**DIABETES PREDICTION**

**Data Pre-processing:**

The dataset contains 0 in many rows. So, we have to remove them first. After removing them we have missing values in that places. So, we have to fill the missing values. It can be done using mean, median and mode. For choosing the best method, we plotted histogram for all the columns and observed how the values are deviated from the middle value. Since some values were too far, we didn’t use mean as it would affect. So median is the best option.

So, all the missing values were filled with median value of that variable. Then we used boxplot to check whether are there any outliers in the variables. Outliers will significantly affect our model. From observing the box plot it is noted that the following variables had many outliers:

1. Blood Pressure
2. Insulin
3. BMI
4. Diabetes Pedigree Function

Now we will have missing values in the place where we removed the outliers. So, we have to remove the rows that have missing values. After this is done row count is reduced from 768 to 656.

Last step in our pre-processing is normalizing the values by performing scaling. And now our dataset is cleaned and ready to be played.

**KNN Algorithm:**

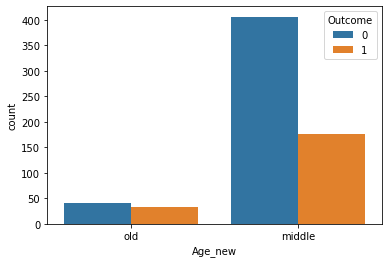
* K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

**ACCURACY:**

The accuracy for this model is 75.34 % approximately. This is a high accuracy, which means we can use this model for predictions which would give results which have the probability of being 75 % correct.

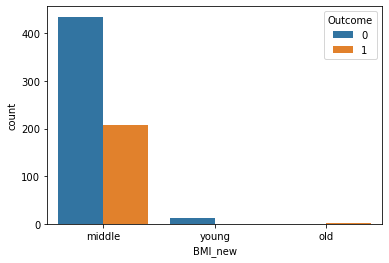
**INFERENCE:**

Age and Diabetes:



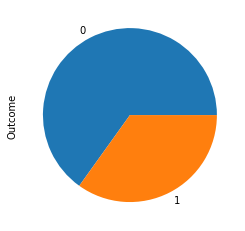
From this, we can note that women who are in middle age are affected most with diabetes.

BMI and Diabetes:



From this plot, we can see that women in under weight are not affected by diabetes. Women who are in correct weight are affected the most with diabetes.

Total women affected:



From this we can infer more than 25 % of women are affected with diabetes.

**SUPPORT VECTOR MACHINES:**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

**Accuracy:**

The accuracy for this dataset using this model is 79.7 %. So this model has higher accuracy than knn classifier.