Gene Expression Analysis for Cancer Classification

Author: Virendrasinh Narendrasinh Chavda
School of Mathematics, Statistics and Actuarial Sciences,
University of Essex, UK

Abstract

Early detection of invasive cancer is crucial for determining effective treatment plans. This project investigates the use of statistical and machine learning methods to classify cancer types based on gene expression data. Given the high dimensionality of gene expression datasets, this study explores various dimensionality reduction techniques including two-sample t-tests, LASSO regression, and variance-based feature selection. Missing data imputation was performed using k-Nearest Neighbors (kNN) and median imputation methods. The reduced datasets were then used to train several supervised machine learning models such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression. Resampling techniques like K-fold cross-validation and bootstrapping were employed to validate model performance and prevent overfitting.

The results demonstrated that combining two-sample t-tests with LASSO regression yielded significant improvements in classification accuracy. Specifically, the KNN and SVM models achieved near-perfect classification performance on the reduced feature set. Unsupervised learning methods, including clustering and Principal Component Analysis (PCA), were also applied to explore the inherent structure of the data. This analysis provides valuable insights into effective dimensionality reduction strategies for high-dimensional biological data and their impact on predictive modeling.

Contents

1. Introduction

The analysis of cancer diagnosis is crucial for understanding the recommended course of treatment for patients. Accurate analysis is key in this task, and innovative approaches to clinical data analysis, such as machine learning and big data tools, have become indispensable. By leveraging supervised and unsupervised learning models, researchers can uncover patterns in data that may aid in determining statistical relevance between genes, thus facilitating cancer type prediction.

In this study, gene expression data from cancer patients serves as predictors, with the invasive and non-invasive cancer as dependent variable. After sampling the data with the largest registration number of the team, 2000 random columns were chosen for further analysis. This high dimensionality in data makes it difficult to identify the genes which are decisive in classifying cancer types. To solve this, dimensionality reduction were employed. Specifically, three dimensionality reduction techniques—Two sampled t-test, LASSO regression, and feature reduction using the genes with more variance—were used to test the models. The aim was to compare the performance of supervised models combined with different reduction techniques.

Section 2 of this report provides a detailed overview of the gene expression data used in the study. Section 3 considers dimensionality reduction strategies, implementation of supervised learning models for classification, resampling methods, and the performance metrics used to evaluate the models. In Section 4, the findings of the study are discussed.

This section highlights cross validated and bootstrap validated performance comparison of different supervised models using data reduced after different reduction methods. Finally, Section 5 summarizes the key findings and implications of these findings in classifying types of cancer using gene expression data.

2. Preliminary Analysis

The dataset includes 4949 columns, 4948 columns represent genes and 1 column represent label "Class" which is divided into classes 1 (invasive) and 2 (non-invasive). The rows are patients and each entry in a row is the expression of a particular gene in each column [3]. First, 2000 columns were selected randomly after setting largest registration number from the group (2315880) as seed value, and a sample is created using these 2000 columns, and column "Class" is added in this sampled data.

*Dimensions of original dataset = 78 rows * 4949 columns*

*Dimensions of the sampled data = 78 rows * 2001 columns*

2.1. Dealing with missing values

It was found that there were 74 missing values in the sampled dataset. Out of these 74 missing values, row 54 had 73 missing values and row 23 had one missing value. An experiment was conducted to determine whether row 54 should be dropped as it is very highly likely that these missing values are due to errors in data collection. First all missing values were replaced using KNN imputation and a logistic regression model was fitted with k-fold cross validation, and it resulted in misclassification error of 0.435. Afterwards, the same model was fitted with the sampled data after dropping row 54, and the error decreased to 0.388. From this experiment, it was decided to drop row 54 with 73 missing values.

Figure 2.1 shows boxplot for this column vs Class. It can be observed that there are outliers in this column, therefore, first the missing value was imputed with median and the above-mentioned experiment with logistic regression was run and got misclassification error 0.38. Again, the same model was fitted after imputation using KNN imputation, which reduced the error to 0.35. Hence, missing value was replaced using KNN imputation method.

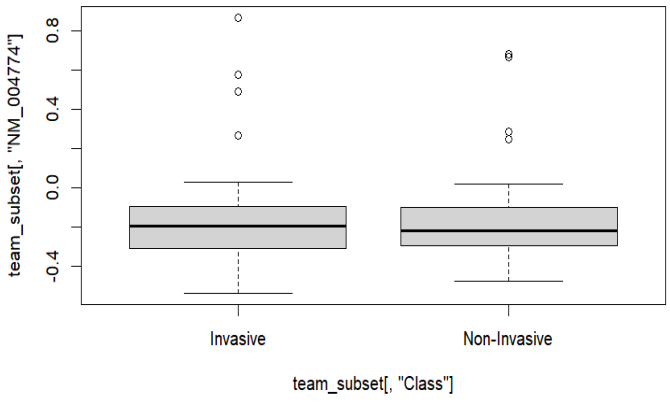

Figure 2.1: Boxplot of "NM_004774" vs "Class"

2.2. Data Scaling

The gene expression values in the dataset are in range 2 to -2. So, no scaling method was implemented.

2.3. Class imbalance

Plotting bar for both classes of cancer, there are 33 observations belonging to type 1 (invasive) and 44 observations belonging to type 2 (non-invasive). This ratio is approximately 42:58, which can be considered a minor imbalance, and hence class imbalance should not hurt supervised models for this dataset. Figure A.1 in appendix shows the bar chart for both Classes.

3. Analysis

3.1. Question 1:

3.1.1 Supervised Dimensionality Reduction: Feature selection using Two sample t-tests.

A t-test is an inferential statistical method used to determine whether there is any significant difference between the mean of two groups in a given numerical variable. Formally defined as a statistical hypothesis test used to compare the means of two population groups [4]. The two-sample t-test assumes that the data is normally distributed, and the variances are equal. It has many advantages such as the easier interpretation of the results, it

identifies the discriminant features, and it is a simple and straightforward test. Significance value (alfa) is the threshold at which we check whether the difference in means between the two groups is statistically significant. The threshold for p-value is considered as 0.05 in conventional statistics.

A function was defined to perform the t-test on each gene which returns an array of p-values. Then the names of the genes and their p-value were extracted from the array and created a data frame. By considering the significant value of p as 0.05, genes were filtered from subset data. After performing the t-test on gene sunset data, the dimension of the dataset has reduced from 2000 to 346 features. Using these reduced features subset, we can train the machine learning models to predict the outcome and evaluate them using standard metrics.

3.1.2. Unsupervised Dimensionality Reduction:

3.1.2.1. Feature Selection using Variance.

There are many methods to perform unsupervised dimensionality reduction such as PCA, t-SNE, dimension reduction using Variance, etc. Initially it was decided to do PCA to extract the significant components from the data. But the columns in the data are more than rows, hence, it is not possible to perform PCA on genes. So, it was then decided to reduce dimensions using variance because it is a simple and intuitive method. In this method, it is assumed that the features with low variance will contribute less to the overall variability of the dataset [5]. Therefore, the features with high variance were selected as they have most of the information.

Advantages of using variance to perform feature extraction are: it simplicity, compresses data, and could prevents overfitting. It is one of the fastest ways to perform dimension reduction by retaining important information in the data.

3.1.2.2. LASSO regression method.

LASSO regression stands for Least Absolute Shrinkage and Selection Operator. It is a regression analysis method used in regularization that penalizes the weights of the variables with L1 regularization. Unlike Ridge, which uses L2 regularization, the LASSO penalty can

reduce some weights to very close to 0 or even 0, avoiding overfitting and removing collinearity. The variables with higher penalties become irrelevant for the model, allowing LASSO to be used as a feature selection method.

$$L_{lasso}(\hat{\beta}) = \sum(y_i - x'_i\hat{\beta})^2 + \lambda\sum|\hat{\beta}_j|$$

In figure 3.1, we see the result of the loss function for different values of the constant α after 10-fold cross-validation is carried out. The minimum value and the optimum value of lambda are stored in the lambda_min and lambda_1se variables.

The accuracy of the logistic regression increased from 0.6428571 to 0.8571429, moving from 347 variables to a total of 29 variables after LASSO penalization. Below mentioned are the variables (genes) that resulted from the LASSO regularization, so we are going to use them to carry out further analysis.

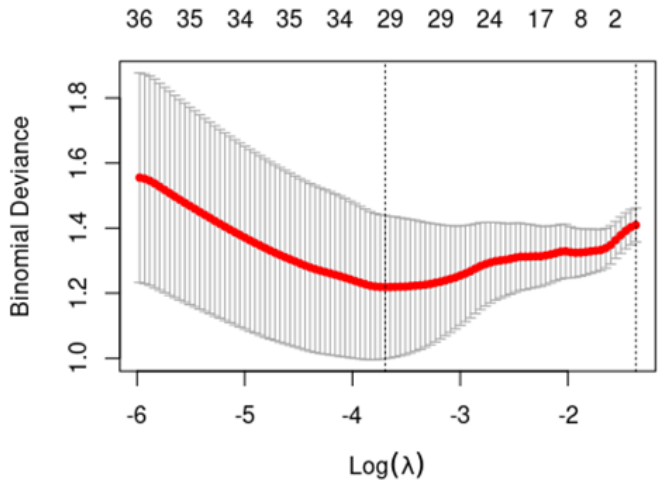

Figure 3.1: LASSO regression: Loss function for different values

3.2. Question 2:

3.2.1. Principal Component Analysis.

The correlation provides an insight into the groupings within the data frame. Here we used the 29 columns (genes) that resulted from the LASSO dimensionality reduction. We observe two distinct regions (one blue and one red). Notably, there is a high correlation of 0.74 between the genes Contig41383_RC and Contig43599_RC, suggesting that these genes are

expressed similarly across the individuals. Figure A.2 in appendix shows correlation matrix for 29 columns selected after Lasso reduction.

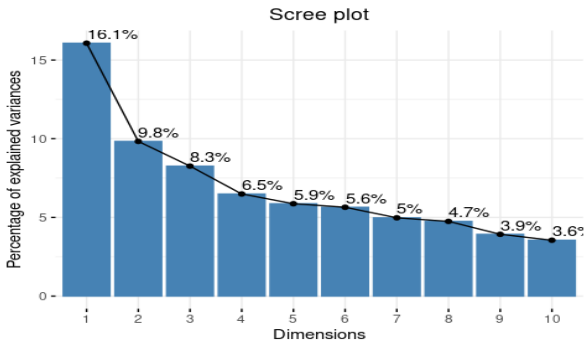

Figure 3.1: Scree Plot for first 10 principal components of genes

In table 3.1, we can see the percentage of explained variances by the first 10 components, which together account for approximately 70% of the dataset's variance. The eigenvalues are arranged from 1 to the total number of features, with the first one explaining the highest variance of the original dataset and the last one the lowest variance of the same dataset. Collectively, they sum up the entire variance. However, the initial components are typically utilized to perform the analysis instead of the complete dataset, which could be highly correlated and excessively large for some algorithms to handle effectively.

Figure A.3 in appendix, the scatterplot shows that with the first two principal components, we can already fairly accurately separate the two groups. There is still some overlap, but it is important to note that we are only using two principal components, which explain only 26% of the total dataset variance; thus, they cannot account for the entire dataset alone.

The second PCA, with the aim to reduce the number of observations rather than the number of genes was performed and its scree plot is shown in Figure A.4 in appendix. Here, the components number are 77, and we observe that the first two components already explain 31.2% of the data variability. The analysis only displays the first 10 components.

3.2.2. Hierarchical Clustering.

The dendrograms shown in Figure A.4 in appendix, explore the relationships between genes, linking pairs of nodes to other nodes or groups according to their distance. In this

instance, all distances were tested without significant differences, so the standard Euclidean distance was used to calculate the distances between genes. The Ward linkage criterion appears to separate the genes clearly into two groups, one with 7 genes and the other with 22 genes. The average and single criteria are quite similar, with almost all genes grouped in one large cluster and one gene (NM_002820) alone in another group/cluster. The complete approach also exhibits two groups; however, there is more distance between the clusters within the larger group. It is important to note that the same gene (NM_002820) forms a cluster on its own even with this linkage criterion.

The clusters in figure 3.2 demonstrate the separation of the observations based on the Canberra distance. The clusters reveal two kinds of distinct groups, with the main variation among them being the order in which the observations are linked. The single linkage criterion appears to separate the groups very poorly, which was anticipated given the overlap of the groups, as evident in the scatterplot of the first two components. Thus, the smallest distances between groups are a poor measure of closeness when such overlaps occur. Both complete and Ward linkage criteria seem to produce identical clusters. The average criterion yields a slightly different result but generally can separate the two main groups effectively.

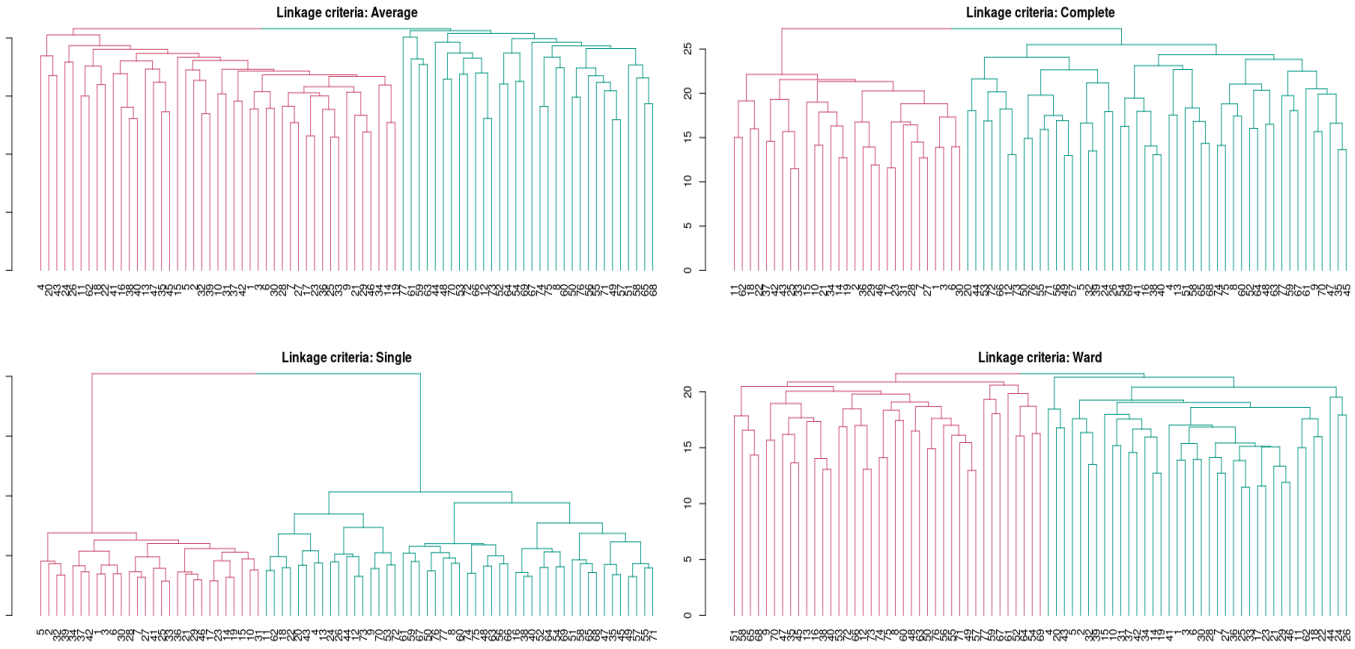

Figure 3.2: Dendrograms showing relationship between patients.

3.2.3. K-means clustering.

Figure A.6 in appendix implies that in this K-means cluster analysis, the elbow method does not provide a clear indication of whether the optimal number of clusters (k) should be 2 or 3. Therefore, we will employ Silhouette analysis. This method calculates the silhouette width, a metric that determines whether a data point is closer to its assigned group or to another group. The measure ranges from -1 to 1: a value of -1 indicates that the data point would fit better in a neighboring cluster, 0 suggests that the data point is on the boundary between two clusters, and 1 signifies that the data point is well assigned to its group.

3.2.4. Silhouette Analysis.

Using this analysis, we observe that the average silhouette width for k=3 is 0.1, whereas for k=2 it is 0.12. This indicates that with k=2, the data points are better assigned to the clusters they belong to, suggesting that two clusters provide a more cohesive and appropriate grouping of the data points compared to three clusters. Figure 3.3 shows cluster separation from silhouette analysis.

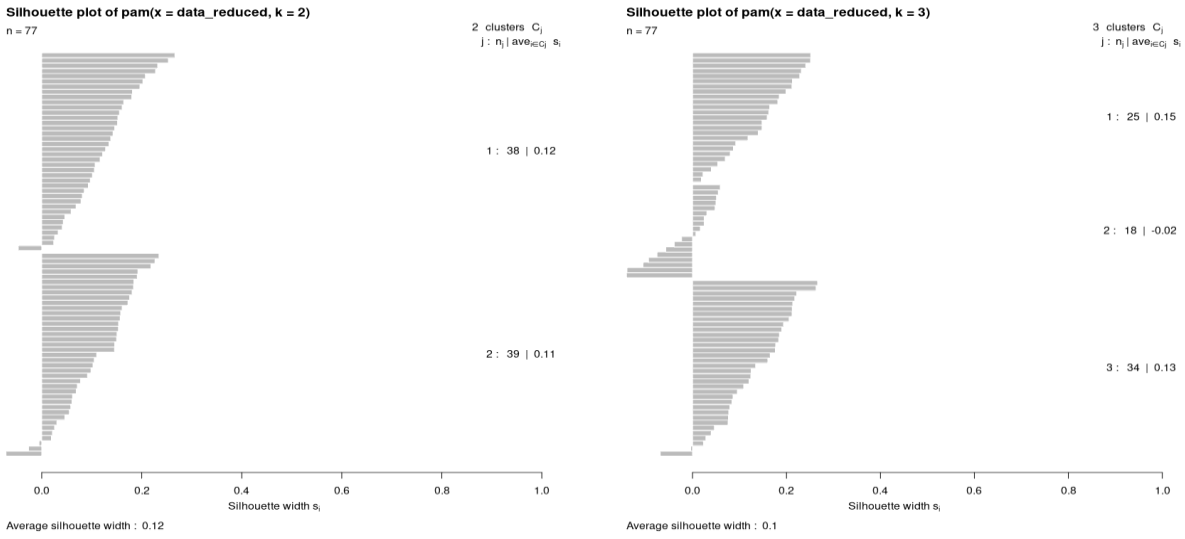

Figure 3.3: Cluster separation silhouette analysis.

3.2.5. Correlation based distance cluster.

Correlation-based distance is another approach used as a distance criterion when it comes to creating clusters, especially in the context of gene expression analysis. As demonstrated in the referenced study [6], using 1 minus the correlation between variables helps to agglomerate groups and provides a result that is congruent with previous clusters. Here, we observe two main clusters of genes: one at the top, predominantly red, featuring a gene that appears to express antagonistically with highly negative correlation, and another at the bottom, which exhibits more variability in terms of correlations compared to the first group. If we were to compare the grouping of observations to the hierarchical cluster, especially the one with complete linkage criteria, we would obtain almost identical results. This indicates that both measures, correlation-based distance, and Canberra distance, can identify similar patterns in the data grouping. Figure A.7 in appendix shows correlation-based clusters.

3.3. Question 3:

For supervised learning, Logistic regression (LR), Poisson regression (PR), LDA, KNN, random forest (RF), SVM, GLM Boost (GLMB) and XG Boost (XGB) algorithms were used. QDA could not be used as the number of observations is only 78, which is smaller compared to the number of columns, i.e. 2001. Even if data is reduced to 20 columns, QDA gives an error of rank deficiency in classes as dataset gets divided in k-fold cross validation.

Misclassification error is used as model evaluation matrix since this is a classification problem. Two types of resampling methods, namely k-fold cross validation and bootstrapping were used to get more robust errors. Experiment to estimate suitable number of k-folds was conducted for this dataset using logistic regression model. It resulted in an optimal value of k as 7, i.e. 7 folds. Figure A.8 in appendix, shows graph for number of folds vs misclassification error.

As discussed in section 3.1, multiple samples from the original dataset were used for model fitting. These samples are listed below: (1) Original dataset with 2001 gene columns. (2) Dataset filtered through two sampled t-tests in question 1 which resulted in 347 columns. (3) Considering the top 100 columns with the most variance from t-test filtered data from question 1. (4) t-test reduced columns further reduced by applying Lasso regression.

Table 1.1 shows errors of all supervised models fitted with above listed samples. Observing errors from all samples, KNN models work best on this gene expression data. The best performing model, KNN model has 0 misclassification error for Lasso reduced genes.

Model	2001 Columns		t-Test Reduced Columns		Top 100 t-Test reduced Columns		t-test + Lasso regression reduced columns	
	k-fold	Boot	k-fold	Boot	k-fold	Boot	k-fold	Boot
LR	0.388	0.518	0.363	0.47	0.55	0.481	0.155	0.243
PR	0.363	0.44	0.428	0.478	0.559	0.428	0.233	0.378
LDA	0.388	0.309	0.28	0.18	0.452	0.418	0.09	0.187
KNN	0.313	0.24	0.155	0.118	0.299	0.268	0	0
RF	0.336	0.231	0.26	0.387	0.298	0.309	0.319	0.411
SVM	0.33	0.29	0.233	0.125	0.415	0.32	0.07	0
GLMB	0.472	0.53	0.44	0.51	0.44	0.45	0.116	0.25
XGB	0.493	0.512	0.33	0.56	0.376	0.45	0.311	0.51

Table 1.1: Errors for supervised models for different samples.

The accuracy of the KNN model is somewhat surprising, as it achieves 100% accuracy, which is atypical for machine learning models, especially with biological data known for its diversity. However, there are two main reasons for this high level of performance. The first is that during the dimensionality reduction process, we performed a t-test that selected the variables where the means between the classes were different. By doing so, the performance of any distance-based algorithm would significantly improve. The second reason is the low number of observations, only 77, with 80% used for training, leaving 15 observations for testing. As a result, the accuracy levels would increase in intervals of approximately 6.67%, and while it may not always be 100%, it could be close, but not exactly 100%. Even when performing K-fold cross-validation, there aren't 100 observations for testing, so the accuracy would not be a precise integer from 1 to 100 but would instead vary between certain values.

Similar results were obtained with the SVM algorithm, where the accuracy was very high, although it did not reach 100%. Like KNN, SVM is also a distance-based algorithm for multiclass classification. In general, achieving 100% accuracy with machine learning algorithms is not entirely unusual, as even simple logistic regression algorithms have reached very high accuracy levels in our dataset.

4. Discussion

Supervised machine learning algorithms like Logistic Regression, Poison Regression, KNN, Decision Tress, XG Boost, Random Forest, GLM Boost and SVM were trained to predict invasive, or non-invasive cancer and misclassifications were validated using k-fold and bootstrapping methods. First, the whole data with 2001 columns was considered for training and KNN model generated 0.313 misclassification error. The error dropped drastically to 0.155 for k-fold and 0.118 for bootstrapping, after reducing the dimensions using two sampled t-tests. Then, the top 100 columns from reduced dataset were chosen using variance for model training and an increase in misclassification error was observed. This implies that when choosing the top 100 columns, important information is lost. Finally, genes identified from two sampled t-tests were further reduced through Lasso regression and it resulted in a perfect KNN model with 0 misclassification and SVM for the same samples came close with 0.07 misclassification error. Lasso regression for dimensionality reduction shows promising implications through this experiment, but it can be misleading as the dataset has only 78 observations.

Further analysis was done using unsupervised clustering algorithms like k-means clustering, hierarchical. It was observed that the two classes were separated fairly by using the first few components of the PCA. Clusters of features and observations were formed using hierarchical clustering and clusters of patients were observed with clear separation. To find the optimal value for number of clusters, elbow method was used and there was uncertainty to decide the optimal value of k as 2 or 3. To decide on that, Silhouette analysis was used. It was concluded that 2 was the optimal value of number of clusters.

Since the KNN model trained on Lasso reduced genes could not be further improved, genes extracted from two sample t-tests, which resulted in k-fold error of 0.155, were chosen to observe the effect of clustering on model performance. Twenty principal components from Lasso reduced genes, clusters from k-means clustering, and hierarchical clustering were used to fit KNN model with k-fold cross validation, and the misclassification error was 0.025, 0.155 and 0.142 respectively. Hence, PCA components improved the model drastically, whereas clusters from hierarchical method improved the model marginally, and clusters

from k-means method showed no improvement.

5. Conclusion

To conclude, the analysis performed on the given gene expression data showed that two sampled t-tests combined with Lasso regression can be a very effective method for dimensionality reduction. Moreover, KNN and SVM can be quite accurate in classifying whether a cancer is of invasive or non-invasive type from this gene expression data. Even a simple logistic model can be stronger when used with Lasso reduction method. Since the data is small with only 78 observations, the errors may amplify with a larger dataset and more conclusions can be drawn.

References

[1] J. Taveira De Souza, A. Carlos De Francisco and D. Carla De Macedo, "Dimensionality Reduction in Gene Expression Data Sets," in IEEE Access, vol. 7, pp. 61136-61144, 2019, doi: 10.1109/ACCESS.2019.2915519

[2] Alan Wee-Chung Liew, Ngai-Fong Law, Hong Yan, Missing value imputation for gene expression data: computational techniques to recover missing data from available information, Briefings in Bioinformatics, Volume 12, Issue 5, September 2011, Pages 498–513, https://doi.org/10.1093/bib/bbq080

[3] Brazma A, Vilo J. Gene expression data analysis. FEBS Lett. 2000 Aug 25;480(1):17-24. doi: 10.1016/s0014-5793(00)01772-5. PMID: 10967323.

[4] Mishra P, Singh U, Pandey CM, Mishra P, Pandey G. Application of student's t-test, analysis of variance, and covariance. Ann Card Anaesth. 2019 Oct-Dec;22(4):407-411. doi: 10.4103/aca.ACA_94_19. PMID: 31621677; PMCID: PMC6813708.

[5] Roberts, Aedan & Catchpoole, Daniel & Kennedy, Paul. (2018). Variance-based Feature Selection for Classification of Cancer Subtypes Using Gene Expression Data.

1-8. 10.1109/IJCNN.2018.8489279.

[6] Glazko G, Mushegian A. Measuring gene expression divergence: the distance to keep. Biol Direct. 2010 Aug 6;5:51. doi: 10.1186/1745-6150-5-51. PMID: 20691088; PMCID: PMC2928186.

Appendix

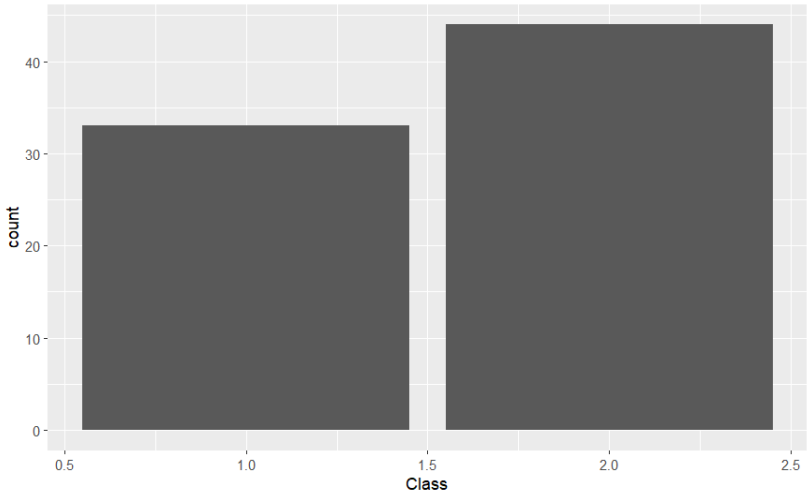

Figure A.1: Bar plot for two types of Cancer

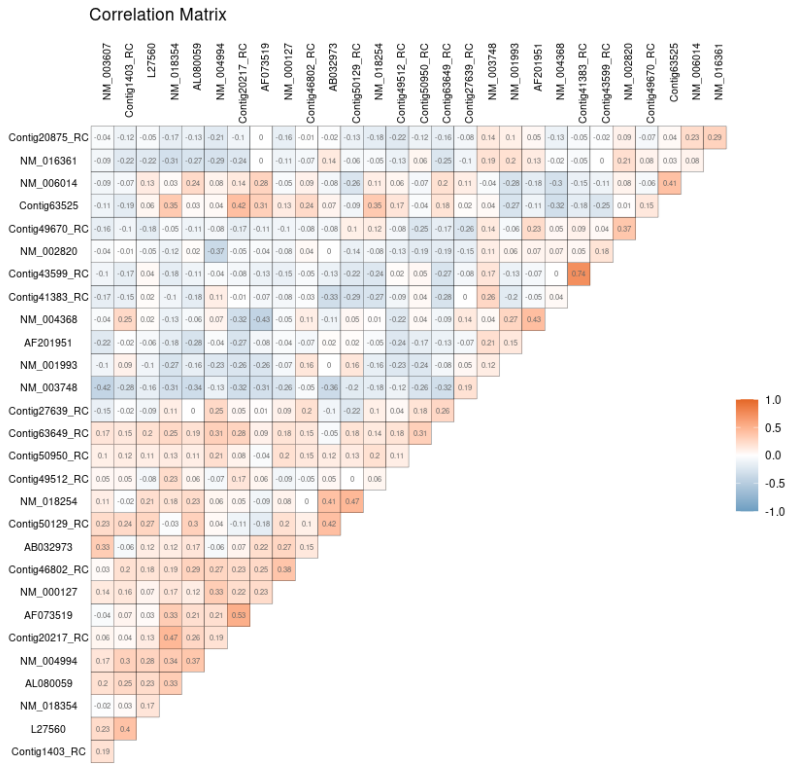

Figure A.2: Correlation matrix for 29 columns selected through Lasso reduction.

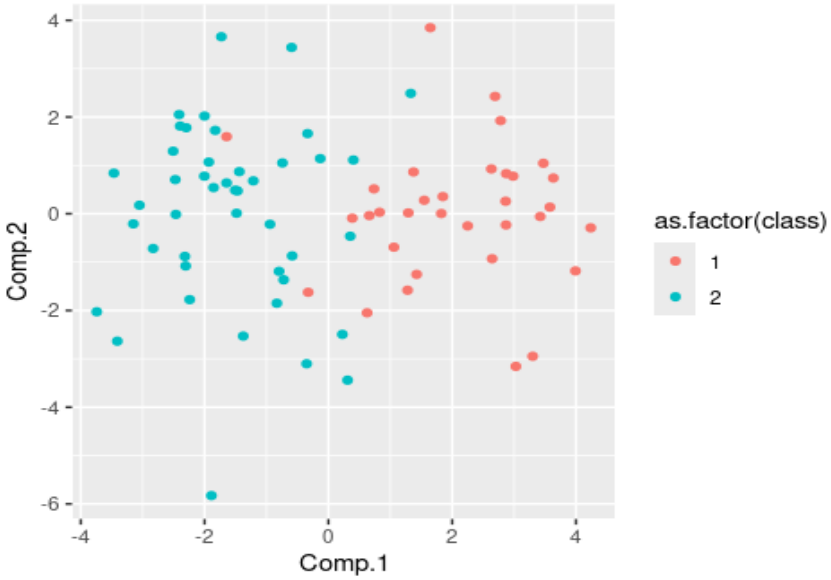

Figure A.3: Class separation by first two principal components of genes.

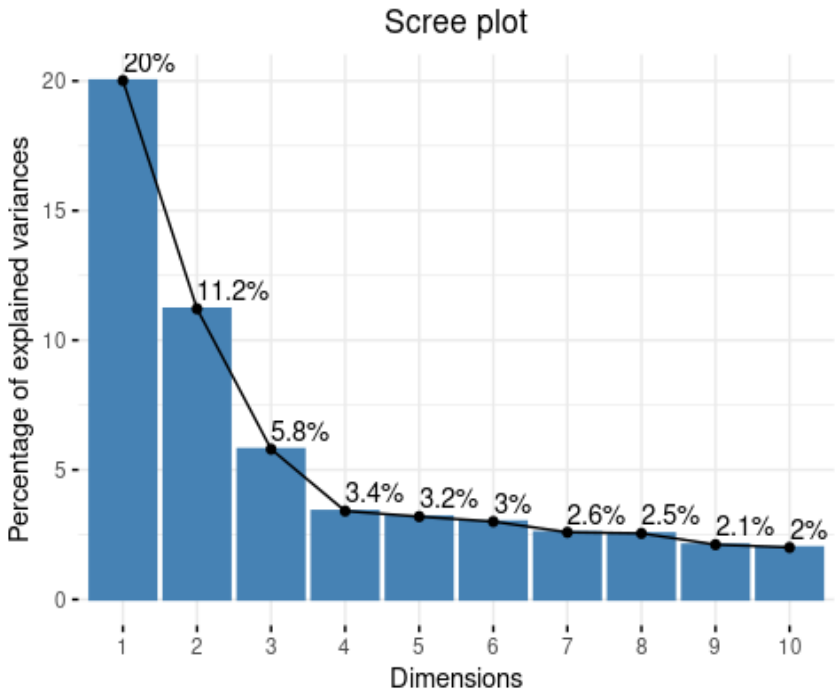

Figure A.4: First ten principal components of patients

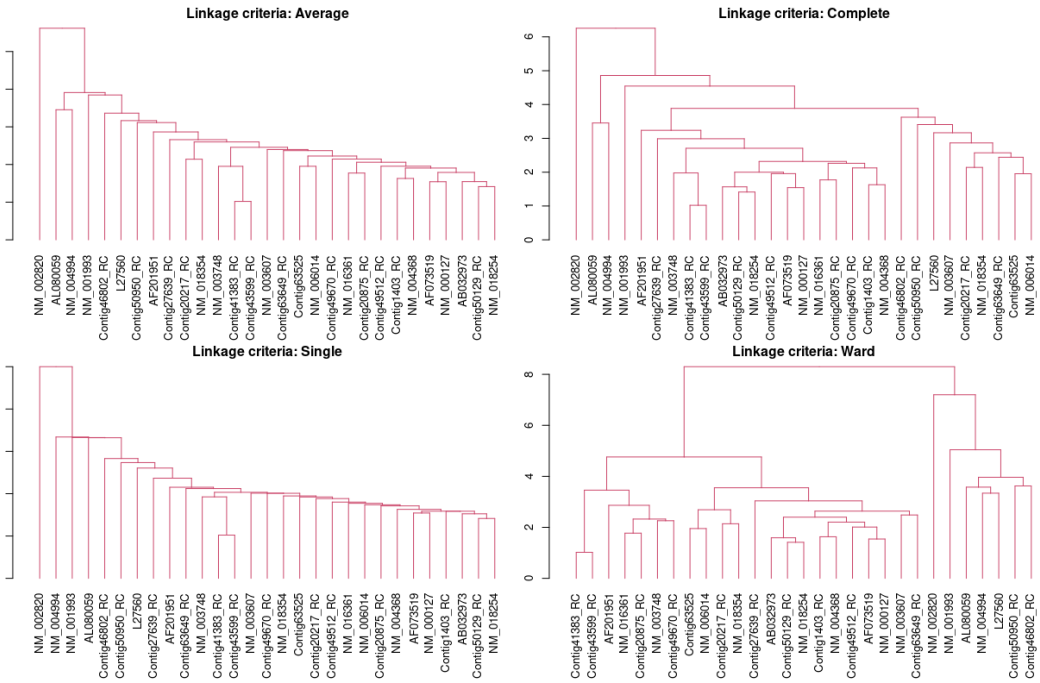

Figure A.5: Dendrograms showing relationship between genes.

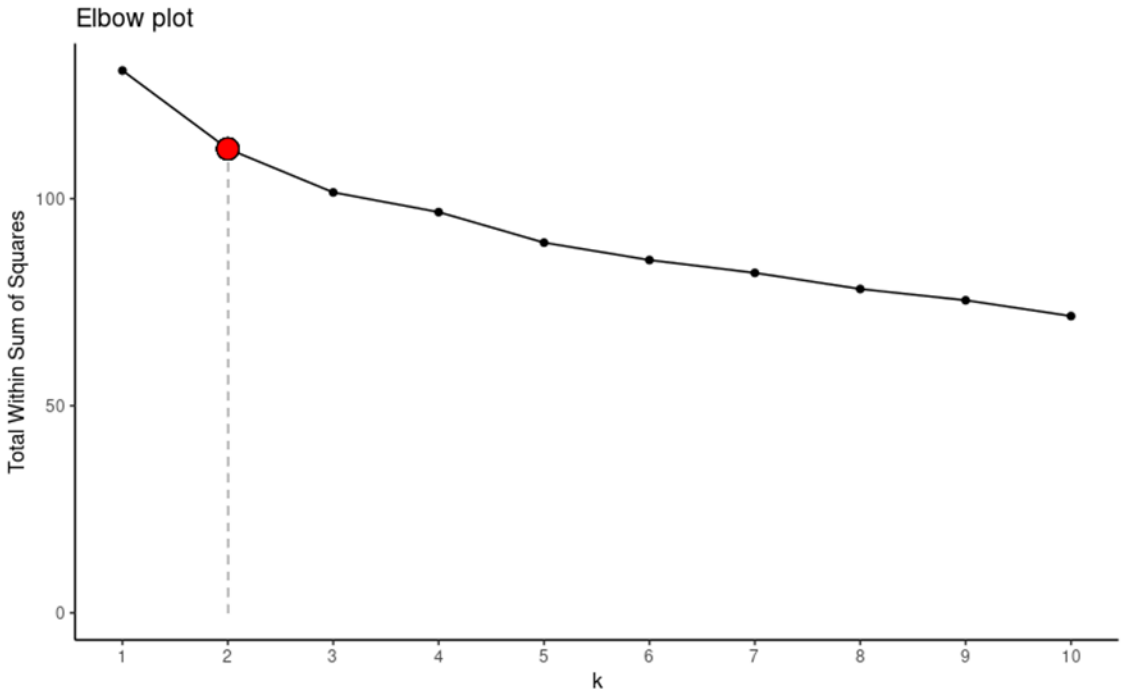

Figure A.6: Elbow graph for k-means clustering.

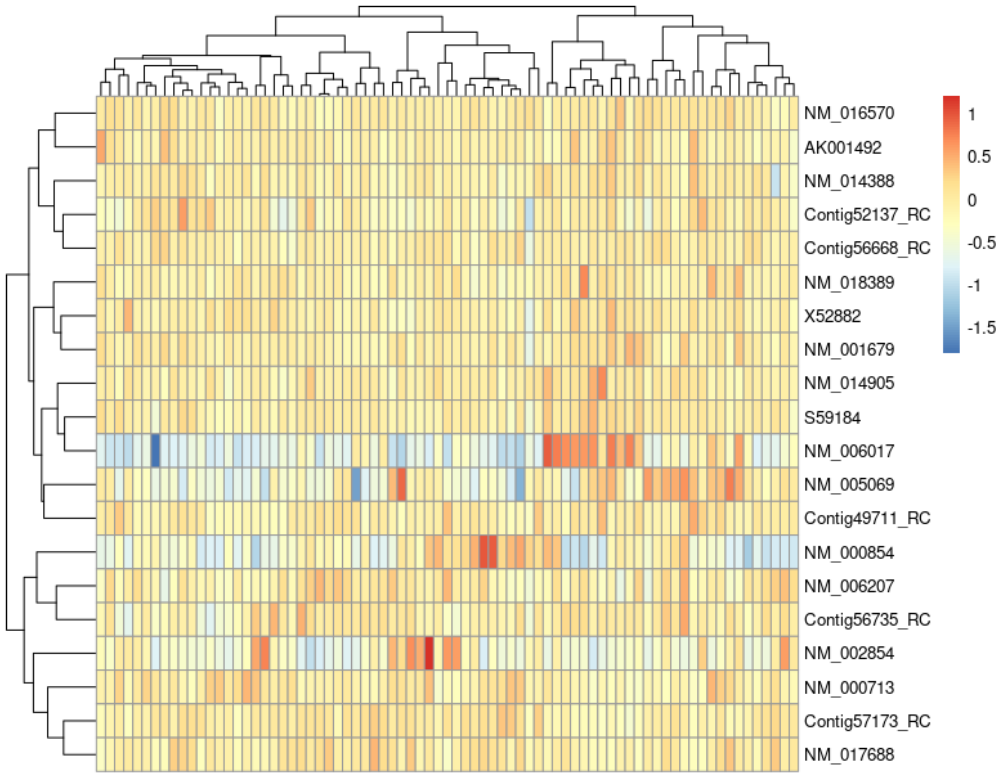

Figure A.7: Correlation-based distance clustering.

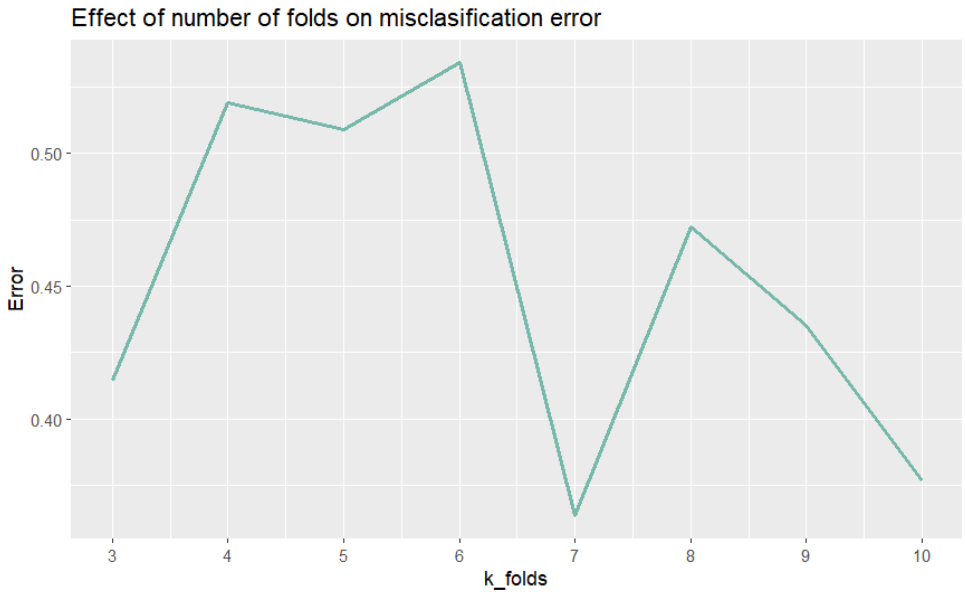

Figure A.8: Number of folds for cross validation vs Misclassification error.