**Introduction of Random Forest:**

A Random Forest consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable. Alternatively, for regression problems, the tree response is an estimate of the dependent variable given the predictors. The Random Forest algorithm was developed by Breiman.

A Random Forest consists of an arbitrary number of simple trees, which are used to determine the final outcome. For classification problems, the ensemble of simple trees vote for the most popular class. In the regression problem, their responses are averaged to obtain an estimate of the dependent variable. Using tree ensembles can lead to significant improvement in prediction accuracy (i.e., better ability to predict new data cases).

**Summary of the Report:**

This report provides the analysis regarding the prediction of breast cancer for the seattle grace hospital. The main goal is to develop a model which predicts the malignant tumors with few categorical variables. This model uses various regression techniques to check the accuracy.

**Model Description:**

Firstly, I would say I am successful in creating a machine learning model to predict the malignant tumors.

From the data set provided by the seattle grace hospital, I have built the model using Random Forest Regression. Using the Random classifiers, I have selected the categorical variables 'uniformity\_of\_cell\_size', 'uniformity\_of\_cell\_shape', 'marginal\_adhesion'. Since, Unnamed:0 variable is same as the serial numbers, I have dropped it out since it has nothing to do with my model. Another variable that can be dropped is the Id\_number, which is of least importance.

The first model was built with default values for the parameter and achieved the c-stat value 0.94.

**Test Parameters:**

* **n\_estimators**: The number of trees in the forest.
* **max\_features**: The number of features to consider when looking for the best split.
* **min\_samples\_leaf**: The minimum number of samples in newly created leaves.
* **n\_jobs**: Determines if multiple processors should be used to train and test the model.

**Model Accuracy:**

1. The model I have built was 95% accurate in predicting the malignancy. However, there are 5% chances that it may be falsely predicted. Considering the uniformity of cell shape and uniformity of cell size, the accuracy that I have obtained was 0.95.
2. The chances that it may miss the malignancy is 5%

**Model Comparison:**

There is not much change in the model after considering the auto, sqrt and log 2 as Max\_features. I have run the model for 3 max features and compared them.

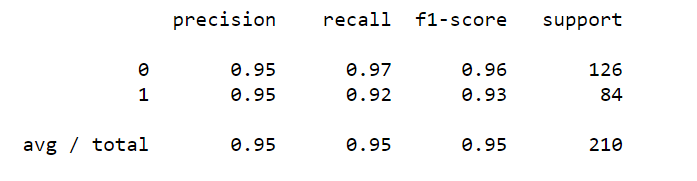
The classification report shows us the precision, recall and f1 score values.

Precision and recall are the basic measures used in evaluating search strategies.

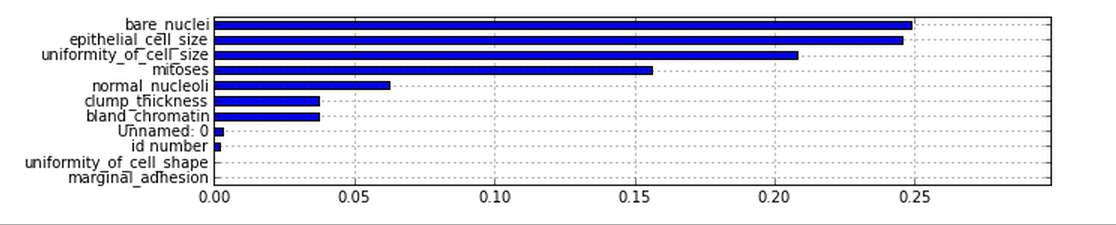
**Recall:** RECALL is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage

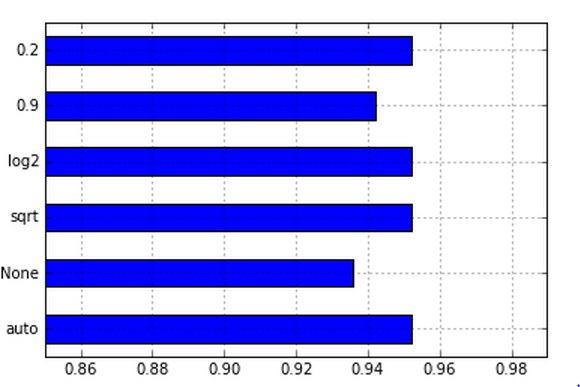
**Precision**: PRECISION is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage

Precision and recall are inversely related.



**Bar Chart:**



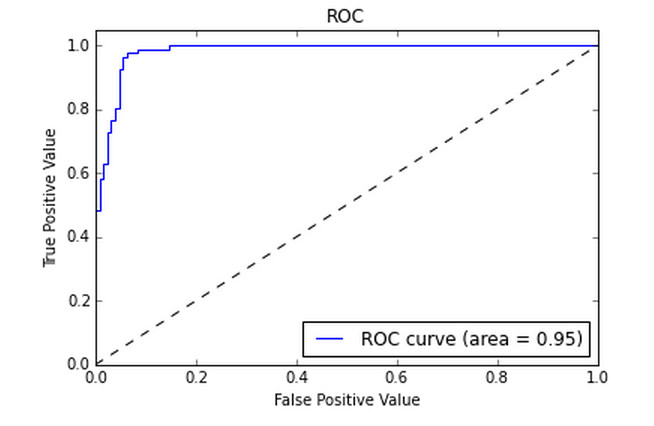


**Explanation to oncologist:**

Bare Nuclei helps in determining whether the tumor is malignant or benign. So, this is the most important variable in predicting the breast cancer. Usually cancer cells have different size and shape and hence these two variables are also important in determining the size of the cell and its intensity. The cells in the human body travel together, but the malignant cells gets separated. So, the marginal adhesion variable plays a major role in showing up the separated cells which would be easier to predict the cancer. In my view, the four variables bare nuclei, uniformity of the cell size, marginal adhesion and epithelial cell size are the most important variables in predicting the cancer.

From the chart, we can understand that the variables bare nuclei, epithelial cell size are the most important variables in predicting the breast cancer. Uniformity of cell size can also be considered and it plays a role in determining the c-stat value.

If we consider, the True positive value and false positive value against each other and plot the ROC curve, it will be like below:



Main Drawbacks of the model:

* 1. This model may miss the malignancy by 5%.
  2. The false malignance report percentage is 5%. This can be reduced by using the regression model instead of classifier model

Source: <https://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching_-_Recall_Precision.pdf>

http://www.statsoft.com/Textbook/Random-Forest