**Predicting Breast Cancer**

**Aim:**

To create machine learning model to predict **malignant** tumors.

**Technique used:**

We are given a datafile breast\_Cancer.csv, contains information 700 tumors which are either malignant or not malignant). This data set is two types 1. train data set 2. test data set. Training set is used to build the model. Test set is used to test the accuracy of the model after it is being built.We need to create multiple models with different samples and finally conclude on a final model. We need to predict malignant tumors.

lets take a random sample of 100 observation and 5 randomly chosen initial variables to predict malignant tumors. Final prediction is a mean of each prediction. Which is more accurate as we consider all possible combination of variables and sample data in the prediction process.

**Success/Failure depends on:**

Our model is tested against “test data set” for accuracy and it is 96% successful in predicting whether the tumor is malignant or benign.

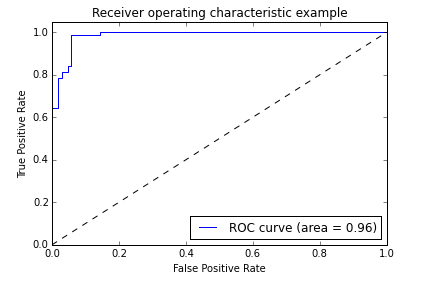
**Performance of the model:**

**Terms used:**

* True positive: Sick people correctly diagnosed as sick.
* False positive: Healthy people incorrectly identified as sick.
* True negative: Healthy people correctly identified as healthy.
* False negative: Sick people incorrectly identified as healthy.

Below diagram is created by plotting the ‘true positive rate’ against the ‘false positive rate’ of the model

Model accuracy is 96%

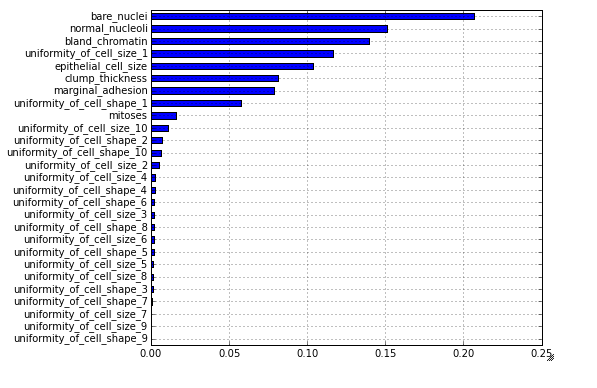
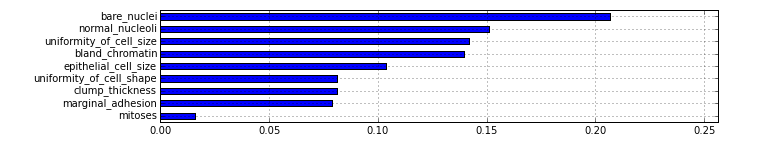
[](https://cloud.githubusercontent.com/assets/10646127/6906783/8bfabf1a-d6f6-11e4-8c65-8075b25614cd.png) Fig 1. ROC curve

Positive predictive values (PPV also called Precision) and negative predictive values (NPV) describe the performance of a diagnostic test:

* Precision (PPV) is calculated as ratio of true positives to all positives:
  + PPV= True positive / (True positive + False positive)
  + This gives people who are actually having cancer out of all the predicted list of people having cancer by the model.
  + Our model gives PPV=0.96. High precision relates to a 'low false positive rate'. This confirms that our model returns more relevant results.
* Recall (True positive rate) is the fraction of relevant instances that are retrieved:
  + Recall= True positive / (True positive + False negative)
  + The percentage of people who are having cancer are correctly identified as having the condition.
  + Our model gives recall as 96%. High recall relates to a 'low false negative rate'. This means that model returned most of the relevant results.

**Variables that are the most important in predicting breast cancer:**

Importance of variables is described below

[](https://cloud.githubusercontent.com/assets/10646127/6906848/e9b4fe04-d6f6-11e4-8b95-312082d7cb85.png) Fig 2. Simple bar chart that shows all of the variables.[](https://cloud.githubusercontent.com/assets/10646127/6906854/f100740e-d6f6-11e4-8d9f-9a3d5acec1b0.png) Fig 3. Bar chart that shows the summary of all the variables.

* Bare nuclei, normal nucleoli and bland chromatin are signs of benignity. So, they have more importance in determining if a tumor is malignant or benign.
* Uniformity of the cell size / shape: Cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not.
* Epithelial cell size: Epithelial cell that are significantly enlarge may be a malignant cell. So, size of the cell matters.
* Clump thickness: Benign cells are tend to be grouped in monolayers. While cancerous cells are often grouped in multilayers. Therefore clump thickness is one of the important variable.
* Marginal adhesion: Normal cells tend to stick together. Cancer cells tend to lose this ability. So, less adhesion is a sign of malignancy.
* Mitoses: The mitotic activity (how much the tumor cells are dividing) or Mitotic count is an important parameter for the prognosis of breast cancer. However, it is difficult to detect mitosis and make an accurate prognosis. So, it has little less importance than other parameters.

**Drawbacks :**

Sometimes the model can be inaccurate.