Breast Cancer Data Analysis

1. Introduction

In this project, characteristics of breast tissue samples taken from 699 Wisconsin women using fine needle aspiration cytology (FNAC) are examined. For each tissue sample, nine easily measurable cytological traits were examined on a one to ten scale, including homogeneity of cell size and shape. The major goal of the clinical investigation is to ascertain how well a tissue sample could be categorised as benign or malignant using only the nine cytological traits. Additionally, exploratory analysis is carried out and the important variables/features are determined based on a selected model. The selection of models is done after discussing different models and their performance levels. The Model showing a close resemblance to the actual data is selected and based on the new model the predictive variables are determined. Also, the report justifies the reason for the model selection as well as for finding the most suitable variables out of the nine variables.

2 Methodology

Data Selected: breastcancer

The **breastcancer** data set is part of the **mlbench** package. It consists of cl.thickness (Clump Thickness), Cell.size (Uniformity of Cell Size), Cell.shape (Uniformity of Cell Shape), Marg.adhesion (Marginal Adhesion), Epith.c.size (Single Epithelial Cell Size), Bare.nuclei (Bare Nuclei), Bl.cromatin (Bland Chromatin), Normal.nucleoli (Normal Nucleoli), Mitoses and Class (Benign and Malignant). Here class is the responsive variable which correlates to the occurrence of Breast cancer and the other 9 variables are predictor variables. For analysis, data pre-processing has been done which includes changing the Benign and Malignant type cancer to 1 and 0 respectively. Even though the independent variable is categorical in nature, it is considered as a Quantitative variable for simplifying the model. Additionally, NA values were eliminated.

The data is divided into two parts, one for modelling (0.7) and the other one for cross-validation.

The following data can be referenced for understanding the meaning of each variable used in this report.

1.	Id	Sample code number
2.	Cl.thickness	Clump Thickness
3.	Cell.size	Uniformity of Cell Size
4.	Cell.shape	Uniformity of Cell Shape
5.	Marg.adhesion	Marginal Adhesion
6.	Epith.c.size	Single Epithelial Cell Size
7.	Bare.nuclei	Bare Nuclei
8.	Bl.cromatin	Bland Chromatin
9.	Normal.nucleoli	Normal Nucleoli
10.	Mitoses	Mitoses
11.	Class	Class

	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
1	1	1	1	2	1	3	1	1	benign
2	4	4	5	7	10	3	2	1	benign
3	10	10	8	7	10	9	7	1	malignant
4	1	1	3	2	1	3	1	1	benign

Table 1 Overview of the Breast Cancer Data

2.1 Exploratory Analysis

The Given Figure 1 depicts the correlation of class with each variable taken from the part of the correlation matrix (table 2)

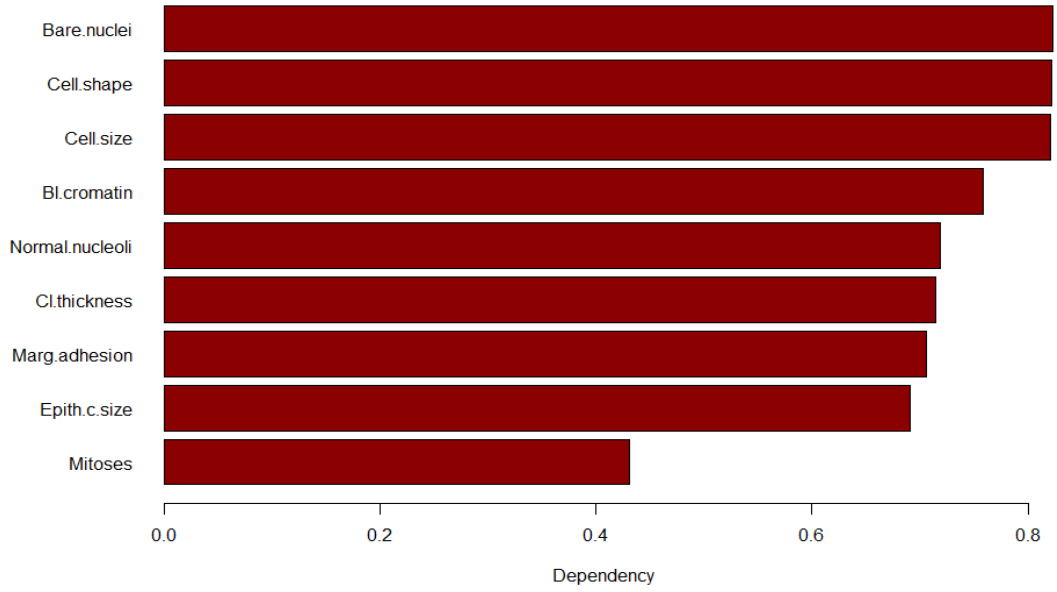

Figure 1 Correlation of each variable with class

	Cell Thickness	Cell size	Cell shape	Marginal adhesion	Epithelial size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
Class	0.71479	0.820801	0.821891	0.706294	0.690958	0.822696	0.758228	0.718677	0.431297	1

Table 2 Part of Correlation Matrix

Summary of the Data

	Variation	Total Number	Mean	Standard deviation	Median	Minimum	Maximum
Cl.thickness	1	683	4.44	2.82	4	1	10
Cell.size	2	683	3.15	3.07	1	1	10
Cell.shape	3	683	3.22	2.99	1	1	10
Marg.adhesion	4	683	2.83	2.86	1	1	10
Epith.c.size	5	683	3.23	2.22	2	1	10
Bare.nuclei	6	683	3.54	3.64	1	1	10
Bl.cromatin	7	683	3.45	2.45	3	1	10
Normal.nucleoli	8	683	2.87	3.05	1	1	10
Mitoses	9	683	1.58	1.64	1	1	9
Class	10	683	0.35	0.48	0	0	1

Table 3 Breast Cancer Summary

	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
Cl.thickness	1	0.6424815	0.65347	0.4878287	0.523596	0.5930914	0.5537424	0.5340659	0.3545301
Cell.size	0.6424815	1	0.9072282	0.706977	0.753544	0.6917088	0.7555592	0.719346	0.4654091
Cell.shape	0.65347	0.9072282	1	0.6859481	0.7224624	0.7138775	0.7353435	0.7179634	0.4468571
Marg.adhesion	0.4878287	0.706977	0.6859481	1	0.5945478	0.6706483	0.6685671	0.6031211	0.4249917
Epith.c.size	0.523596	0.753544	0.7224624	0.5945478	1	0.5857161	0.6181279	0.6289264	0.4811836
Bare.nuclei	0.5930914	0.6917088	0.7138775	0.6706483	0.5857161	1	0.6806149	0.5842802	0.3490108
Bl.cromatin	0.5537424	0.7555592	0.7353435	0.6685671	0.6181279	0.6806149	1	0.6656015	0.3536683
Normal.nucleoli	0.5340659	0.719346	0.7179634	0.6031211	0.6289264	0.5842802	0.6656015	1	0.4370424
Mitoses	0.3545301	0.4654091	0.4468571	0.4249917	0.4811836	0.3490108	0.3536683	0.4370424	1

Table 4 Correlation Matrix

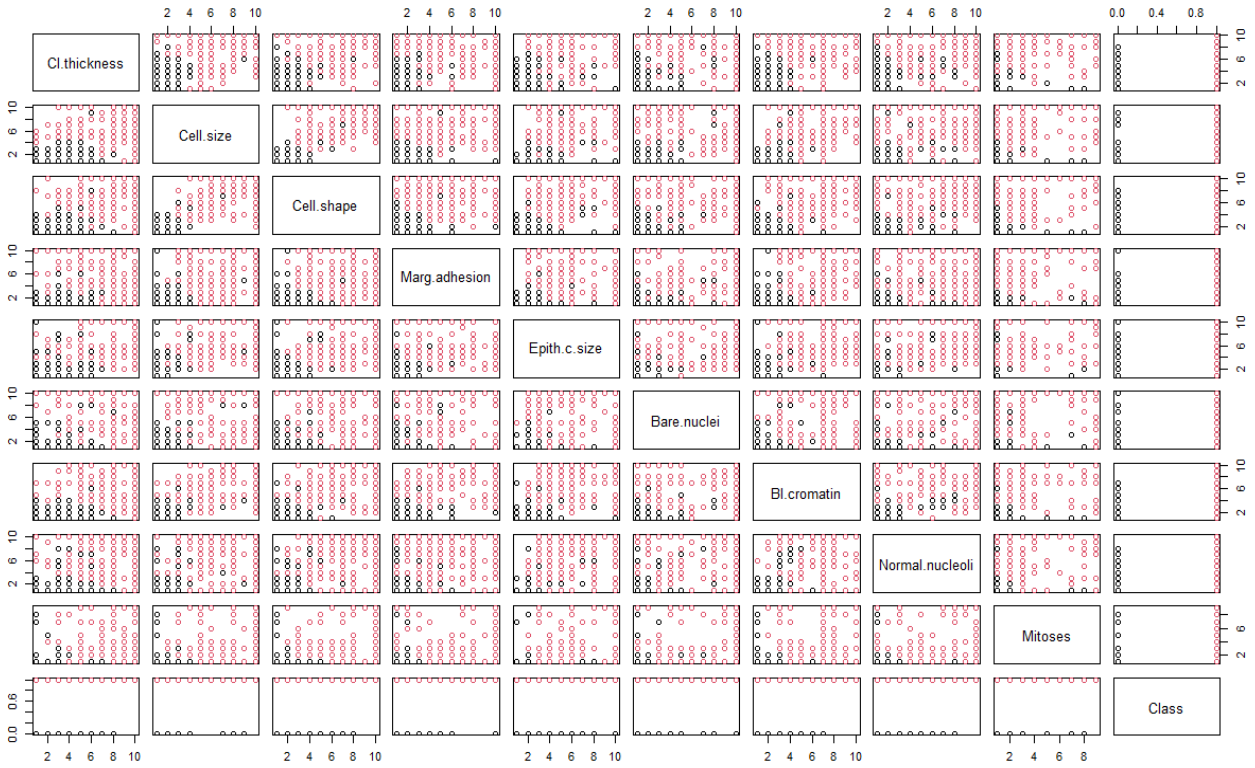

Figure 2 Relationships within all the Variables

Comments:

- **Figure 2 and Table 4 show an approximately linear relation between Uniformity of Cell Size and Uniformity of cell shape. (Highlighted in table 4)**
- **Figure 1 shows that Mitoses is relatively less significant variable.**
- **Also, benign observations (444) are seen more than the malignant (239) in the data set.**

Principle Component Analysis

PCA1 and PCA 2 are selected since their cumulative percentage is more than 74 % (using the summary function in R). Figure 3 shows that there is clustering occurring based on the variable *class*. The pc1 account for 65% of the variation of the data, whereas pc2 represents around 8 % variation. Observation of table 5 shows that the pc1 axis takes all the variables equally in terms of variations. PCA1 and PCA2 equations are:

$$y_1 = -0.3018371z_1 - 0.380523z_2 - 0.3773482z_3 - 0.3325843z_4 - 0.3358808z_5 - 0.3350417z_6 - 0.34561z_7 - 0.3353311z_8 - 0.23265z_9$$

$$y_2 = -0.1481572z_1 - 0.0495318z_2 - 0.0844956z_3 - 0.04715193z_4 + 0.1588028z_5 - 0.2553437z_6 - 0.22651z_7 + 0.03131929z_8 + 0.9074834z_9$$

	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
PC1	-0.3018371	-0.380523	-0.3773482	-0.3325843	-0.3358808	-0.3350417	-0.34561	-0.3353311	-0.23265
PC2	-0.1481572	-0.0495318	-0.0844956	-0.04715193	0.1588028	-0.2553437	-0.22651	0.03131929	0.9074834

Table 5 Weightage of variables in each pc

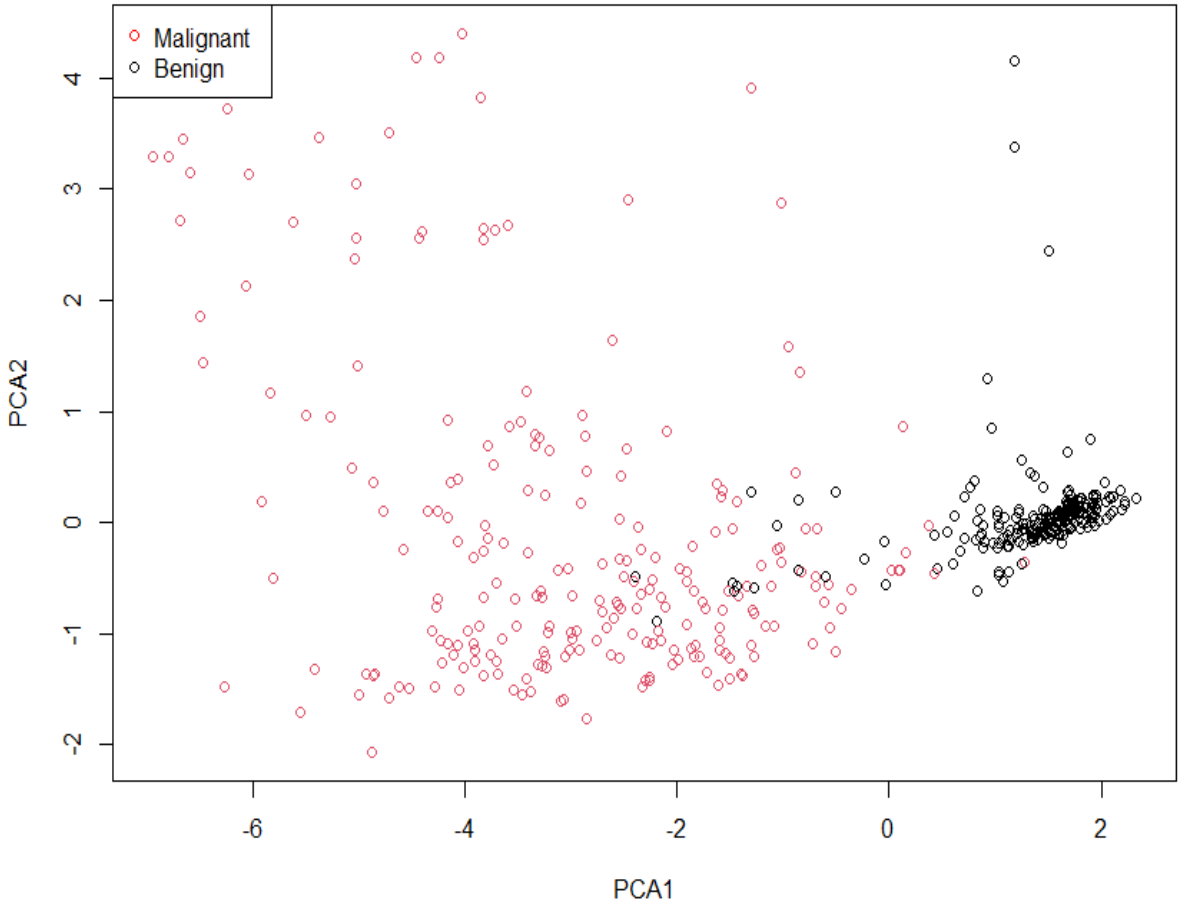

Figure 3 Pc1 vs Pc2

2.2 Logistic regression

All the variables are considered for this logistic regression. The predicted variables are standardized for the easiness of analysis.

The result of logistic regression is shown in table 6, from it the coefficients for all the predictors can be obtained. The P- values associated with each predictor variable indicate **Uniformity of Cell Size, Single Epithelial Cell Size, cell shape, Normal Nucleoli and Mitoses are not statistically significant based on a 95% confidence level**. Hence, we need to improve the model for finding the best predictors to understand whether the cancer is benign or malignant.

Also, the Column estimate in this table refers to the weightage of each predictor in this first Model.

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-1.3833	0.4222	-3.276	0.00105
Cl.thickness	1.5933	0.5274	3.021	0.00252
Cell.size	-0.2137	0.8476	-0.252	0.80095
Cell.shape	1.8756	0.9533	1.967	0.04913
Marg.adhesion	0.8352	0.3938	2.121	0.03391
Epith.c.size	0.1618	0.4565	0.354	0.72305
Bare.nuclei	0.8921	0.3957	2.254	0.02417
Bl.cromatin	1.4092	0.5551	2.539	0.01112
Normal.nucleoli	0.5058	0.4155	1.217	0.22351
Mitoses	0.7831	0.702	1.116	0.26463

Table 6 Logistic Regression Result

2.3 Improving the model

AIC and BIC methods are used for finding the best predictors for improving the model. These can be done using the R library *bestglm* and with the aid of the negative log-likelihood statistical technique. The two tables for AIC and BIC are given in tables 7 and 8 respectively. With the help of figure 4, the best model was found to be fifth and they are highlighted in red font in each table. That is, the model with five predictors is a good compromise.

In this improved model, we are ignoring **uniformity of Cell Size, Single Epithelial Cell Size, Normal Nucleoli and Mitoses** variables as they are statistically less significant.

Comments: The Selected model uses Clump Thickness, Uniformity of Cell Shape, Marginal Adhesion, Bare Nuclei and Bland Chromatin as a predictor and the coefficient of this predictor in the 5-predictor model is significantly different from zero. The coefficients of these variables are also positive, and the p-value associated with it is smaller too. Hence, the result suggests that these 5 variables are good in classifying data into benign and malignant cancer types. These results can be observed in table 9.

	Intercept	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	logLikelihood	AIC
0	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	-309.91057	619.82114
1	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	-84.73728	171.47456
2	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	-58.619	121.238
3	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	-43.61996	93.23993
4	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	-38.13487	84.26974
5	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	-35.69012	81.38024
6*	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	-34.54962	81.09925
7	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	-33.59922	81.19844
8	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-33.53611	83.07223
9	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-33.50495	85.00991

Table 7 AIC Data

	Intercept	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	logLikelihood	BIC
0	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	-309.91057	619.8211
1	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	-84.73728	175.6442
2	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	-58.619	129.5772
3	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	-43.61996	105.7488
4*	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	-38.13487	100.9482
5	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	-35.69012	102.2283
6	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	-34.54962	106.1169
7	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	-33.59922	110.3857
8	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-33.53611	116.4291
9	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-33.50495	122.5364

Table 8 BIC Data

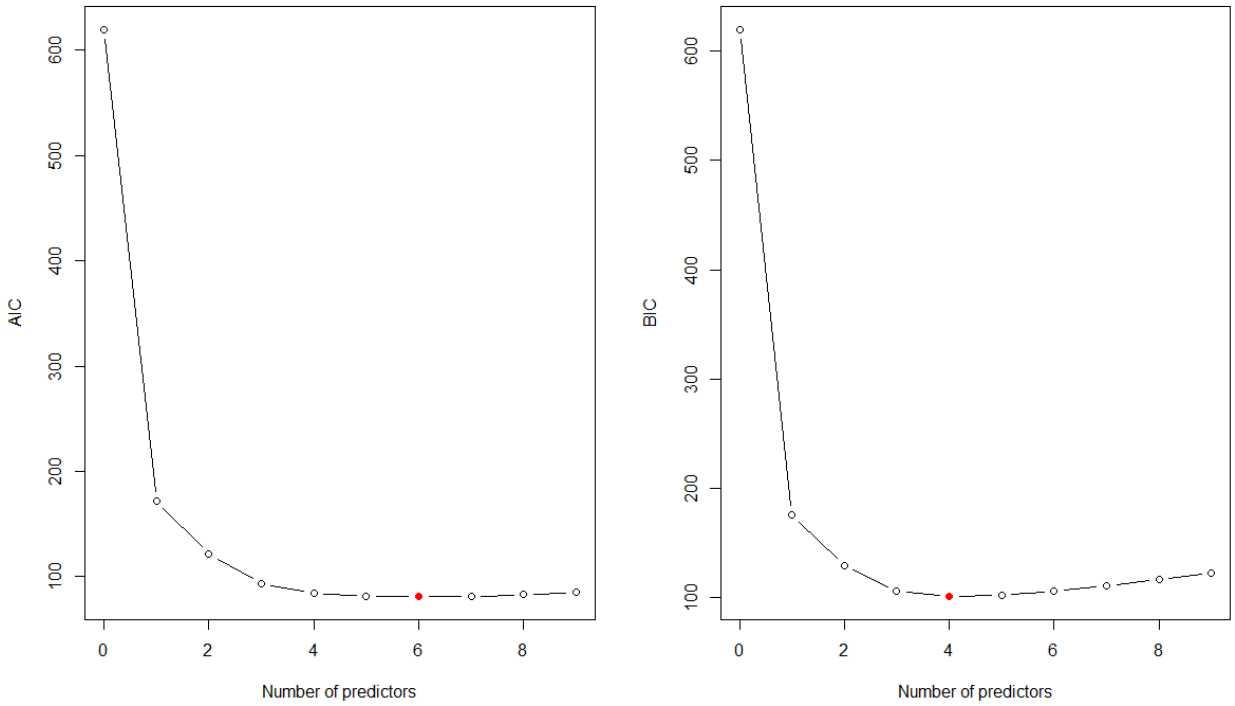

Figure 4 Plot comparing AIC and BIC Model

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-1.4506	0.3632	-3.994	6.49E-05
Cl.thickness	1.9174	0.5064	3.787	0.000153
Cell.shape	2.2163	0.6372	3.478	0.000505
Marg.adhesion	0.8076	0.3605	2.24	0.025083
Bare.nuclei	0.8621	0.3777	2.283	0.022451
Bl.cromatin	1.5552	0.5268	2.952	0.003157

Table 9 Modified Model Data

The predictive variables in the revised model can be the best subsets for this data as the p-values are less than 0.05.

The selected variables are **Clump Thickness, Uniformity of Cell Shape, Bare Nuclei, Marginal Adhesion and Bland Chromatin.**

2.4. Regularization method - LASSO and RIDGE

2.4.1 LASSO MODEL

It is done by scaling a penalty with a certain form and adding it to the loss function, which in this case is the negative log-likelihood function. They can be fitted in R by making use of the *glmnet* package. Here, we'll merely take the LASSO into account. The tuning parameter is added as a grid of values, and each value in the grid is used to fit the model with a LASSO penalty. This method reduces complete dependency on training data thereby increasing the predictive performance for the future observations.

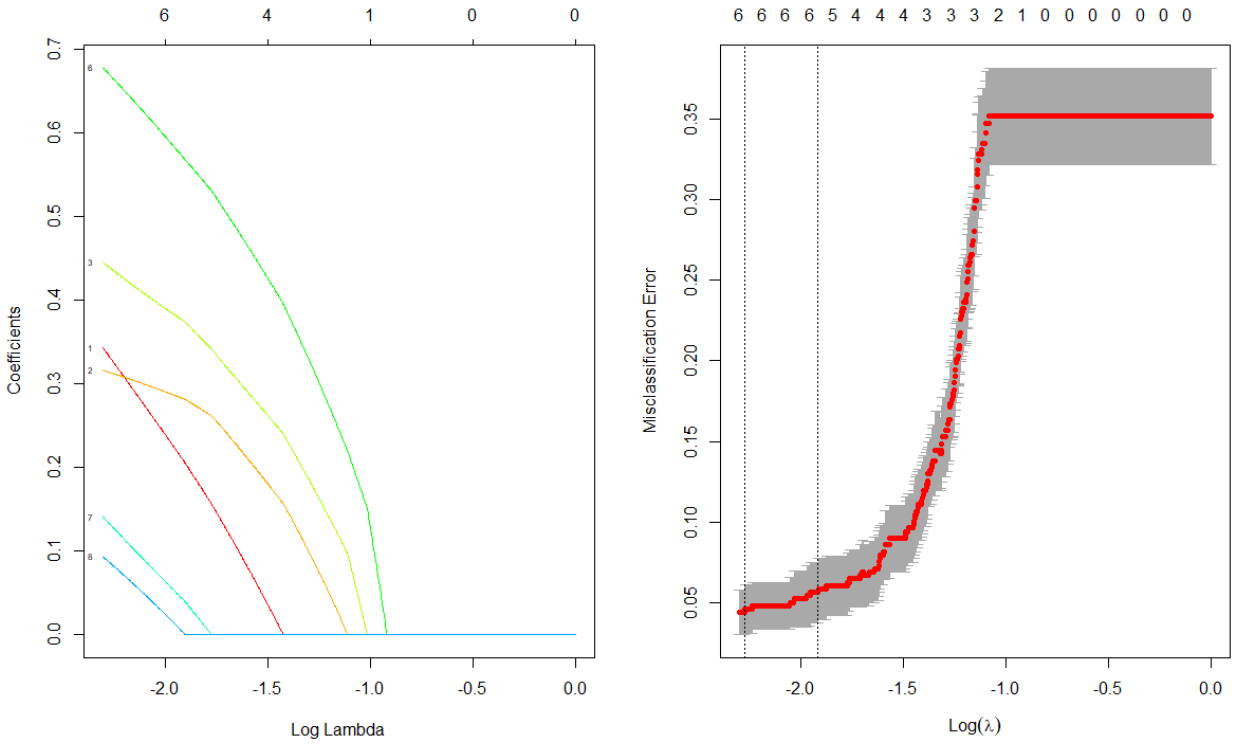

Figure 5 The effect of varying the tuning parameter in the logistic regression model with LASSO penalty for the Breast cancer data using k-fold cross-validation.

The lasso regularization gives the coefficients of the predictor variable as given in table 10, for the minimum lambda value 0.1028045 (considering k-fold cross-validation with k=10). Here, all the predictor variables are considered and it can be seen that the variables **Mitoses, Single Epithelial Cell Size** and **Marginal Adhesion are shrunk to zero**. Also, from figure 5 the last variable to be shrunk is **Bare Nuclei. Hence, Bare Nuclei is more statistically significant in this case.**

(Intercept)	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
-0.8322221	0.3335547	0.3139972	0.4399109	*	*	0.6707421	0.1336669	0.087179	*

Table 10 LASSO Model results

2.4.2 RIDGE MODEL

Analysis of the data through the ridge model gives the following results (figure 6 and table 11). The Ridge regularization gives the predictor variable values as given in table 11, for the minimum lambda value 0.1787526 (considering k-fold cross-validation with k=10). This method reduces complete

dependency on training data like LASSO. Here, all the predictor variables were considered, and a low coefficients is associated with the variables **Mitoses, Single Epithelial Cell Size and Marginal Adhesion**. Additionally, from the figure 6 the last variable to be shrunk (not to zero) is **Bare Nuclei.**

(Intercept)	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
-0.8645871	0.3966016	0.3330388	0.3590463	0.2733175	0.244523	0.4370575	0.3340433	0.2868954	0.15976

Table 11 Ridge Model Result

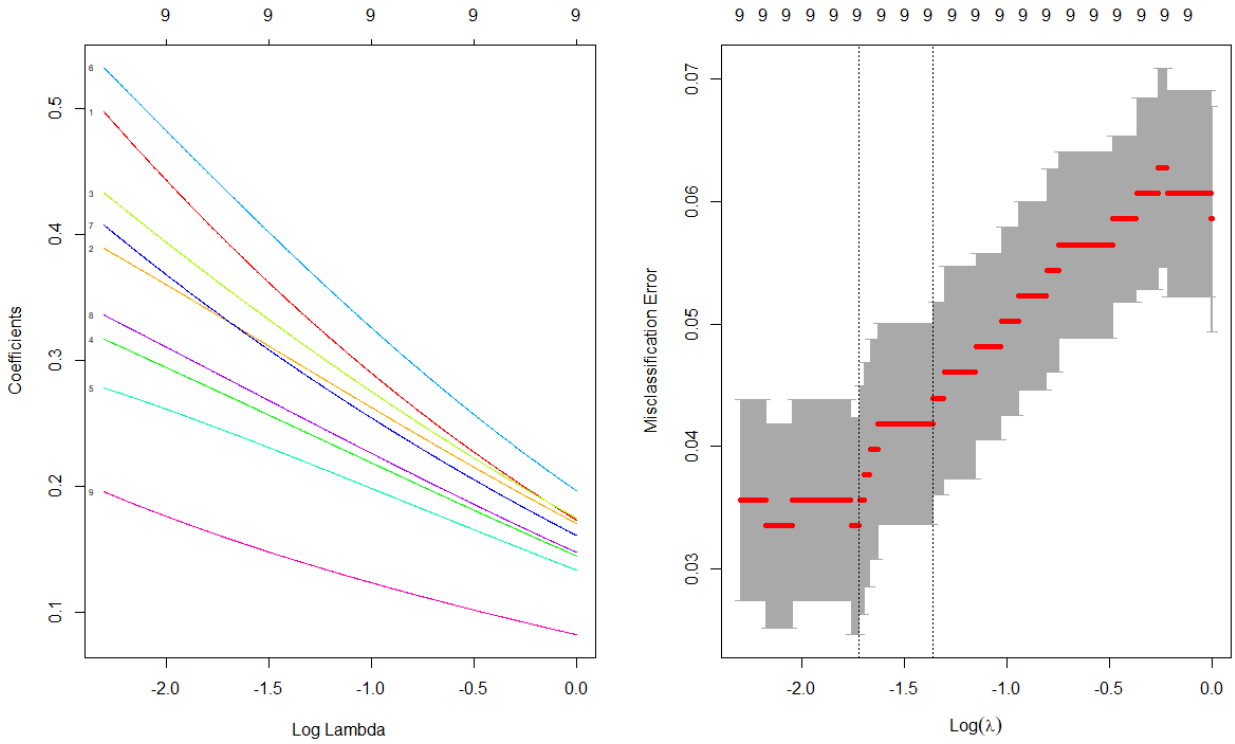

Figure 6 The effect of varying the tuning parameter in the logistic regression model with RIDGE penalty for the Breast cancer data using k-fold cross-validation.

2.5 Linear discriminant analysis

All the predictor variables are considered for the LDA analysis, and we obtained the following test results from the LDA model.
0 means benign and 1 means malignant

	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
0	-0.5375713	-0.598561	-0.5976732	-0.5060166	-0.5029638	-0.5903381	-0.5556705	-0.5226575	-0.3028916
1	1.0122984	1.086845	1.1010997	0.9694263	0.8906653	1.1606255	1.0113595	0.9532877	0.6115016

	Cl.thickness	*Cell.size*	*Cell.shape*	*Marg.adhesion*	*Epith.c.size*	*Bare.nuclei*	*Bl.cromatin*	*Normal.nucleoli*	*Mitoses*
LD1	*0.5611038*	*0.33828*	*0.3854692*	*0.1152415*	*0.10967931*	*0.90754572*	*0.26902377*	*0.32990994*	*0.03397*

Table 12 group Mean based on LDA

Table 13 LD1 values which correlate to the weightage of each predictive variable

Comments: In this Model, the linear discriminant value of variables is relatively high for Bare. nuclei, Clump Thickness and very low for Mitoses, Single Epithelial Cell Size and Marginal Adhesion (table 13). Additionally, the absolute value of the difference between the two groups for different variables can be found in the group mean table (Table 12). Since Mitoses, Single Epithelial Cell Size and Marginal Adhesion have the minimum values, classification based on these variables is difficult.

2.6. Quadratic discriminant analysis

All the predictive variables are given as input to this model, and it generated the group mean data as per table 14.

	Benign	Malign	Difference
Bare.nuclei	-0.5903381	1.1606255	1.7509636
Cell.shape	-0.597673	1.1010997	1.6987727
Cell.size	-0.598561	1.086845	1.685406
Bl.cromatin	-0.5556705	1.0113595	1.56703
Cl.thickness	-0.5375713	1.0122984	1.5498697
Normal.nucleoli	-0.5226575	0.9532877	1.4759452
Marg.adhesion	-0.5060166	0.9694263	1.4754429
Epith.c.size	-0.5029638	0.8906653	1.3936291
Mitoses	-0.3029	0.6115	0.9144

Table 14 Group Mean data based on QDA

Comments: The group means in this table 14 can be used to categorize a new observation into malignant or benign. The *difference* field takes the absolute value of the difference between each group. Since **Mitoses, Single Epithelial Cell Size and Marginal Adhesion have the minimum value, classification based on these variables is difficult.**

2.7. Model comparison using cross-validation techniques

The test data are used for generating 10 test errors based on cross-validation techniques for each model. The minimum mean test error is used for comparing the performance of each model. This process is given in the steps below.

1. Divide the test data into two parts, the first part is used to train the model and the second part will generate test errors.
2. Data for splitting is randomly chosen and the mean test error is stored in matrix form.
3. The above step is carried out on each model and repeats the whole process again 10 times and a matrix is obtained.
4. The matrix is converted to a data frame with suitable variable names and the best model is selected based on the minimum mean error

model_lda	-	Model based on Linear Discriminant Analysis
model_qda	-	Model based on quadratic Discriminant Analysis
model_lasso	-	Logistic Regression Model using LASSO with penalty
model_ridge	-	Logistic Regression Model using RIDGE with penalty
model_log_red_reg	-	Logistic Regression Model using subset selection without penalty

	model_lda	model_qda	model_lasso	model_ridge	model_log_red_reg
1	0.09836066	0.03278689	0.1147541	0.04918033	0.04918033
2	0.08196721	0.04918033	0.08196721	0.04918033	0.06557377
3	0.06557377	0.1147541	0.06557377	0.04918033	0.04918033
4	0.06557377	0.04918033	0.06557377	0.03278689	0.03278689
5	0.03278689	0.03278689	0.03278689	0.03278689	0.03278689
6	0.06557377	0.01639344	0.13114754	0.09836066	0.04918033
7	0.04918033	0.04918033	0.08196721	0.03278689	0.03278689
8	0.04918033	0.04918033	0.08196721	0.04918033	0.03278689
9	0.04918033	0.03278689	0.04918033	0.01639344	0.01639344
10	0.06557377	0.04918033	0.09836066	0.09836066	0.04918033

Table 15 Cross-validation results for each model obtained after 10 iterations

	variation	**Total Count**	**Mean**	**Standard deviation**	**median**	**minimum**	**maximum**
model_lda	1	10	0.06	0.02	0.07	0.03	0.1
model_qda	2	10	0.05	0.03	0.05	0.02	0.11
model_lasso	3	10	0.08	0.03	0.08	0.03	0.13
model_ridge	4	10	0.05	0.03	0.05	0.02	0.1
model_log_red_reg	5	10	0.04	0.01	0.04	0.02	0.07

Table 16 Summary of table 15

Comments- Table 16 shows that the mean error is minimum for the logistic regression model with reduced variables (0.04), this model is considered as the best-performing model for this data.

More details about the **logistic regression** model are provided in section 2.2. In this model, variables **Uniformity of Cell Size, Single Epithelial Cell Size, cell shape, Normal Nucleoli and Mitoses are not statistically significant**.

Mean accuracy can be determined by subtracting the mean error from 1. Therefore, The mean accuracy for The selected model is 96%.

The important point that has to be noted here is that we assumed the independent variable as continuous so that it can be analyzed by the LDA and QDA models. However, some of the actual variables are categorical in nature and logistical regression might give more accurate results but it will make the model more complicated. Also, since random data were selected for cross-validation the accuracy level can fluctuate. So, increasing the number of iterations (cross-verifications), results in good approximate for cross-validations.

Hence, it is not fair to input the same type of data to each model and expect a good accuracy level. However, in logistic regression reduced model performs best because of the optimal test data dependency. Additionally, logistic regression models with a penalty (RIDGE/LASSO) might perform well if more test data are included.

Confusion Matrix based on Test data

	Predicted			Accuracy	Test eror
		Benign	Malignant		
Model LDA	Benign	38	0		
	Malignant	4	19	0.93442623	0.06557377
Model QDA	Benign	35	3		
	Malignant	0	23	0.950819672	0.049180328
Model LASSO	Benign	38	0		
	Malignant	6	17	0.901639344	0.098360656
Model RIDGE	Benign	38	0		
	Malignant	6	17	0.901639344	0.098360656
logistic regression with subset	Benign	38	0		
	Malignant	3	20	0.950819672	0.049180328

Table 17 Confusion matrix of all the model from the last iteration cross-validation

Comments:

- Model LASSO and Model RIDGE predicted 6 observations as benign instead of malignant.
- Only QDA model predicted 3 observations as malignant even though they are benign.

2.8. Best predictor selection

Table 18 shows the insignificant variables from each model. Since the model chosen is logistic regression with subset, the interested predictive variables are **Clump Thickness, Uniformity of Cell Shape, Bare Nuclei, Marginal Adhesion and Bland Chromatin.**
The variables Single Epithelial Cell Size, Mitoses, Uniformity of Cell Size and Normal Nucleoli are not considered which means that out of 9 predictive variables only 5 are being focused. This reduction in variables can be justified Since the selected model with the above-mentioned 5 predictive variables alone gives around 96% accuracy which is comparatively more than other models' accuracy. Hence, by only using these 5 variables new observations can be classified based on cancer type accurately as benign or malignant. (The highlighted row in table 18 indicates the details of the selected model)

	Cl.thickness	**Cell.size**	**Cell.shape**	**Marg.adhesion**	**Epith.c.size**	**Bare.nuclei**	**Bl.cromatin**	**Normal.nucleoli**	**Mitoses**
model_lda				*	*				*
model_lasso				*	*				*
model_ridge				*	*				*
model_log_red_reg		*			*			*	*

Table 18 Insignificant variables given by each Model

3. Conclusion

The simple exploratory analysis provided an overall idea about the data, and it showed that the relation of the variable Mitoses on classifying the data is not significant. This helped in understanding the data and generating various modelling methods.

The data were analyzed mainly using the method like subset selection in logistic regression, regularized form of logistic regression, i.e., with a ridge and LASSO penalty and analysis (LDA) or quadratic discriminant analysis (QDA). For each model, the appropriate weightage/coefficients of all the predictor variables with the dependent variable is obtained and discussed. With the help of cross- Verification techniques most suitable model is selected which is found to be regularized form of **logistic regression with reduced variable by subset**. This Model is then compared and discussed whether it is a good fit for the data or not. However, the assumption of treating the data as continues will have a significant effect on the interpretation but it makes the model more complicated.

Finally, the most statically significant variables were found, and they **are Clump Thickness, Uniformity of Cell Shape, Bare Nuclei, Marginal Adhesion and Bland Chromatin, and the best model is selected using cross-verification based on test error.**

4.Reference

1. Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning. *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2018*, 1–4. https://doi.org/10.1109/EBBT.2018.8391453

2. Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms and Computational Technology*, *12*(2), 119–126. https://doi.org/10.1177/1748301818756225

3. Ray, R., Abdullah, A. A., Mallick, D. K., & Ranjan Dash, S. (2019). Classification of Benign and Malignant Breast Cancer using Supervised Machine Learning Algorithms Based on Image and Numeric Datasets. *Journal of Physics: Conference Series*, *1372*(1). https://doi.org/10.1088/1742-6596/1372/1/012062

4. Sankari, M. L., & Rajbharath, M. R. (n.d.). *Predicting Breast Cancer using Novel Approach in Data Analytics*. www.ijert.org

5.Appendix

R program for the analysis

```
#loading relevant libraries
library(mlbench)
library(dplyr)
library(psych)
library(bestglm)
library(glmnet)
library(nclSLR)
library(MASS)
library(ggplot2)
library(klaR)
## Load the data
data(BreastCancer)
?BreastCancer
## Print first few rows
head(BreastCancer)
BreastCancer=BreastCancer[-1]

################################Cleaning the data ##########################
#Cl.thickness,Cell.size,Epith.c.size,Cell.shape (scores),should be numeric with reason for choosing it #as a continouus data
#resopnsevariables changed to numeric
BreastCancer$Cl.thickness=as.numeric(BreastCancer$Cl.thickness)
BreastCancer$Cell.size=as.numeric(BreastCancer$Cell.size)
BreastCancer$Cell.shape=as.numeric(BreastCancer$Cell.shape)
BreastCancer$Marg.adhesion=as.numeric(BreastCancer$Marg.adhesion)
BreastCancer$Epith.c.size=as.numeric(BreastCancer$Epith.c.size)
BreastCancer$Bare.nuclei=as.numeric(BreastCancer$Bare.nuclei)
BreastCancer$Bl.cromatin=as.numeric(BreastCancer$Bl.cromatin)
BreastCancer$Normal.nucleoli=as.numeric(BreastCancer$Normal.nucleoli)
BreastCancer$Mitoses=as.numeric(BreastCancer$Mitoses)


#changing class values to either 1 or 0 (predictable variable)

BreastCancer$Class <- as.character(BreastCancer$Class)
BreastCancer$Class[BreastCancer$Class == "benign"] = 0
BreastCancer$Class[BreastCancer$Class == "malignant"] = 1
BreastCancer$Class=as.numeric(BreastCancer$Class)



BreastCancer <- BreastCancer %>% na.omit()
summary(BreastCancer)
count(BreastCancer)
```

```
# the relationships between the response variable and predictor variables and about the #relationships
between predictor variables?
par(mfrow=c(1,1))
hist(BreastCancer$Cl.thickness, main=" Cell Thickness", xlab="x", ylab="Frequency")
hist(BreastCancer$Cell.size, main="Cell Size", xlab="x", ylab="Frequency")
hist(BreastCancer$Cell.shape, main=" Cell Shape", xlab="x", ylab="Frequency")
hist(BreastCancer$Marg.adhesion, main="", xlab="x", ylab="Frequency")
hist(BreastCancer$Epith.c.size, main="", xlab="x", ylab="Frequency")
hist(BreastCancer$Bare.nuclei, main="", xlab="x", ylab="Frequency")
hist(BreastCancer$Bare.nuclei, main="", xlab="x", ylab="Frequency")
hist(BreastCancer$Bl.cromatin, main="", xlab="x", ylab="Frequency")
hist(BreastCancer$Mitoses, main="", xlab="x", ylab="Frequency")
pairs(BreastCancer[,-1])
summary(BreastCancer)
describe(BreastCancer)

BreastCancer_scaled=scale(BreastCancer[1:9])
BreastCancer_scaled_1=scale(BreastCancer[1:10])
BreastCancer_scaled_co = cor(BreastCancer_scaled_1)


######################### dependency checking#################################

par(mar=c(11,4,4,4))
barplot(sort(BreastCancer_scaled_co[c(1:9),c(10)]),
      main = "",
      xlab = "",
      ylab = "Dependency",
      las=2,
      names.arg =
c("Mitoses","Epith.c.size","Marg.adhesion","Cl.thickness","Normal.nucleoli","Bl.cromatin","Cell.size","
Cell.shape","Bare.nuclei"),
      col = "darkred",
      horiz = 0)
# the graph show that Mitoses dependency is relatively low.

#############################################################################################
#########################################################################

############################## PCA Analysis#################################
pca2 = prcomp(x=BreastCancer_scaled)
summary(pca2)
pca2$pc1
class(pca2)
#variable reduction table using pca analysis and plotting them

#plot
plot(pca2$x[,1], pca2$x[,2], xlab="First PC", ylab="Second PC")
# Add labels representing the cancer types
text(pca2$x[,1], pca2$x[,2], labels=round(BreastCancer[,10]), cex=0.7, pos=3)
```

```
legend(x="topleft", pch=1, legend = c("0-benign", "1-malignant"))
#from the plot 0 means the benign and 1 means malignant

##############################################################################################
#########################################################################

##################################data preparation ##################################
X1=scale(BreastCancer[1:9])
y=BreastCancer[,10]
BreastCancer_data=data.frame(X1,y)
n=nrow(BreastCancer_data)
p=ncol(BreastCancer_data)-1
cv=trainTestPartition(BreastCancer_data,trainFrac = 0.7)
y=cv$yTr
BreastCancer_data_tr=data.frame(cv$XyTr)
BreastCancer_data_te=data.frame(cv$XyTe)
y_te=cv$yTe
X1=data.frame(cv$XTr)
##################################logistic regression ##################################

logreg_fit_1 = glm(y ~ ., data=data.frame(BreastCancer_data_tr), family="binomial")
summary(logreg_fit_1)

###############subset selection ######################################33

best_fit_aic=bestglm(BreastCancer_data_tr,family=binomial,IC='AIC')
best_fit_bic=bestglm(BreastCancer_data_tr,family=binomial,IC='BIC')
best_fit_aic$Subsets
best_fit_bic$Subsets


## Identify best-fitting models
(best_AIC = best_fit_aic$ModelReport$Bestk)
(best_BIC = best_fit_bic$ModelReport$Bestk)
par(mfrow=c(1,2))
plot(0:p, best_fit_aic$Subsets$AIC, xlab="Number of predictors", ylab="AIC", type="b")
points(best_AIC, best_fit_aic$Subsets$AIC[best_AIC+1], col="red", pch=16)
plot(0:p, best_fit_bic$Subsets$BIC, xlab="Number of predictors", ylab="BIC", type="b")
points(best_BIC, best_fit_bic$Subsets$BIC[best_BIC+1], col="red", pch=16)

##########new model with updated variables###############
pstar = 1
## Check which predictors are in the 1-predictor model
best_fit_aic$Subsets[pstar+5,]
```

```
(indices = as.logical(best_fit_aic$Subsets[pstar+5, 2:(p+1)]))
BreastCancer_data_red = data.frame(X1[,indices], y)
## Obtain regression coefficients for this model
logreg1_fit = glm(y ~ ., data=BreastCancer_data_red, family="binomial")
summary(logreg1_fit)

##################Regularization lasso############################

grid = 10^seq(-5,2, length.out=5000)
## Fit a model with LASSO penalty for each value of the tuning parameter
lass_fit = glmnet(X1, y, family="binomial", alpha=1, standardize=FALSE, lambda=grid)
## Examine the effect of the tuning parameter on the parameter estimates
plot(lass_fit, xvar="lambda", col=rainbow(p), label=TRUE)

lass_cv_fit = cv.glmnet(as.matrix(X1), y, family="binomial", alpha=1, standardize=FALSE,
lambda=grid,
                type.measure="class")
plot(lass_cv_fit)
## Identify the optimal value for the tuning parameter
(lambda_lass_min = lass_cv_fit$lambda.min)

which_lambda_lass = which(lass_cv_fit$lambda == lambda_lass_min)
## Find the parameter estimates associated with optimal value of the tuning parameter
coef(lass_fit, s=lambda_lass_min)


#################################################################################

####################Regularization Ridge#################################
ridge_fit = glmnet(X1, y, family="binomial", alpha=0, standardize=FALSE, lambda=grid)
## Examine the effect of the tuning parameter on the parameter estimates
plot(ridge_fit, xvar="lambda", col=rainbow(p), label=TRUE)

#par(mfrow=c(1,2))
ridge_cv_fit = cv.glmnet(as.matrix(X1), y, family="binomial", alpha=0, standardize=FALSE,
lambda=grid,
                type.measure="class")
plot(ridge_cv_fit)
## Identify the optimal value for the tuning parameter
(lambda_ridge_min = ridge_cv_fit$lambda.min)

which_lambda_ridge = which(ridge_cv_fit$lambda == lambda_ridge_min)
## Find the parameter estimates associated with optimal value of the tuning parameter
coef(ridge_fit, s=lambda_ridge_min)


######################LDA ANALYSIS ######################
#k=linDA(variables=data.frame(X1),group=y)
#k
model <- lda(BreastCancer_data_tr$y~., data = data.frame(BreastCancer_data_tr),CV=FALSE)
```

```
model
######################## Qda Analysis#################################################
#k=linDA(variables=data.frame(X1),group=y)
#k

model_q <- qda(BreastCancer_data_tr$y~., data = data.frame(BreastCancer_data_tr),CV=FALSE)
model_q

############Cross validation based on test error#################################################
#model selection algorith
m=matrix(data=NA,nrow=10,ncol=5)
grid_1 = 10^seq(-5,0, length.out=500)
for (x in 1:10)
{
  cv_2=trainTestPartition(BreastCancer_data_te,trainFrac = 0.7)


  model_lda=lda(cv_2$yTr~., data = data.frame(cv_2$XTr),CV=FALSE)
  model_qda=qda(cv_2$yTr~., data = data.frame(cv_2$XTr),CV=FALSE)
  model_lasso=glmnet(data.frame(cv_2$XTr),cv_2$yTr, family="binomial", alpha=1,
standardize=FALSE, lambda=grid_1)
  lass_cv_fit_1 = cv.glmnet(as.matrix(cv_2$XTr), cv_2$yTr, family="binomial", alpha=1,
standardize=FALSE, lambda=grid_1,
                      type.measure="class")
  lambda_lass_min_1 = lass_cv_fit_1$lambda.min

  ridge_fit_1 = glmnet(as.matrix(cv_2$XTr), cv_2$yTr, family="binomial", alpha=0,
standardize=FALSE, lambda=grid_1)
  ridge_cv_fit_1 = cv.glmnet(as.matrix(cv_2$XTr), cv_2$yTr, family="binomial", alpha=0,
standardize=FALSE, lambda=grid_1,
                       type.measure="class")
  lambda_ridge_min_1 = ridge_cv_fit_1$lambda.min

  reg_red_mod = glm(cv_2$yTr ~ Cl.thickness + Epith.c.size + Bare.nuclei + Bl.cromatin +
Normal.nucleoli, data=data.frame(cv_2$XyTr), family="binomial")



  lda_test=predict(model_lda,cv_2$XyTe)
  qda_test=predict(model_qda,cv_2$XyTe)
  lasso_test = predict(model_lasso, cv_2$XTe, s=lambda_lass_min_1, type="response")
  ridge_test = predict(ridge_fit_1, cv_2$XTe, s=lambda_ridge_min_1, type="response")
  reg_red = predict(reg_red_mod, cv_2$XyTe, type="response")

  yhat_test_1=lda_test$class
  yhat_test_2=qda_test$class
  yhat_test_3 = ifelse(lasso_test > 0.5, 1, 0)
  yhat_test_4 = ifelse(ridge_test > 0.5, 1, 0)
  yhat_test_5 = ifelse(reg_red > 0.5, 1, 0)
```

```
  confusion_lda=table(observed=cv_2$yTe,predicted=yhat_test_1)
  confusion_qda=table(observed=cv_2$yTe,predicted=yhat_test_2)
  confusion_lasso=table(observed=cv_2$yTe,predicted=yhat_test_3)
  confusion_ridge=table(observed=cv_2$yTe,predicted=yhat_test_4)
  confusion_reg_red = table(Observed=cv_2$yTe, Predicted=yhat_test_5)

  m[x,1]= 1-mean(cv_2$yTe == yhat_test_1)
  m[x,2]= 1-mean(cv_2$yTe == yhat_test_2)
  m[x,3]= 1-mean(cv_2$yTe == yhat_test_3)
  m[x,4]= 1-mean(cv_2$yTe == yhat_test_4)
  m[x,5]= 1-mean(cv_2$yTe == yhat_test_5)

}
m_dt_fr=data.frame(m)
#renaming
names(m_dt_fr)[names(m_dt_fr) == 'X1'] <- 'model_lda'
names(m_dt_fr)[names(m_dt_fr) == 'X2'] <- 'model_qda'
names(m_dt_fr)[names(m_dt_fr) == 'X3'] <- 'model_lasso'
names(m_dt_fr)[names(m_dt_fr) == 'X4'] <- 'model_ridge'
names(m_dt_fr)[names(m_dt_fr) == 'X5'] <- 'model_log_red_reg'
m_dt_fr
describe(m_dt_fr)
summary(m_dt_fr)
```