**SD03Q03-1832003-Adithya R**

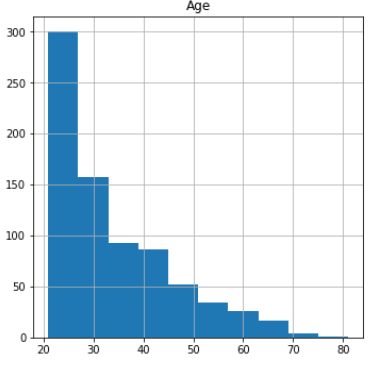
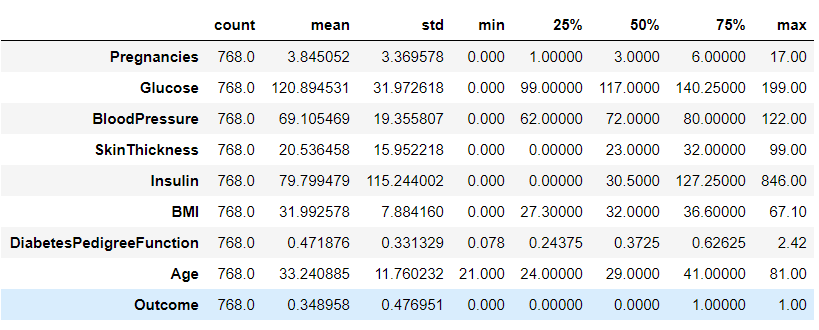
**Diabetes Patient Prediction**

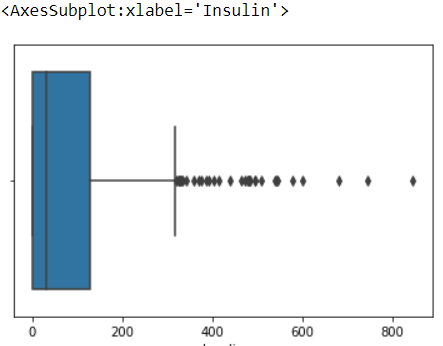
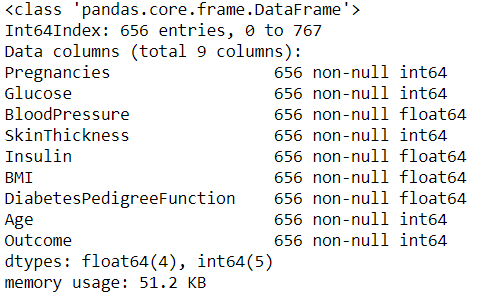
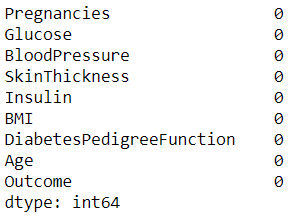
**Dataset:**

The given dataset consists of 11 rows and 768 columns. In this diabetes dataset, “Pregnancies”, “Glucose”, “BloodPressure”, “SkinThickness”, “Insulin”, “BMI”, “DiabetesPedigreeFunction”, “Age” are the columns as feature variables and “Outcome” as Predictor variable.

**Data Pre-Processing:**

The objective is to find the best machine learning model to classify and predict the outcome. So, dataset must be understood by the handler. So the box plots for each and every columns to outcome were found out. The outliers were occurred in boxplots of columns such as [Insulin,BloodPressure,BMI,DiabetesPedigreeFunction,Age]. So they were removed from the columns using quartiles, which results in the occurrence of missing values (NaN). The missing values row were removed from the dataframe using droppingna library, where the dataset is stored as dataframe using numpy & pandas.





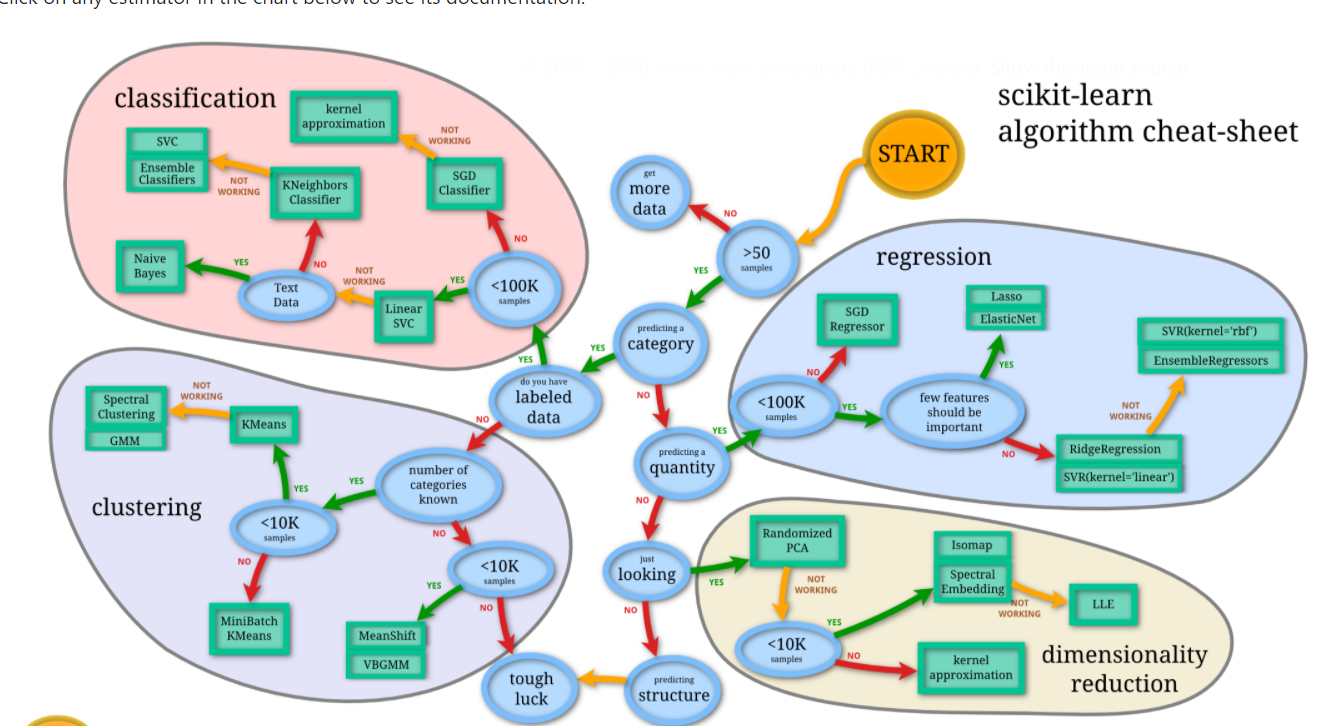
**Dataset Snippet, Histogram of Age, Missing values, Info of Dataframe and BoxPlot of Insulin**

**(Clockwise Direction)**

**Building the Model:**

**The Model I chose: KNN Classifier**

**Why I Chose:**

****

Since the dataset convinced the following factors in above image, I chose KNN Classifier to solve this problem.

**What is KNN Classifier?**

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in [Machine Learning for regression](https://quantra.quantinsti.com/course/trading-with-machine-learning-regression) and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function).

Using Sklearn\_model\_selection library, the implementation of splitting of train and test data set is done. The test and train sets are scaled using Standardscaler library.

To find the neighbours, Euclidean Distance is found out for the variables. And the splitted datasets such as train and test sets are fitted into KNN classifier model. After finding out the best K value for the model, it was also fitted into model for predicting the outcome using parameters(feature variable and predictor variable). The model is created and accuracy of 75.3 out of 100 is obtained using sklearn.metrics library - accuracy\_score function



It is observed from the dataset that the insulin level and glucose level of an individual person can result in the diabetes, also with the level of BMI. With 75.3% of accuracy, we can predict the person diagnosed with diabetes.

***GitHub Link for Code:*** [***https://github.com/adiram10/Curneu-Assessment/blob/main/Diabetes.ipynb***](https://github.com/adiram10/Curneu-Assessment/blob/main/Diabetes.ipynb)