**Predicting Breast Cancer**

Creating Machine Learning Model for to predict malignant tumors or benign.

**Concept:**

In the data file provided, there is information about 699 tumors examined and their class as either malignant or not malignant.

The data set is divided as “train data set” which is used to build the model and “test data set” which is used to test the accuracy of the model.

The Concept lies from building multiple models using various variables, Usually the Final prediction would be the mean of every prediction & attributes provide information to predict malignancy & which is accurate by considering all possible combination of variables and sample data in the prediction process

This is how our model predicts the best possible outcome from the given set of data.

**How good is your model?**

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

**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.

ROC curve is created by plotting the ‘true positive rate’ against the ‘false positive rate’ of the model.

Graph suggests that “true positive rate” is much higher than the “false positive rate” making model accuracy is 0.96 (i.e. 96%).

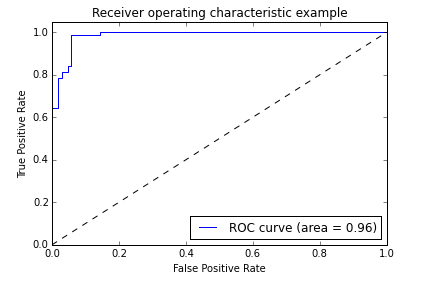
[](https://cloud.githubusercontent.com/assets/10646127/6906783/8bfabf1a-d6f6-11e4-8c65-8075b25614cd.png)

Figure: 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.

The below bar charts indicates the importance of the variables in descending order.

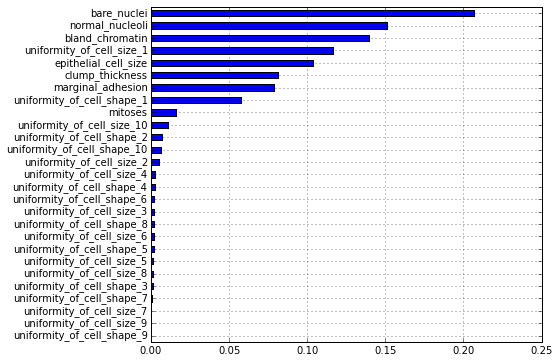
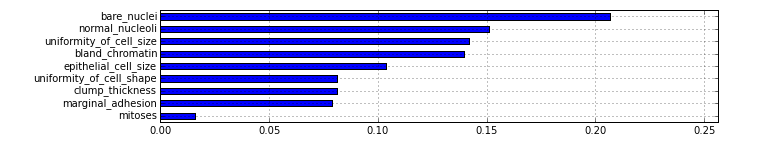


  Figure : Simple bar chart showing all the variables.

[](https://cloud.githubusercontent.com/assets/10646127/6906854/f100740e-d6f6-11e4-8d9f-9a3d5acec1b0.png)

How likely is it falsely predict breast cancer?

There are chances that model may falsely predict breast cancer. From the precision (0.96) we can interpret that 0.04 of the times model may falsely predict breast cancer.

How likely is it to miss a malignant case?

Model may also miss a malignant case while predicting. From the recall (96%) we can say that 4% of the times model may miss a malignant case.

**Source:**

* <http://en.wikipedia.org/wiki/Precision_and_recall>
* <https://answers.yahoo.com/question/index?qid=20101204013824AAWTufG>