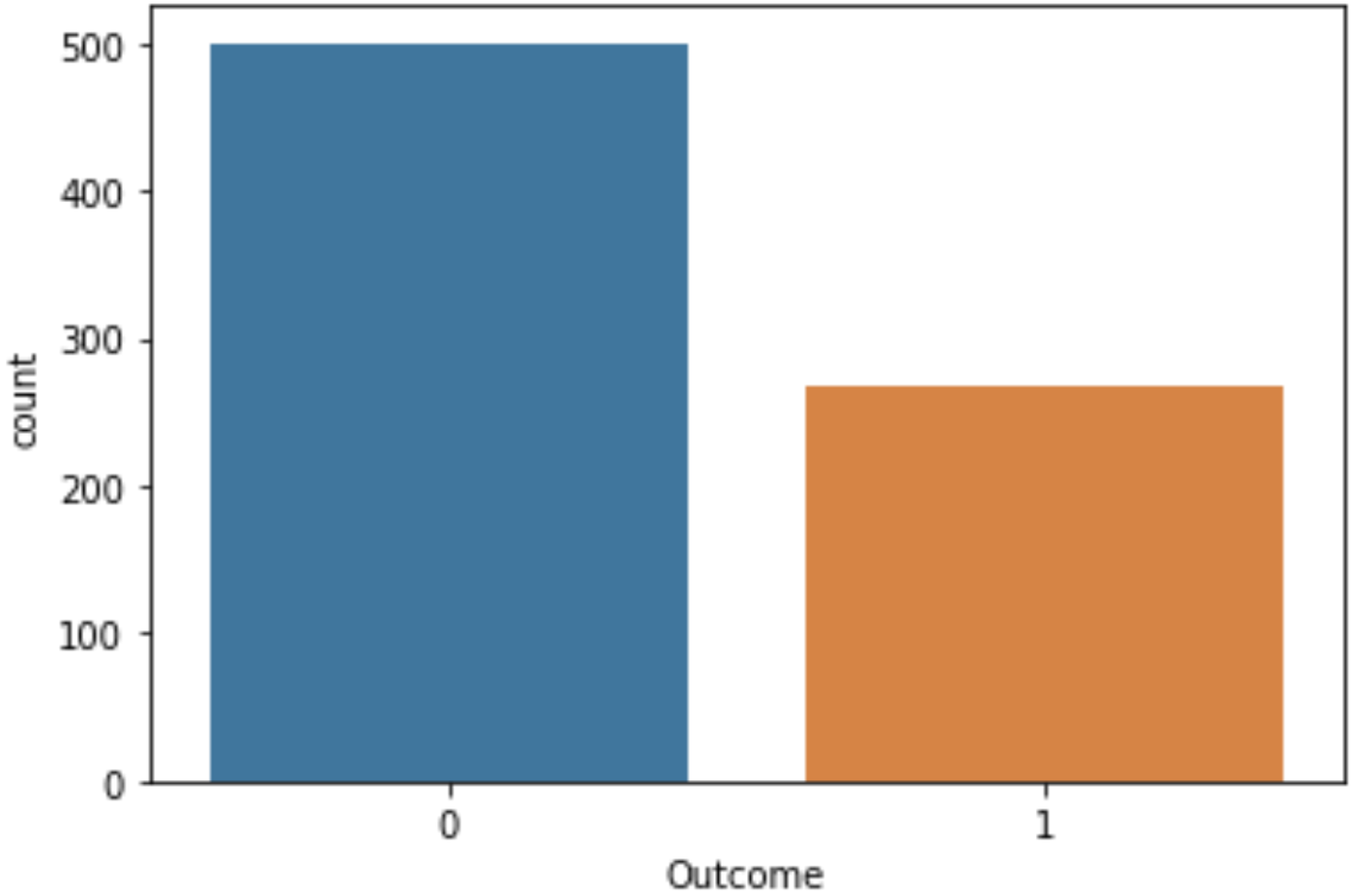
Report - Prediction of diabetes

By :
Dheeravath Dharma Raj



ABSTRACT

Diabetes is an illness caused because of high glucose level in a 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 affects 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 result for prediction by con- structing 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 Neighbor (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 capa- ble of predicting diabetes effectively. Our Result shows that Random Forest achieved higher accuracy compared to other machine learning techniques.

INTRODUCTION

1. Diabetes is noxious diseases 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 level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering 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 major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the 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 result 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 can capable to do prediction, however its tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

PROPOSED METHODOLOGY

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

1. Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Da- taset. The dataset have many attributes of 768 patients.
   Table 1: Dataset Description.

2. The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

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

1. **Data Preprocessing**- Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effective- ness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively the process is essential for accurate result and successful prediction. For 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 diamentionality of data and help to work faster.

    2. Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm 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 aim of normalization is to bring all the attributes under same scale.

MODEL BUILDING

This is most important phase which includes model build- ing 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 percentage split of 80% to divide dataset as Training set and 20% to Test set.
Step4: Select the machine learning algorithm i.e. K- Nearest Neighbor, Decision Tree, Logistic regression, Random Forest and.
Step5: Build the classifier model for the mentioned ma- chine learning algorithm based on training set.
Step6: Test the Classifier model for the mentioned ma- chine learning algorithm based on test set.
Step7: Perform Comparison Evaluation of the experi- mental performance results obtained for each classifier.
Step8: After analyzing based on various measures con- clude the best performing algorithm.

CONCLUSION

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which

. Naive Bayes
· Logistic Regression
· K Nearest Neighbour
· Decision Tree Classifier
· Random Forest Classifier
· Support Vector Classification (SVC) are used. And 77% classification accuracy has been achieved. The Exper- imental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.

Thank you