**DIABETES PREDICTION**

**Problem Statement:**

Here we have a real life problem of people who are having cardiovascular diseases due to various factors .The dataset contains information of people like age,sex,Pregnancies,Gulcose,BloodPressure,BMI,Insulin,DiabetesPedigreeFunction,outcome. The outcome column contains labels for the people who were alive are not.

## **Exploratory Data Analysis (EDA):**

Exploratory data analysis is a complement to [inferential statistics](https://medium.com/analytics-vidhya/basic-statistics-in-data-science-38245e9b32bf), which tends to be fairly rigid with rules and formulas. At an advanced level, EDA involves looking at and describing the data set from different angles and then summarizing it.

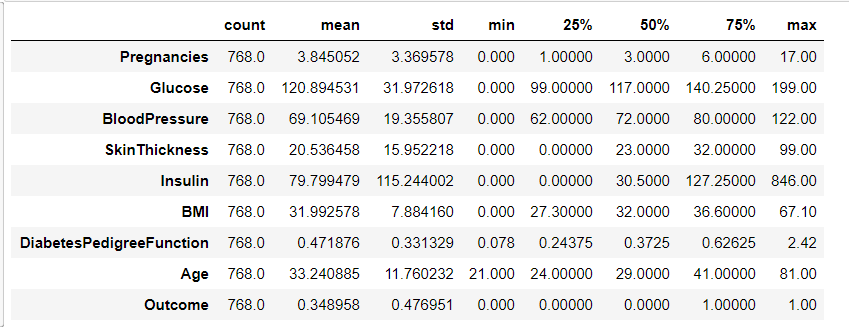
**Pre-processing:**

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

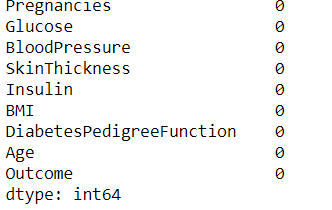
When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this we use data pre-processing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models.

Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

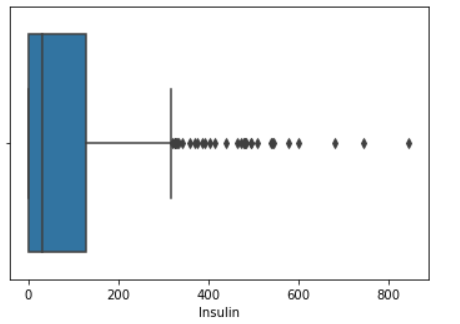


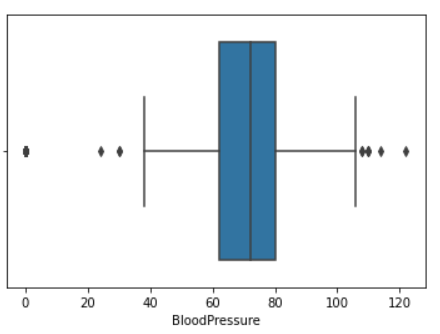
**Handle Missing value**  
 – Removing duplicates  
 – Outlier Treatment  
 – Normalizing and Scaling (Numerical Variables)  
Handling missing values:

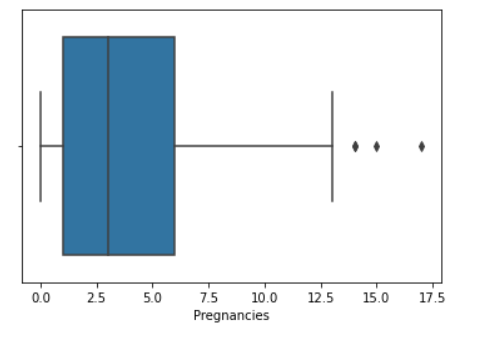


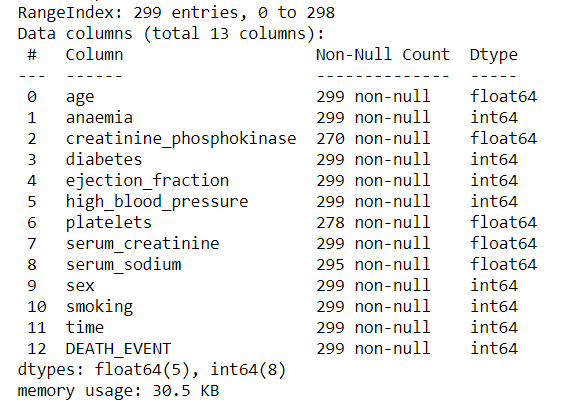
Handling Outlier:

Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations. We generally identify outliers with the help of boxplot, so here box plot shows some of the data points outside the range of the data.









## Normalizing and Scaling**:**

Often the variables of the data set are of different scales i.e. one variable is in millions and others in only 100. For e.g. in our data set Income is having values in thousands and age in just two digits. Since the data in these variables are of different scales, it is tough to compare these variables.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale. StandardScaler normalizes the data using the formula (x-mean)/standard deviation. We will be doing this only for the numerical variables.

**ML Model Implemented:**

K-nearest neighbours (KNN):

**KNN algorithm** is one of the simplest classification **algorithm** and it is one of the most used learning **algorithms**. **KNN** is a non-parametric, Easy and user friendly learning **algorithm**. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

**Algorithm and Implementation:**

• Load the data.

• Initialize K to your chosen number of neighbours.

• For each example in the data,

1. Calculate the distance between the query example and the current example from the data.

2. Add the distance and the index of the example to an ordered collection.

• Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances.

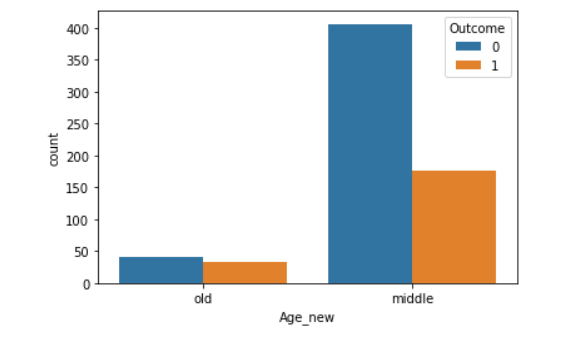
• Pick the first K entries from the sorted collection.

• Get the labels of the selected K entries.

• If regression, return the mean of the K labels.

• If classification, return the mode of the K labels.

**Prediction and Accuracy:**

****

