Final Assignment

Analysis of Chowdary Dataset
Applied Statistical Modelling CS7DS3

Submitted by:

Karan Dua (21331391)

Objective

The primary objective of this statistical modelling project is to analyse the Chowdary dataset. This dataset measures the gene expression levels of tissue samples taken from the lymph node-negative breast tumours and Dukes' B colon tumours. There are 104 observations and 182 variables, plus the tumour variable (which is to be predicted), in the dataset. The project aims to determine which genes significantly impact cancer type designation and whether their influence is increasing or decreasing.

I conducted a series of statistical analyses on the dataset to achieve the aforementioned goals. Firstly, I performed Data-prepressing on the data. Next, I applied **Normalisation Test** on the dataset to assess the normal distribution of the data. Next, I followed the **Feature Selection process** on the dataset using **Lasso Regression, Elastic-Net regression and Correlation test** to reduce the dimensionality of the dataset and identify the features that primarily influence the cancer type designation. Finally, I have analysed the results using a **Logistic Regression model** to identify the influence of each of the selected genes, and whether it is increasing or decreasing. The **Results and Discussion** section of this report presents a comprehensive analysis of the findings and conclusions of this study.

Methodology

The below flow diagram represents the methodology that I have followed to analyse this dataset:

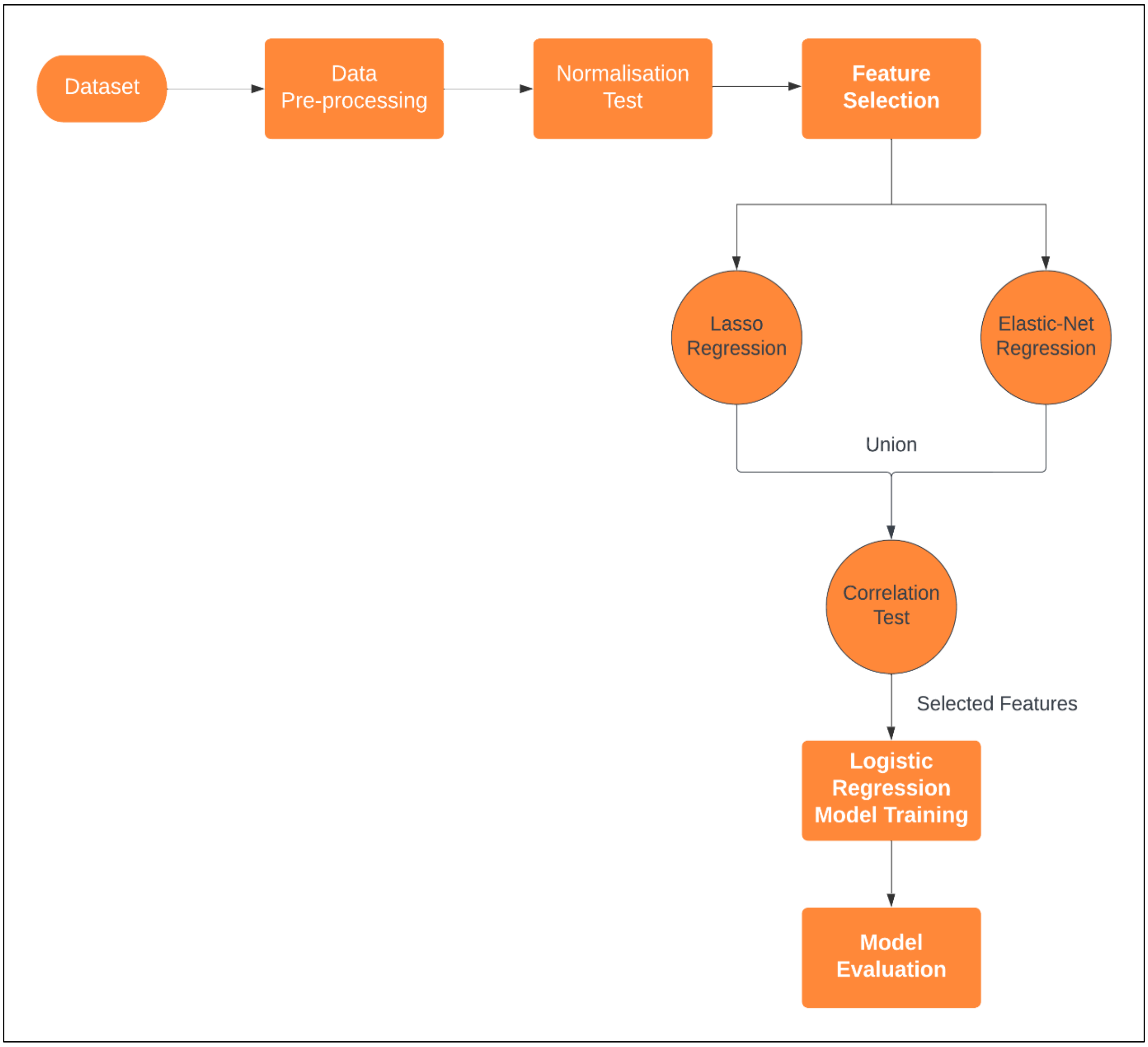


Figure 1: Flow Diagram for analysing the Chowdary dataset

Dataset

The Chowdary dataset comprises of **104 observations, with a total of 182 predictor variables** representing gene expression levels that potentially influence cancer type designation. Additionally, the dataset contains **a predicted variable, "tumour",** which corresponds to lymph node-negative breast tumours and Dukes' B colon tumours. **The variable "B" denotes the former, while "C" represents the latter, as indicated by the tumour variable in the dataset.**

Data Pre-processing

To pre-process the data, I first loaded the dataset into a data frame in R. **The first column in the dataset contained a value that represents an ID for each observation. Therefore, I removed this column from the data frame.**

Next, I checked if the dataset has any **Null values** and if it required any imputations. As per my analysis, this **dataset does not contain any null values in any observation, therefore, imputation of the data is not required.**

Then, I performed **data scaling for all the numeric columns in the dataset, except for the tumour column**. The primary reason for performing data scaling is to **bring all the variables to a common scale**. If we do not scale the data, the **variables with higher ranges and units may dominate the model, leading to biased results**. Therefore, we perform data scaling in this study to ensure that all the variables have equal importance in the analysis and to improve the performance of the statistical models.

Additionally**, I converted the tumour column to a numeric column with 0/1 values, where "0" represent tumour type "B" and "1" represents tumour type "C"**

Finally, I split the dataset into a training dataset and a test dataset with 80% of the data used for training and 20% data used for testing.

Normalisation Test

I have performed normalisation test on the dataset to analyse the normal distribution of the data. The primary reason for performing this test is that if the data is not normally distributed, it can lead to biased results and incorrect conclusions.

To perform this test, I have first executed the **Shapiro-Wilk test** on the data to identify if the variables come from a normally distributed population. This test computes the correlation between the dataset and the corresponding values that would be estimated from a normal distribution that has mean and variance equivalent to the observed data. The result of this test is a p-value in the dataset. A p-value less than 0.05 indicates that the dataset is not normally distributed, whereas a p-value greater than 0.05 suggests that the dataset follows a normal distribution. **I have identified that none of the 182 variables in the dataset followed the normal distribution.**

To further validate this, I have **generated Q-Q plots** for some of the features in the data as it is not possible to the plots for all 182 features. Figure 2 represents the Q-Q plot for 6 features in the dataset. **I can clearly analyse from the figure that these features are not normally distributed as the data is not following the straight line. This confirms the results that were analysed using the Shapiro-Wilk test.**

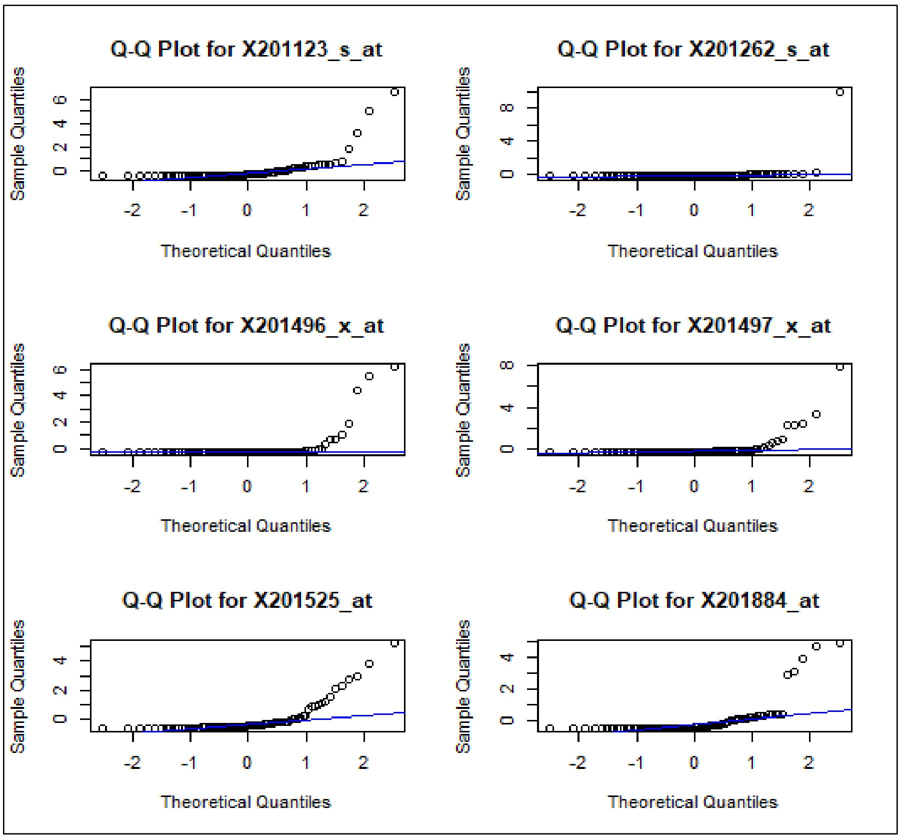


Figure 2: Q-Q plots for the features

Since our data is not normally distributed, I have decided to use multiple feature selection methods for the variables in the dataset to ensure that our model is robust to different sources of variation in the data. This process will help me to identify the most important features in the data while minimizing the impact of any potential biases caused by non-normality.

Feature Selection

The Chowdary dataset contains 182 variables that can be used to predict the type of cancer tumour. However, not all features contribute significantly to prediction. Therefore, **I have performed a Feature Selection process to identify the most significant features in the dataset that contribute in the overall accurate predictions.** Another advantage of using feature selection process is that it helps to reduce the complexity of the model, thereby reducing the risk of overfitting, which leads to better generalisation of the model for unseen data.

Lasso Regression

Lasso regression is a linear regression technique that reduces the coefficients of less important features towards zero while retaining the coefficients of the most important features. **The primary reason for using Lasso Regression for feature selection is that it leads to a simpler model with fewer features, which in turn reduces the risk of overfitting and improves its generalization to new data**. By penalizing the magnitude of the coefficients, Lasso regression encourages the model to select only the most relevant features, thus helping to identify the key predictors of the target variable.

I utilized the training dataset to train a Lasso regression model, employing the **10-fold cross-validation technique** to identify the optimal alpha and lambda values for the model. The optimized alpha value is determined by selecting the value at which the cross-validation error is minimal, while the optimized lambda value is determined by selecting the value at which the cross-validation deviance is minimal. Figure 3 highlights the cross-validation plot for Lasso regression.

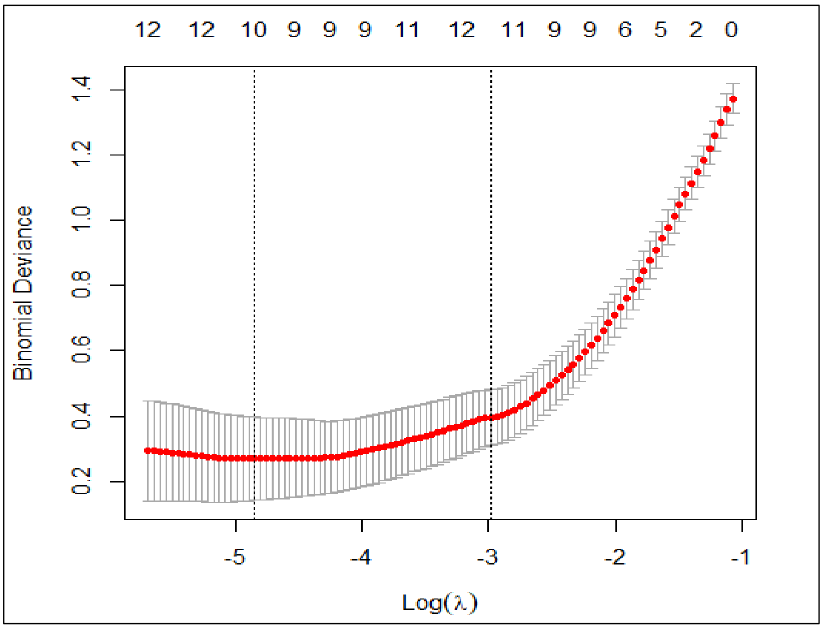


Figure 3: Cross-validation plot for Lasso Regression

After conducting the analysis, **I found that the optimal alpha value for the Lasso model is 1, while the optimal lambda value is 0.007875139**. I used these optimised values to train the Lasso model and identify the non-zero coefficients in the model which represent the most significant features. **I have identified 10 features with non-zero coefficients.** These features are:

Lasso Features	
X204653_at	X202575_at
X209016_s_at	X212236_x_at
X209604_s_at	X209351_at
X218502_s_at	X202831_at
X202859_x_at	X201496_x_at

Table 1: Significant features identified by Lasso Regression

Elastic-Net regression

Elastic Net regression is a linear regression technique that combines both Lasso and Ridge regression. **The primary reason for using Elastic Net regression for feature selection is that it aims to overcome the limitations of Lasso and Ridge regression by reducing the coefficients of less important features towards zero like Lasso and shrinking the magnitude of the coefficients towards zero like Ridge**. This approach leads to a model that is both sparse and stable, with a balance between bias and variance. Another benefit of Elastic Net regression is that it is particularly useful when dealing with datasets that contain a large number of features, some of which are highly correlated which is essentially suitable for our dataset.

I utilized the training dataset to train an Elastic Net regression model, using the **10-fold cross-validation technique** to identify the optimal alpha and lambda values for the model. The optimized alpha value is determined by selecting the value at which the cross-validation error is minimal, while the optimized lambda value is determined by selecting the value at which the cross-validation deviance is minimal. Figure 4 highlights the cross-validation plot for Elastic Net regression.

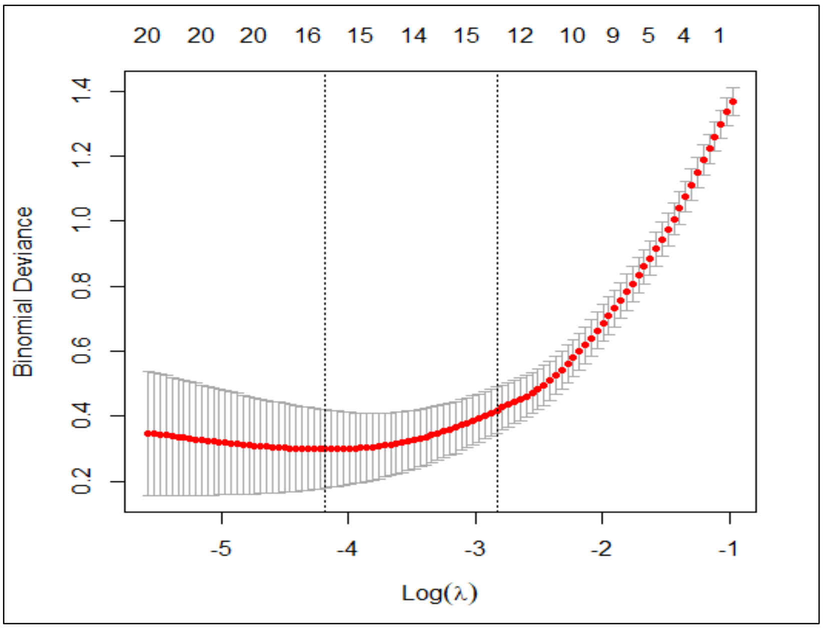


Figure 4: Cross-validation plot for Elastic Net Regression

After conducting the analysis, **I found that the optimal alpha value for the Elastic Net model is 0.5, while the optimal lambda value is 0.01529114**. I used these optimised values to train the Elastic Net model and identify the non-zero coefficients in the model which represent the most significant features. **I have identified 16 features with non-zero coefficients.** These features are:

Elastic Net Features	
X204653_at	X202831_at
X209016_s_at	X209343_at
X209604_s_at	X205044_at
X202575_at	X205225_at
X202859_x_at	X201909_at
X218502_s_at	X201525_at
X212236_x_at	X201496_x_at
X209351_at	X202018_s_at

Table 2: Significant features identified by Elastic Net Regression

Comparison of Lasso and Elastic Net Regression

I can clearly analyse that Elastic Net identified more significant features than Lasso Regression. Therefore**, I have decided to use union of the features identified by Lasso and Elastic Net regression for further analysis**. The primary reason for using the union of features identified by Lasso and Elastic Net regression is that **it combines the strengths of both methods**. **Lasso tends to select a subset of features, while Elastic Net tends to include groups of highly correlated features.** By taking the union of the selected features, we can retain the best of both worlds - **including highly correlated groups of features while also selecting only the most relevant ones**. This can lead to a more robust and accurate model with improved generalization performance on new data.

Correlation Test

Ultimately, to select the final set of features to train our model, I have performed Correlation test on the union of features selected by Lasso regression and Elastic Net regression.

Correlation test is a statistical method to analyse the association between the two continuous variables in the dataset. It is used to identify how strongly two continuous variables are related to each other in the dataset. The test provides coefficients for each pair of variables in the dataset in range from -1 to +1, with negative values indicating a negative correlation, which signifies that as one variable increases, the other decreases, and positive values indicating a positive correlation, which signifies that as one variable increases, the other also increases, and zero indicating no correlation. Figure 5 highlights the correlation plot generated for the 16 significant features.

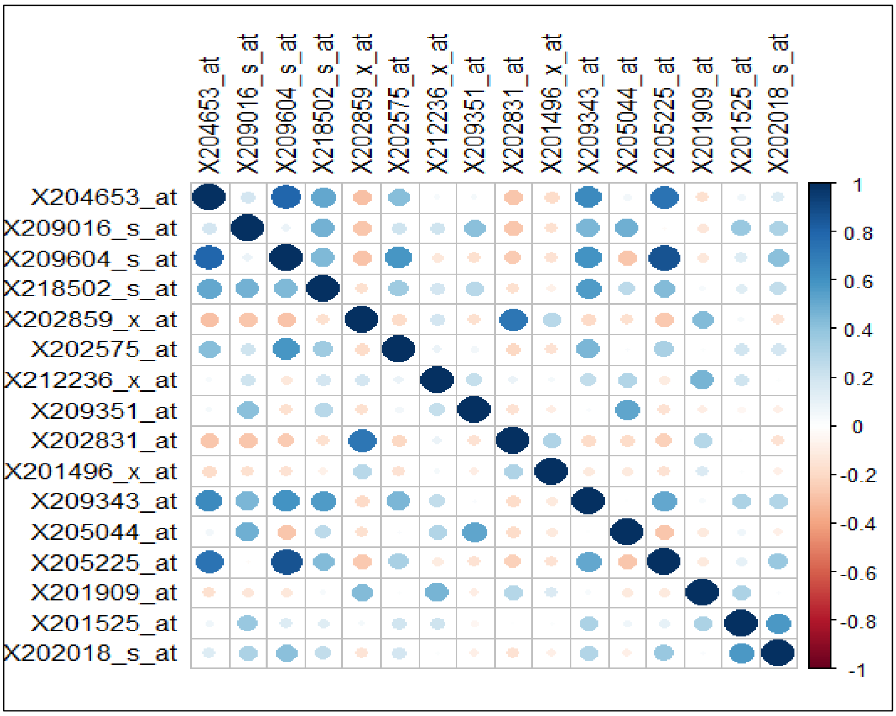


Figure 5: Correlation plot for Significant Features

Using the above figure, I have tried to identified the features that are highly correlated. These features were selected using the correlation score. **If the correlation score was found to be greater than 0.7 or less than -0.7, then two variables were considered strongly related. I have removed these variables from the final features set which is used to build the model as these can lead to biased results.** Table 3 shows features with high correlation:

Highly Correlated Features
X209604_s_at
X205225_at
X202859_x_at

Table 3: Highly Correlated features identified using Correlation Test

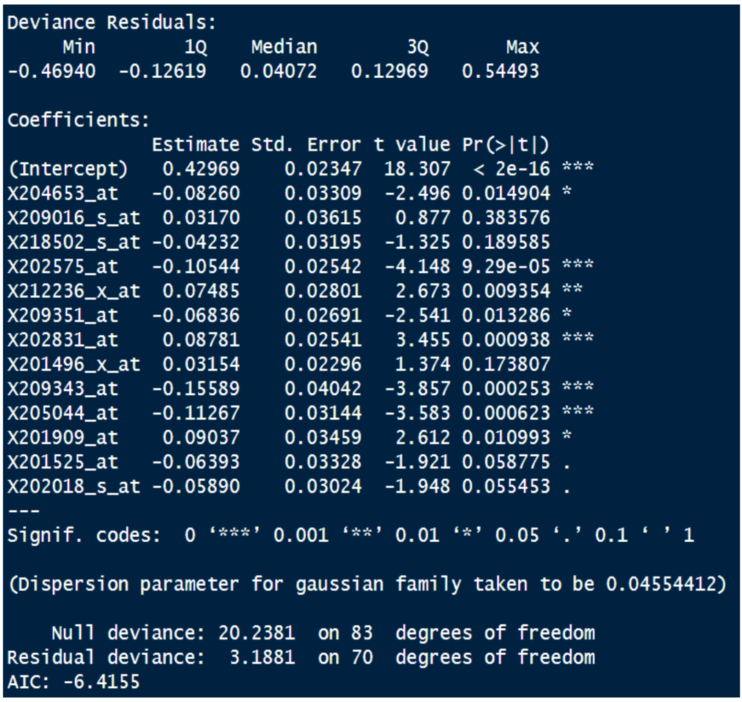
Table 4 highlights the set of features selected using Feature Selection process. **By performing the feature selection process, the total number of features in the dataset was reduced significantly from 182 to only 13, resulting in a more manageable dataset that can be used to train and test machine learning models more efficiently.**

Final Features Set			
X204653_at	X202831_at	X212236_x_at	X201525_at
X209016_s_at	X201496_x_at	X209351_at	X202575_at
X218502_s_at	X209343_at	X202018_s_at	X205044_at
X201909_at			

Table 4: Features Selected after Feature Selection Process

Logistic Regression Model

Finally, I have trained a Logistic Regression model using the features identified using Feature Selection Process. The model is trained using the training data for the selected features only. Figure 5 highlights the summary of the Logistic Regression Model.



```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.46940  -0.12619   0.04072   0.12969   0.54493

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.42969    0.02347  18.307  < 2e-16 ***
X204653_at   -0.08260    0.03309  -2.496 0.014904 *
X209016_s_at  0.03170    0.03615   0.877 0.383576
X218502_s_at -0.04232    0.03195  -1.325 0.189585
X202575_at   -0.10544    0.02542  -4.148 9.29e-05 ***
X212236_x_at  0.07485    0.02801   2.673 0.009354 **
X209351_at   -0.06836    0.02691  -2.541 0.013286 *
X202831_at    0.08781    0.02541   3.455 0.000938 ***
X201496_x_at  0.03154    0.02296   1.374 0.173807
X209343_at   -0.15589    0.04042  -3.857 0.000253 ***
X205044_at   -0.11267    0.03144  -3.583 0.000623 ***
X201909_at    0.09037    0.03459   2.612 0.010993 *
X201525_at   -0.06393    0.03328  -1.921 0.058775 .
X202018_s_at -0.05890    0.03024  -1.948 0.055453 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.04554412)

    Null deviance: 20.2381  on 83  degrees of freedom
Residual deviance:  3.1881  on 70  degrees of freedom
AIC: -6.4155
```

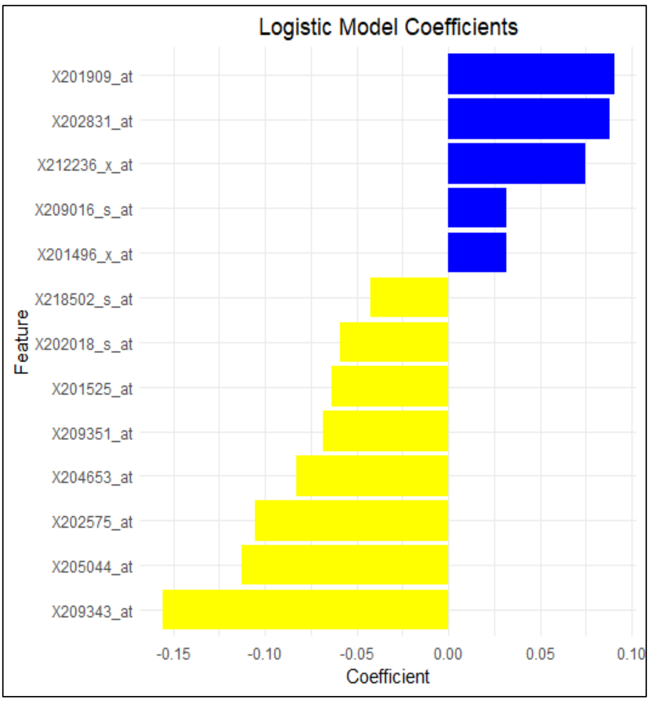


Figure 5: Summary of the Logistic Regression model and Plot of the magnitude of coefficients

From Figure 5, I can clearly analyse that the **minimum and maximum values of Deviance Residuals are -0.46940 and 0.54493 respectively**, which means that this is the range of values for the difference between the observed response variable and the predicted response variable, after accounting for the number of model parameters. It indicates the amount of variability that the model is unable to explain and can be used to assess the overall goodness-of-fit of the model. **A smaller range of Deviance Residuals indicates a better fit of the model to the data.**

Furthermore, **I can also analyse that the model has very low Std. Error for each coefficient**. The standard error in the coefficients table shows the precision with which the estimated regression coefficients are measured and **lower standard error indicates that the estimated coefficients are consistent and reliable**.

Furthermore, a p-value of less than 0.05 indicates that the coefficient is significantly different from zero at the 95% confidence level, suggesting that the corresponding feature is statistically significant and has a significant effect on the outcome variable. The significance codes in the coefficient table show the level of significance of each coefficient.

Looking at the coefficient table, I can analyse that some features have p-values less than 0.05, indicating that they are statistically significant predictors of the tumour. These include **X204653_at, X202575_at, X212236_x_at, X209351_at, X202831_at X209343_at, X205044_at, and X201909_at**. Furthermore, **X201525_at, and X202018_s_at are also significant at 0.1 level.** The remaining variables in the model are not statistically significant predictors of the outcome variable.

Lower AIC values indicate better model fit, and the negative AIC value suggests a strong model fit in this case.

Now, during the Data pre-processing step, I have masked Tumour type B as 0 and Tumour type C as 1. **This means that coefficients that are less than 0 have higher influence on Tumour type B and coefficients that are greater than 0 have higher influence on Tumour type C. Furthermore, I can analyse that there are 5 Gene expressions that are positive and 8 Gene expressions that are negative.**

Model Evaluation

I have evaluated the performance of trained Logistic Model on test data. I was able to achieve an **accuracy of 90% on the test data and F1 Score of 89%**. Table 5 highlights the performance of the Logistic Regression model on various parameter and Figure 6 shows the Confusion Matrix for the generated model.

Model Evaluation	
Accuracy	0.9
F1 score	0.89
Precision	1
Recall	0.8

Table 5: Model Evaluation matrix

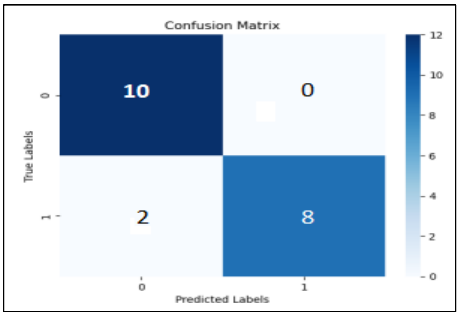


Figure 6: Confusion Matrix

Results and Discussion

The primary objective of this assignment was to analyse the Chowdary dataset, perform Feature Selection to identify features that significantly impacts the predictions of tumour type and build a model using these parameters to identify the genes that increases the chances of tumour type B and C.

Using the results above, we can now identify the genes and if their influence is increasing or decreasing on tumour type. From Figure 5, I can clearly identify that **Gene expressions: X209016_s_at, X201909_at, X202831_at, X212236_x_at, and X201496_x_at have positive coefficient and therefore, will influence the increase of the tumour type C. Out of these 5 genes X201909_at, X202831_at, X212236_x_at are statistically significant at 0.05 level and therefore, have strong association with tumour type C.**

Meanwhile Gene Expressions: X204653_at, X218502_s_at, X202575_at, X209351_at, X209343_at, X205044_at, X201525_at, X202018_s_at have negative coefficient and therefore, will influence the increase of the tumour type B. Out of these 8 genes X204653_at, X202575_at, X209351_at, X209343_at, X205044_at are statistically significant at 0.05 level and therefore, have strong association with tumour type B.

To recapitulate, I was able to reduce the dimensionality of the dataset from 182 variables to 13 variables using the Feature Selection Process. This process has allowed me to identify the features that significantly impact the predictions. Next, I trained a Logistic Regression model using the selected features and I was able to achieve an accuracy of 90% on the test data. Furthermore, using the logistic Regression model, I was able to analyse the features that are strongly associated with each of the tumour type B and C. Moreover, I was also able to analyse the influence of each of the gene expressions on these tumour types and understand how increase in any one these can lead to change in overall prediction of the model for tumour type.