# Random Forest Classification Model:

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms.

A random forest is a machine learning method for tackling classification and regression issues. It makes use of ensemble learning, a method for solving complicated issues by combining a number of classifiers.

Ensemble methods:

which combines several decision trees to produce better predictive performance than utilizing a single decision tree. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner.

Let’s talk about few techniques to perform ensemble decision trees:

1. Bagging

2. Boosting

Bagging(Bootstrap Aggregation):

Bagging is used when our goal is to reduce the variance of a decision tree.*Here idea is to create several subsets of data from training sample chosen randomly*with replacement*. Now, each collection of subset data is used to train their decision trees. As a result, we end up with an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree.*

Random Forest is an extension over bagging. It takes one extra step where in addition to taking the random subset of data, it also takes the random selection of features rather than using all features to grow trees. When you have many random trees. It’s called Random Forest

## Advantages of using Random Forest technique:

* Handles higher dimensionality data very well.
* Handles missing values and maintains accuracy for missing data.

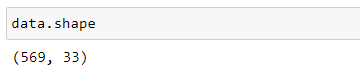
## Disadvantages of using Random Forest technique:

* Since final prediction is based on the mean predictions from subset trees, it won’t give precise values for the regression model.

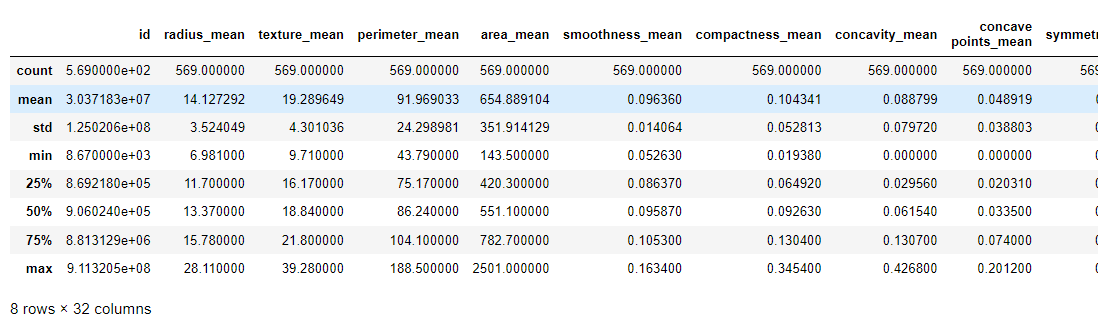
Boosting *is another ensemble technique to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. In other words, we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.*

# Exploratory Data Analysis:

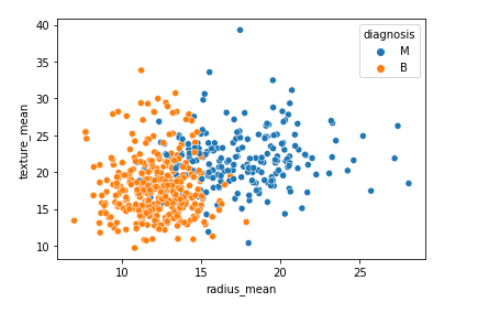
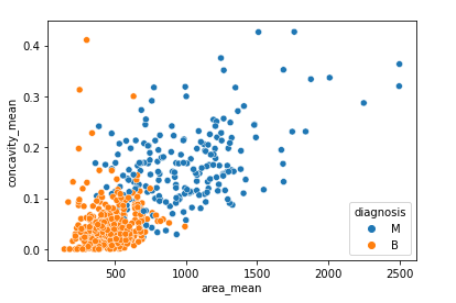
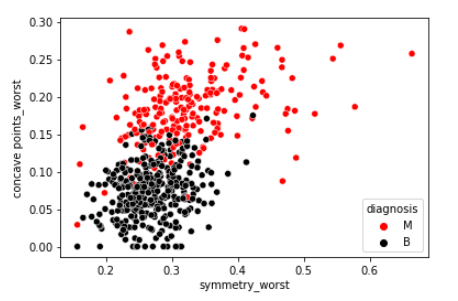
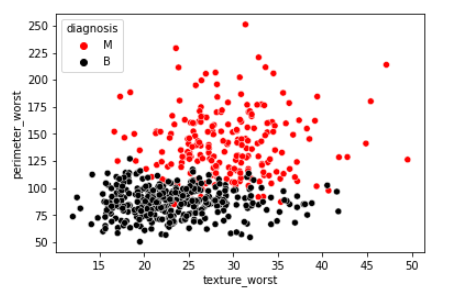
Shape of the data



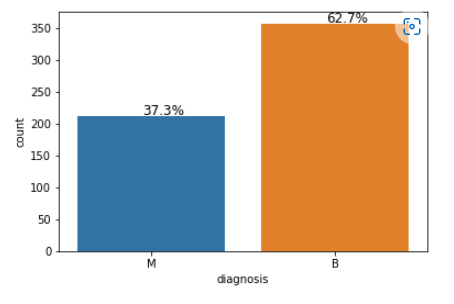
Statistical Descriptive of each feature in the dataset



## Bivariate Analysis

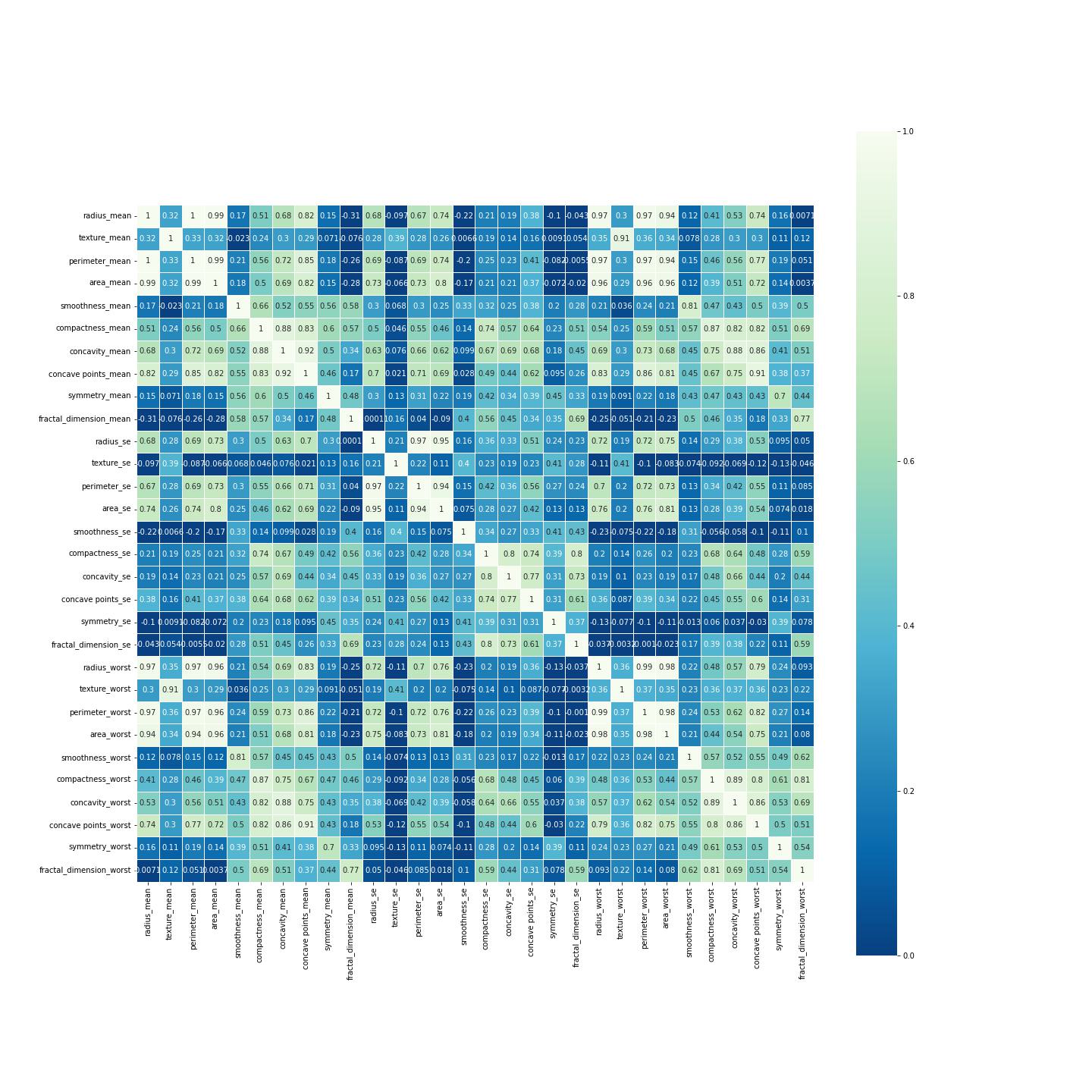
****  ****

## Ratio of Malignant and Benign



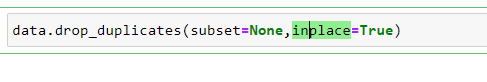
## Heatmap:

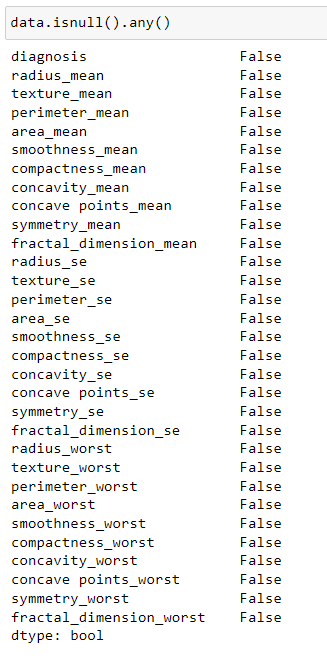
It shows the correlation between each variable in dataset



# Data Cleaning:

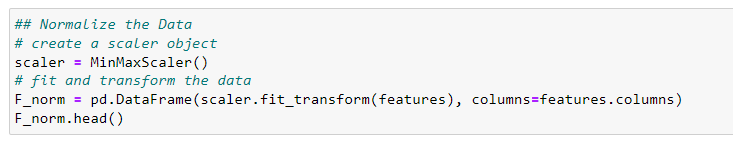
Data preprocessing is very important in machine learning as to remove any unwanted any data before feeding it to model so it won’t confuse model. Here we first check for null values in dataset and the remove the duplicates rows.

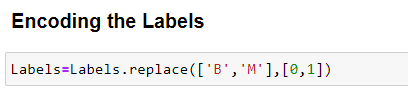




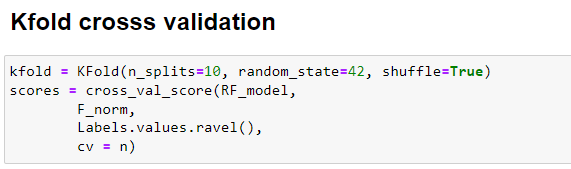
## Data Normalization

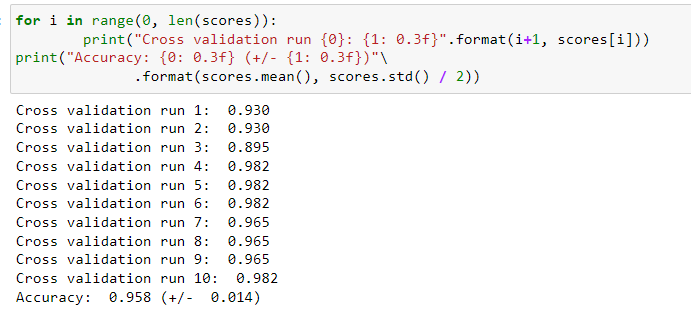
The goal of normalization is to transform features to be on a similar scale. This improves the performance and training stability of the model.











# Confusion Matrix for each Classification:

