This project focuses mainly on the exploratory data analysis (EDA) of Diabetes data taken from Kaggle, which consist of 253680 rows and 22 columns. Our primary aim is to find the relation between various features and to predict the occurrence of Diabetes with respect to these features.

The target variable in the dataset is Diabetes012 which consist of three categories 0-non diabetic,1-pre diabetic and 2- diabetic.

This project also focuses on finding any type of relation between the features of the dataset other than the target feature.

For the prediction of diabetes, which is a classification problem, we apply algorithms like

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. Naïve Bayes Classifier
5. Linear SVM

Since the pre-diabetic class is only 1.83% of the whole dataset, we applied all the Machine Learning algorithms mentioned above, by dropping the pre-diabetic patients from the dataset.

This was to check whether the imbalance in the dataset is highly affecting the prediction output.