Newcastle University

Statistical Learning for Data Science (MAS8404)

Dataset: BreastCancer	**Student Name**: Akash Mohandoss
	Student ID: 220488118

1. Introduction:

Breast cancer is a common type of malignant tumour in women caused by abnormal growth of tissue in certain areas of the breast. Many diagnostic methods are used to diagnose breast cancer, including MRI, ultrasound, X-ray, and biopsy. From my insight through web, breast cancer is usually divided into 3 types Benign (non-cancerous), pre-Malignant (pre-cancerous), Malignant (cancerous).

The purpose of this analysis is to generate a report containing exploratory data analysis on a specific BreastCancer dataset generated from Wisconsin using fine needle aspiration cytology (FNAC), creating graphs and numerical summaries that extract relationships between predictor and response variables and provide us with some relationship between variables. The next part of the report uses logistic regression with one of the discriminant analysis methods to build a classifier model that classifies benign and malignant classes of breast cancer based on the given data and identifies the best-fitted model to classify the tumor.

2. Exploratory data analysis:

The Given Dataset contains 699 observations and 10 variables including Null values. Excluding the NA value there are 683 observations. Class is the reponse / categorical variable in this dataset and the rest of the variables are considered to be predictors

Numerical and Graphical summaries for the given dataset is applied and is displayed. Summarization of dataset delineates the quantitative and categorical variables. Class is a categorical variable that contains 458 Benign data and 241 Malignant data. The ID column in the BreastCancer dataset is unnecessary so it has been removed.

Summaries	cl.thickness	cell.size	cell.shape	Marg.adhesion	Epith.c.size
Minimum	1.000	1.000	1.000	1.00	1.000
1st Quartile	2.000	1.000	1.000	1.00	2.000
Median	4.000	1.000	1.000	1.00	2.000
Mean	4.442	3.151	3.215	2.83	3.234
3rd Quartile	6.000	5.000	5.000	4.00	4.000
Maximum	10.000	10.000	10.000	10.00	10.000
Standard Deviation	2.820	3.065	2.988	2.864	2.223

Summaries	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
Minimum	1.000	1.000	1.00	1.000
1st Quartile	1.000	2.000	1.00	1.000
Median	1.000	3.000	1.00	1.000
Mean	3.545	3.445	2.87	1.583
3rd Quartile	6.000	5.000	4.00	1.000
Maximum	10.000	10.000	10.00	9.000
Standard Deviation	3.643	2.449	3.052	1.636

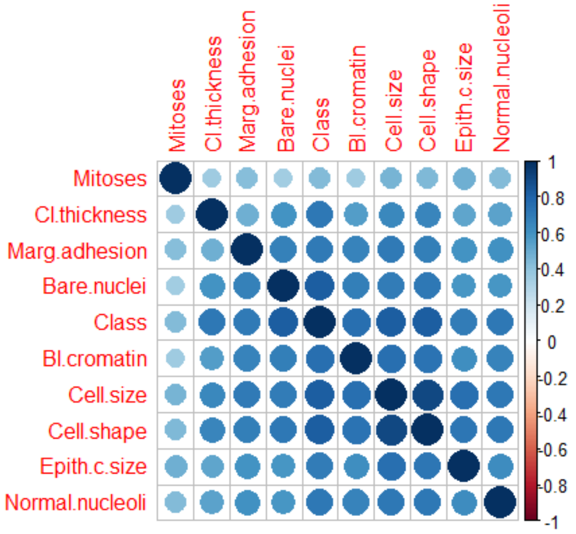
[Table 1.1]

Null values have been removed from the given dataset and the summary table that exhibits the mean, median, and quartiles are displayed in above table 1.1. it can be clearly viewed from the above table that the Mean and Median of the above feature don't have that much difference. So most of the features are symmetrically distributed without much skewness.

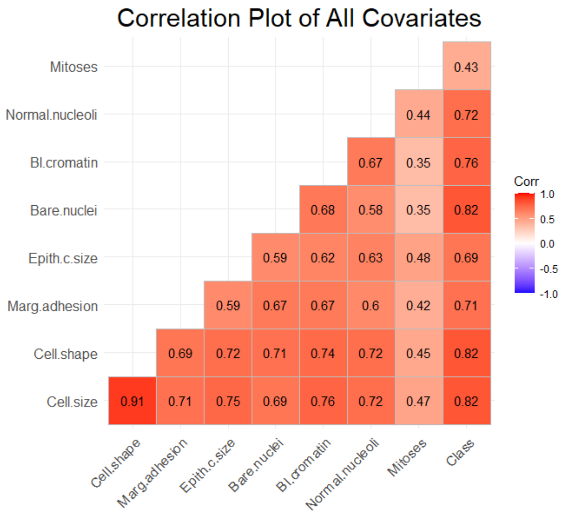
Among all the variables Cell.size, Cell.shape and Bare.nuclei have much difference between mean and median, thus showing us that there's little skewness in the given data of that variable.

Graphical summaries:

The below-plotted Correlation matrices clearly identifies that the Cell.size and Cell.shape predictor variables have much correlation (0.91) when compared to the other predictor variables.



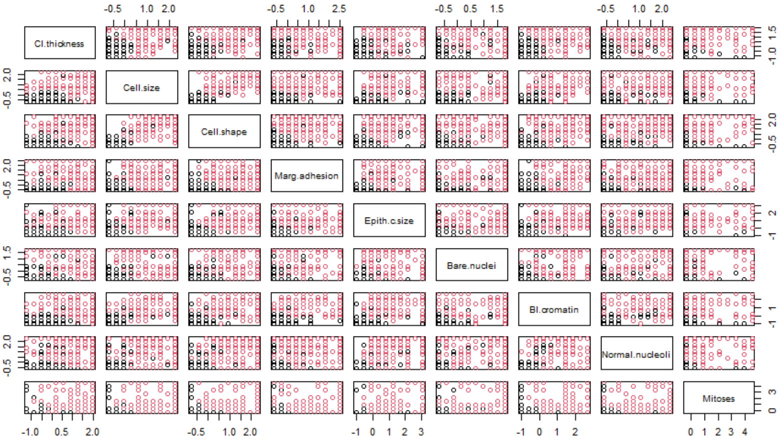
[Plot 1.1a]



[Plot 1.1b]

The darker the colour of the circle the greater the correlation of the variables. Bare.nuclei do have a little higher correlation (0.82) but when comparing it is lower in value to them. we see that the mitoses features have a lower correlation compared to the other features so it can be used in future analysis since it is statistically significant. Practically Cell.size and Cell.shape are physical features describing the BreastCancer and have too high correlation for future analysis they won't be able to help us give good accuracy.

The below scatterplot clearly shows us the distribution of data. We can easily predict the flow whether it has a normal distribution or not. It is coherent that the benign and malignant cancer classes have been clearly separated and is visible in this graph with benign as black and malignant classes being the red circles.



[Plot 1.2]

3. Applying Logistic regression Classification for the BreastCancer Dataset

Logistic regression is a classification method to best fit regression model/curve, y=$f(x)$, where y is a categorical variable. In our case, we have a set of 9 predictor variables and only 1 response variable. The predictor variables considered in this regression fit model at the initial phase are **Cl.thickness, Cell.size, Cell.shape, Marg.adhesion, Epith.c.size, Bare.nuclei, Bl.cromatin, Normal.nucleoli**, and **Mitoses**. The response variable **Class** has either Benign or Malignant in the dataset. For regression analysis purposes we convert the given all factor values to quantitative values. Instead of Benign and malignant it is replaced with binary values of 0’s and 1’s.

So, it is Binomial logistic regression that we are going to apply since we are going to predict variables with binary values 0’s and 1’s. First, we are splitting our pre-processed dataset into train and test datasets respectively. Then we try to fit a model for the training dataset without applying any regularisation method and trying to predict the accuracy of the model that has been fitted.

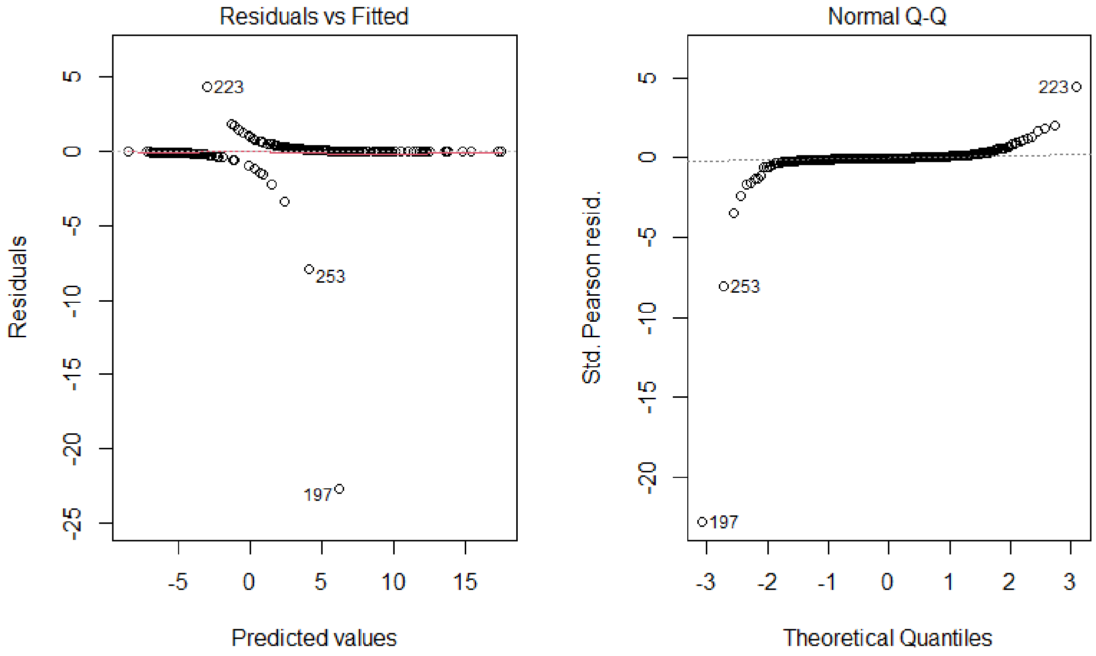
Deviance Residuals:

Min	**1Q**	**Max**	**3Q**	**Max**
-3.5349	-0.1094	-0.0585	0.0249	2.4547

Null deviance: 608.852 on 478 degrees of freedom

Residual deviance: 59.935 on 469 degrees of freedom

AIC: 79.935



[Plot 1.3]

Further, we will use the glm(general linear model) function to fit the logistic regression model to the pre-processed dataset. it can clearly be viewed that the deviance residuals are closely centered on 0 and are roughly symmetrical.

Plot 1.3 shows the logistic regression model fit for the dataset and majorly all the residuals have been fitted to the zero line for both the predicted and the theoretical values. The errors have constant variance since the residuals are fitted along the zero line. If for example the data is small and the residuals grow or descend with the fitted values in a pattern, then the errors may not have a constant variance.

	Estimate	**Standard Error**	**z-value**	**Pr(>|z|)**	
(intercept)	-9.83728	1.34916	-7.291	3.07e-13	***
Cl.thickness	0.48075	0.14792	3.250	0.001153	**
Cell.size	0.03556	0.21826	0.163	0.870568	
Cell.shape	0.30024	0.24496	1.226	0.220313	
Marg.adhesion	0.36950	0.12928	2.858	0.004263	**
Epith.c.size	-0.16464	0.18932	-0.870	0.384492	
Bare.nuclei	0.36528	0.10378	3.520	0.000432	***
Bl.cromatin	0.40113	0.18425	2.177	0.029467	*
Normal.nuclei	0.37553	0.14154	2.653	0.007975	**
Mitoses	0.66790	0.32506	2.055	0.039909	*

[Table 1.2]

Table 1.2 shows us the average change in the log odds of the Class variable with respect to the predictor variables. The summary tells us that based on the p-value the predictor variables

Cl.thickness, Marg.adhesion, Bare.nuclei, Bl.cromatin, Normal.nuclei, and mitoses are statistically significant with the response variables in the model that has been fitted. For instances, it can be stated that one unit increase in Cl.thickness is associated with an average increase of 0.48075 in the log odds of the Class.

Testing Accuracy
0.9509804

Fitting null model for pseudo-r2 McFadden
0.8745649

In linear regression, we use R^2 value to predict the accuracy of the model fitted and there is no such value to predict the accuracy of the model fitted using logistic regression so we use MCFadden's R^2 value which ranges from 0 to 1 to foretell the predictive power of the model. As shown above test accuracy of the model fitted is predicted to be 95 %. The test error calculated is said to be **0.0490196**.

In addition, we have McFadden's value which is 0.8745649 which is quite high, which speculate that out model fits the data well and high predictive power.

Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
1.238208	2.994298	2.897187	1.245197	1.696254	1.249938	1.280846	1.383461	1.063390

[VIF values]

Furthermore, Variance Inflation Factor (VIF) values are calculated for each predictor variable in the model to check if multicollinearity is a problem or not. Based on the above table it can be inferred that all the variables have a value less than or away from 5 which indicates that multicollinearity does not exist and also there is a low correlation among the variables under ideal conditions.

3.1 Subset Selection Technique:

To reduce the number of predictor variables to find the best fit model subset selection methods are applied such as Best subset selection, automated(stepwise) selection, and model comparison criterions. Here, we consider the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) model comparison criterions to predict the best k value for both AIC and BIC models.

	logLikelihood	AIC
0	-442.17509	884.3502
1	-127.37980	256.7596
2	-83.15598	170.3120
3	-67.77778	141.5556
4	-61.37155	130.7431
5	-56.13177	122.2635
6	-53.57186	119.1437
7*	-51.63998	117.2800
8	-51.45031	118.9006
9	-51.44991	120.8998

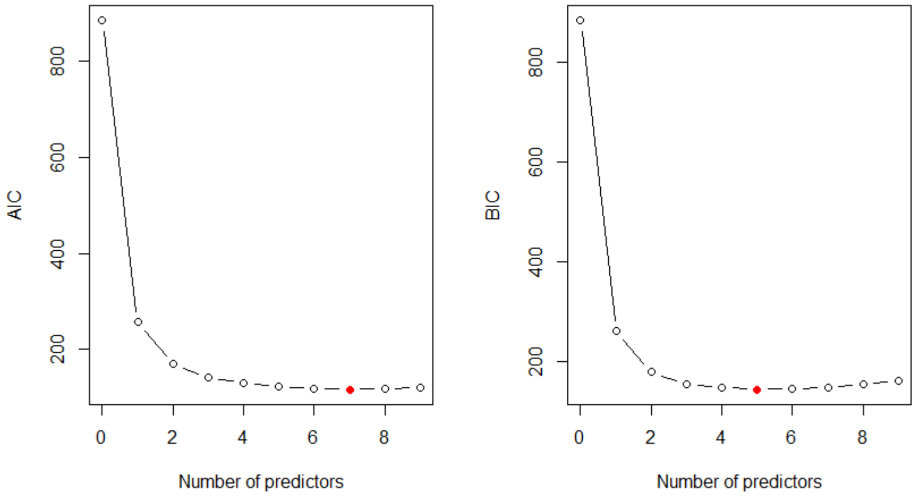
[AIC subset table with loglikelihood and AIC]

	logLikelihood	AIC
0	-442.17509	884.3502
1	-127.37980	261.2861
2	-83.15598	179.3649
3	-67.77778	155.1351
4	-61.37155	148.8491
5*	-56.13177	144.8960
6	-53.57186	146.3027
7	-51.63998	148.9654
8	-51.45031	155.1126
9	-51.44991	161.6383

[BIC subset table loglikelihood and AIC]

AIC model's Bestk value is 7 and BIC Bestk value is 5. These values indicate the number of predictors in the models individually which is a perfect fit to find the accuracy of the response variable.

Based on plot 1.4, it seems like a model with 6 predictors will be a good fit for obtaining good accuracy from logistic regression. We can further extract the variables from the best-fitting 6-predictor model from the output. We receive the reduced dataset in a variable named as BC_data_red with only 6 predictors obtained from the subset selection method. Then logistic regression is applied to the obtained reduced subset data and a model is fitted using the glm function.



[Plot 1.4]

The below coefficient table is extracted from the summary of the model fitted using reduced data. We see that the selected model uses Cl.thickness, Marg.adhesion, Bare.nuclei, Bl.cromatin, Normal.nuclei and Cell.shape as the predictors and the coefficients of these predictors are significantly apparent from zero. So, one unit increase in any of the predictors is associated with an average increase in the log odds of the Class variable. The variable that that drop out of the table are Cell.size, Mitoses, and Epith.c.size. This means that all the variable are statistically significant and is effective at predicting the probability of the response variable ‘Class’.

	Estimate	Standard Error	z-value	Pr(>|z|)	
(intercept)	-9.7357	1.2883	-7.557	4.13e-14	***
Cl.thickness	0.5902	0.1530	3.856	0.000115	***
Cell.shape	0.4231	0.2002	2.113	0.034563	*
Marg.adhesion	0.3756	0.1259	2.983	0.002857	**
Bare.nuclei	0.2948	0.1051	2.806	0.005019	**
Bl.cromatin	0.5187	0.1833	2.830	0.004656	**
Normal.nucleoli	0.1866	0.1310	1.425	0.154158	

[Table 1.3]

Test Accuracy
0.98039022

Test error: 0.019609
AIC: 96.996

McFadden R^2 value
0.8662832

	Value	Degree of freedom
Null deviance	620.686	478
Residual deviance	82.996	472

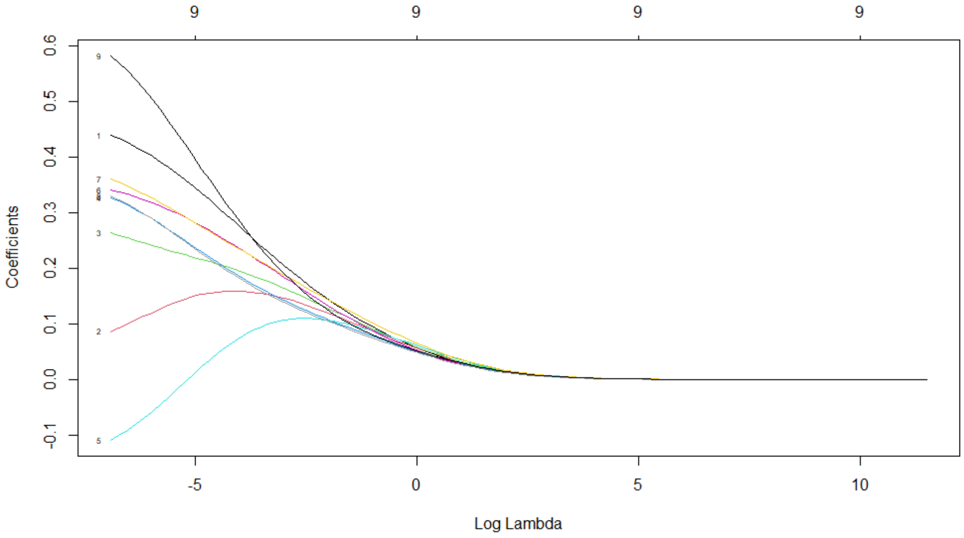
The model chosen predicted about 0.98039022 (97%) accuracy from train data obtained from the subset data and obtained 0. 8642736 McFadden R-square value which adds on that the model is very good fit and has a high predictive power when compared to the previous model. All the variable from the chosen model is statistically significant apart from Normal.nucleoli since its p-value is higher than 0.05.

3.2 Regularisation Method (Ridge Regression):

Ridge regression is an L2 regularisation method that adds a penalty square of the magnitude of regression coefficients and tries to shrink them. To apply ridge regression to our data we convert the predictor variables and its data in the form of matrix and the response variable as a vector. It is then passed to the glmnet function to fit a model. we have chosen a grid value of vector for tuning parameter and force the function to use it. In our case the range is $\lambda = 10^5$ (lots of shrinkage) to $\lambda = 10^{-3}$ (very little shrinkage).

When the λ is 10^5, the regression coefficients for the predictor variables are shrunk to zero. Further moving towards the lambda value of 10^{-3} the coefficients are diverged from the zero. Now we predict the minimal value of the lambda using ridge regression and predict accuracy and find Area under the ROC Curve (auc) value to justify the found using full data.

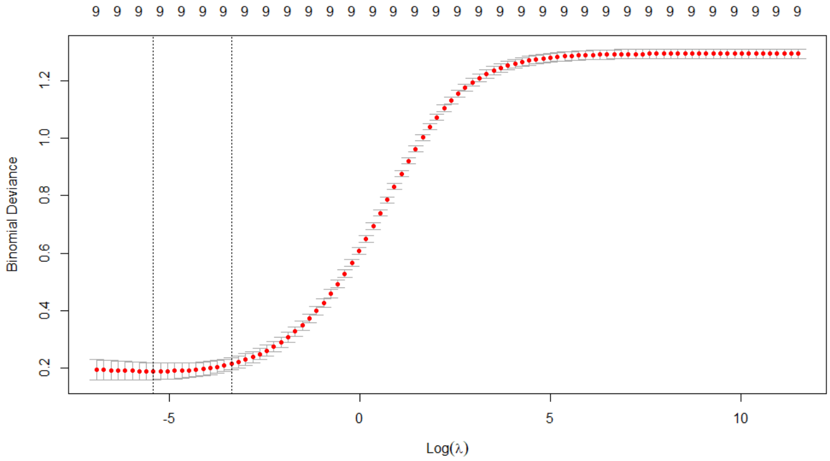
Plot 1.5 illustrates how the regression coefficient values of the regression parameter shrink towards the zero and each number in the graph indicated a predictor in its order. It tells us that the all the regression coefficients of the predictor variable reach zero when it comes to 5 due to the grid range we have specified. The last predictors to drop and reach the zero point are number 9, number 1, and number 7 indicating predictors Mitoses, Cl.thickness, and Bl.cromatin and these variable are statistically significant with the Class variable as mentioned in previous observations. They can be used to predict the Class of the disease with high accuracy.



[Plot 1.5]

Using cv.glmnet function the optimal lamda value is predicted by performing a k-folds cross validation on the data. By default, it performs 10-folds cross validation to the dataset.

The Lambda value that minimizes the test MSE is predicted to be **0.004430621.**



[Plot 1.6]

The cross-validation score varying with lambda values is plotted above. The mean MSE value is plotted against each lambda value along with error bars which cover the mean plus or minus one standard error. The number at the top of the plot indicate the number of coefficients which are non-zero.

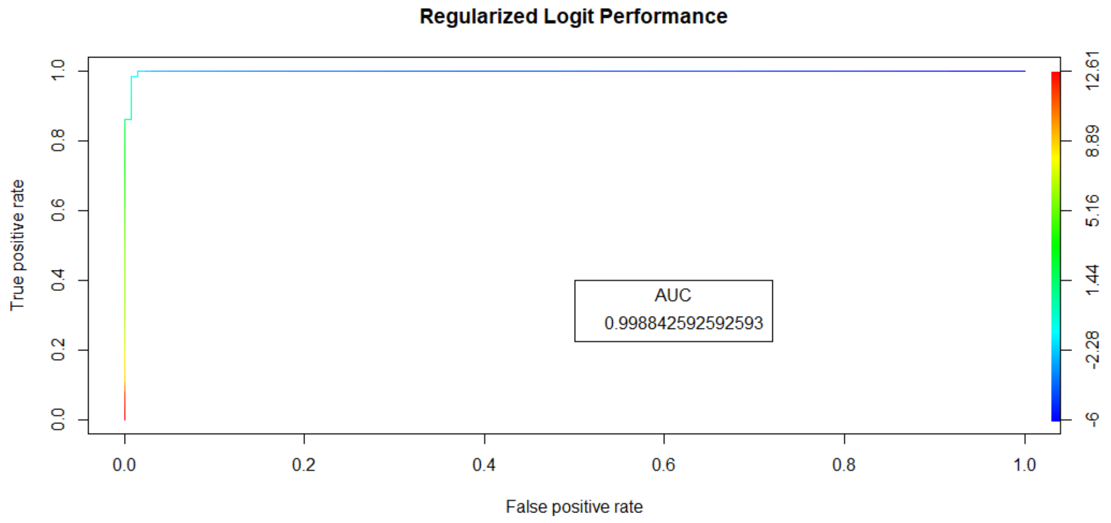
Minimum tuning parameter	Mean MSE
92	0.1746352

We perform ridge regression with the chosen value of λ from the ridge *BC_ridge* object that we fitted using full dataset. The prediction of the fitted model is calculated and it is calculated to be **0.9988426** which I quite a high value to obtain accuracy of finding Class of the disease. The confusion matrix is tabulated below.

BC_predictions	**0**	**1**
0	131	9
1	1	63

Accuracy: *0.951*

The accuracy obtained from the above is 95% and the test error calculated is 0.049. to add advantage to the model we calculate auc value to represent the strength of the model fit.



[Plot 1.7]

The obtained auc value is 0.9988426 and and the corresponding auc curve plot is showed in plot 1.7. it is high value and clearly shows us that the model is a excellent fit with excellent statistical significance between the predictors and the response variable and will be a good fit to accurate the class of the disease.

3.3 Linear Discriminant Analysis [*LDA*]

Linear Discriminant Analysis is a dimension reduction technique used to reduce the number of variables in the dataset while trying to maintain as much as information in the data. The lda function in the MASS package is used to fit a model using the dataset. the model gives us prior probabilities of the group, Group means, and Coefficients of the linear discriminants.

From the below obtained data we can clearly see that **65.13%** of the data are **Benign** and **34.86%** of them are **Malignant.** Then the group means table defines the mean value for the each predictor for disease class. The more is the difference between the means of the class for the particular variable, the easier it will be to classify the observation.

Prior probabilities of group:

0	1
0.651357	0.348643

Group Means:

	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei
0	2.983974	1.358974	1.442308	1.394231	2.137821	1.403846
1	7.299401	6.664671	6.598802	5.497006	5.317365	7.419162

	Bl.cromatin	Normal.nucleoli	Mitoses
0	2.096154	1.298077	1.067308
1	5.892216	5.934132	2.491018

Coefficients of linear discriminants:

Predictor variables	LD1
Cl.thickness	0.17411864
Cell.shape	0.13852975
Cell.size	0.10310763
Marg.adhesion	0.05294571
Epith.c.size	0.02096944
Bare.nuclei	0.24085770
Bl.cromatin	0.09796167
Normal.nucleoli	0.11591039
Mitoses	0.04284702

The coefficients of linear discriminants display the coefficient of the linear equation that is used to predict the response class. Since there are only two response classes there will be only one coefficient (LD1).

Accuracy = 0.9657

	Reference	
Predicted	**0**	**1**
0	131	6
1	1	66

[*Confusion Matrix*]

The accuracy for the above predicted model using Linear discriminant analysis is 0.9657 and its confusion is tabulated and is displayed above. The test error obtained is **0.03431**.

3.4 Conclusion

To conclude, all the classifier modelled using the BreastCancer data are exhibiting very good accuracy in terms of predicting the disease Class whether it is Benign or Malignant. After analysis through the above classifier and based on their accuracy shown, I prefer subset selection to have good accuracy rate in predicting the Class of the BreastCancer disease.

	Logistic Regression (With Pre-processed Data)	Best Subset Selection	Ridge Regression	Linear Discriminant Analysis
Model Accuracy	0.9509804	0.9803922	0.951	0.9657
Test Error	0.0490196	0.01960978	0.049	0.03431

[Table 1.3]

I have chosen this classifier since its accuracy predicted from the model fitted is highest when compared to other classifier's accuracy and it uses a good number of predictor variables to predict the accuracy instead of omitting much of the predictors. As shown in the above table 1.3 the Best Subset Selection has the least error (0.0196) compared to other classifiers. the predictors used in this model are statistically significant with the response variable 'Class' and tend to exhibit more positive results while predicting the Class of the disease. the next to the Best Subset method is Ridge regression having low test error (0.049) and then followed by Linear Discriminant Analysis with an error calculated as 0.034. To conclude, each model fitted using different classifier predicts the Class of the disease with good accuracy and in comparison, with the above used classifier I choose best subset selection to be the best one among others. Various techniques are available outside the box to get excellent accuracy in predicting the Class of the disease and can be used in future to build an excellent model to predict the class of the Breast cancer disease.

3.5 References

https://www.statology.org/linear-discriminant-analysis-in-r/

https://rpubs.com/Moon-C/breast-cancer-project

3.6 Appendix

R File included