**Diabetes Prediction using Random Forest Machine Learning Algorithm**

1. Prof. Sudeshna Roy

Professor at Bharati Vidyapeeth’s Institute of Management and Information Technology,

Navi Mumbai.

2. Mr. Sourabh Suhas Mokal

MCA Student at Bharati Vidyapeeth’s Institute of Management and Information Technology, Navi Mumbai.

**I. Abstract**

Diabetes is an illness caused because of high glucose levels in the human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affect other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better results for prediction by constructing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are K-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). The accuracy is different for every model when compared to other models. The Project work gives the accurate or higher accuracy model shows that the model is capable of predicting diabetes effectively. Our Result shows that Random Forest achieved higher accuracy compared to other machine learning techniques.

**II. Introduction**

Diabetes is a noxious disease in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves the level of sugar in the blood. Diabetes occurs when the body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffer from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is the major cause of death in the world. Early prediction of diseases like diabetes can be controlled and save human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient results to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning are capable of prediction, however it’s tough to choose the best technique. Thus for this purpose we apply popular classification and ensemble methods on the dataset for prediction.

**III. Literature review**

K.VijiyaKumar et al. [11] proposed the Random Forest algorithm for the Prediction of diabetes to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly. Nonso Nnamoko et al. [13] presented predicting diabetes onset: an ensemble supervised learning approach they used. Five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Tejas N. Joshi et al. [12] presented Diabetes Prediction Using Machine Learning Techniques aimed to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of diabetes disease. Deeraj Shetty et al. [15] proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilising diabetes patient’s database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbour) to apply on diabetes patient’s databases and analyse them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar et al. [10] proposed a study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify if the patient is diabetic or not by applying a proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is an important area in computers, to handle the issues identified based on previous research.

**IV. Proposed system**

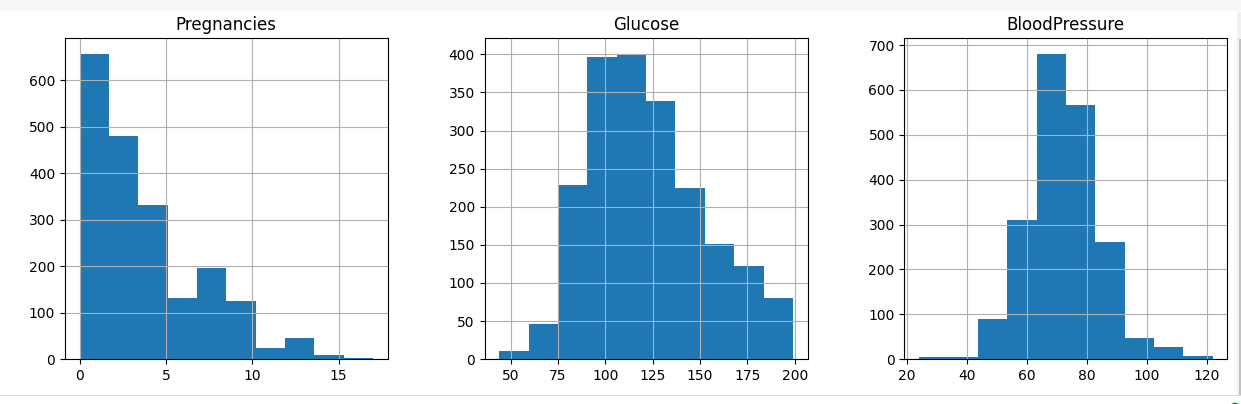
The machine learning model is classifying the result of diabetes based on the values entered by the user whether the entered inputs are showing you have diabetes or not. It will help all the open source users to check if they have diabetes with seating at the comfort of their house. And the model has 98% accuracy which is enough to be trustworthy.

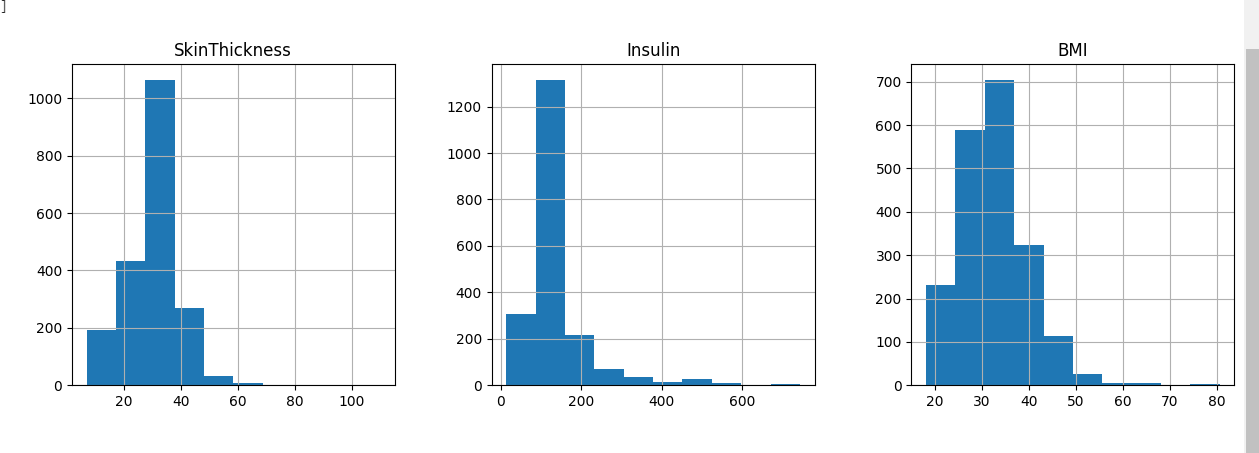
Goal of the paper is to investigate a model to predict diabetes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

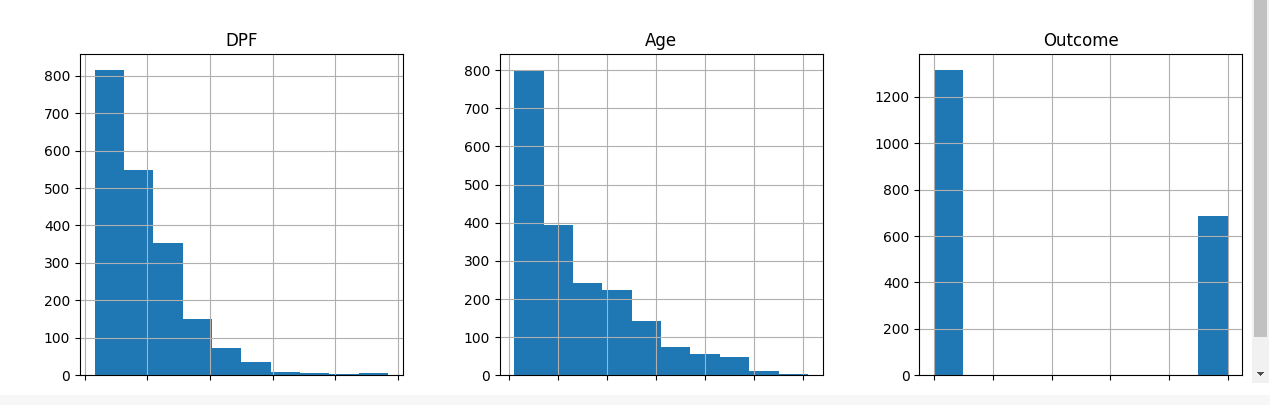
**Dataset Description:-**

The data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset has many attributes of 768 patients.

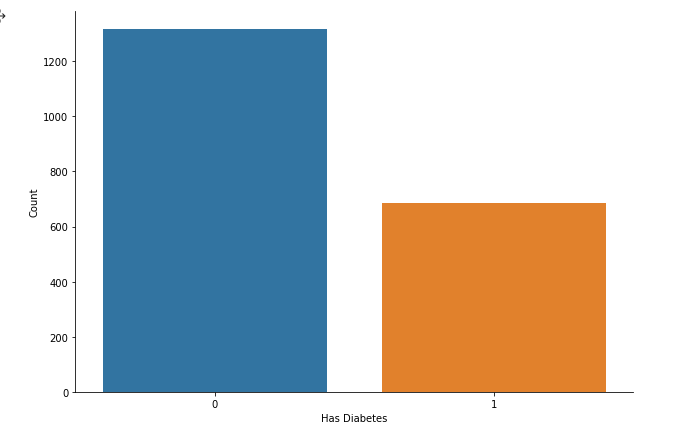
|  |  |
| --- | --- |
| Serial No. | Attributes |
| 1 | Pregnancy |
| 2 | Glucose |
| 3 | Blood Pressure |
| 4 | Skin Thickness |
| 5 | Insulin |
| 6 | BMI(Body Mass Index) |
| 7 | Diabetes Pedigree Function |
| 8 | Age |







The 9th attribute is class variable of each data points. The 9th attribute is the class variable of each data point. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics. Distribution of Diabetic patients- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labelled as 0 means negative means no diabetes and 268 labelled as 1 means positive means diabetic.



**Figure 1: Ratio of Diabetic and Non Diabetic Patient**

**Data Preprocessing-** Data preprocessing is the most important process. Most healthcare related data contains missing value and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after the mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For the Pima Indian diabetes dataset we need to perform pre-processing in two steps.

1). Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and helps to work faster.

2). Splitting of data- After cleaning the data, data is normalised in training and testing the model. When data is spitted then we train algorithms on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically the aim of normalisation is to bring all the attributes under the same scale. C. Apply Machine Learning- When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective is to apply Machine Learning Techniques to analyse the performance of these methods and find accuracy of them, and also be able to figure out the responsible/important features which play a major role in prediction.

**V. Algorithm Used :-  Random Forest Classifier**

It is a type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is greater than compared to other models. This method can easily handle large datasets. Random Forest was developed by Leo Bremen. It is a popular ensemble Learning Method. Random Forest Improves Performance of Decision Trees by Reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

The first step is to take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and store the anticipated outcome at intervals at the target place. Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. Some of the options of Random Forest do correct predictions results for a spread of applications.

Research methodology.

**VI. Model Flow:- Data Analysis and visualization.**

This is the most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction. Procedure of Proposed Methodology

**Step1**: Import required libraries, Import diabetes dataset.

**Step2**: Pre-process data to remove missing data.

**Step3**: Perform a percentage split of 80% to divide the dataset as Training set and 20% to Test set.

**Step4**: Select the machine learning algorithm i.e. KNearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

**Step5**: Build the classifier model for the mentioned machine learning algorithm based on the training set.

**Step6**: Test the Classifier model for the mentioned machine learning algorithm based on the test set.

**Step7**: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

**Step8**: After analysing based on various measures conclude the best performing algorithm.

**VII. Conclusion.**

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of those methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning methods in which SVM, Knn, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. And 77% classification accuracy has been achieved. The Experimental results can assist health care to take early predictions and make early decisions to cure diabetes and save humans life.

**VIII. References**

1. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

2. 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.

3. 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.

4. 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.

5. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.