Prediction of Breast Cancer Using Data Mining and Multi Variate Techniques

Maryam Soltanpour Gharibdousti, Dieudonne N Ouedraogo, Syed M Haider, Susan Lu

Abstract

Breast cancer is a serious disease that affect females around the globe. The goal of our study is to use the statistical tools and multivariate techniques plus some machine learning models to try to predict the disease based on some relevant information. We will be using the Breast Cancer Wisconsin Diagnostic Dataset obtained from the UCI Irvine Machine Learning Repository. The dataset contains 699 observations with 11 Features that were obtained from digitized image of fine needle aspirates (FNA) of breast masses; Those features are, Sample ID, Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, Single epithelial cell size, Number of bare nuclei, Bland chromatin, Number of normal nuclei, Mitosis and Classes (diagnosis). The dataset requires some cleaning, after this step we will explore the correlation between features and attempt to use principal component analysis; We will then use appropriate statistical methods learned in class to refine our study, then we will explore different machine learning techniques outside the scope of the course to build suitable models for predicting the disease, at the evaluate the accuracy of the prediction models are and it is concluded that, Discriminate Analysis and a hybrid selection of Discriminate Analysis and Logistic Regression are best methods for feature selection and also Naïve Bayes and Super Vector Machines outperform other methods.

1. INTRODUCTION

Breast cancer is one of the leading cancer among women It diagnosis involves series of medical tests including initial breast exam, mammograph, biopsy, ultrasound and MRI. The correct diagnosis at early stage significantly increases the survival rate of the patient. The biomedical researchers are exploring ways to determine the automatic data driven diagnosis. The procedure of finding hidden and unidentified patterns and trends in big datasets, extracting information from them and building predictive models is defined as data mining. In another word, it's the process of collection and exploration of data sets and building models by huge data stores to expose previously unknown outlines.

Due to complexity and vagueness of data engendered by healthcare transactions, it is impossible to analyze them with traditional tools. In order to make the decision-making process easier and more trustable, data mining techniques are provided to transmute these data into useful information and makes it feasible to get useful results and patterns and trends out of these huge amounts of data. Data mining has been widely used in many areas. One of these areas which is using it even more and more as an essential tool is healthcare management. All agents in a healthcare industry can significantly benefit Data mining applications. Data mining is not new.it has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling.

The purpose of this paper is to apply data mining techniques in predicting the breast cancer tumor by reducing the number of features using various multivariate analysis including Principle Component Analysis, Discriminant Analysis. The model created with reduced variables is tested by performing classification using support vector machine, Naive Bayes, Decision Tree and Artificial neural network. The performance of classifiers then is compared based on their accuracy and ROC and the best prediction model is selected.

2. LITERATURE REVIEW

Data mining is extensively used in the diagnosis of the breast cancer. The medical facilities have an enormous amount of data available. The information hidden in data can be utilized to help the doctors in correctly diagnosing the disease at an early age. Mammography is one of the most extensively used method in detecting breast tumor. If the tumor is detected by the mammography, further diagnostic or invasive technique is required to determine whether the tumor is malignant or benign. The breast cancer data consist of large number of features [1]. Some features are irrelevant or multicollinear that may cause the classification model to decrease its precision [2]. Feature selection is a very important preprocessing step in data mining and machine learning [3].Feature selection is performed to reduce the number of variables and determine the significant factors in the diagnostic step.

Statistically only 20-30 percent biopsy cases are found to be actual cancerous. Hence a lot of research has been conducted in last years to classify the tumor correctly. Vural and Wang proposed classification of breast cancer thru somatic mutation profiles by examining the exome sequencing data [4]. The dataset is tested with five different machine learning algorithms and the best 10-fold cross validation accuracy of 70.86% was achieved by random forest.

Artificial intelligence techniques such as artificial neural network has been used in breast cancer diagnosis with a great success. [5] [6] [7]. HA applied evolutionary multi-objective approach to artificial neural network and achieved 98.1 % accuracy with 50 % reduced computational cost as compared to traditional back propagation [8] . H.A Abbas and M.Towsey uses traditional back propagation algorithm and achieved an accuracy of %97.5 [9]

Sou Jin et al concluded to have better results using two binary classifiers with Naïve Bayes and Functional Trees (FT) as compared to multiclass classifier (one-step classifier) for predicting diagnosis and prognosis of Breast cancer [10]. Hasan and Mediha proposed hybrid GSA (Genetic Algorithms and Sim.ulated Annealing) and accomplished an accuracy of %98 [11]. Sahan and Polat introduces hybrid KNN algorithm along with data reduction using artificial immune system (AIS) technique. Their proposed Fuzzy-AIS-KNN achieved an accuracy of % 99.

Carlos and Moshe Sipper [12] uses fuzzy-genetic approach to automatically produce systems for diagnosis Singh. Selvi [13] worked on new classification approach for breast cancer tumors in n digital mammograms using Particle Swarm Optimized Wavelet Neural Network (PSOWNN). Detlef and Rudolf [14] applied rule based neuro fuzzy classification approach. Alic and Subasi

[15] perform genetic algorithm based feature selection and classify the data using logistic regression, Bayesian network, multilayer perceptron, random forest and support vector machines. Fatih [16] perform support vector machine method coupled with feature reduction.

3. Data Description

The data used in this research is from UC Irvine Machine Learning Repository. The data is collected by researchers in University of Wisconsin, Clinical Sciences Center between at 1995-11-0. The data set includes nine features and two classes as Bengine and Malignant. 699 observations are available which some cells are missing values. Next, data will be preprocessed and cleaned to be ready to apply statistical and data mining techniques on it.

4. Experimental Analysis

In order to clean the data set, first missing values should be filled with appropriate values. Dropping the missing values is doable, but since the dataset is not that large, it can affect the final results. Also, imputing raise uncertainty. There are different ways to cover missing values like using the mean or mode of each column. Since, fulfilling the missing cells is not accurate by using mean or mode, the distribution of each variable in the dataset is identified the missing data are filled based on those distributions. By using this technique, effect of bias in the results will be minimized. This method is considered as a heuristic algorithm which imputes missing values in a dataset without inserting much bias. The next step is to standardize the data set which is not needed here, because all variables are in the same scale from 1 to 10. Figure 1 shows how the data set in unbalanced the class of Bengine, which easily lead to prediction bias, because the prediction model will tend to predict the class with more observations and accuracy measure in that case could not be fully trusted. The cases are grouped it is known that the prior probabilities by then number of benign and malignant in the dataset. There are total of 444 benign cases and 239 malignant cases, hence the data is not balanced.

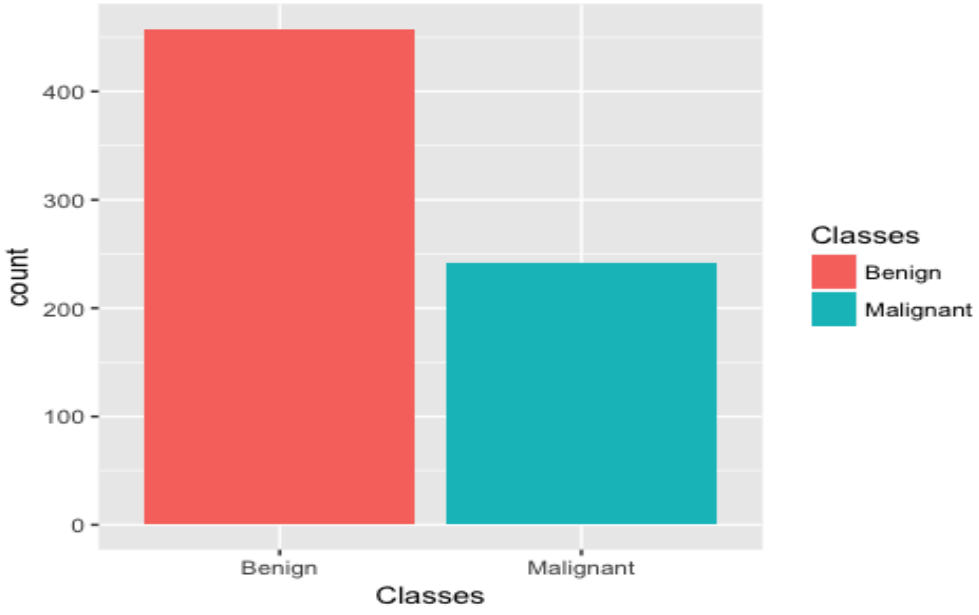

Figure 1. The distribution of data based on their classes

4.1. Correlation Between Features

Before starting building prediction models, always reading the correlation matrix is useful. If the variables are dependent, it is better to remove independent variables and reduce the dimension of the data set. Table 1-4 shows the correlativity of variables regarding each other. Table 1 shows the correlativity of all variables regarding Clum Thickness and Uniformity of Cell Size. Tables 2 shows the correlativity regarding Uniformity of Cell Shape and Marginal Adhesion. Table 3 represents the correlativity in regard of Bland Chromatin, Normal Nucleoli and Mitosis.

As it is seen, Except the two variables, uniformity of cell size and uniformity of cell shape, other variables are not highly correlated. Still, since there is %50 correlativity among some variables, feature selection methods are applied at next step to see if prediction models perform better with reduced number of variables or not.

Table 1. Correlativity among variables

	ClumpThickness	UniformityOfCellSize
ClumpThickness	1	0.6449
UniformityOfCellSize	0.6449	1
UniformityOfCellShape	0.6546	0.9069
MarginalAdhesion	0.4864	0.7056
SingleEpithelialCellSize	0.5218	0.7518
BareNuclei	0.5943	0.6952
BlandChromatin	0.5584	0.7557
NormalNucleoli	0.5358	0.7229
Mitosis	0.35	0.4587

Table 2. Correlativity among variables

	UniformityOfCellShape	MarginalAdhesion
ClumpThickness	0.6546	0.4864
UniformityOfCellSize	0.9069	0.7056
UniformityOfCellShape	1	0.6831
MarginalAdhesion	0.6831	1
SingleEpithelialCellSize	0.7197	0.5996
BareNuclei	0.7159	0.671
BlandChromatin	0.7359	0.6667
NormalNucleoli	0.7194	0.6034

Mitosis	0.4389	0.4176

Table 3. Correlativity among variables

	SingleEpithelialCellSize	BareNuclei
ClumpThickness	0.5218	0.5943
UniformityOfCellSize	0.7518	0.6952
UniformityOfCellShape	0.7197	0.7159
MarginalAdhesion	0.5996	0.671
SingleEpithelialCellSize	1	0.5875
BareNuclei	0.5875	1
BlandChromatin	0.6161	0.6826
NormalNucleoli	0.6289	0.5926
Mitosis	0.4791	0.3359

Table 4. Correlativity among variables

	BlandChromatin	NormalNucleoli	Mitosis
ClumpThickness	0.5584	0.5358	0.35
UniformityOfCellSize	0.7557	0.7229	0.4587
UniformityOfCellShape	0.7359	0.7194	0.4389
MarginalAdhesion	0.6667	0.6034	0.4176
SingleEpithelialCellSize	0.6161	0.6289	0.4791
BareNuclei	0.6826	0.5926	0.3359
BlandChromatin	1	0.6659	0.3442
NormalNucleoli	0.6659	1	0.4283
Mitosis	0.3442	0.4283	1

In order to visually see the correlation between variables the correlation plot is shown in Figure Darker means variables are more correlated.

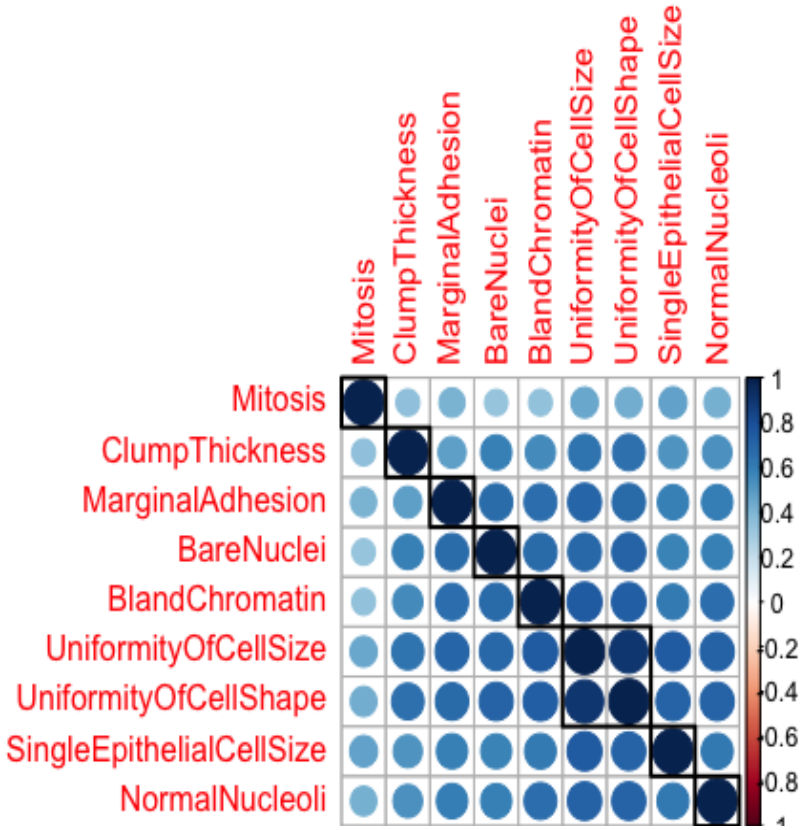

Figure 2. Correlation plot of variables

The breast cancer patient can either have malignant or benign tumor and the multivariate analysis can be used to reduce the number of independent variables. Therefore, the next step is to identify the variables that contribute most toward the classification or identification of diseases type In order to explore the variable reduction, various multivariate techniques are applied including principle component analysis, discriminant analysis, logistic regression and cluster analysis. So, variables that have correlation above 0.85 will be removed to reduce the dimension.

4.2 Principal Component Analysis

In this section, Principal Component Analysis is used to represent some new variables that are linear combination of original variables. As it is shown in Figure 3-4, two principal components could represent almost %67 of the total variance. Overall performance of PCA is not acceptable since one principal component is representing %69 and another one %7 of total variance. Figure 3 and Figure 4 show how original variables are covered with two new variables visually.

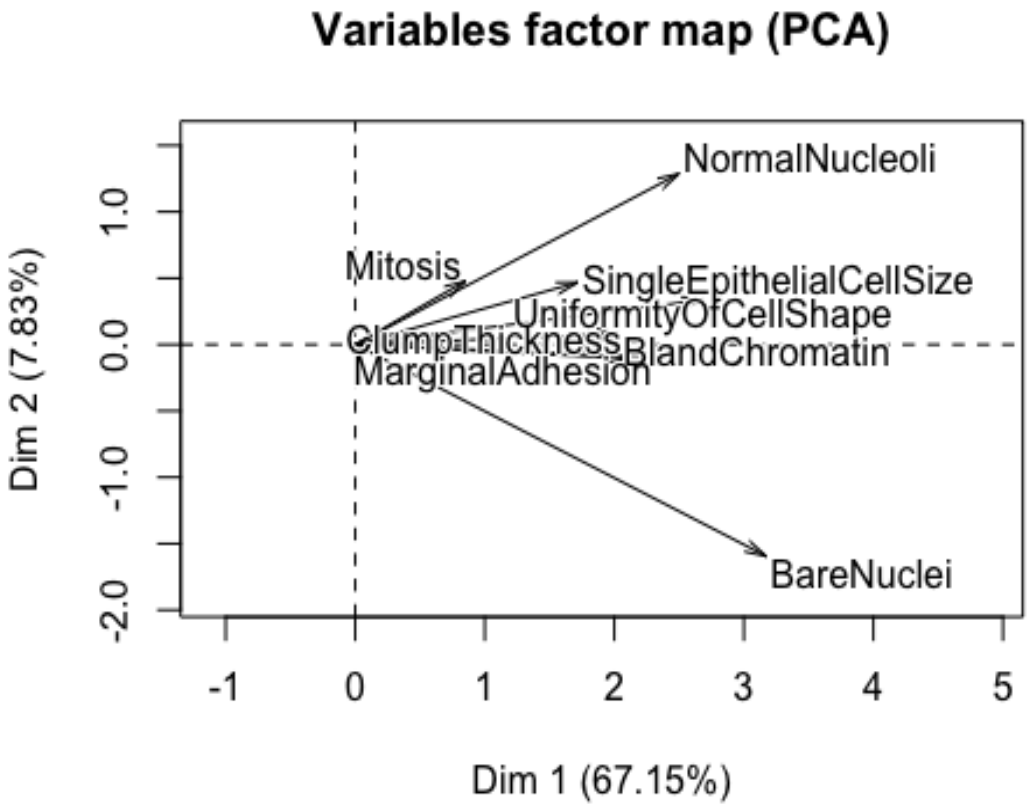

Figure 3. Coverage of riginal variables by PC1 and PC2

The data is well suited for PCA. The independents variables could be reduced to essentially two dimensions. In this paper, PCA will be applied on the original data, and also the filtered (where the highly-correlated variable is removed). From the output above in regard to the correlation between the new dimensions and the original variables it is concluded that (2 first) dimensions are:

- prin1(%69 of the variance). It represents theses variables: uniformity of cell size, uniformity of cell shape, bare nuclei, bland chromatin, normal nuclein, marginal adhesion, single epithelisl cell size, clumpthicness). Variables with correlation greater than 0.5 are retained.
- Prin2 (%7 of the total variance). it represents theses variables: the opposite of bare nuclei. since the correlation are pretty low, variables with correlation of -0.47 are retained.

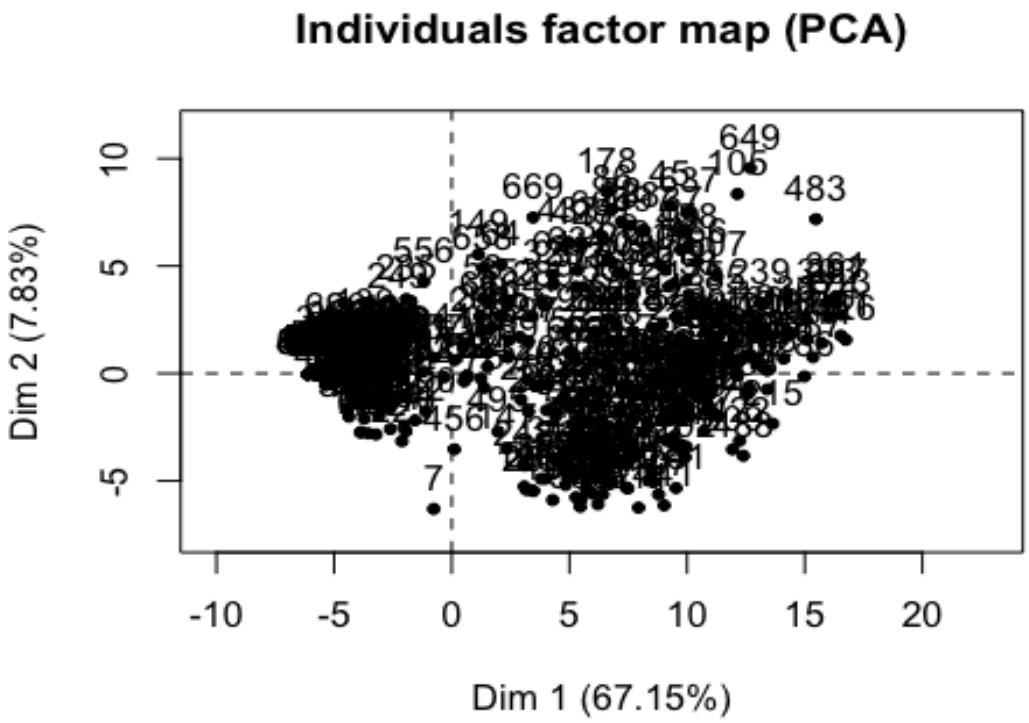

Figure 4. Representing the total variance

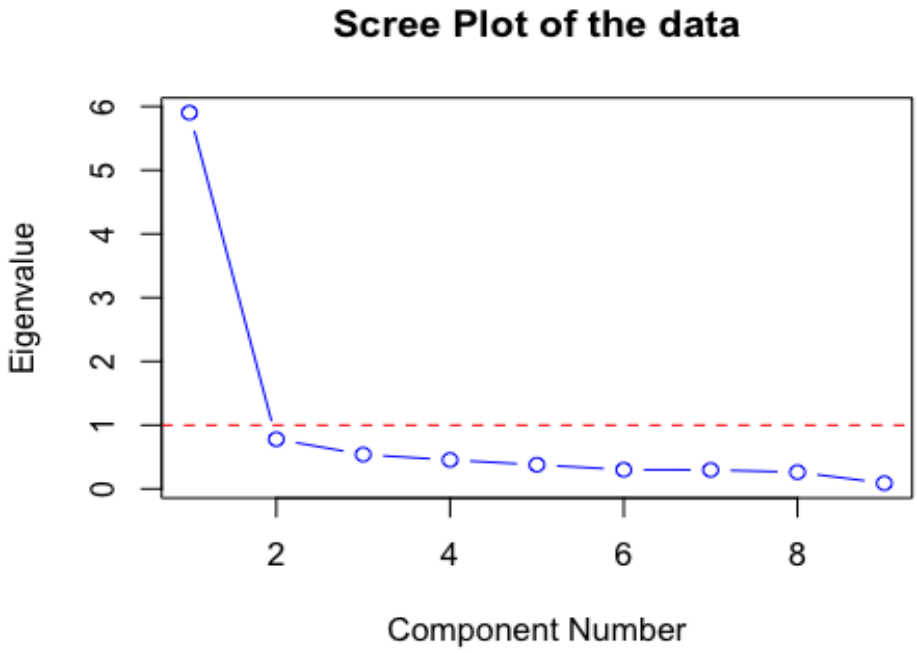

Figure 5. The loadings and the PCs

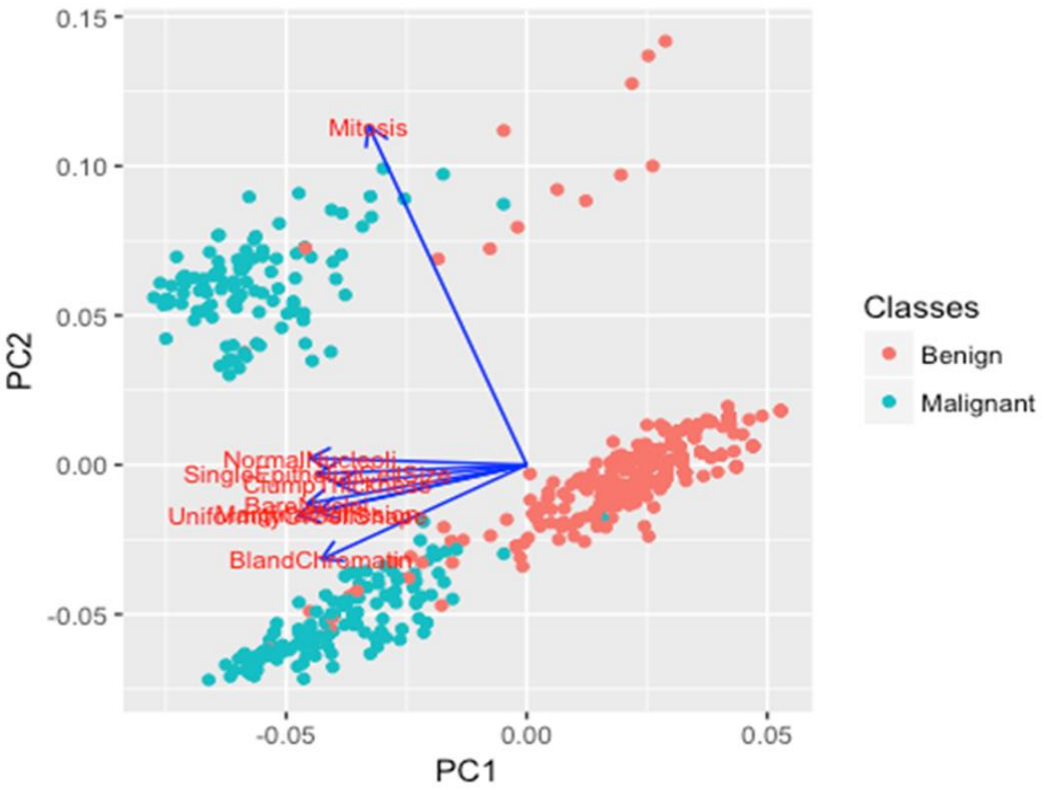

Figure 6. Coverage of original variables based on their classes

Figure 7 shows that the number of points above the red line determine the right number of PC, which means that two PCs mainly explain the variance of the dataset.

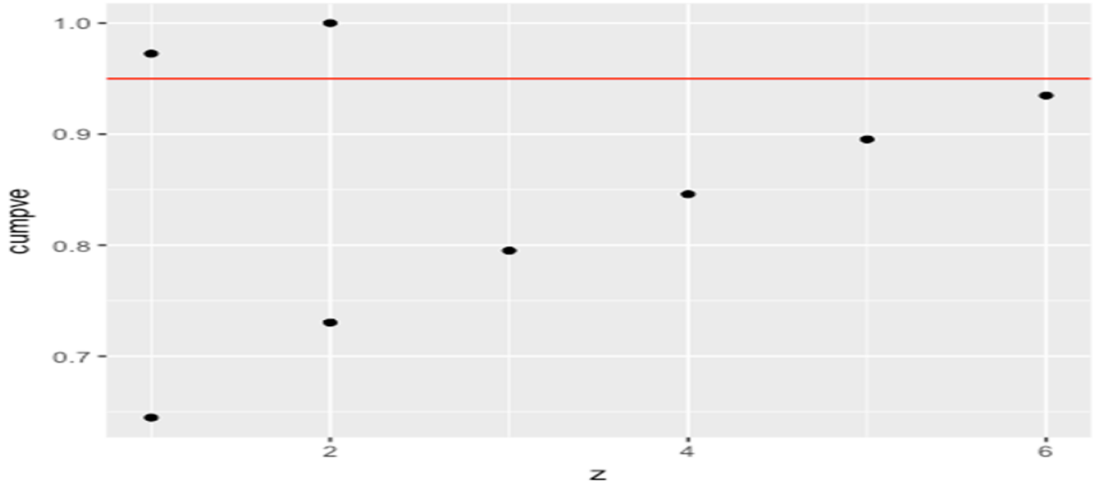

Figure 7. Number of PCs selected

The objective of the principle component analysis is to discover or reduce the dimensionality of the dataset. Therefore, now that two PCs are suitable for the dataset, it means,

any data point in our set could be projected into a 2-dimension orthogonal space with PC1 and PC2 as components and be clearly clustered (separated). Figure 8 also shows clustering the variables based on their classes of Benign and Malignant.

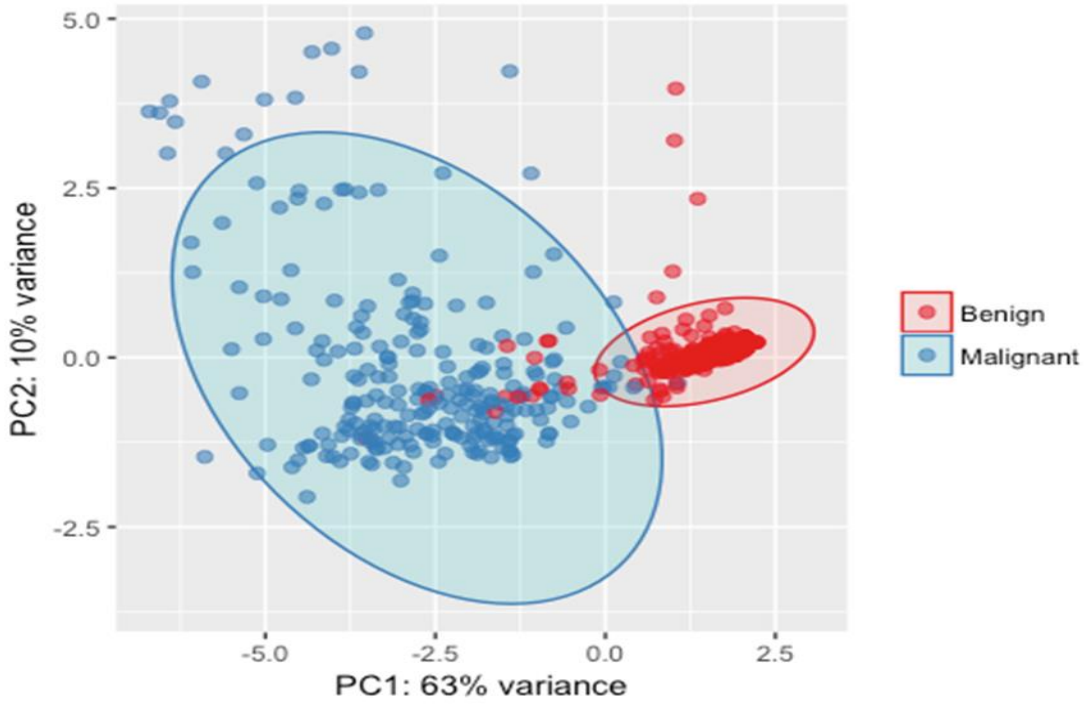

Figure 8. Clustering the variables based on classes

The principle component consists of a portion of all variables. The Eigen vectors also known as loading, demonstrate how much each variable contribute in the principle component. Higher the Eigenvector (loading), higher will be the contribution of that variable in the principle component.

Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
ClumpThickness	0.302063	-.140801	0.866372	0.107828	0.080321	-.242518	-.008516	0.247707	-.002747
UniformityOfCellSize	0.380793	-.046640	-.019938	-.204255	-.145653	-.139032	-.205434	-.436300	-.733211
UniformityOfCellShape	0.377583	-.082422	0.033511	-.175866	-.108392	-.074527	-.127209	-.582727	0.667481
MarginalAdhesion	0.332724	-.052094	-.412647	0.493173	-.019569	-.654629	0.123830	0.163434	0.046019
SingleEpithelialCellSize	0.336234	0.164404	-.087743	-.427384	-.636693	0.069309	0.211018	0.458669	0.066891
BareNuclei	0.335068	-.261261	0.000691	0.498618	-.124773	0.609221	0.402790	-.126653	-.076510
BlandChromatin	0.345747	-.228077	-.213072	0.013047	0.227666	0.298897	-.700417	0.383719	0.062241
NormalNucleoli	0.335591	0.033966	-.134248	-.417113	0.690210	0.021518	0.459783	0.074012	-.022079
Mitosis	0.230206	0.905557	0.080492	0.258988	0.105042	0.148345	-.132117	-.053537	0.007496

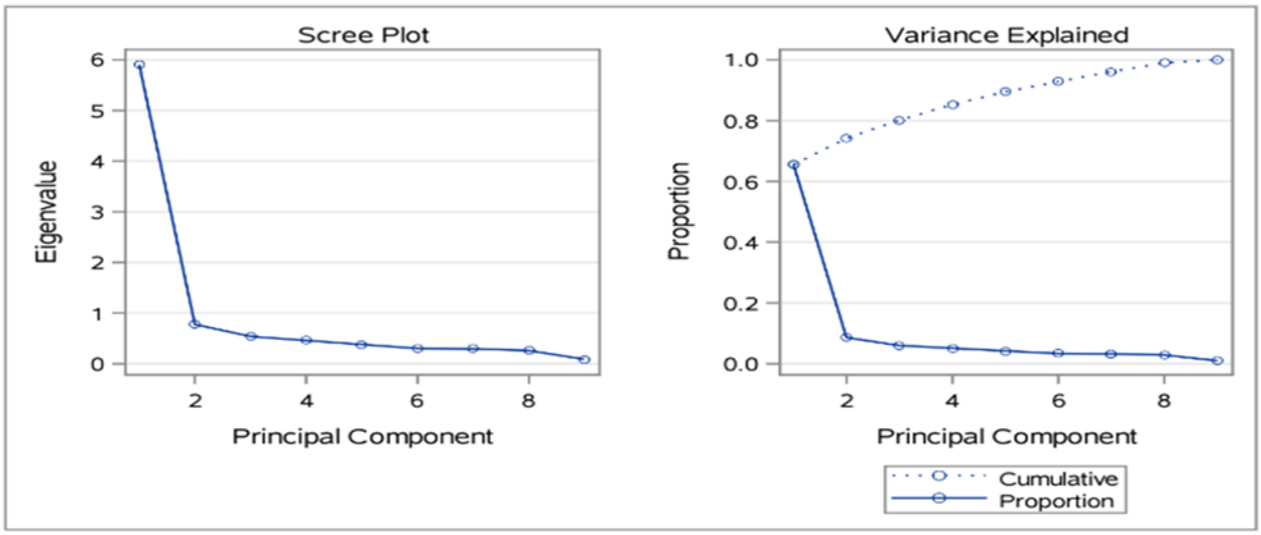

Figure 9. Loading of PCs

The PCA is not helpful in reducing the number of independent variables, though a result does indicate the presence of multicollinearity. The presence of multicollinearity interferes the precise effect of each predictor and makes the estimates very sensitive to minor changes in the model.

4.3 Stepwise Discriminant Analysis

Here it is attempted to use stepwise discriminant analysis to classify observations. The discriminant analysis has the multivariate normality assumption. If the data is the mixture of independent and dependent variables, the multivariate normality assumption will not hold. The objective of discriminant analysis is to identify the variables that discriminate best between the two groups.

The stepwise discriminant analysis is applied with significance level of entry and stay is set to 0.01. The variable with the largest F-statistics is selected first. There are total of six variables entered and at each step Wilks' lambda is calculated that determines how well each function separates cases into groups. Smaller values indicate greater discriminatory ability of the function. With each variable entered, the Wilk Lambda statistics value is evaluated and is found to be decreased as shown below. The variables are entered in the following order shown in Figure 10.

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	BareNuclei		0.6768	1426.24	<.0001	0.32317150	<.0001	0.67682850	<.0001
2	2	UniformityOfCellSize		0.3760	409.72	<.0001	0.20166426	<.0001	0.79833574	<.0001
3	3	ClumpThickness		0.1199	92.47	<.0001	0.17749135	<.0001	0.82250865	<.0001
4	4	NormalNucleoli		0.0700	51.01	<.0001	0.16507150	<.0001	0.83492850	<.0001
5	5	BlandChromatin		0.0276	19.23	<.0001	0.16051248	<.0001	0.83948752	<.0001
6	6	UniformityOfCellShape		0.0107	7.35	0.0069	0.15878713	<.0001	0.84121287	<.0001

Figure 10. Order of entering variables

The DA analysis reported that 6 variables meet the 0.01 significance level of entry. If the 7th variable is entered, the model will not meet the significance level of entry. It should be noted that 0.01 is pretty strict - We can increase up to 0.15 but then it will include all variable and will not justify variable reduction.

The discriminant analysis suggests that six variables discriminate best between the malignant and benign cases. The order of the variables is determined by the F-statistic score. The variables entered in the stepwise discriminant analysis will stay if their p-value is less than the significance level of entry. Similarly, the variables entered in the model will stay if their p-value of the overall model is less than the significance level of stay.

4.4 Logistic regression

In this section, full model and stepwise selection is explained. Here is the logistic regression approach, we start with a full model approach (all Variable include and later we use forward and backward method).

The logistic regression is recommended when the independent variables do not satisfy the multivariate normality test. The model converges with relative gradient convergence criterion (GCONV) with default precision in SAS. The stepwise logistic regression is applied with 0.05 significance level for entry. The model fit statistics Akaike Information and Schwarz Criterion are report. The chi-square test for likelihood ratio, Score and Wald p-value should be within the acceptable significance value which are shown in Figure 11 and Figure 12.

Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	BareNuclei		1	1	462.2739		<.0001
2	UniformityOfCellShap		1	2	180.4102		<.0001
3	ClumpThickness		1	3	30.0228		<.0001
4	BlandChromatin		1	4	16.6428		<.0001

Figure 11. Score and Wald p-value

Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
5	MarginalAdhesion		1	5	10.2061		0.0014
6	NormalNucleoli		1	6	5.2962		0.0214

Figure 12. Score and Wald p-value

The logistic regression analysis suggests that six variables are important that classify best between malignant and benign cases. If 0.01 significance level for entry is used. Then four variables will be selected in the logistic regression model. Unlike discriminant analysis, the logistic regression didn't include the variable "Uniformity of Cell Size", instead it selected the "Uniformity of Cell Shape". This could be explained by the fact that the both variables are highly collinear (0.9) so only one variable is needed to classify the dependent variable

5. Classification Methods

This section describes the proposed methodology for data mining from breast cancer dataset. As explained in earlier sections, the very first and important step is preprocessing and cleaning the data. Cleaning process is the process of filling the missing values base on either there are categorical or nominal data. second, five different machine learning methods, namely SVM, LR,

NN, DT and NB are applied on both original and normalized data set. Next, feature selection using LR and PCA feature selection based methods are used and at last performance measurement criteria are used to compare the utility of different techniques to each other "Logistic Regression" and "SVM" based feature selection technique is used for variable selection. As the parameter control ‘c’ is decreased fewer features are selected. Classification techniques are applied on all features and selected features for both original and normalized data. Data is split into training set and testing set. 70 % of the data is used for training the model and rest 30 % is used for testing. Different classification algorithms such Decision Tree, Logistic regression, Support Vector Machine, Naïve Bayes and Artificial neural networks are applied on the original data (with and without feature selection). Performance measures such accuracy, sensitivity, specificity and Area Under Curve (AUC) for Receiver Operating Characteristic (ROC) Curve are used to evaluate various techniques. Keeping the split percentage constant, training and testing data are selected five times randomly from the dataset and the average performance measures are reported. To visually look into the distribution of each variable and separate the case of training and testing (70% training and 30% testing). The data set is split into training and testing and you build the model on training and validate the results using testing dataset. Figure 13 shows how variables are distributed based on test and train data.

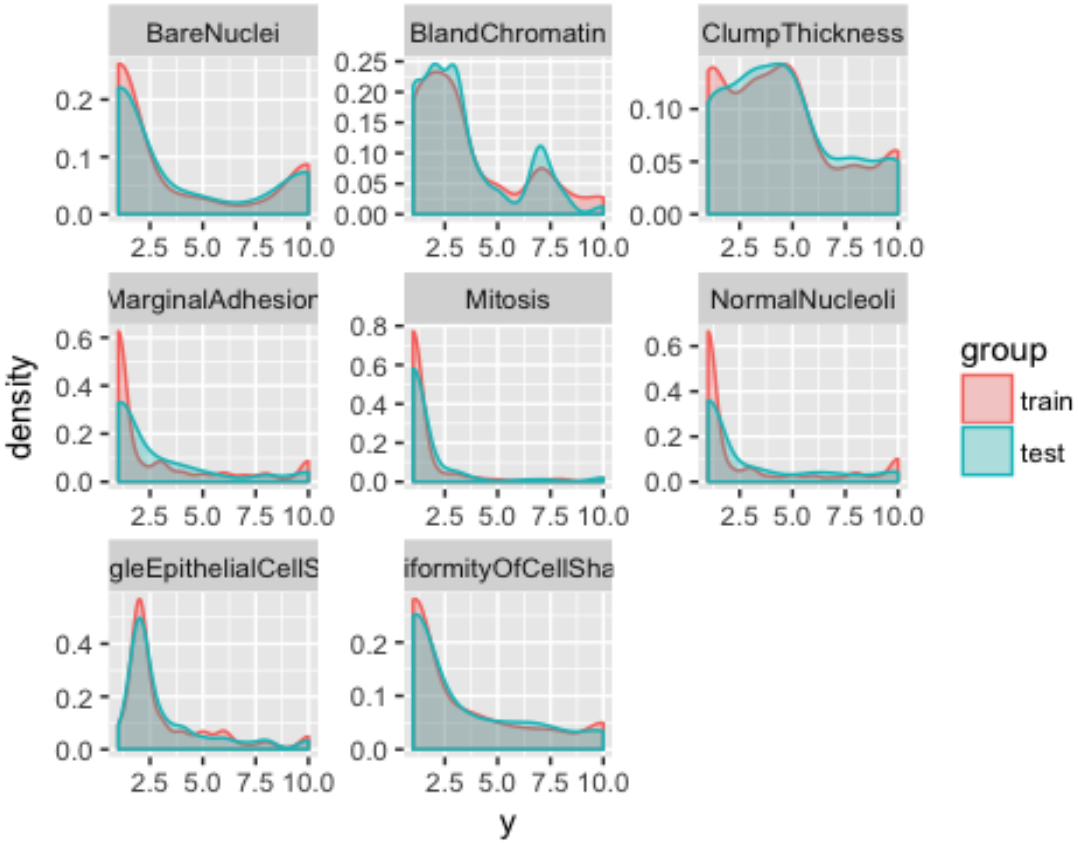

Figure 13. Distribution of variables

To explore the importance of the features, Random Forest is used and it is seen that Uniformity of Cell Shape and Bare Nuclei are most important features.

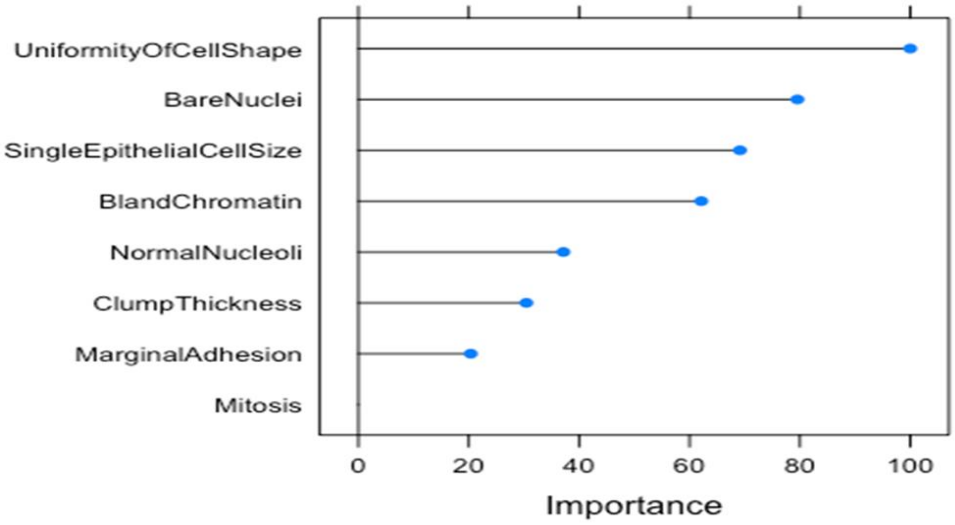

Figure 14. Importance of features

Figure 15. shows the misclassification results beside the corrected classification. As it is seen, Random Forest is performing good in classification and its accuracy, sensitivity and specificity are 0.9617, 0.9562 and 0.9722 respectively shown in Figure 14.

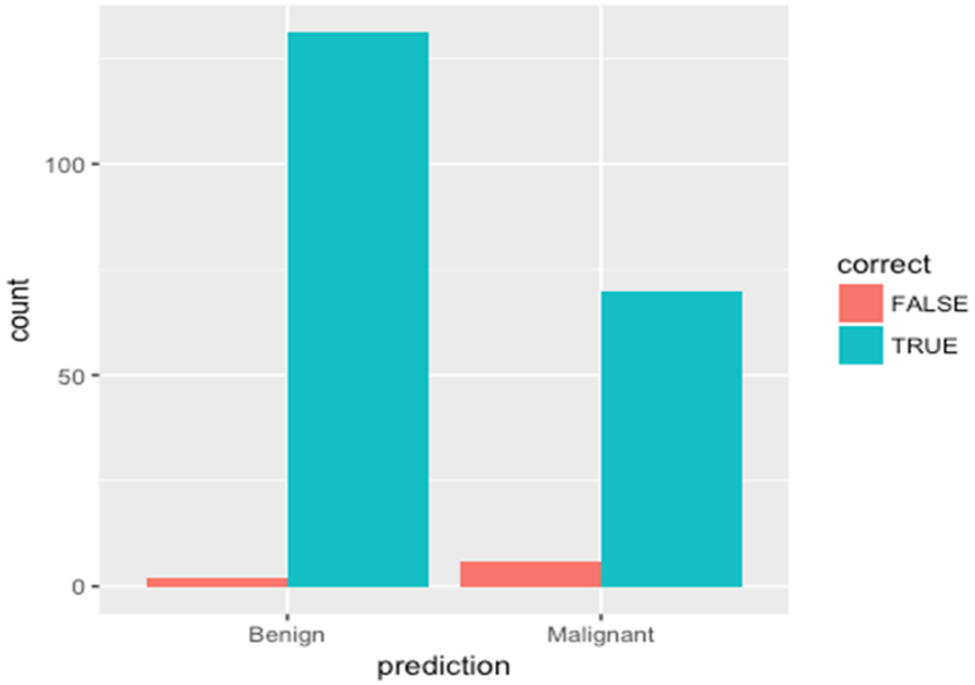

Figure 14. The misclassification results

5.1. Classification keeping all features

Next, classification and prediction is done holding all features with methods named in prior section and their performance is compared in Table 5.

Table 5. Performance of classifiers for all features

Method	Accuracy	Sensitivity	Specificity	AUC
NB	0.961	0.970	0.958	0.964
DT	0.952	0.865	0.993	0.933
LR	0.966	0.910	0.993	0.951
SVM	0.971	0.955	0.979	0.967
ANN	0.680	0.0	1.0	0.5

Figure 15 compares AUC of classifiers visually. As it is seen, keeping all features, NB is performing the best as a classifier to predict the class of breast cancer data.

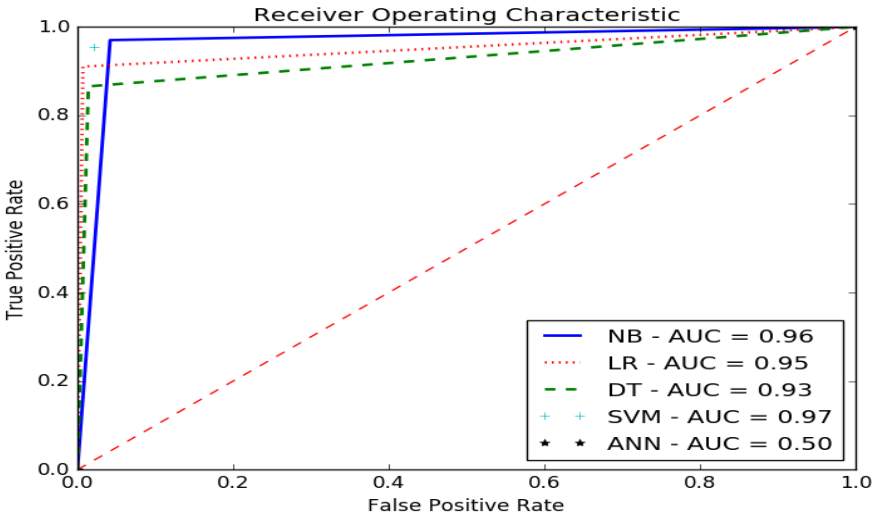

Figure 15. AUC of classifiers for all features

5.2. Classification with LASSO Features

By applying LASSO as a feature selection method, just one feature which is Single Epithelial Cell Size is selected and Table 6 shows the comparison of performance of classifiers.

Table 6. Performance of classifiers for LASSO features

Method	Accuracy	Sensitivity	Specificity	AUC
NB	0.866	0.701	0.944	0.822
DT	0.890	0.910	0.881	0.895
LR	0.866	0.701	0.944	0.822
SVM	0.890	0.910	0.811	0.895
ANN	0.890	0.910	0.811	0.895

As is it seen in Table 6, DT, SVM and ANN have the higher AUC. Results of LASSO, will not be used to classify the data. Because, first of all, LASSO has selected just one of the features which contradicts the fact that features are not highly correlated so it's not right to remove them and second of all, comparing the AUC and accuracy of classifiers with and without using LASSO features, it is obtained that classifiers have better performance not using LASSO as a feature selection method.

5.3 Classification with LR Selected Features

From logistic regression following variables are selected:

- Bare Nuclie
- Uniformity of Cell Shape
- Clump Thickness
- Bland Chromotin
- Marginal Adhension

Table 7 shows the Performance of classifiers for LR features. As it is seen, NB and SVM perform better using LR as a feature selection technique.

Table 7. Performance of classifiers for LR features

Method	Accuracy	Sensitivity	Specificity	AUC
NB	0.961	0.940	0.972	0.960
DT	0.923	0.805	0.979	0.903

LR	0.961	0.910	0.986	0.933
SVM	0.976	0.985	0.972	0.967
ANN	0.938	0.865	0.720	0.907

Figure 16 compares AUC of classifiers visually. As it is seen, keeping LR features, NB and SVM are performing the best as a classifier to predict the class of breast cancer data. Comparing the results with classification when all features are kept shows that NB is performing better when all features are kept and SVM is performing better when features selected by LR are kept.

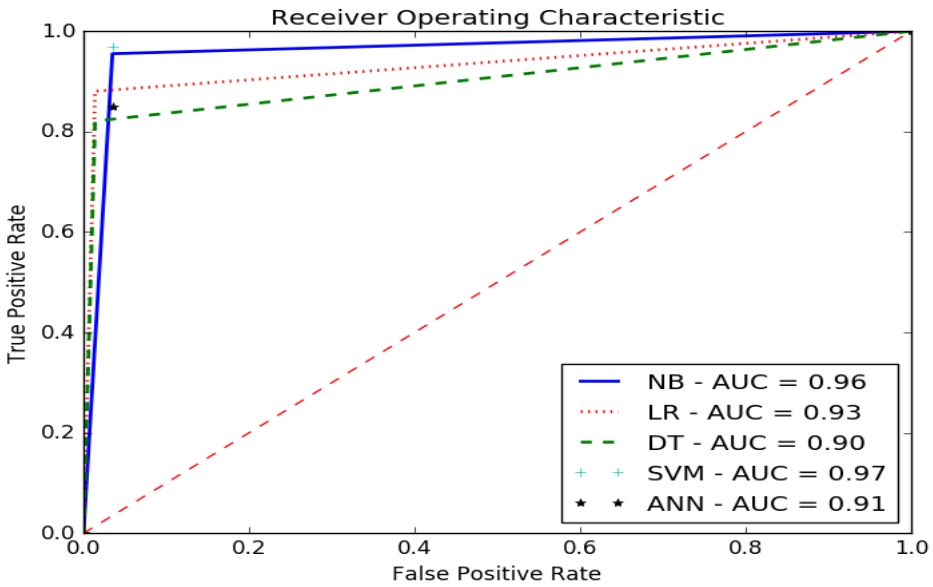

Figure 16. AUC of classifiers for LR features

5.4 Classification with DA Selected Features

By applying Discriminant analysis, the following variables are selected:

- Bare Nuclie
- Uniformity of Cell Size
- Clump Thickness
- Normal Nuclie

Table 8 shows the results of classification for features selected by DA. As it is seen, again NB and SVM are performing better.

Table 8. Performance of classifiers for DA features

Method	Accuracy	Sensitivity	Specificity	AUC
NB	0.961	0.985	0.951	0.968
DT	0.928	0.850	0.965	0.900
LR	0.957	0.895	0.986	0.940
SVM	0.971	1.0	0.958	0.979
ANN	0.323	1.0	0.006	0.503

Figure 17 compares AUC of classifiers visually. As it is seen, keeping DA features, NB and SVM are performing the best as a classifier to predict the class of breast cancer data. Comparing the results with classification when all features are kept shows that both NB and SVM have higher accuracy and AUC when applying on features selected by DA. It is concluded that, between LR, DA and LASSO, DA is more suitable for feature selection.

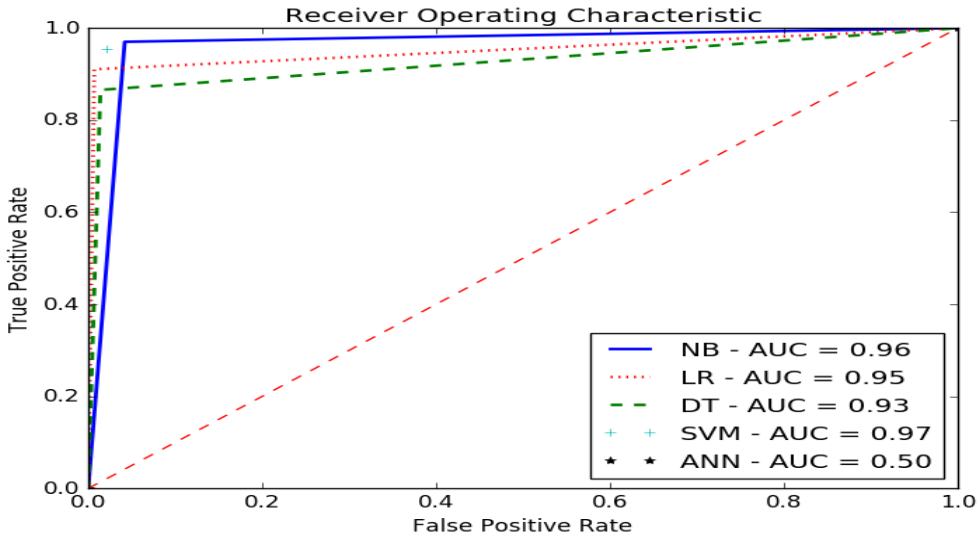

Figure 17. AUC of classifiers for DA features

5.5 Classification with Hybrid Selected Features

Table 9 shows the results of classification for features selected by a hybrid algorithm based on DA and LR. As it is seen, again NB and SVM are performing better.

Table 9. Performance of classifiers for hybrid method

Method	Accuracy	Sensitivity	Specificity	AUC
NB	0.966	0.985	0.985	0.971
DT	0.923	0.805	0.958	0.911
LR	0.961	0.910	0.986	0.948
SVM	0.961	1.0	0.986	0.972
ANN	0.680	0.0	1.0	0.5

Figure 18 compares AUC of classifiers visually. As it is seen, keeping hybrid features, NB and SVM are performing the best as a classifier to predict the class of breast cancer data. Comparing the results with classification when all features are kept shows that both NB and SVM have higher accuracy and AUC when applying on features selected by hybrid method. Comparing the accuracy and AUC of NB and SVM when using features selected by hybrid method and DA, it is concluded that NB performs better using DA features and SVM performs better when using hybrid method features.

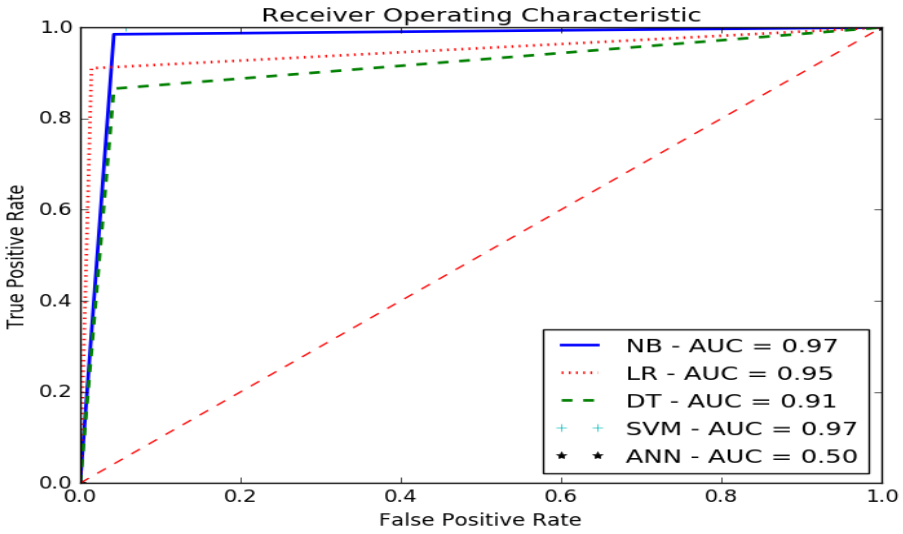

Figure 18. AUC of classifiers for hybrid method

6. Conclusion and Future Work

In this paper first the data set containing 700 samples and 9 features was selected from UCI data base and preprocessing was done to remove noisy and unreliable data. In order to do so, first missing value is filled via using the distribution of data set. Since, all data were in the scale of 1-10 no normalization or standardization is done. The correlation matrix of features is obtained and it is observed that except size and shape of cells, other features are not highly correlated to each other. Still, PCA, LASSO, LR And DA are applied n data set to reduce the dimension and remove correlated features.

Classification with NB, DT, SVM, LR and ANN has been done in five stages. In the first stage, classification is done using all features. In the second stage, classification is done using LASSO features. Respectively, in stage three to five, classification is done using features selected by LR, DA and a hybrid of DA and LR which takes into account common factors of LR and DA.

At each stage, performance of five different techniques, namely, DT, NN, LR, SVM and NB in classifying both based is compared based on accuracy and AUC. Results show that. NB and SVM outperforms other methods. When keeping all features NB performs best. When keeping features selected by LR, SVM outperforms NB and when using features selected by LR and DA, SVM outperforms NB and at last, using common features of LR and DA NB outperforms SVM. It is concluded that, SVM, using hybrid feature selection performs the best comparing to results of all stages. Therefore, based on the results of this paper, SVM is the method to classify breast cancer data while dimension is reduced using hybrid feature selection method. In order to extend this study, it is suggested to use a data set with more number of observations and also try different classification methods.

References

1. JD1, S. RC, H. E and G. RE, "Follow-up of abnormal screening mammograms among low-income ethnically diverse women: findings from a qualitative study.," Patient Educ Couns, vol. 72(2), pp. 283-92, 2008 Aug.
2. N. Abe, M. Kudo, J. Toyama and a. M. Shimbo, "A Divergence Criterion for Classifier-Independent Feature Selection.," Springer, pp. 668-676, 2000.
    A. E. Guyon, "An introduction to variable and features election," J. Mach. Learn. Res, vol. 3, pp. 1157-1182, 2003.
3. S. Vural, X. Wang and C. Guda, "Classification of breast cancer patients using somatic mutation profiles and machine learning approaches," The International Conference on Intelligent Biology and Medicine (ICIBM), 2015.
    A. P and M. S. e. a. Buadu LD, "Neural network analysis of breast cancer from MRI findings," Radiat Med, vol. 15(5), p. 283–293, 1997.
4. P, B. LD and N. H, "Feature extraction and classification of breast cancer on dynamic magnetic resonance imaging using artificial neural network," Cancer Lett, vol. 171(2), p. 183–191, 2001.
5. HB, G. PH and R. D. e. a, "Artificial neural networks improve the accuracy of cancer survival prediction," Cancer , vol. 79(4), p. 857–862, 1997.
6. HA, "An evolutionary artificial neural networks approach for breast cancer diagnosis," Artif Intell Med, vol. 25(3), p. 265–281, 2002.
7. M. T. G. F. H.A. Abbass, "C-net: a method for generating non-deterministic and dynamic multivariate decision trees," Knowledge Inf. Syst, p. 184–197, 2001.
8. S.-Y. Jin, J.-K. Won, H. Lee and H.-J. Choi, "Construction of an automated screening system to predict breast cancer diagnosis and prognosis," Basic & Applied Pathology, vol. 5, no. 1, pp. 15-18, 2012.
9. H. Hasan, M. İsa and Mediha, "A Hybrid Applied Optimization Algorithm for Training Multi-Layer Neural Networks in Data Classification.," Gazi University Journal of Science, vol. 28, no. 1, pp. 115-132, 2015.
10. A. Pen˜a-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," Artificial Intelligence in Medicine, p. 131–155, 1998.
11. N. S. J. Dheeba and S. Selvi, "Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach.," Journal of Biomedical Informatics,, vol. 49, p. 45–52, 2014.
12. R. K. Detlef Nauck *, "Obtaining interpretable fuzzy classification rules from medical data," Artificial Intelligence in Medicine, vol. 16, p. 149–169, 1999.
13. E. Alicˇkovic´ and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," Neural Comput & Applic, vol. 28, p. 753–763, 2017.
14. M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," Expert Systems with Applications, p. 3240–3247, 2009.