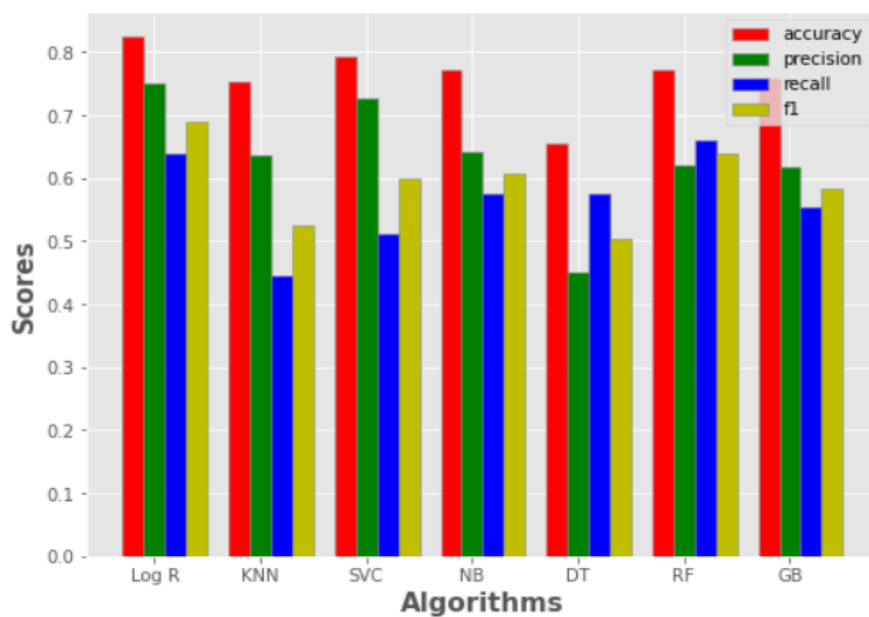
$$F1 = 2 * ((P * R) / (P + R)) \quad (4)$$

3.6 Comparative Analysis

A performance evaluation of various classifiers is performed. We have found that, “Logistic Regression Algorithm” are achieving the highest accuracy among all the algorithms. Here table 1 shows the scores of all algorithms and figure 3 is comparative analysis of each algorithm.

Table 1. Comparative Analysis of Algorithms

Algorithm	Accuracy	Precision	Recall	F1 score
Log R	82.4675	75	63.8298	68.9655
KNN	75.3247	63.6364	44.6809	52.5
SVC	79.2208	72.7273	51.0638	60
NB	77.2727	64.2857	57.4468	60.6742
DT	65.5844	45	57.4468	50.4673
RF	77.2727	62	65.9574	63.9175
GB	76.6234	62.7907	57.4468	60

**Figure 3.** Comparative Analysis Plot

3.7 Graphical User Interface

For a simple and interactive interface for user we built the GUI application using Tkinter package.

Tkinter: It is a standard Python package for creating graphical user interfaces (GUIs). It provides a set of tools and widgets that allow developers to build desktop applications with interactive windows, buttons, menus, and other GUI elements.

Creation of Tkinter GUI involves the following steps:

Step 1: Installing the Tkinter package and importing it into python script, which provides required classes and functions for creating a GUI.

Step 2: Creating the main window and customizing it by modifying the attributes like the title, size.

Step 3: After Creating the main window, now create and add widgets to it. It involves widgets like labels, buttons, and entry fields, etc.

Here we created 6 labels and 6 entry fields for the user's name, phone number and features and Predict, Report buttons.

Step 4: Now positioning the created widgets using a layout manager (Grid Manager).

Step 5: Binding functions to widget events or defining event handlers.

After clicking on the predict button it displays the result like "Diabetic" or "Non-Diabetic" and similarly after clicking on the report button it sends report to the user's mobile number through WhatsApp.

Step 6: Finally run the main loop

3.8 Sending Reports through WhatsApp

We are using PyWhatKit package for sending report through WhatsApp.

PyWhatKit: It is a Python library that provides various functionalities related to WhatsApp automation.

Sending reports using PyWhatKit involves the following steps:

Step 1: Installing the PyWhatKit package and importing it into python script.

Step 2: Use the `sendwhatmsg()` function to send a WhatsApp message. This function takes the phone number of the user, the message content (Report), the hour (in 24-hour format), and the minute when you want to send the message.

syntax: `pwk.sendwhatmsg(ph_no, msg, hour, minute)`

where;

ph_no: The phone number of the user entered in GUI, including the country code.

msg: Detailed text report.

Hour and minute: These are set for two minutes i.e.00:02 from the point of time when user clicks the report button.

4. Results

One of the major illnesses that might lead to several consequences is diabetes. According to all of the previous experiments, we found the accuracy of Logistic Regression is higher than other algorithms for the taken data set. At most we achieve 82% accuracy for Logistic Regression. After linking the model to the GUI application, the user enters new data in the Graphical User Interface, and it will predict whether the person as diabetic or non-diabetic and sends the Report and Suggestions to the user through WhatsApp message. Below figures are related to Diabetes Prediction using GUI. Figure 4 shows the user interface, figure 5 shows the details entered by the user and in figure 6 when user clicks on predict it shows diabetic or non-diabetic and user clicks on report it sends message as shown in figure 7.

Diabetes Prediction Using Machine Learning

Diabetes Prediction Using Machine Learning

Name

Glucose

Enter Value of BMI

Enter Value of Insulin

Enter Value of DiabetesPedigreeFunction

Enter phone number

Predict

Report

Figure 4. Graphical User Interface

Diabetes Prediction Using Machine Learning

Name	Ramu
Glucose	148
Enter Value of BMI	48
Enter Value of Insulin	75.8
Enter Value of DiabetesPedigreeFunction	0.66
Enter phone number	6305258073

Predict Report

Figure 5. Prediction for the Given Data

Diabetes Prediction Using Machine Learning

Name	Ramu
Glucose	148
Enter Value of BMI	48
Enter Value of Insulin	75.8
Enter Value of DiabetesPedigreeFunction	0.66
Enter phone number	6305258073

Predict Report

Diabetic

Figure 6. Displaying Output and Sending Report to User

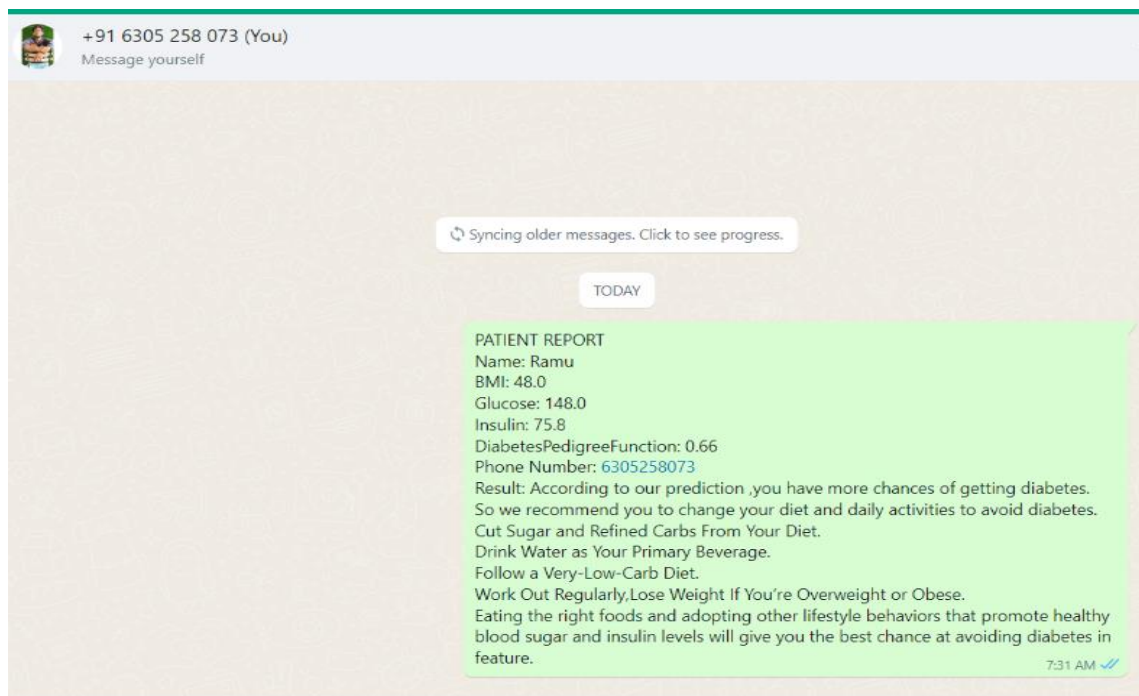


Figure 7. Detailed Report was Sent to the Patient Through Whatsapp

5. Conclusion

One of the most important real-world health problems is the detection of diabetes at an early stage. In this study, various procedures and implementation of various algorithms are done in building the prediction system which results in the prediction of diabetes. During this work, the research has used the seven machine learning classification algorithms. These algorithms are studied and evaluated on different parameters. And the tasks are done on Diabetes Dataset that are collected from Kaggle website. This study results determine that system achieves an accuracy of 82% using Logistic Regression algorithm.

References

- [1] Arwatki Chen L, Nurul Amin and Soumen Moulik's "Diabetes Disease Prediction Using Machine Learning Algorithms" published in IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) in 2020.

- [2] Le, T. M., Vo, T. M., Pham, T. N., & Dao, S. V.T. (2021). A Novel Wrapper Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic. *IEEE Access*, 9,7869–7884. doi:10.1109/access.2020.3047942.
- [3] P, Anirudh Hebbar; M V, Manoj Kumar; H A, Sanjay (2019). [IEEE 2019 1st International Conference on Advances in Information Technology (ICAIT) - Chikmagalur, India(2019.7.25- 2019.7.27)] 2019 1st International Conference on Advances in InformationTechnology (ICAIT) - DRAP: Decision Tree and Random Forest Based Classification Model to Predict Diabetes. 271–276. doi:10.1109/icaity47043.2019.8987277
- [4] Sivaranjani, S., Ananya, S., Aravindh, J., & Karthika, R. (2021). Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction. 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). doi:10.1109/icaccs51430.2021.9441935
- [5] Barhate, Rahul; Kulkarni, Pradnya (2018). [IEEE 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) - Pune, India (2018.8.16- 2018.8.18)] 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) - Analysis of Classifiers for Prediction of Type II Diabetes Mellitus.1–6. doi:10.1109/ICCUBEA.2018.8697856
- [6] Chaudhuri, A. K., & Das, A. (2020). Variable Selection in Genetic Algorithm Model with Logistic Regression for Prediction of Progression to Diseases. 2020 IEEE International Conference for Innovation in Technology(INOCON) doi:10.1109/inocon50539.2020.9298372
- [7] Ahmed, Hager; Younis, Eman M.G.; Ali, Abdelmgeid A. (2020). [IEEE 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE) - Aswan, Egypt (2020.2.8-2020.2.9)] 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE) - Predicting Diabetes using Distributed

Machine Learning based on ApacheSpark*,44–49.
doi:10.1109/ITCE48509.2020.9047795

- [8] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.
- [9] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.
- [10] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.
- [11] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques". Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13
- [12] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.
- [13] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.
- [14] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.
- [15] A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.