1. Introduction:

A Tumor is defined as an abnormal body mass of body tissue. It could be two types, they are malignant (cancerous tumor) and benign (non-cancerous tumor). A patient diagnosis in breast cancer when the tumor is highly malignant type. This tumor originates in mainly in the woman breast cells area. In general, our body constantly grow new cells to replace the old or damaged cell. This is a healthy replacement of our body cell function; however, a tumor may form if the balance of death cell is disturbed. Unfortunately, the scientist could not yet find the exact reasons for this cell disturbance and breast cancer to grow.

Furthermore, according to the report, on average worldwide breast cancer is the second leading cause of death in women and this number is increasing (Yixiao Feng, Published online 2018 May 12). It possible to prevent this cancer to grow, if a patient can be diagnosed at an earlier stage. In reality, this early diagnosis cases are very low, around 10%. The aim of this project is proposing a better model which could predict the presence of malignant tumor based on a set of clinical measurements. Once we establish the best model, it can be used to predict the class of tumor either the Benign or Malignant when X's the measurements are known.

The essay will follow the 1st explaining the different model definition, explaining the data set, exploratory data analysis, explaining the best-subset selection, lastly comparing the model based on the testing error. In our analysis we compare two models between Logistic regression and Linear discriminate analysis (LDA).

2. Modelling

Since our target variable i.e Y takes two values such as Benign and Malignant and categorical, we can not use linear regression model for prediction. We only use linear regression when Y is always a continuous variable. If we use linear regression where response variable is binary, the prediction Y can exceed 0 and 1 range. Furthermore, there are two ways we could consider this classification problems; one is partitioning multivariable observations into groups i.e discriminant based approach (Linear discriminate analysis (LDA) and Quadratic discriminant analysis (QDA)) and secondly we could consider classification as a means of predicting a categorical response using a collection of predictor variables i.e regression-based approach (Logistic regression)

a. Logistic regression

Logistic regression is a classic predictive model technique when the target variable Y is binary categorical. However, if Y variable has more than 2 classes, it will be multi class classification. It will not possible to model the data in vanilla logistic regression. Here we use Log function as a probability function where we assume that log odd has linier relationship with input X. therefore it is possible to

$$\text{Log}[\frac{P(X)}{1-P(X)}] = \beta_0 + \beta_1 X + \varepsilon$$

$$\text{Log}[\frac{P(X)}{1-P(X)}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$$

b. Linear discriminate analysis (LDA) & Quadratic discriminant analysis (QDA)

These are discriminate analysis based on Bayes theorem and it is different analysis comparing from Logistic regression. There are two discriminate type analysis LDA and QDA. The different between LDA and QDA are there assumptions. LDA assumes normality distribution among the predictive variables and common covariance matrix to all classes. In QDA, although it does assumes normality distribution similar to LDA, the only difference is that for each classifier has its own covariance matrix. The observation of LDA and QDA follow is two sets,

First identify the distribution of predictive variable X for each of the class (Y=K1,K2…)

Next step is to introduce Bayes theorem to calculate the probability Pr(Y=k|X=x)

$$\Pr(Y=k \mid X=x) = \frac{\Pr(X=x|Y=k) * \Pr(Y=k)}{\sum_{p=1}^{p=k} \Pr(X=x|Y=p) * \Pr(Y=p)}$$

Nonetheless, it is ideal to analyse the data in LDA, when the requirement is to find linier boundary between the classifiers. However, QDA tries to find a non-linear boundary between classifiers.

3. Dataset:

In this project we collect the data from the Wisconsin Breast Cancer Database, collected by Dr. William H. Wolberg, University of Wisconsin Hospitals, Madison. It comprises measurements from n=699 women on 10 variables. On this database they have collected Fine needle aspirantion cytology (FNAC) data sample. FNAC is one of the popular methods for the breast cancer diagnosis, where they collect the blood sample through a needle. There are 9 clinical measurements i.e Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses. Those are the independent or predictive variables for the analysis. There is one target variable binary (benign or malignant), which is the dependent or responsive variable of our analysis.

a. Assumptions:

Every individual or item in our dataset or could observe values have two sets of quantities:

Firstly, observe value Y (response variable): a categorical (random) variable that can take one of K possible values, because of the regression-based approach classification. Y is the breast cancer class status

Secondly, X1,…….,Xp: p (random) variables, often called the predictor variables. The number of clinical measures

b. Cleaning the data:

It is a difficult task to analyse the data if the data set is not in a better format. Therefore, our first step would be to preparing the data in a format which could avoid all the problems during analysis such as overfitting, handling missing data, changing the data categories for better interpretation etc.

Table 1: Wisconsin Breast Cancer Database

Clinical Measurements features	Sample code number	Character	Id
	Clump Thickness	Ord.factor	1-10
	Uniformity of Cell Size	Ord.factor	1-10
	Uniformity of Cell Shape	Ord.factor	1-10
	Marginal Adhesion	Ord.factor	1-10
	Single Epithelial Cell Size	Ord.factor	1-10
	Bare Nuclei	Factor	1-10
	Bland Chromatin	Factor	1-10
	Normal Nucleoli	Factor	1-10
	Mitoses	Factor	1-10
	Class	Normal	Benign, Malignant
Class Distribution		Benign:458(65.5%)	
		Malignant:241(34.5%)	
Number of Missing values (Bare Nuclei)		16	
Number of Instances		699	

Each of the clinical measurement feature is presented in ordered categorically of 1 to 10, where 1 is closest to benign and 10 is closest to malignant. **Firstly**, in the sample code number column explains as id for each data sample row. Since it is not a clinical measurement for breast cancer diagnosis, this column is redundant for the analysis and deleted accordingly.

Secondly, except the ID column our predictive variables present in ordered categorically such as factor with 10 levels. It will be difficult to analyse, if I introduce to dummy variable for each of the category. Following other empirical research papers is this dataset and better explanation, I transform all the predictor variables into numeric form (Borges, October 2015). This normalization will help to transform the inconsistent data to consistent data. It also helps to transform the highly skewed dataset into less skewed dataset.

Since R store factor variable as an integer format, in our target variable class column Benign & Malignant represented as 1 & 2 accordingly. It is a popular method among the data analyst to introduce binary value when the target variable is binary categorical. Therefore, I have adopted the standard numerical labels 0/1 for Benign/Malignant. I have also changed the label for class variable to Y for easy referencing.

Thirdly, in the Bare Nuclei column in the data set have missing values in 16 different instances. Table-3 summarise the replacing the missing values using Multivariate Imputation by Chained Equations(MICE). Several methods are considered for managing the missing values in the dataset, they are filtering replacement of missing values, removing row with the missing value and replacing missing values with a dummy value.

The problem with removing the row will reduce the sample size to 683 instances. Since the limitation of the sample size and better predictability of our result, this paper echoes with other empirical analysis paper and rejects the idea of removing rows in the dataset (Sumathi, 2018). A key issue of proposing a better model and a good generalization of the model would achieve by avoiding 'overtraining' the classifier. The reducing the number of input features in the model would be the ideal way for this generalization. It would not be possible for better prediction if we reduce the sample size before even starting the analyse.

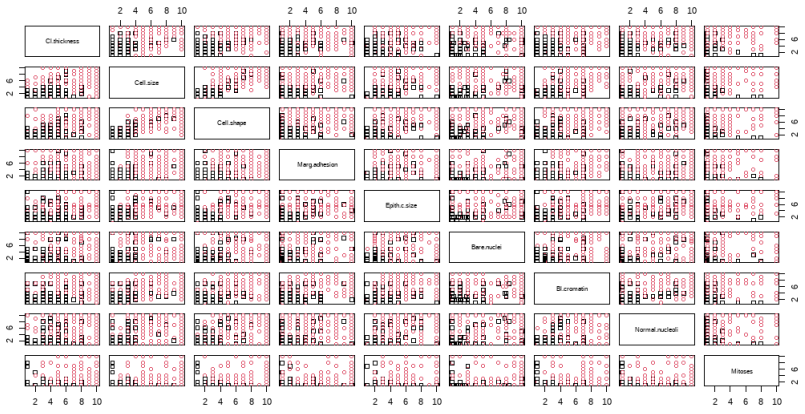
Update the missing value with mean or Multivariate Imputation is very popular in filter preplacement process. I have updated the missing value in Bare Nuclei with Multivariate Imputation. Since using mean value as a replacement is not a good option, each individual has different size of cell and it is not equal to the other cell mean size. Multivariate imputation by chained equations is undoubtedly a better method to deal with missing values in a dataset. It fills the missing value through an iterative series of predictive models. The ideal way is to create a 5 copies of original data set and replace the missing values by applying MICE method. Later report on combining result. The scope of this research would fall outside of this project, I implement imputate_na() in R to apply MICE procedure (Felix Neutatz, 13 May 2022).

4. Exploratory data analysis

There are various ways to explore our data analysis. In this analysis we analyse our data in three different ways such scatterplot matrix, graphical correlation plotting and correlation hypothesis testing.

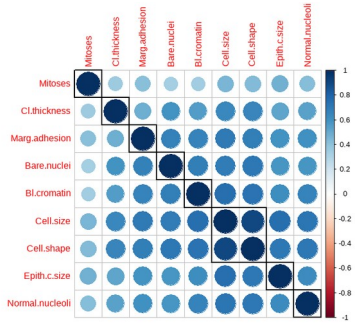
a. Scatterplot matrix analysis

Exhibit 1.1 Scatterplot matrix relationships



Above scatterplot matrix shows the relationships among the predictor variables and relationships between the predictor variables and the categorical response variables. I colour the plot in Black(Benign) and Red(Malignant) points according to the individuals tumor class with the different clinical measurements. Here, each label in the column value represents in the x axis whereas same label in the row represents in Y axis. Looking at this plot in a row, I can assert that most of the black points are line up between value 1 to 5. Although there are few red points between the value 1 to 5, the black points are highly visible. It can be interpreted that the chances of patient diagnosis in Benign class tumor increases when their clinical measurements are below 5 levels whereas chances diagnosis in Malignant tumor increases when the clinical measurements level increases.

Exhibit 1.2 Pearson multivariate analysis



Since the unbalance of responsive categorical variable such as 65.5% and 34.5% patient with Benign and Malignant accordingly. It would be beneficial to analyse the correlation between the predictor variables.

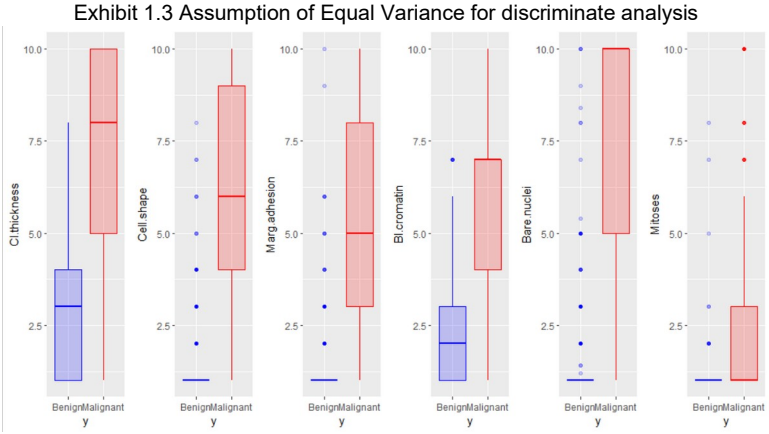
In generalised linear model algorithm assumes that predictor variable is independent from each other. Exhibit 1.2 represents the graphical view of correlation matrix whereas correlations are represented in colour gradient. It can be asserted cross the matrix that each predictive variable has a positive correlation. The highest positive correlation value is traced between cell size and cell shape. This correlation expected because tumor cell size increases along with cell shape. For robust analysis to remove multicollinearity for our further analysis, we could remove this cell.

Similar result shows at Exhibit 1.3.1 (appendix), all the correlation coefficient matrix are positive. The T-test, hypothesis H_0:ρ=0 (no linear correlation) against H_1:ρ>0 (positive correlation). All the coefficient is Significantly not equal to zero. Since the purpose of this project is to predict the cancer tumor class rather than the interpret the variable, we will keep all the predictive variable for further analysis.

b. Checking the Assumption of Equal Variance

In exhibit 1.3 shows us the different length box plot and clear that plots are different from each other. Now although our data did not follow the assumptions for LDA and QDA analysis, similarly to the correlation analysis we keep all the variables for further analysis.

Exhibit 1.3 Assumption of Equal Variance for discriminate analysis



5. Application of Best subset selection approach with different models

a. Setting a benchmark

It is essential to understand the estimation with all predictable variable, before we select all the best-subset in logistic regression.

Exhibit 5.1 Model with all of the predictor variables with logistic regression

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -9.507921   1.033162  -9.203  < 2e-16 ***
Cl.thickness    0.518906   0.129717   4.000 6.33e-05 ***
Cell.size       0.006012   0.184591   0.033  0.97402
Cell.shape      0.342802   0.205868   1.665  0.09588 .
Marg.adhesion   0.239287   0.113323   2.112  0.03472 *
Epith.c.size    0.074871   0.148650   0.504  0.61449
Bare.nuclei     0.370551   0.086425   4.288 1.81e-05 ***
Bl.cromatin     0.406405   0.156204   2.602  0.00927 **
Normal.nucleoli 0.141329   0.101084   1.398  0.16207
Mitoses         0.545793   0.297522   1.834  0.06659 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Above exhibit shows the estimation of the predictor variables for $\widehat{\beta_0}, \ldots\ldots, \widehat{\beta_9}$. Considering the z-value(hypothesis test) with the corresponding probability column, 6 predictable variables i.e cl.thickness, cell.shape, Marg, Bare, Bl and Mitoses; are significantly different from zero when testing at the 10% level. In other words, I would only reject the null hypothesis at 10%. Also, there are high coefficient estimation among those significant predictable variables, i.e as the level of cl.thickness goes up, the probability of individual diagnosis in Malignant cancer tumar. It is also consistent with our intuition.

However, in our dataset there are a number of predictable variables have large p value, which means that those variable does not contribute much to our model, it contains all the predictor variables. Similarly, the disadvantageous to include more predictive variables in a logistic regression model, than it actually need. Unnecessary predictive variable will have higher variances in estimate of the regression coefficients. Also, high predictive variances is regarded as a poor predictive performance.

Therefore it is necessary to use only predictive variables that is relevant to the model. I will consider best subset selection i.e AIC (slight variation of Mallow's Cp statistic), Bayes Information criterion (BIC) and k-fold cross validation to eliminate predictors. In all three tests small values of the information criterion consider a "better" model.

b. Best-subset selection

To select the best-subset, I will consider 2^p possible models. we will use R to estimate the parameters (maximum likelihood estimation technique).

c. Logistic regression:

i. AIC and BIC

AIC adds penalty for model complexity and penalty increases as the model complexity increases. Because this statistics is based upon SSE(sum of square error) so we need to minimise it as small as possible. we pick the model which has AIC as smallest as it possibly can be. In the appendix, exhibit 1.1 shows the AIC scores with the number of predicted variables. From the regression, model with 1 predicted variable has the higher AIC score whereas model with 7 predicted variables has the lowest AIC score around 134. Furthermore 5 predicted variables were selected in BIC. It is consistent with BIC definition that it considers penalise model complexity a little bit more than AIC. In the appendix exabit 1.2 shows that BIC score around 282 in 1 predictor variable whereas it scores lowest in 5 predictor variables. There is a patten emerges in both AIC and BIC sub-selection that it does not recommend a simple model i.e a model with one or two predictor variables. This result also suggests that model with 5 or 7 predictor variables are to be preferred.

Table 5.1: Model selection

	Model selection	Score
AIC	7 predictor variables	134.1104
BIC	5 predictor variables	161.1085
K-fold	7 predictor variables	0.03433476

Nonetheless, searching for the best subset selection, both AIC and BIC do exhaustive search over and each possible model. The result of best subset selection becomes computationally infeasible, when the number of predictor variable increases. One of popular method in classification method is K-fold cross validation for selecting the best subset.

ii. K-fold subset selection

In this project 10-folds cross validation is considered, where the data is divided into 10 folds of approximately equal size.

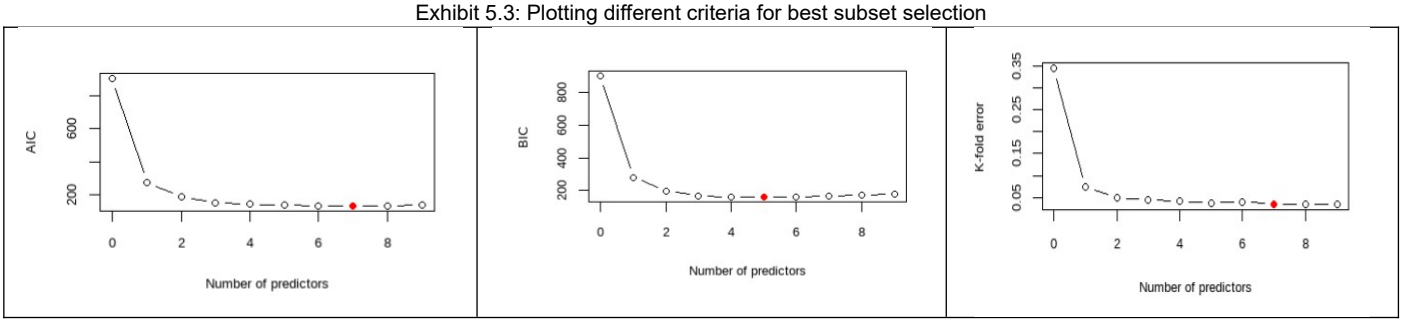
Table 5.2: K-fold subset selection best predictor variable

Predicted variable	k-fold error
1	0.07296137
2	0.04864092
3	0.04434907
4	0.04148784
5	0.03719599
6	0.04005722
7*	0.03433476

8	0.03433476
9	0.03433476

Since the criteria of K-fold cross validation is to select the best subset based on the less tasting error, above table 5.2 shows 7 predictor variables with a model has less testing error.

Exhibit 5.3: Plotting different criteria for best subset selection



Therefore, the above plot shows the best subset selection with different subset criteria, where AIC and K-fold validation suggests the models with 7 predictors and BIC suggests the model with 5 predictors respectively. Since both subset selector AIC and K-fold suggest the 7 predictors, it is a good compromise for 7 predictors for my further analyse.

Furthermore, for the future analysis for LDA and QDA we will keep the 7 predictor variables.

6. Misclassification rate

In all these kinds of methods where we are trying to predict an output on the basis of some input (supervised learning), predictive performance is the way to judging the best model. To improve the performance of a model it is necessary to reduce the number of predicted variables in a model. Therefore, we often compare different supervise learning methods (logistics and Discriminate analysis), based on the model prediction.

Furthermore, any rule of classification method that we come with is never going to be perfect. It is because we select the best-subset selection bast on the logistic regression. Following exhibit:5.3 shows we compromise the subset selection base on visually compare the criteria and so on. In reality, our data are much more complicated than any model that we can come up with.

Therefore, most obvious way to measure the performance of a classification method is according to it's degree of misclassification rate. The preferred method is to identify with a small degree of misclassification, among the different methods. I use empirical method in this project to estimate these misclassification probabilities.

The data used to construct the classifier in training data and the data use to estimate the misclassification rate are the validation data or test data. We will fit the model to the training data in pairs $(\underline{X}_1, Y_1), \ldots\ldots\ldots, (\underline{X}_n, Y_n)$ (x = values of our predictive variables of observation for individual 1…n; y= value of the categorical response variable of individual 1…n). This misclassification rate present in *K x K* matrix, which is often called confusion matrix.

If we use exactly the same data to train our classifier to fit the model and also to estimate the misclassification rate, generally it is referred to as in sample validation. The reason is the data that we use to assess the performance of our classifier is within the sample that we use to fit the data. If we are doing in sample validation, then the confusion matrix that we come up with is called the Training confusion matrix and the estimate of misclassification is referred to as training error.

Similarly, if the training datasets and the validation datasets are different then this is called out of sample validation with associated test confusion matrix and test error.

Training error is almost of approximation of test error, it tends to be underestimate, because we are using the same data to fit the model and then to test it. So typically, the training error will be less than the test error. However, we are interested in test error because this gives us more rigorous assessment of the performance of the classifier by testing on data that haven't seen before.

a. Logistic regression model (LR)

i. Training error

Table 6.1: Confusion training matrix with entire data set

	Predicted	
Observed	0	1
0	447	11
1	12	229

Classification rule is if our predictive probability is bigger than half, then we would say our prediction is that $\overline{y}$ = 1 and if our probability is less than a half 0.5 then $\overline{y}$=0. There is be one probability for every row in our dataset. If that probability is >0.5 we put 1(Malignant) and if it is not, we put 0(Benign).

Observed values are listed here zero and one, and the possible predicted values are listed here zero and one. The 447 number is telling us the when the observe value is 0, we predicted 0 458 times. When the observed value is 1, we predicted 0 12 times. When the observed value is 0, we predicted 1 11 times when the observed value is 1, we predicted 1 229 times.

Table 6.2 Normalising the classification probabilities full dataset

	Predicted	
Observed	0	1
0	0.975983	0.024017
1	0.049793	0.950207

Normalise the vector(matrix of classification probabilities**)** here we simple divide each of the rows by the row sum. We applied the fun normalised to every row.

Above table 6.2 help us to identify the number mistakes have been done on estimation. It's immediately clear from this matrix that we actually make quite a few mistakes when you observe values one. So, when the observed value is 1, 4% of the time we actually predict zero. We also find the average training error rate table6.3(appendix) around 3%.

ii. Testing error (out of sample validation)

There are two very popular ways of estimating the test error. Validation set approach and k-fold cross validation. K-fold cross validation is the superior procedure for estimating the misclassification rate. We divide the data into k-fold, in our analysis we divided the data into 10 folds. Here first fold treated as a test data and all of the folds are treated as training data. Then fit the classifier using the training data, next compute the test error over the test data and repeat the procedure in each fold give. It gives us the k estimates of the test error which we need to average to get overall estimate of test error or misclassification rate.

Table 6.3: Misclassification rate	
Training error	Testing error
0.032904	0.034335

As explained above, it is clear from the above table that training error of the logistic model is showing an optimistic estimation of the performance. Exhibit 6.4(appendix) showing the permitter estimations of 7 predicted model, where all the estimations are statistically significant at a 5% level except Normal and Mitoses. This could be because I have selected higher subset selection out of three best-subset choice.

b. LDA

It indicates from the exhibit 6.5(appendix) that when the class=0(Benign) & 1(Malignant) the means of predictor values all are positive accordingly. Since the object is a list, (see the appendix Exhibit 6.6) we also calculate the prior membership probabilities $\pi_1 \wedge \pi_2$ 65% and 34% accordingly. This suggests that around 34% of our data analysis include woman, who has been diagnosed with Malignant type breast cancer. Similarly, 65% of time our dataset did not include woman who has not been diagnosed.

Exhibit: 6.4: group means of Benign and Malignant

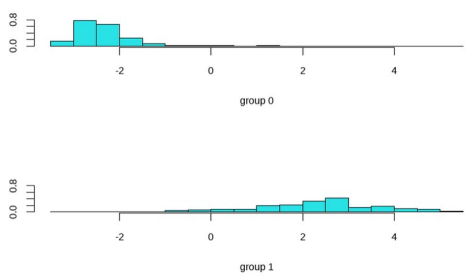


Exhibit:6.4 also shows us the group means of each predictor within the class. It suggests that on average 40% of woman diagnosis with cancer, have Bare.nuclei level of around 7, Mitoses is around 7, cl.thickness is around 7 etc.

Table 6.3 confusion matrix and training error with 7 predicted variables				
	Predicted		Testing error (10-fold cross validation)	
Observed	0	1	LDA	0.3447783
0	448	10		
1	20	221		

Observed values at table 6.3 shows the similar result with LR. However, there are slightly variation with the LDA prediction. when the observe value is 1, it predicts 0 20 times and 1 221 times. Similarly, our testing error for LDA almost identical with LR model keeping the 7 predicted variables. Nevertheless, LDA testing is slightly higher 0.3447783, whereas LR is 0.034335.

7. Conclusion:

It is clear from the above analysis that Logistic regression produces less testing error comparing the Linear discriminant analysis. Due to time limitation this project did explore the Quadratic analysis. However, it can presume that QDA testing error will be higher comparing LDA because of the QDA draw back. It is higher dimension of the data set would lead to high variance. In conclusion, though there are very little difference between LR and LDA testing error, LR has less error and it is easy to calculate the subset selection. Furthermore, for future prediction of Breast cancer I would recommend the Logistic regression.

Appendix

Exhibit 1.1: Logistic regression AIC Best subset selection

```
> bss_fit_AIC$Subsets
   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    -450.26372 900.5274
1       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE       FALSE       FALSE          FALSE   FALSE    -137.77717 277.5543
2       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE        TRUE       FALSE          FALSE   FALSE     -93.29491 190.5898
3       TRUE         TRUE      TRUE      FALSE         FALSE        FALSE        TRUE       FALSE          FALSE   FALSE     -75.89018 157.7804
4       TRUE         TRUE     FALSE       TRUE         FALSE        FALSE        TRUE        TRUE          FALSE   FALSE     -68.70794 145.4159
5       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE        TRUE        TRUE          FALSE   FALSE     -65.20969 140.4194
6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE        TRUE        TRUE          FALSE    TRUE     -62.31797 136.6359
7*      TRUE         TRUE     FALSE       TRUE          TRUE        FALSE        TRUE        TRUE           TRUE    TRUE     -61.20928 136.4186
8       TRUE         TRUE     FALSE       TRUE          TRUE         TRUE        TRUE        TRUE           TRUE    TRUE     -61.07900 138.1580
9       TRUE         TRUE      TRUE       TRUE          TRUE         TRUE        TRUE        TRUE           TRUE    TRUE     -61.07896 140.1579
```

Exhibit 1.2: Logistic regression AIC Best subset selection

```
> bss_fit_BIC$Subsets
   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    -450.26372 900.5274
1       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE       FALSE       FALSE          FALSE   FALSE    -137.77717 282.1040
2       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE        TRUE       FALSE          FALSE   FALSE     -93.29491 199.6891
3       TRUE         TRUE      TRUE      FALSE         FALSE        FALSE        TRUE       FALSE          FALSE   FALSE     -75.89018 171.4293
4       TRUE         TRUE     FALSE       TRUE         FALSE        FALSE        TRUE        TRUE          FALSE   FALSE     -68.70794 163.6145
5*      TRUE         TRUE     FALSE       TRUE          TRUE        FALSE        TRUE        TRUE          FALSE   FALSE     -65.20969 163.1676
6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE        TRUE        TRUE          FALSE    TRUE     -62.31797 163.9338
7       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE        TRUE        TRUE           TRUE    TRUE     -61.20928 168.2661
8       TRUE         TRUE     FALSE       TRUE          TRUE         TRUE        TRUE        TRUE           TRUE    TRUE     -61.07900 174.5552
9       TRUE         TRUE      TRUE       TRUE          TRUE         TRUE        TRUE        TRUE           TRUE    TRUE     -61.07896 181.1048
```

Exhibit 1.3: finding the AIC and BIC recommended predictor variable

```
> ## Identify best-fitting models
> (best_AIC = bss_fit_AIC$ModelReport$Bestk)
[1] 7
> (best_BIC = bss_fit_BIC$ModelReport$Bestk)
[1] 5
```

Exhibit 1.4: finding the K-fold subset selection best predictor variable

```
> ## Identify the number of predictors in the model which minimises test error according to K-fold subset selection
> (best_cv = which.min(cv_errors) - 1)
[1] 7
```

Exhibit 1.3.1 Correlation hypothesis testing

```
Call:psych::corr.test(x = sf[, 1:9])
Correlation matrix
              Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses
Cl.thickness          1.00      0.64       0.65          0.49         0.52        0.60        0.56           0.54    0.35
Cell.size             0.64      1.00       0.91          0.71         0.75        0.69        0.76           0.72    0.46
Cell.shape            0.65      0.91       1.00          0.68         0.72        0.72        0.74           0.72    0.44
Marg.adhesion         0.49      0.71       0.68          1.00         0.60        0.67        0.67           0.60    0.42
Epith.c.size          0.52      0.75       0.72          0.60         1.00        0.59        0.62           0.63    0.48
Bare.nuclei           0.60      0.69       0.72          0.67         0.59        1.00        0.68           0.59    0.34
Bl.cromatin           0.56      0.76       0.74          0.67         0.62        0.68        1.00           0.67    0.34
Normal.nucleoli       0.54      0.72       0.72          0.60         0.63        0.59        0.67           1.00    0.43
Mitoses               0.35      0.46       0.44          0.42         0.48        0.34        0.34           0.43    1.00
Sample Size
[1] 699
Probability values (Entries above the diagonal are adjusted for multiple tests.)
              Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses
Cl.thickness             0         0          0             0           0           0           0              0       0
Cell.size                0         0          0             0           0           0           0              0       0
Cell.shape               0         0          0             0           0           0           0              0       0
Marg.adhesion            0         0          0             0           0           0           0              0       0
Epith.c.size             0         0          0             0           0           0           0              0       0
Bare.nuclei              0         0          0             0           0           0           0              0       0
Bl.cromatin              0         0          0             0           0           0           0              0       0
Normal.nucleoli          0         0          0             0           0           0           0              0       0
Mitoses                  0         0          0             0           0           0           0              0       0

 To see confidence intervals of the correlations, print with the short=FALSE option
```

Exhibit 6.4: estimation of best subset with 7 predictable variables.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.43440    0.99228  -9.508  < 2e-16 ***
Cl.thickness     0.52176    0.12872   4.053 5.05e-05 ***
Cell.shape       0.36652    0.15914   2.303  0.02127 *
Marg.adhesion    0.25133    0.10853   2.316  0.02057 *
Bare.nuclei      0.37450    0.08613   4.348 1.37e-05 ***
Bl.cromatin      0.41650    0.15387   2.707  0.00679 **
Normal.nucleoli  0.15260    0.09829   1.552  0.12055
Mitoses          0.54795    0.29306   1.870  0.06152 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Exhibit 6.5 Linear discriminate best-subset with 7predicted model and estimation of $\pi_1 \wedge \pi_2$

```
Prior probabilities of groups:
        0         1
0.6552217 0.3447783

Group means:
  Cl.thickness Cell.shape Marg.adhesion Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
0     2.956332   1.443231      1.364629    1.365066    2.100437        1.290393 1.063319
1     7.195021   6.560166      5.547718    7.628216    5.979253        5.863071 2.589212

Coefficients of linear discriminants:
                    LD1
Cl.thickness    0.18905498
Cell.shape      0.18947159
Marg.adhesion   0.05084074
Bare.nuclei     0.25839982
Bl.cromatin     0.13682060
Normal.nucleoli 0.10747128
Mitoses         0.03826441
```

Exhibit 6.6 Finding list in LDA

```
> View(BreastCancer)
> is.list(lda_fit)
[1] TRUE
> lda_fit$prior
        0         1
0.6552217 0.3447783
> table(sf$y)/ nrow(sf)

        0         1
0.6552217 0.3447783
> is.list()
```

Exhibit: 6.6 Estimation in QDA

```
Call:
qda(y ~ ., data = sf_red)

Prior probabilities of groups:
        0         1
0.6552217 0.3447783

Group means:
  Cl.thickness Cell.shape Marg.adhesion Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
0     2.956332   1.443231      1.364629    1.388210    2.100437        1.290393 1.063319
1     7.195021   6.560166      5.547718    7.626556    5.979253        5.863071 2.589212
```

Bibliography

Borges, L. R. (October 2015). Analysis of the Wisconsin Breast Cancer Dataset and. *ResearchGate*, Conference: Workshop de Visão Computacional.

Felix Neutatz, B. C. (13 May 2022). Data Cleaning and AutoML: Would an Optimizer Choose to Clean? *Schwerpunktbeitrag*, Datenbank-Spektrum volume 22, pages121–130 (2022).

Male Breast Cancer Causes, Risk Factors for Men, Symptoms and Treatment on. (n.d.). Retrieved 12 6, 2022, from Medicinenet.com: http://www.medicinenet.com/male_breast_cancer/article.htm

Sumathi, B. M. (2018). Feature selection using Linear Discriminant Analysis for breast cancer dataset. *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1-5, doi: 10.1109/ICCIC.2018.8782399.

Yixiao Feng, a. M.-C. (Published online 2018 May 12). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *PMCID: PMC6147049*, PMID: 30258937.

R-code