INDE 7397 - Engineering Analytics

Fall 2017

Breast Cancer Diagnostics

Department of Industrial Engineering

University of Houston

Instructor: Dr. Ying Lin

Submitted on December 8th, 2017

Project Report By

ARAVIND PALEMPATI - 1520220

JISNU PRABHU - 1530969

Contents

1. Abstract

Three machine learning methods are compared in this paper based on the overall accuracy, misclassification error, and the area under the ROC curve. These classification methods are used to predict the diagnosis result of the breast cancer which can be either benign or malignant in nature. To create the classifier, the WBCD (Wisconsin Breast Cancer Diagnosis) dataset is used. The dataset is widely utilized for this kind of application because it has many instances, it is virtually noise-free and has no case of missing values. Prior to the analysis, a large fraction of this work will be dedicated for cleaning and pre-processing the data to optimize the classifier. The highly correlated features are identified and dimension reduction is done on the dataset. The first part of this work is to overview the database, what information does it contains, when and how it was created, if it is noisy, if it has missing values. This section is important to understand what are the issues that will need to be processed while preparing the data to create the classifier. The second part is to propose the machine learning methods and algorithms to optimize the training set and different solutions are proposed in this paper. The results are presented in tables, which contains the accuracy of the classifier, and the area under the ROC curve. The analysis is done using R-programming, which is an open source programming language and software environment for statistical computing and graphics that can perform machine learning techniques like pre-processing, classification, regression, clustering and association rules. The best accuracy in this paper was achieved by the Support Vector Machines (SVM) algorithm, which had, in its best configuration, 97.65% of accuracy. The second algorithm tested was the KNN, which had 95.88% of accuracy.

2. Introduction

Breast cancer is the most common cancer among women and is one of the major causes of death among women worldwide. It is also seen in men, but most pre-dominant in women. As per statistical data on breast cancer, a man’s lifetime risk of breast cancer is about 1 in 1000, and is about 1 in 8 for women. More women are diagnosed with breast cancer than any other cancer. Every year approximately 124 out of 100,000 women are being diagnosed with breast cancer, and the estimation is that 23 out of the 124 women will die of this disease. When the cancer is detected in early stages, there is a 30% chance that the cancer can be treated effectively, but the late detection during advanced stage tumor makes the treatment more worse. The early detection of the breast cancer is vital and effective in prolonging the patient’s life and helping them to take proper treatment before the tumor gets worse. However, the process of invasive biopsy is time consuming costing huge amount of money and are subjected to human errors. This is where machine learning can help.

Currently, there are many techniques used to detect breast cancer in early stages. Some of the most used techniques are Mammography (77% sensitivity), Breast Ultrasound combined with Mammography (89% sensitivity), Fine Needle Aspiration (FNA) with visual interpretation (90%

sensitivity) and surgical biopsy (approximately 100% correctness). Though surgical biopsy is reliable, this treatment is invasive and costly and mammography and FNA with visual interpretation correctness varies widely.

This paper discusses the dataset collected by using FNA (Fine Needle Aspiration) technique with computational interpretation via machine learning and aims to create a classifier that provides a high level of accuracy, with a low rate of misclassification error. Several papers were published during the last 25 years trying to achieve the best performance for the computational interpretation of FNA samples, and in this paper three well-known machine learning techniques are tested: Random Forest, Supply Vector Machine (SVM), and K-Nearest Neighbors (KNN). It is a difficult task to build a classifier using machine learning techniques if the dataset used is not clear and interpreted correctly. Therefore, a considerable portion of this work will be spent to clean the dataset and avoid problems such as over-fitting. To pre-process the dataset, the caret package in R-programming will be explored to identify highly correlated features and separating them to prepare the training set before it can generate the classifier.

3. Dataset Overview

The Breast Cancer (Wisconsin) dataset used in this paper is publicly available in the UCI machine learning repository and was created by Dr. William H. Wolberg, Dr. W. Nick Street, and Olvi L. Mangasarian. The doner is W. Nick Street. The dataset is created by Dr. Wolberg by taking suspected tumor samples via a thin needle from patient's solid breast masses and the samples are placed on dent-shaped glass slides as shown in the Figure 1. The slide is dent shaped in order to distinguish the nearby cells with the tumor cells. The collected tissue samples are then examined under microscope and features are computed from the digitized image of a fine needle aspirate (FNA) of breast mass. This is done via a graphical computer program, which is capable of performing the analysis of cytological features based on a digital scan.

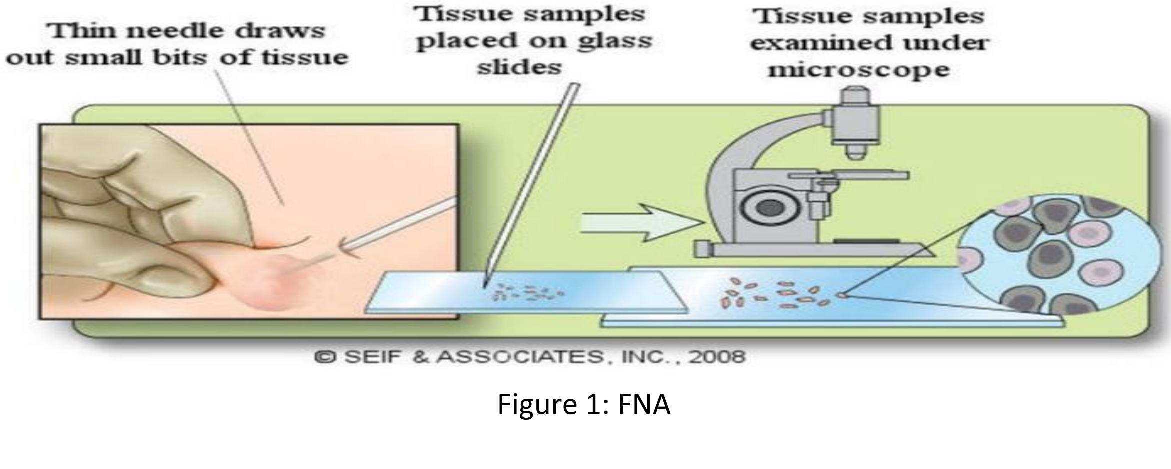

Thin needle draws out small bits of tissue
Tissue samples placed on glass slides
Tissue samples examined under microscope
© SEIF & ASSOCIATES, INC., 2008

Figure 1: FNA

The program uses a curve fitting algorithm to compute ten real-valued features for each cell nucleus and then, it calculates the mean value, extreme value(worst) and standard error of each feature for the image. The contrast between the benign and malignant diagnosis is shown is Figure 2.

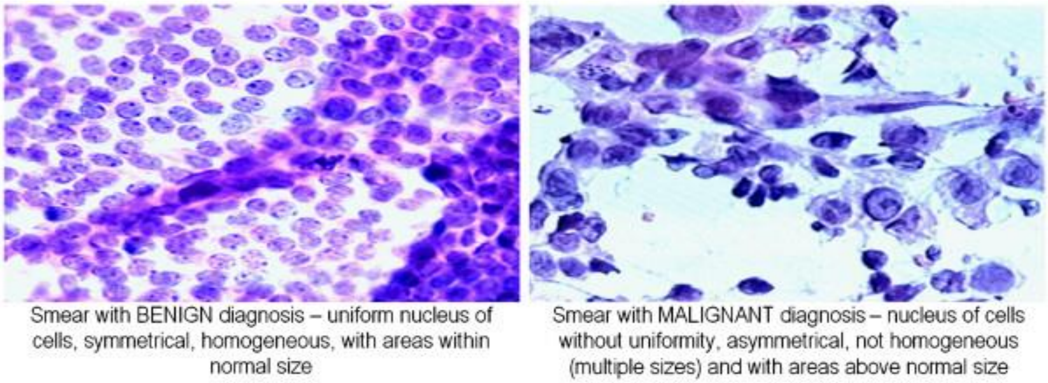

Smear with BENIGN diagnosis – uniform nucleus of cells, symmetrical, homogeneous, with areas within normal size
Smear with MALIGNANT diagnosis – nucleus of cells without uniformity, asymmetrical, not homogeneous (multiple sizes) and with areas above normal size

Figure 2: Images taken using FNA test (a) Benign and (b) Malign

The dataset is a classification problem consisting of 569 instances of which 357 of the observations are benign and 212 are malignant, where each one represents FNA test measurements for one diagnosis case. In this dataset, there are 32 attributes for each instance, where the first two attributes correspond to a unique identification number and the diagnosis status (benign / malignant). The rest 30 features are computations for ten real-valued features, along with their mean, standard error, and the mean of the three largest values ("worst" value) for each cell nucleus respectively. Table 1 summarizes the current state of the dataset used in the problem.

Cell Nuclei Characteristics (Features)	
1. Radius [mean of distances from center to points on perimeter]	Numeric
2. Texture [standard deviation of grey-scale values]	Numeric
3. Perimeter	Numeric
4. Area	Numeric
5. Smoothness [local variation in radius lengths]	Numeric
6. Compactness [((perimeter)2/area) – 1]	Numeric
7. Concavity [severity of concave portions of the contour]	Numeric
8. Concave points [number of concave portions of the contour]	Numeric
9. Symmetry	Numeric
10. Fractal dimension [“coastline approximation”-1]	Numeric

Table 1: Breast Cancer Wisconsin Dataset

4. Experiments

Cleaning and preparing data is a critical first step in any machine learning project. The objective is to structure the data to facilitate the data analysis you set out to perform. The data is examined and the features are examined. Two columns featured in the dataset named as id number, which represents a randomly generated field for unique identification of the patients, and X33 column, which consists of null values only are removed from the dataset, as they do not lay emphasis on predicting the diagnosis of the breast cancer. There are no missing values in the dataset, therefore there is no need to drop any observations. All the attributes are numerical data type apart from the diagnosis column which is the target column for modelling to predict the outcome as either benign(B) or malign(M). The goal of this paper is to predict which of the two values the diagnosis falls under (benign(B) or malign (M)), the problem can be treated as binary classification. The benign (B) values are transformed as 0 and the malign(M) values are transformed as 1.

The next part is pre-processing phase where feature scaling is done on the cleaned dataset using scale function in from R-programming language. This reduces the impact of data scale on the result. The next step is to identify and remove multi-collinearity; this is done using caret package in R-programming language. The findcorrelation function from caret package is used to identify the highly correlated features. To decide which attributes are more relevant for classification, a cut off value of 0.9 is given. After identifying the correlated features, dimension reduction is done on the dataset thereby reducing the attributes from 30 in the cleaned dataset to 21. The correlation plot is shown in Figure 3. Figure 3 is the correlation plot of 21 features which shows less amount of correlation among the attributes.

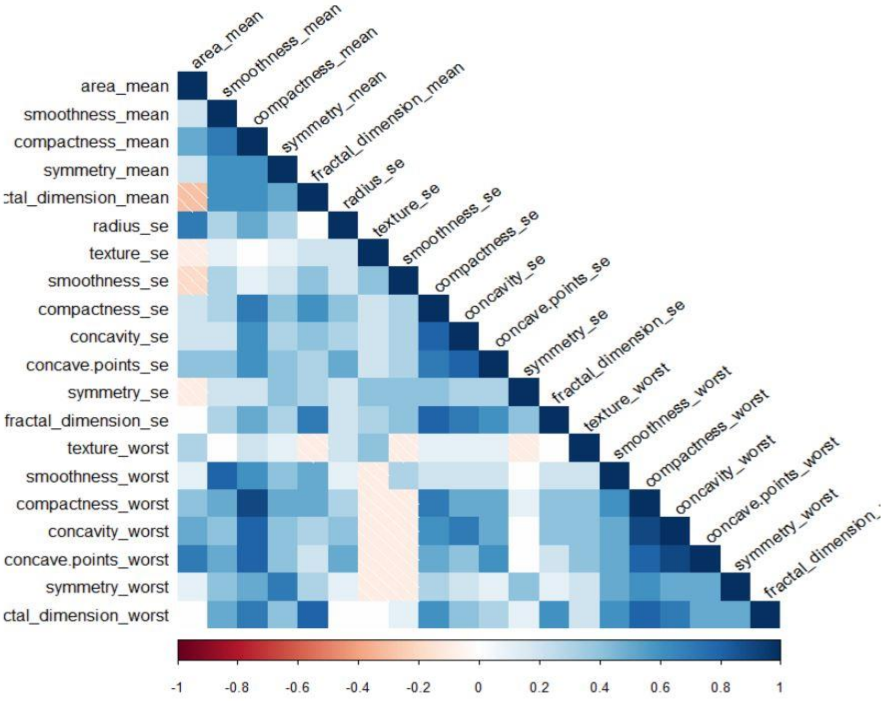


Figure 3: Correlation plot of 21 attributes

The data is slightly imbalanced with benign showing 63% (358 instances) and malign showing 37% (211 instances) on the diagnosis outcome. In this paper, this problem is ignored since there are only two outcomes which are classified either as benign (0) or malign (1). In order to implement machine learning methods on the new dataset, it is then divided into training and testing datasets where the training dataset contains 70% (399 instances) and the testing dataset consists of 30% (170) instances. Several machine learning classification models like the Random Forest, Supply Vector Machine (SVM), and K-Nearest Neighbors (KNN) are then implemented and are compared based on the overall accuracy, sensitivity, misclassification error, and area under the receiver operating characteristic (ROC) curve to select the best model that fits the dataset.

5. Experimental Results

To correctly analyze the results, it is important to keep in mind that for this application of machine learning, having an accurate classifier is as important as having a low rate of false-negative when classifying a malignant lump, because each instance miss-classified as a benign lump can delay the correct diagnosis and turn the treatment even more difficult.

5.1 Supply Vector Machine (SVM)

SVM for breast cancer detection utilized heuristics SVM approaches such as the smooth SVM, the linear SVM and general non-linear SVM. The first set of tests was made using the SVM algorithm, under SVM algorithm we have used three different kernel methods SVM Radical, SVM Linear and SVM Polynomial.

SVM Linear is the newest extremely fast machine learning algorithm for solving multiclass classification problems from ultra large data sets that implements an original proprietary version of a cutting plane algorithm for designing a linear support vector machine. Linear SVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time which scales linearly with the size of the training data set. The Radial basis function kernel, is a kernel that is in the form of a radial basis function (more specifically, a Gaussian function). The RBF kernel represents this similarity as a decaying function of the distance between the vectors (i.e. the squared-norm of their distance). The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized.

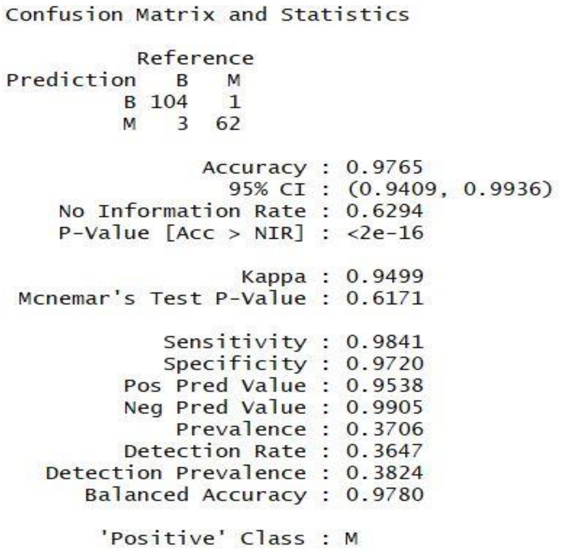
We trained the model and predicted the result using the test data. The observation of the results of each of the model is recorded is shown in Table 2.

SVM Methods	Accuracy	Sensitivity	Specificity
SVM Linear	0.9824	0.9841	0.9813
SVM Radical	0.9765	0.9841	0.9720
SVM Polynomial	0.9706	0.9524	0.9813

Table 2

Sensitivity: The probability of a positive test result if the disease is present
Specificity: The probability of negative test if the disease is not present
Probability of False Alarm = 1 - Specificity

The confusion matrix and statistics of all the different kernels is generated for the test set and the misclassification error rate for each of the kernels was noted. Figure 4 shows the confusion matrix of SVM Radial.



```
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 104   1
         M   3  62

               Accuracy : 0.9765
                 95% CI : (0.9409, 0.9936)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9499
 Mcnemar's Test P-Value : 0.6171

            Sensitivity : 0.9841
            Specificity : 0.9720
         Pos Pred Value : 0.9538
         Neg Pred Value : 0.9905
             Prevalence : 0.3706
         Detection Rate : 0.3647
   Detection Prevalence : 0.3824
      Balanced Accuracy : 0.9780

       'Positive' Class : M
```

Figure 4: SVM Radial confusion matrix

It can be observed from Figure 4, the SVM Radial model when implemented on the test set produced an accuracy of 97.65%, sensitivity of 98.41%, Kappa value of 0.9499, and the misclassification errors rate is found to be 2.35%.

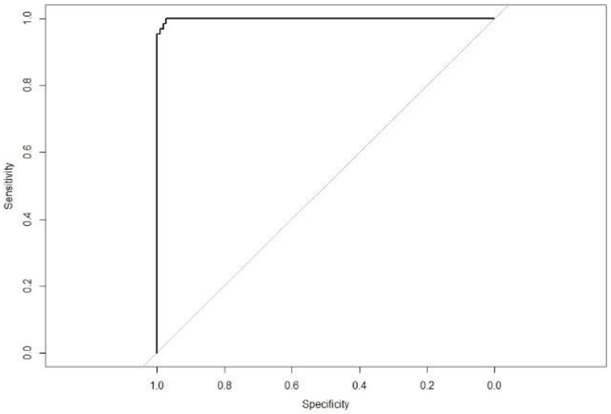


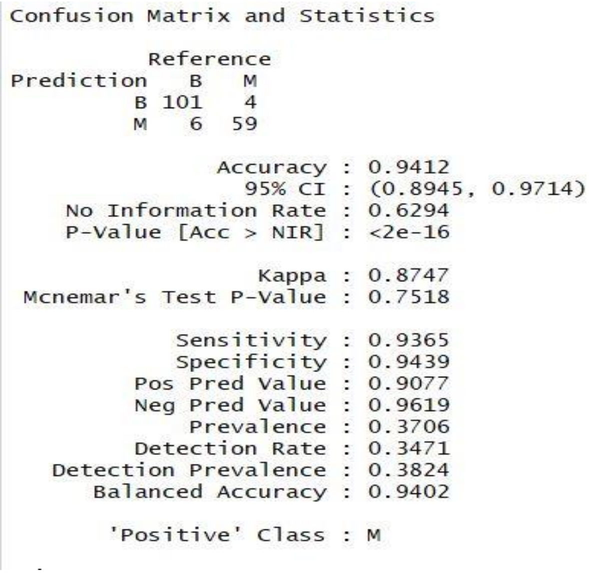
Figure 5: SVM ROC curve

The ROC curve has been plotted for each of the three models of which ROC curve of SVM Radial is included here as shown in Figure 5. The area under the curve in this ROC curve is 0.9991.

5.2 Random Forest (RF)

RF brings together many decision trees to ensemble a forest of trees. One of the primary reasons behind using RF in cancer detection is its ability to handle data minorities. For example, a tumor can be classified as either benign or malignant, despite the latter class is only 10% of the input data set. The RF method is based on a recursive approach in which every iteration involves picking one random sample of size N from the data set with replacement, and another random sample from the predictors without replacement. Then the data obtained is partitioned. The out-of-bag data is then dropped, and the above steps repeated many times depending on how many trees are needed. Finally, a count is made over the trees that classify the observation in one category and in the other. Cases are then classified based on a majority vote over the decision trees.

The confusion matrix and statistics of random forest is generated for the test set and the misclassification error rate was noted. Figure 4 shows the confusion matrix of Random Forest model.



```
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 101   4
         M   6  59

               Accuracy : 0.9412
                 95% CI : (0.8945, 0.9714)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8747
 Mcnemar's Test P-Value : 0.7518

            Sensitivity : 0.9365
            Specificity : 0.9439
         Pos Pred Value : 0.9077
         Neg Pred Value : 0.9619
             Prevalence : 0.3706
         Detection Rate : 0.3471
   Detection Prevalence : 0.3824
      Balanced Accuracy : 0.9402

       'Positive' Class : M
```

Figure 5: RF confusion matrix

It can be observed from Figure 5, the Random Forest model when implemented on the test set produced an accuracy of 94.12%, sensitivity of 93.65%, Kappa value of 0.8747, and the misclassification errors rate is found to be 5.88%.

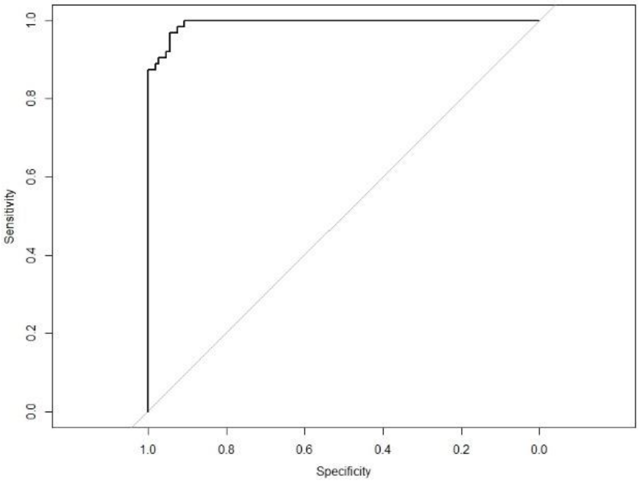


Figure 6: RF ROC curve

The ROC curve for Random Forest(RF) model is plotted as shown in Figure 6 and the area under the curve (AUC) is known. It is observed that the AUC value of this model is 0.9932.

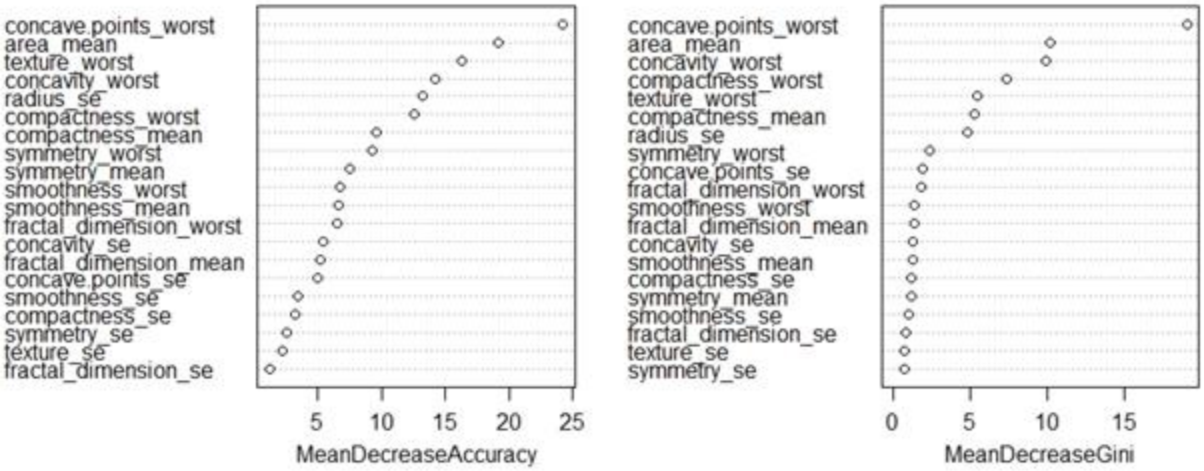


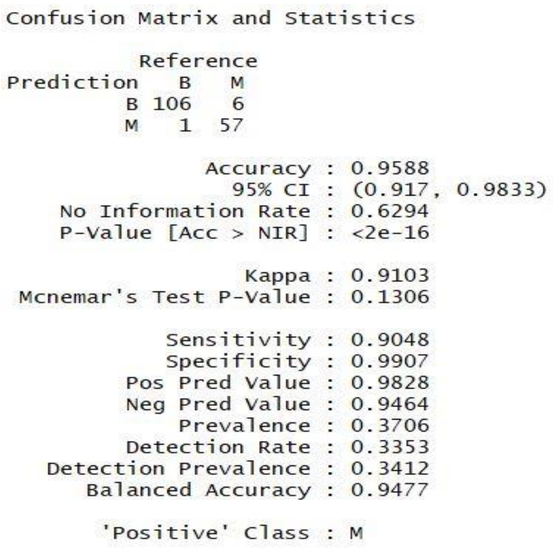
Figure 7: Feature Importance

Figure 7 shows the feature importance graph. As shown in the figure, the top 6 features having suffix as worst is considered important in predicting the diagnosis of breast cancer as Benign (B) or Malign (M). Usually the malignant cells have extreme values of the nuclei with area way above the normal size and therefore it makes sense that the attributes with suffix worst (contains extreme values) is considered as important. So concave.points_worst is the most important feature and the features with suffix as se (standard error) has least importance.

5.3 K-Nearest Neighbors (KNN)

K Nearest Neighbors (KNN) is a non-parametric method used for classification and regression analysis. KNN is an instance-based learning or lazy learning type, where the function is approximated locally and all computation is deferred until classification is done. The KNN algorithm is one of the simplest of all machine learning algorithms. The KNN algorithm finds a group of K nearest neighbors in the training set which are the closest to the test object, and the assignment of the class to the test object is based on a majority vote from the K nearest neighbors.

The confusion matrix and statistics of KNN model is generated for the test set and the misclassification error rate was noted. Figure 8 shows the confusion matrix of K Nearest Neighbors model



```
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 106   6
         M   1  57

               Accuracy : 0.9588
                 95% CI : (0.917, 0.9833)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9103
 Mcnemar's Test P-Value : 0.1306

            Sensitivity : 0.9048
            Specificity : 0.9907
         Pos Pred Value : 0.9828
         Neg Pred Value : 0.9464
             Prevalence : 0.3706
         Detection Rate : 0.3353
   Detection Prevalence : 0.3412
      Balanced Accuracy : 0.9477

       'Positive' Class : M
```

Figure 8: KNN confusion matrix

It can be observed from Figure 8, the KNN model when implemented on the test set produced an accuracy of 95.88%, sensitivity of 90.48%, Kappa value of 0.9103, and the misclassification errors rate is found to be 4.12%.

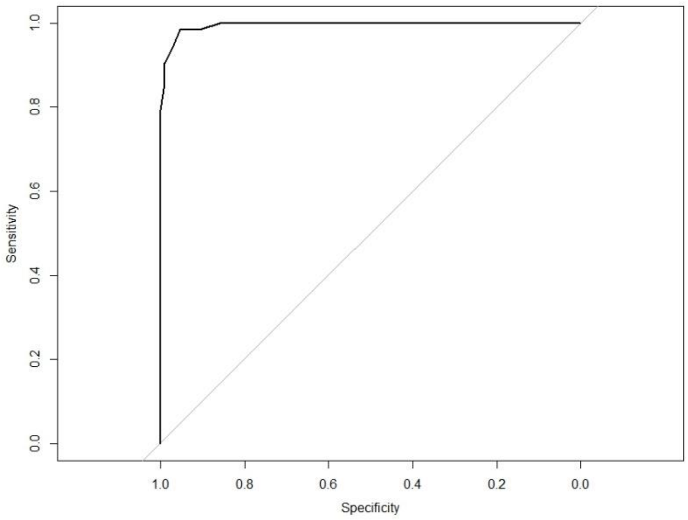


Figure 9: KNN ROC curve

The ROC curve for K Nearest Neighbors (KNN) model is plotted as shown in Figure 9 and the area under the curve (AUC) is known. It is observed that the AUC value of this model is 0.995.

6. Conclusion:

Early detection of breast cancer cells can be predicted accurately by the use of machine learning techniques. This may result in the decrease of health cost and may enhance time required for a patient to receive treatment. In this project the SVM, KNN and Random Forest have been discussed in providing diagnostic assessment for breast cancer. The SVM has been determined to be more superior compared to the other models since it provides higher prediction accuracy, higher sensitivity, higher Kappa value, and lower misclassification error rate as shown in Table 3. The prediction accuracy of the networks discussed in this project emphasizes the need of employing the machine learning techniques not only on the prediction of breast cancer data but on other medical conditions in which predictions of conditions are difficult to diagnosis.

Classification	**Accuracy**	**Misclassification**	**Sensitivity**	**Kappa**
SVM	97.65%	2.35%	98.41%	94.99%
RF	94.12%	5.88%	93.65%	87.47%
KNN	95.88%	4.12%	90.48%	91.03%

Table 3: Summary of classification methods

7. Reference:

1. Hsu, C., Chang, C., and Lin, C., “A practical guide to support vector classification”, Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003.http//www.csie.ntu.edu.tw/~cjlin/libsvm/.
2. Akay, M., “Support vector machines combined with feature selection for breast cancer diagnosis”, Expert systems with applications, Vol.36, 2009, pp.3240-3247.
3. UCI Irvine machine learning repository, “Breast Cancer Wisconsin (Diagnostic) Data Set”, http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), Nov. 1995.
4. Olvi, L.M., Street, W.N., “Breast cancer diagnosis and prognosis via linear programming”, Operations Research, Vol.43, No.4, 1995, pp. 570-577.
5. Gunn, S., “Support vector machines for classification and regression, Technical paper, 1998.
6. Ubyeli. E., “Implementing automated diagnostic systems for breast cancer detection, Expert systems with application”, Expert systems with applications, Vol.33, 2007, pp.1054-1062.
7. McMorran, J., Crowther., D.C., “Fine needle aspiration cytology (breast)”, General Practice Notebook – a UK medical reference on the world wide web, Feb 2009.

8. Appendix:

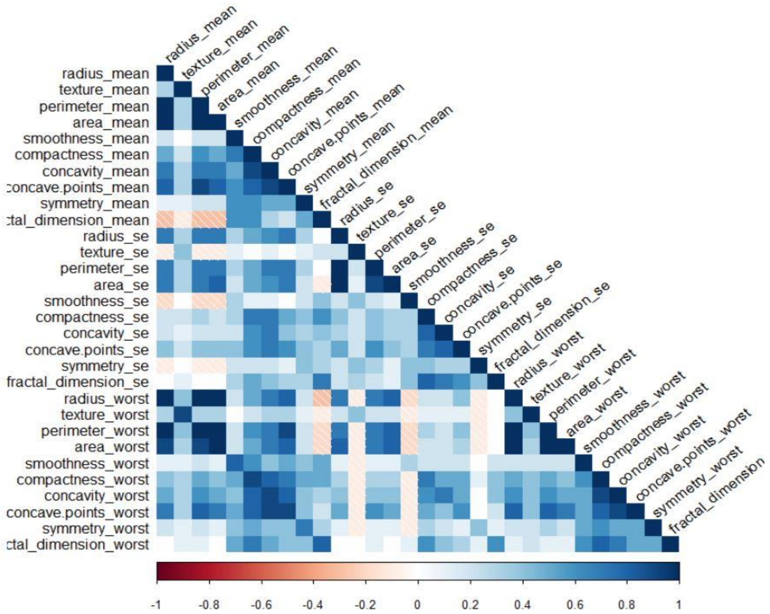
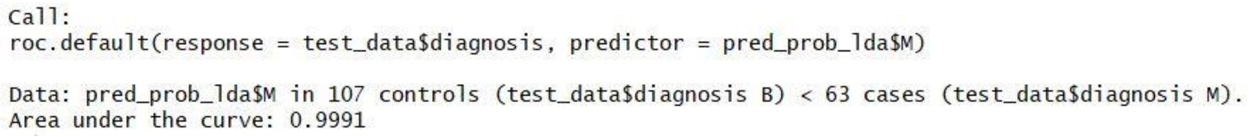


Figure 10: Correlation plot

The above figure is the correlation matrix plotted against the original dataset. From the figure, we can see that there exists higher amount of correlation among some features which is represented as dark blue shaded blocks in the figure. This features are removed and the new correation plot is shown in Figure 3.

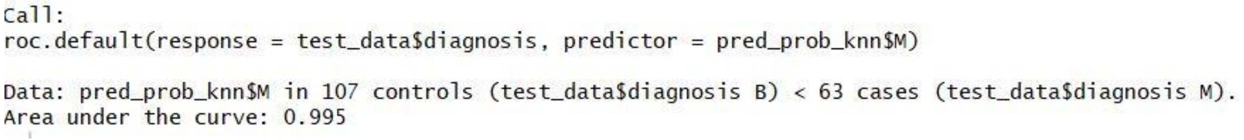
SVM Summary:



```
Call:
roc.default(response = test_data$diagnosis, predictor = pred_prob_lda$M)

Data: pred_prob_lda$M in 107 controls (test_data$diagnosis B) < 63 cases (test_data$diagnosis M).
Area under the curve: 0.9991
```

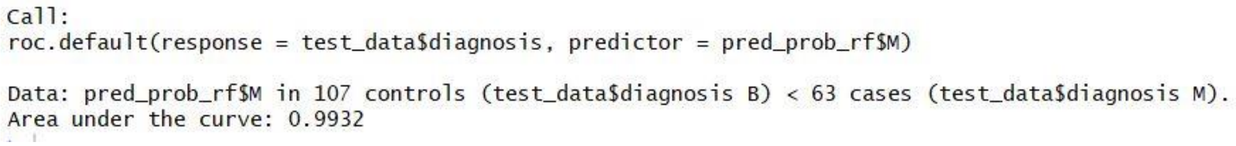
KNN Summary:



```
Call:
roc.default(response = test_data$diagnosis, predictor = pred_prob_knn$M)

Data: pred_prob_knn$M in 107 controls (test_data$diagnosis B) < 63 cases (test_data$diagnosis M).
Area under the curve: 0.995
```

Random Forest Summary:



```
Call:
roc.default(response = test_data$diagnosis, predictor = pred_prob_rf$M)

Data: pred_prob_rf$M in 107 controls (test_data$diagnosis B) < 63 cases (test_data$diagnosis M).
Area under the curve: 0.9932
```

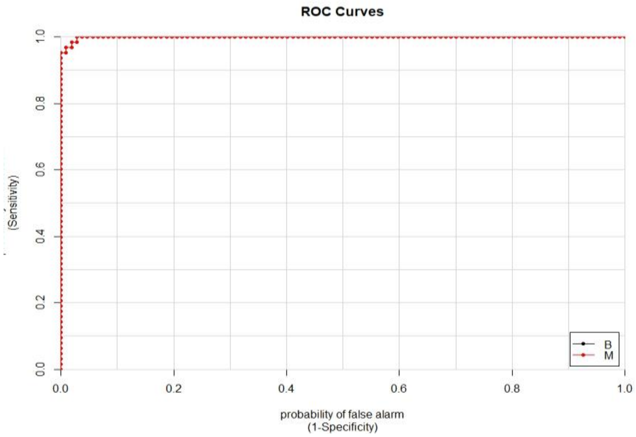


Figure 11: SVM AUC

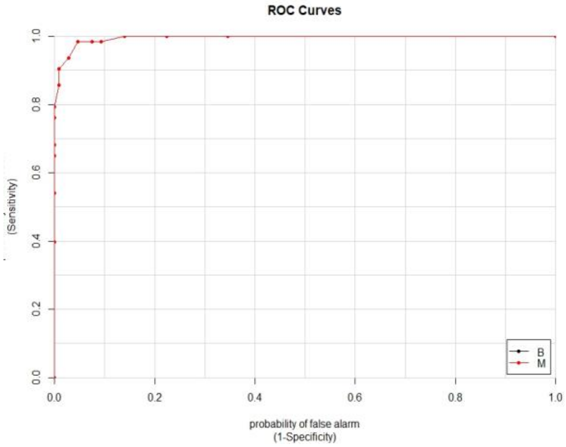


Figure 12: KNN AUC

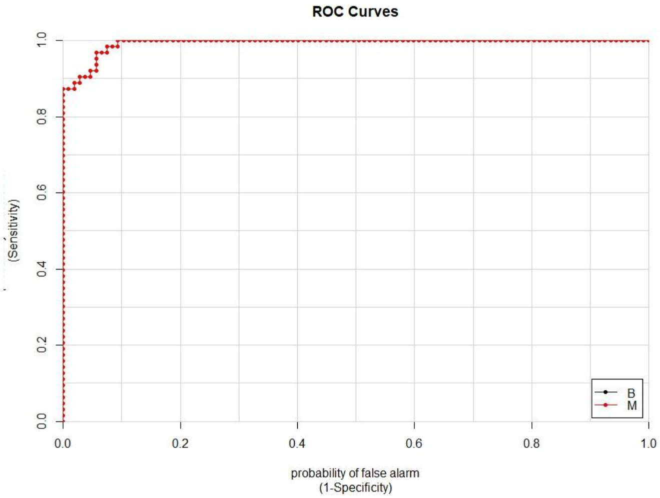


Figure 13: RF AUC

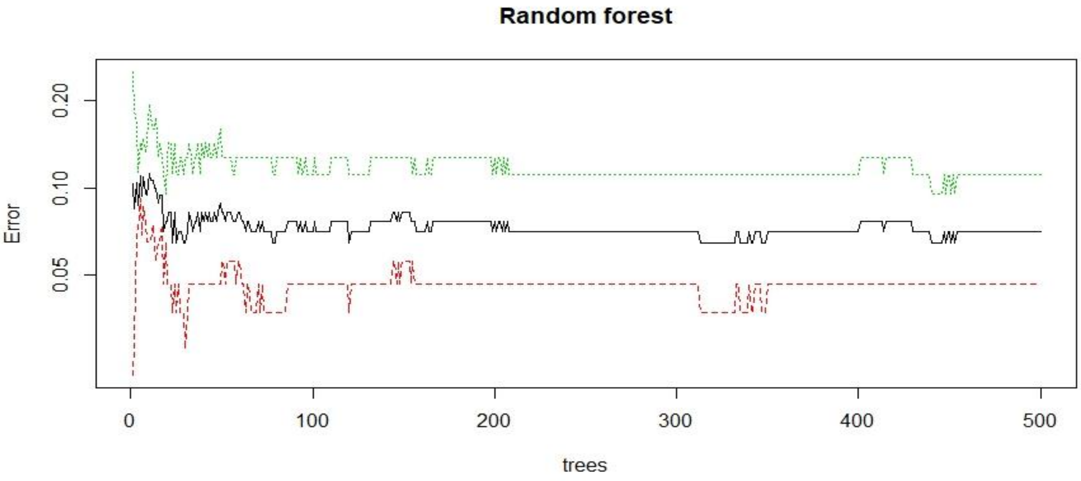


Figure 14: RF error