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.

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.

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

The data set is a 768 by 9 data set that has 768 instances which has the following attributes pregnancy, glucose , Blood pressure, skin thickness, insulin, BMI( Body Mass Index) , Diabetes Pedigree Function and Age the last label which is Outcome which tell if the person has diabetes or not. Using 1 as True and 0 as False. The dataset contains details of 500 negetive results and 268 positive results.

**Data preprocessing**

This stage involves checking for missing values, removing all instances where any attribute except outcome and pregnancy is 0. Checking for outliers and removing outliers. The next step is using standard scalar method for normalization

Splitting data set

This involves dividing the data set into 60, 20 and 20 percent which is training, validation and testing set.

And further more we train models based on the following algorithms

A Lasso regression which has a changing hyper parameter which helps to find out the important features which play major role for the prediction and use those features when training the models

The project involves

A logistics regression

Support Vector Machine

K- Nearest Neighbor

Random Forest

Gradient Boosting

Decision tree

­­­­­­­­