**Detection of Diabetes Mellitus with KNN algorithm using Machine Learning**

Diabetes Mellitus is a word for several conditions involving how your body turns food into energy. It is due to metabolic disorders and the reduction of blood glucose levels in the body. There are stages in diabetes based on the severity and it is an emergency task to identify diabetes at an early stage.

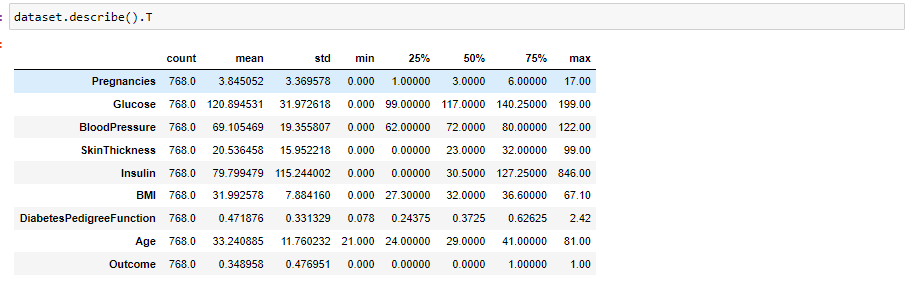
**Keywords**: Diabetes, Machine Learning, Prediction, Feature subset selection, Recursive partition algorithm

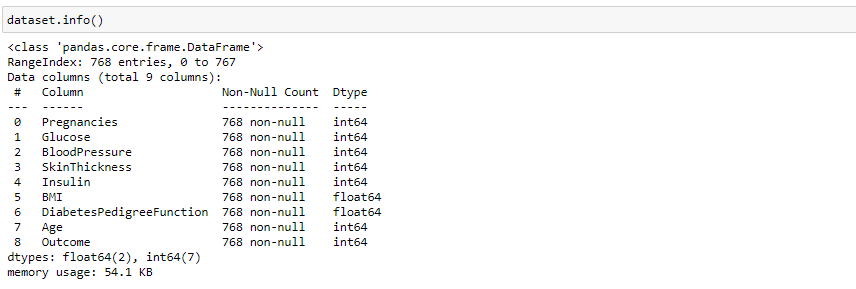
In this project the author is uses **Pima Indians Diabetes data set**, we also used the same data set with different machine algorithm called **recursive partitioning algorithm (using decision tree)** to improve the accuracy in predicting diabetes at the early stage of human life. As a first step we have performed pre-processing of data to improve the quality of dataset. Next we have performed feature subset selection to select only important features from given data set.

In this project we have used Naïve Bayes, Linear regression, SVM, Decision Tree, Random Forest Classification, Knn.

Data Set Source - <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

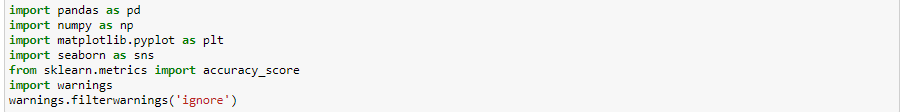
Dataset Description-





To implement this project author has used following modules

1. **Importing Libraries**



Pandas and numpy for data pre-processing, Matplotlib and seaborn for data visualization and sklearn for model building.

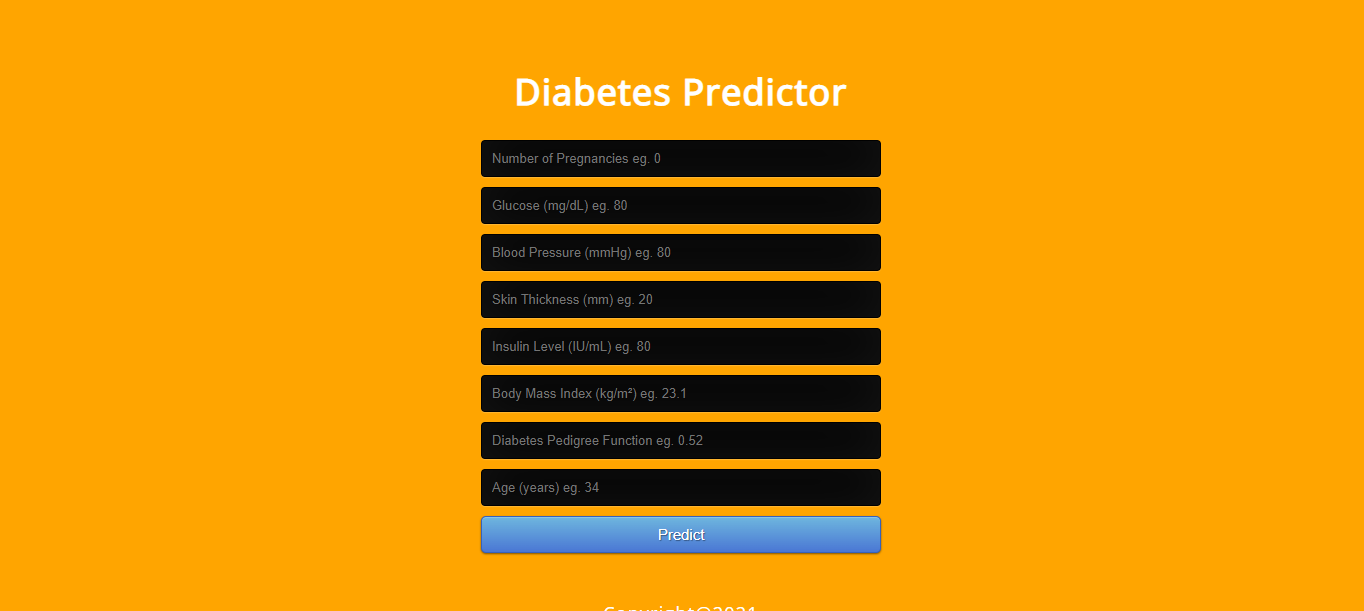
1. **Data Pre-processing** - checking for the null values. Then apply Recursive partition algorithm after that PCA is applied on the dataset.
2. **PERFORMANCE ANALYSIS** - The preliminary analysis discloses the subsequent insights of data. The given data set comprises of female patient data and their ages ranging from 21-81. The feature set used in the dataset includes the features A1-A8. Now the data set is classified into 70 percentages of training data and 30 percentages of test data. Here we are representing the results feature subset selection to improve the accuracy of the results**.**

Then we applied algorithm on different sub-sets ((A1, A2, A3, A4), (A1, A2, A3, A4, A6), (A1, A2, A3, A4, A5, A6), (A2, A4, A5, A6, A8), (A1, A2, A3, A7), (A1, A2, A3, A4, A5, A6, A7, A8), Full attribute set, (A8), (A7, A8), (A6,A7,A8), (A5,A6,A7,A8), (A4,A5,A6,A7,A8), (A3,A4,A5,A6,A7,A8), (A2,A3,A4,A5,A6,A7 ,A8)) to get the proper insight of the attributes.

1. **Model Building-** Then we have applied Naïve Bayes, Linear regression, SVM, Decision Tree, Random Forest Classification, Knn. From this, Knn is most effective, so, we have used for model building.

**ScreenShots-**

To run the project double click on ‘run.bat’ file to get below screen



Enter the following details to predict diabetes.



