**Dataset – Cancer Detection**

This dataset contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. We have to develop a model that learns from the features and predicts whether the cancer is Benign or Malignant.

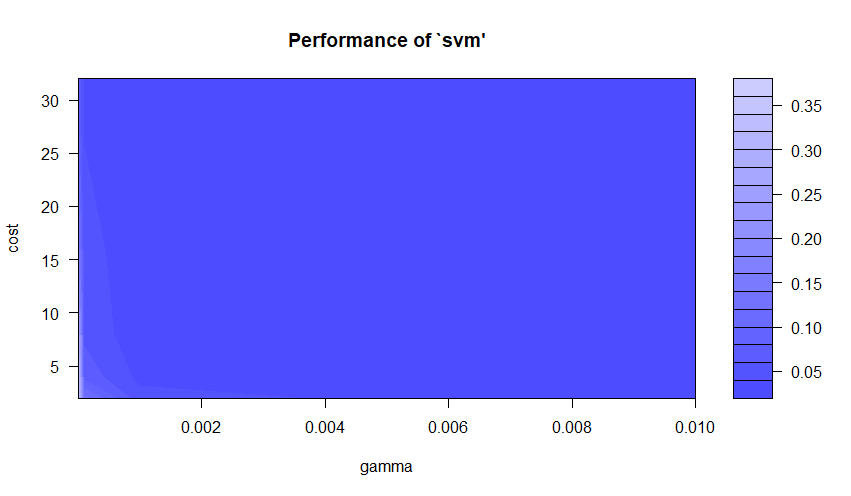
This problem is interesting as it helps us understand what are the important features that makes cancer Benign or Malignant.

We shall build 3 types of models – SVM, Decision Tree, Gradient Boosting.

**Support Vector Machines**

Using Support vector machines, we aim to classify employees utilizing 3 different kernels – Radial, Linear, Polynomial.

We first tune the SVM model based on different gamma and cost values using a grid.



We see from the above graph that the best tuning hyperparameters for SVM are gamma = 0.01, cost = 4 for a performance: 0.02

We shall use the above Hyperparameters to build our SVM models for different kernels.

**Radial Kernel**

Using the Radial Kernel, we get the following Confusion Matrix

actual

fitted B M

B 105 2

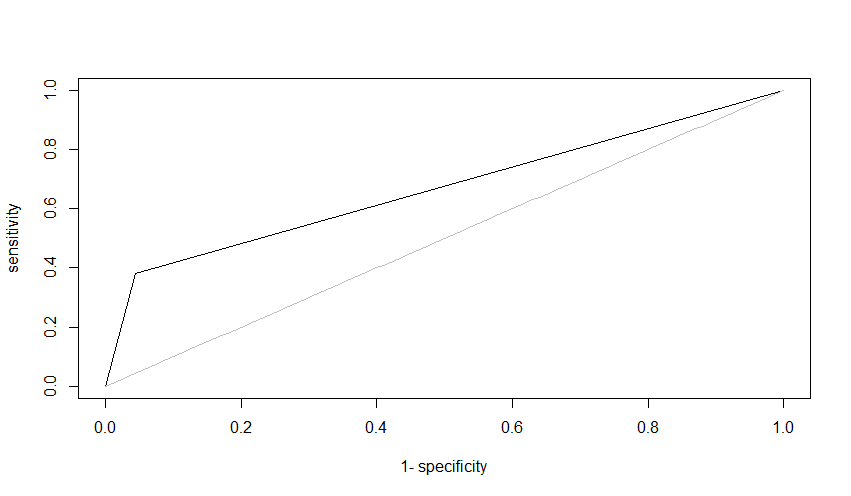
M 2 61

Accuracy is 97.64%

True Positive Rate - When cancer is benign, and how often our model predicts that: 98.13%

Specificity – when cancer is malignant, and how often our model predicts that: 96.82%

Next, we build the ROC to find the Area under the curve metric.



AUC = 0.9747812

**Linear Kernel**

Using the Linear Kernel, we get the following Confusion Matrix

actual

fitted B M

B 102 2

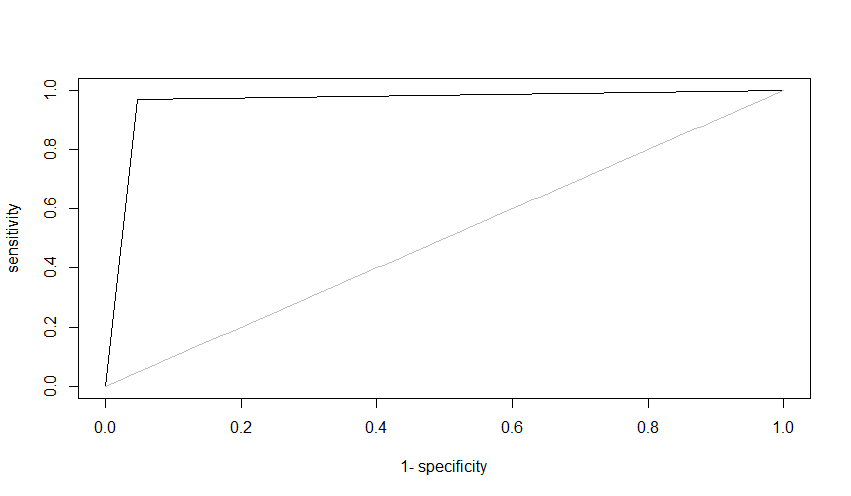
M 5 61

Accuracy is 95.88%

True Positive Rate - When cancer is benign, and how often our model predicts that: 98.07%

Specificity – when cancer is malignant, and how often our model predicts that: 92.42%

Next, we build the ROC to find the Area under the curve metric.



AUC = 0.9607625

**Polynomial Kernel**

Using the Polynomial Kernel, we get the following Confusion Matrix

actual

fitted B M

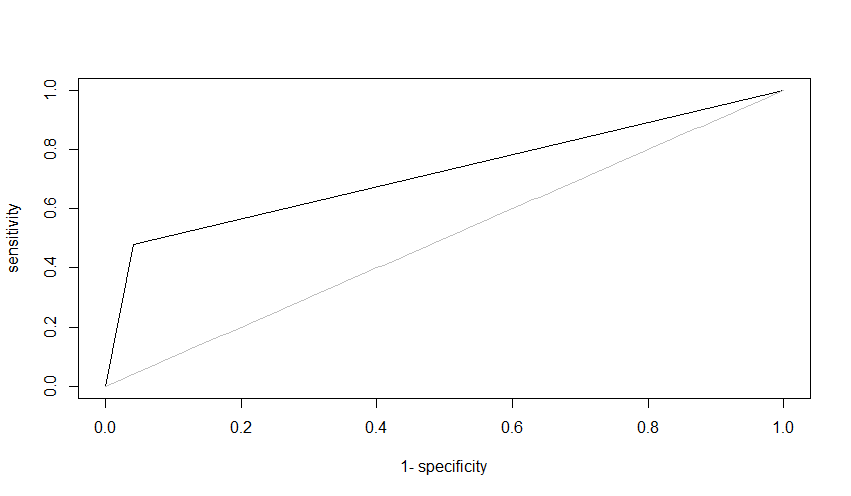
B 106 4

M 1 59

Accuracy is 97.05%

True Positive Rate - When cancer is benign, and how often our model predicts that: 96.36%

Specificity – when cancer is malignant, and how often our model predicts that: 98.33%

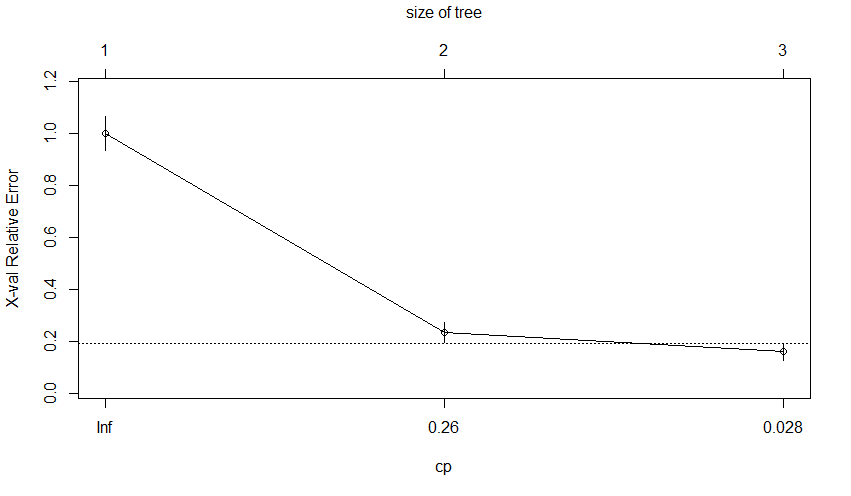


AUC = 0.9635811

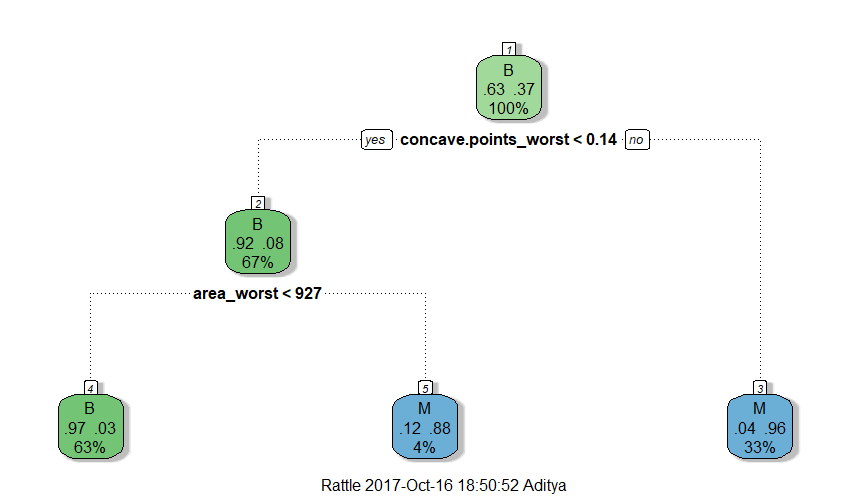
Radial Kernel performs the best among all the kernels with Accuracy of 97.64 and AUC = 0.9747812

**Decision Tree**

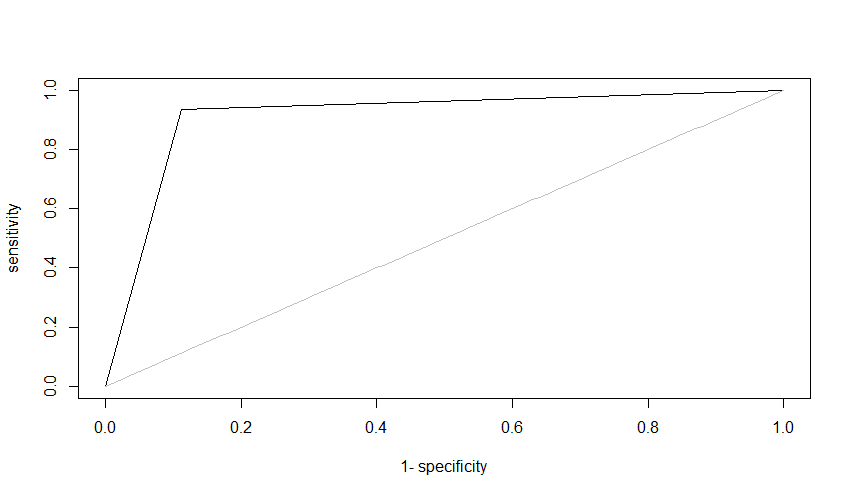
Next, we use a Decision tree to build our model.



We use pruning and check for what size of the tree the cross-validation error is the least. Using the size of the tree with lowest cross validation error we build our decision tree.



Main variables affecting breast cancer: area\_worst, concave.points\_worst .



actual

fitted B M

B 95 4

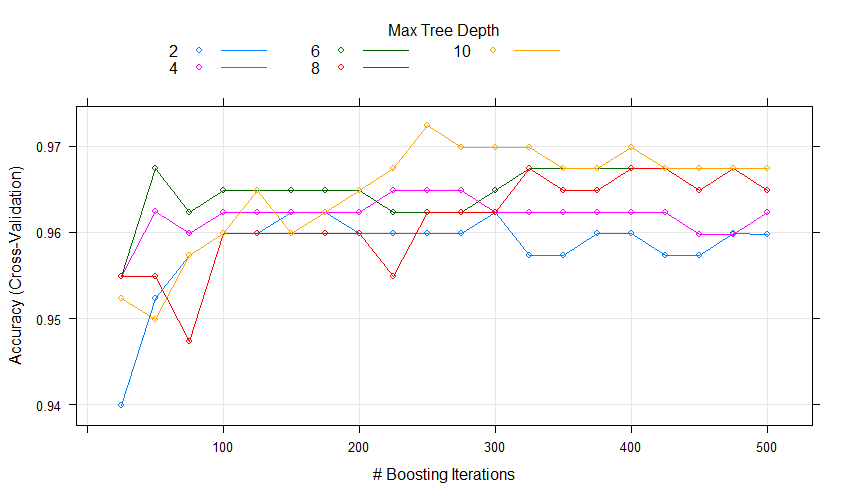
M 12 59

Accuracy = 90.58%

AUC = 0.9121792

**Boosting**

Next, we utilize Boosting to build our model. We build a gbm grid having all our test values for tree depth and Boosting iterations.



We plot the cross-validation error for different tree depths and boosting iterations to find the best parameters.

We then use the fitted model to get predictions and find the below top variables affecting Diagnosis.

**var rel.inf**

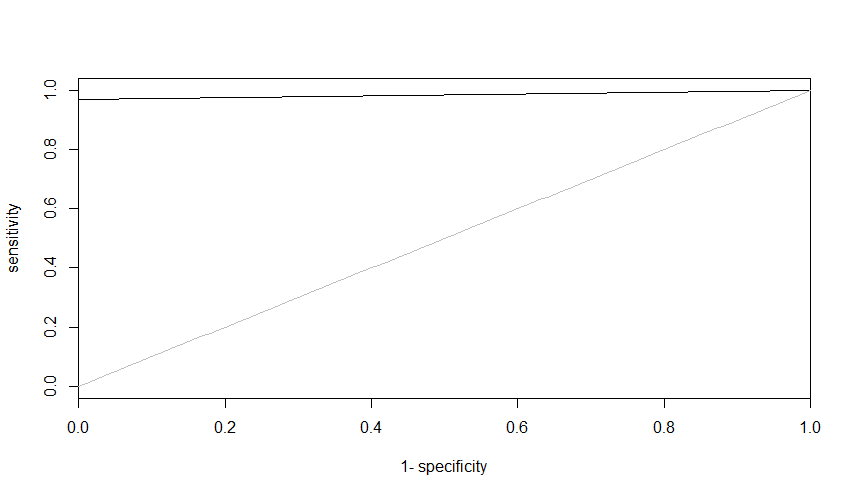
concave.points\_mean 23.51382018

area\_worst 19.63253806

perimeter\_worst 18.35509132

concave.points\_worst 13.50902538

area\_se 3.71962349



The boosting model has AUC = 0.984127

actual

fitted 0 1

0 107 2

1. 0 61

Accuracy = 98.82

**Comparison of the three Models**

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | AUC |
| SVM | 97.64% | 0.9747812 |
| Decision Tree | 90.58% | 0.9121792 |
| Boosting | 98.82% | 0.984127 |

Gradient Boosting provides the best result as it has the highest AUC