Classification Problem of Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) by Gene Expression Monitoring

Albert Chen

Executive Summary:

This study aims to find implement a classification model correctly determine if the cancer patient has Acute Myeloid Leukemia (AML) or Acute Lymphoblastic Leukemia (ALL) using results from gene expression testing. The dataset includes results from 38 patients training set and 34 patients for the testing set with gene expression results that include the numerical readout from the assay, and the associated categorical determination from the assay preformed – providing insight of presence or absence of the gene given the readout.

To avoid multicollinearity related results, the categorical and numerical data is split due to their direct correlation and the separate datasets and both used and compared for final testing results. In the data, 7129 genes are tested. Therefore, data reduction is performed. Two methods are used. First – PCA is preformed on both numerical and categorical data's training set and applied to the testing set. Second – The highest correlated genes regarding the classes are sorted and the highest correlated genes regarding the training data is used. With the preprocessing completed – four datasets are formed, PCA vs. Correlation, and Numerical vs. Categorical.

The four datasets are implemented into various models. The models tested are Log Regression, Random Forest, K-Nearest-Neighbor, Linear Discriminant Analysis, Support Vector Machine, Neural Network. To implement each of the models with tuning parameters, a cross fold validation method is used for each with accuracy as the metric. Finally, each of the models using the four datasets are used to predict the testing dataset the final accuracy value.

Overall, it was observed that mapping correlation associated with each of the genes for numerical and categorical data was greatly effective at predicting the correct class of Leukemia. Comparatively, it was observed performing PCA on the numerical dataset had performed poorly across all models. Between the numerical and categorical datasets in which correlated genes were chosen, the results are within variance, but numerical dataset performed slightly better.

Given the results, it is recommended that the top twenty genes used for the tests can accurately determine the class of Leukemia. The dataset best suited is the numerical dataset and implementing either Neural Network, SVM or random forest, however other models like Log Regression and Naive Bayes work well.

I. Background and Introduction

a. Background Information

One of the big challenges of cancer treatment has been the classification of cancer, because specific therapies should be targeted to specific tumor types. Without clear classification, even though there are plenty of therapies to choose, the treatment can be inefficient or causing serious toxicity to patients when the appropriate therapy is not used (Dwivedi, 2018).

In the past 50 years, although the classification of cancer has improved a lot, there has been no systematic and unbiased approaches to recognize tumor subtypes before Golub et al. mentioned "Gene Expression Monitoring" for cancer class discovery and class prediction (Golub et al., 1999).

Before monitoring gene expression, morphological appearance of tumors is mostly used to classify cancer, which is not accurate, because tumors showing similar morphological appearance do not exactly follow the same clinical courses or show similar responses to therapies. In this case, the right therapy still can not be decided to the specific subtype of tumor.

Among all types of cancers, leukemia has been one of the most mortal cancer, which can be classified to acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

At first, people were observing the variability in clinical outcome and subtle differences in nuclear morphology (Farber et al., 1948; Frei et al., 1961). In the 1960s, Enzyme-based histochemical analyses were introduced to demonstrate that leukemias could be periodic acid-Schiff positive or myeloperoxidase positive (Quaglino & Hayhoe, 1959; Bennett & Dutcher, 1969; Graham, Lundholm & Karnovsky, 1965). Around 2000s, it has been found that particular subtypes of acute leukemia are associated with specific chromosomal translocations (Barker et al., 1995; Shurtleff et al., 1995; Romana et al., 1995).

Till then, the distinction between AML and ALL has been clear, but there had been not enough tests to establish the diagnosis. When wrong therapies are used, for example when ALL therapy is used for AML, although symptom can be slightly relieved, cure rates are significantly diminished, and toxicities are encountered.

Therefore, a more systematic and unbiased approach must be used to classify patients with AML and ALL. And to do that, DNA microarrays have been important for being used to measure the gene expression of thousands of gene simultaneously.

b. Data Collection

A gene expression dataset from a proof-of-concept study by Golub et al is used for this study. There are two datasets containing the initial (training, 38 samples) and independent (test, 34 samples) datasets used in the paper. And each sample is form of gene expression profile of 7129 genes.

Dataset download link: https://www.kaggle.com/crawford/gene-expression

c. The Problem

Given a dataset of gene expressions from a specified patient – the goal is to successfully classify if they have a specific condition. In the case of the current leukemia dataset – the patients have either Acute lymphocytic leukemia (ALL) or Acute myeloid leukemia (AML). The goal is to be able to separate the two classes given their gene expressions.

d. Possible Solution

There are multiple models and methods which can be applied to solve this classification problem with a structed dataset. Some methods are as follows:

1. **Dimensional Reduction:** Reduction methods such as PCA or EFA can be applied to reduce the variables imputed into the model. Especially since there are thousands of genes presented in the dataset – it could be beneficial to reduce over fitting the data. However, the reduction of variables could make it more difficult to specify which genes are better indicators.

2. **Model Selection:** There are multiple models which can be selected to solve a classification problem with a structed dataset. Some interesting ones that can be selected are a type of **(1)** neural network which can be helpful for trying to classify the model without dimension reduction, **(2)** logistic regression as a standard method for two class classification problem, **(3)** KNN or Random Forest can also be interesting methods however might be prone to overfitting if dimensions are not reduced before.

3. **Performance Evaluation:** Given that this is a two-class problem – % accuracy would be a simple method of grading the model (using the test data). However, there is some class imbalance as well, so using a confusion matrix alongside accuracy would also be beneficial.

II. Data Exploration and Visualization

Initial exploration of the data included specific aggregation and statically methods to determine trends among the two cancer classes. Initially to explore the data a basic summarization of the data was completed. The gene expression dataset from a proof-of-concept study by Golub et al. There are two datasets containing the initial (training, 38 samples) and independent (test, 34 samples) datasets used in the paper. And each sample contains both numerical and categorical status of 7129 gene expressions.

The dataset was separated into the numerical data which contained the results of the assay – which differed per gene, and the categorical data which are predetermined metrics for each assay specifying if the expressed gene was present, marginal, or absent. With both numerical and categorical data, the averages of each class were aggregated against each gene to also view if there were clearer indicators of each class.

To support reducing the number of genes observed and to view any correlations between the genes and the class, correlation matrix was derived to determine which genes had the highest correlation between the classes were used to narrow down the selection of genes to test out.

Due to the large number of genes in the database – only a few of the top correlated variables were visualized to see differences between the two classes for the most correlated genes in terms of classes.

III. Data Preparation and Preprocessing

a. Assay Output – Categorical and Numerical

The data was separated into categorical and numerical data which had no missing values. For each gene, there was a single numerical and categorical value that correlated with the readout of an assay.

For the categorical data – the values were converted from P (present), M (marginal), A (absent) to 1, 0, -1 respectively. This decision was made since while categorical; the data is still ordinal is respect to the specific assay used to detect the gene expression. Therefore, there can be important data saved from keeping the ordinal nature of the data. Furthermore – the dataset would be duplicated to fourteen thousand columns which would become even more unmanageable.

b. Principle Component Analysis

To reduce the data further, PCA was performed on the training data to reduce the number of dimensions needed to support the analysis. Given that 7000+ gene's variance needed to be captured and the constraints of only 38 samples in the training data, it was difficult to capture the full variance in a few components. This reduction however is still significant to the reducing the original dataset of seven thousand gene expression. The figures below show the percent variance of components using both numerical and categorical data.

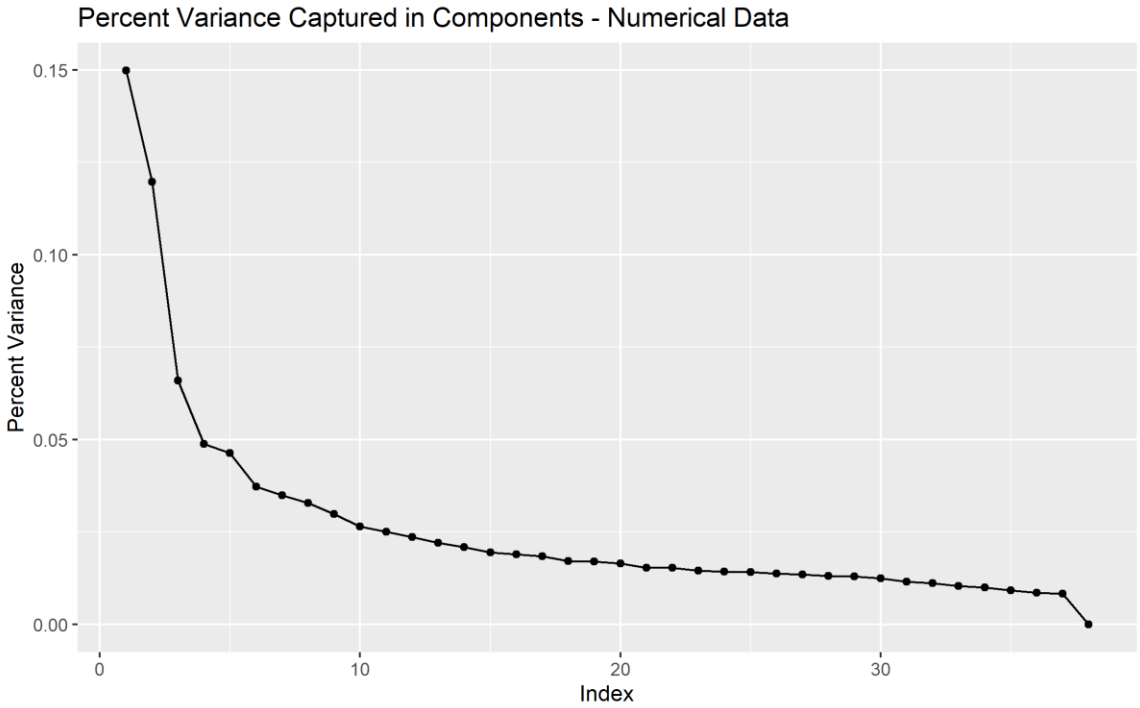


Figure 1: Percent Variance Captured per Components in Numerical Training Data

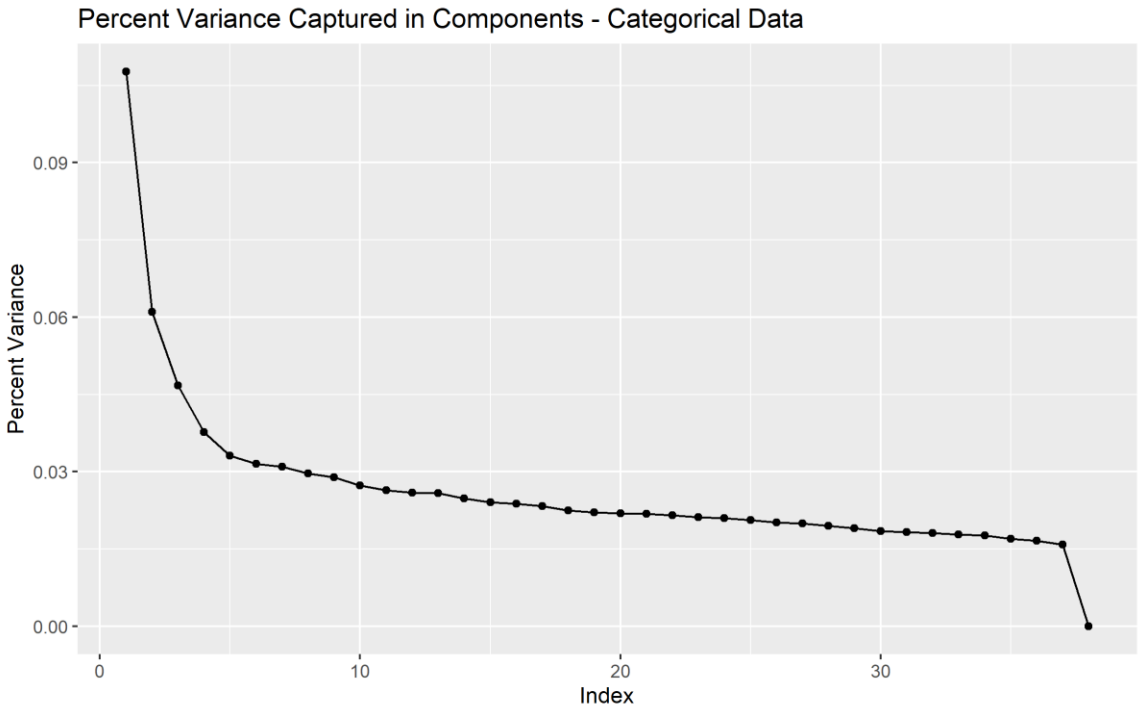


Figure 2: Percent Variance Captured per Components in Numerical Training Data

The scaling and weights of the PCA from the training data was then also applied to the testing data to provide an unbiases final testing score when implementing the models. To determine how well the principle components were at splitting the classes after capturing

the variance, the testing data was used to compare the two classes against the components extracted. The figure below shows a matrix plot of the first six components with samples colored by their respective classes. As observed, the numerical data show some separation present in PC1 vs. PC3/4/5 and some separation in PC4 vs. PC6. As for the categorical data, almost all the components show separation; most significantly PC1 vs. PC2/3 show almost perfect separation.

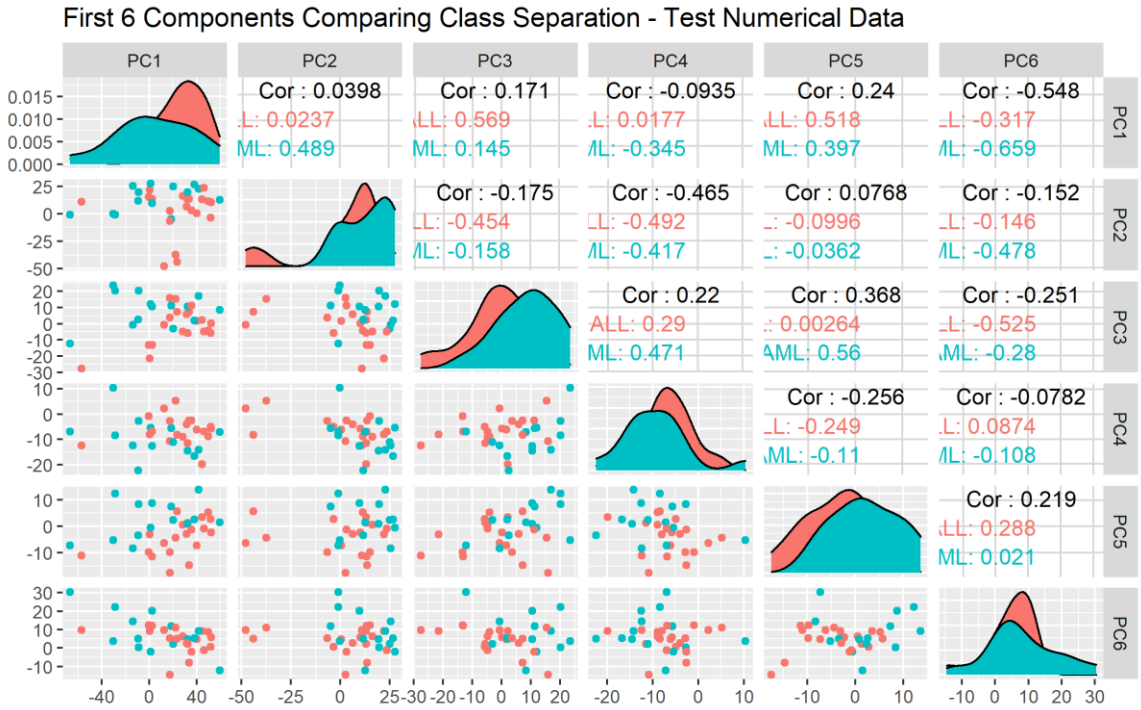


Figure 3: Matrix Plot of First 6 Components and Class Separation Using Numerical Test Data

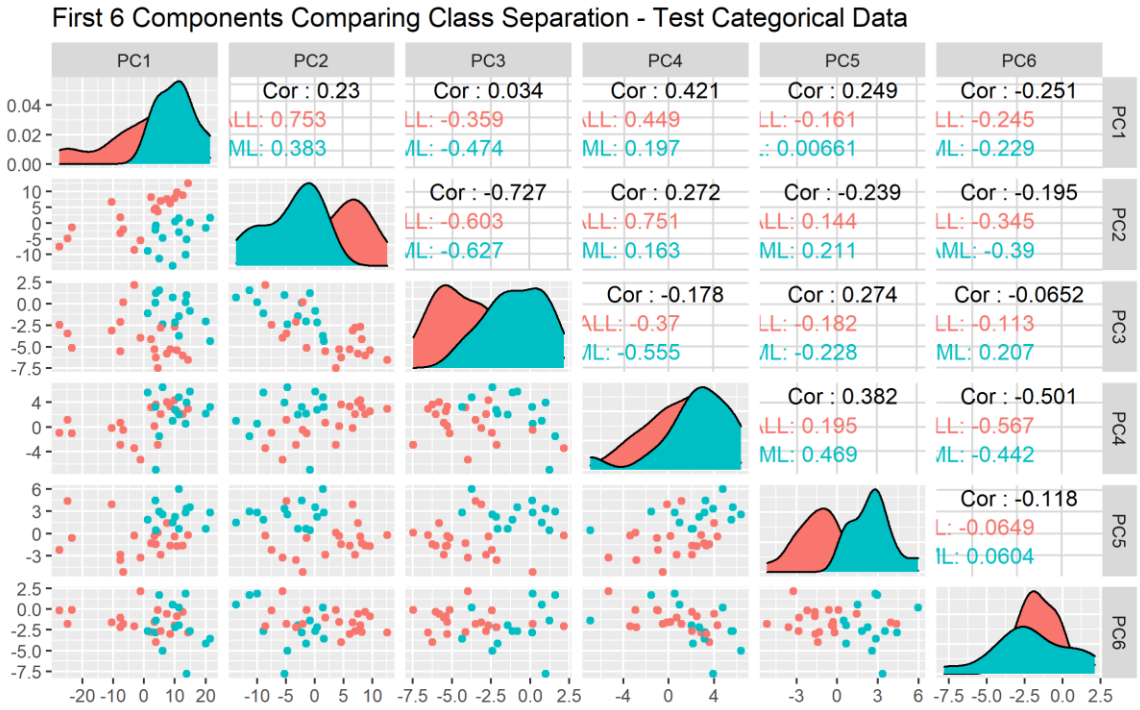


Figure 4: Matrix Plot of First 6 Components and Class Separation Using Categorical Test Data

Overall, for PCA, the separation between classes seem to be significant for the categorical data, comparatively, the numerical data don't seem to have components which separate as well. Therefore, during model implementation it's hypothesized that the categorical data will perform better than numerical for PCA based reduction.

c. Highly Correlated Genes

To determine which genes are more correlated with each class, a correlation matrix was derived across every gene and the classes which were mapped to 1 and 0. This allowed both the numerical and categorical data (post transformation to 1, 0, -1) to be compared how well they are correlated with the classes. To determine which genes were picked, genes with the highest absolute correlation value for the combination of numerical and categorical correlations with classes.

The following table was derived showing the top twenty genes and their respective correlation. As seen, some gene are negatively correlated while others are more positively correlated. This difference in correlation must be observed when visualizing how well each gene is correlated with the classes.

Gene Description	Gene Accession Number	Numerical Correlation	Categorical Correlation
Zyxin	X95735_at	-0.82228	-0.74411
DF D component of complement (adipsin)	M84526_at	-0.74349	-0.80904
Liver mRNA for interferon-gamma	D49950_at	-0.75781	-0.78639

inducing factor(IGIF)			
LEPR Leptin receptor	Y12670_at	-0.7765	-0.74846
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891_at	-0.7216	-0.71542
Azurocidin gene	M96326_rna1_at	-0.6916	-0.7363
CHRNA7 Cholinergic receptor; nicotinic; alpha polypeptide 7	X70297_at	-0.60912	-0.78649
MYL1 Myosin light chain (alkali)	M31211_s_at	0.596166	0.791899
CD33 CD33 antigen (differentiation antigen)	M23197_at	-0.7707	-0.6127
Putative cyclin G1 interacting protein mRNA; partial sequence	U61836_at	-0.58775	-0.79355
CTSD Cathepsin D (lysosomal aspartyl protease)	M63138_at	-0.7106	-0.6491
CYSTATIN A	D88422_at	-0.56102	-0.77972
PTX3 Pentaxin-related gene; rapidly induced by IL-1 beta	M31166_at	-0.55272	-0.78639
Inducible protein mRNA	L47738_at	0.538805	0.800163
INTERLEUKIN-8 PRECURSOR	Y00787_s_at	-0.73211	-0.59271
PLCB2 Phospholipase C; beta 2	M95678_at	-0.63488	-0.67274
PFC Properdin P factor; complement	M83652_s_at	-0.68504	-0.60984
Interleukin 8 (IL8) gene	M28130_rna1_s_at	-0.7092	-0.57189
Epican; Alt. Splice 11	HG2981-HT3127_s_at	-0.63352	-0.62644
GB DEF = Neutrophil elastase gene; exon 5	M20203_s_at	-0.51344	-0.74448

For genes with positive correlation, they are more present for the ALL class while negatively correlated genes should show higher presence with the AML class. This is due to how the ALL and AML classes were mapped numerically.

To further visualize the results of the highest correlated genes, numerical and categorical data were visualized different. For the numerical data, the data was scaled and centered to standardize the scale. Furthermore, to ensure the positively and negatively correlated genes are correctly ranked, positively correlated genes were flipped by multiplying by -1. With the scaled and correctly oriented numerical data, the top 100 correlated genes were graphed by rank vs. scaled numerical value. The classes were clearly labeled by color and

as observed the top 100 gene clearly separate out the two classes with some outliers here and there. This figure is observed below.

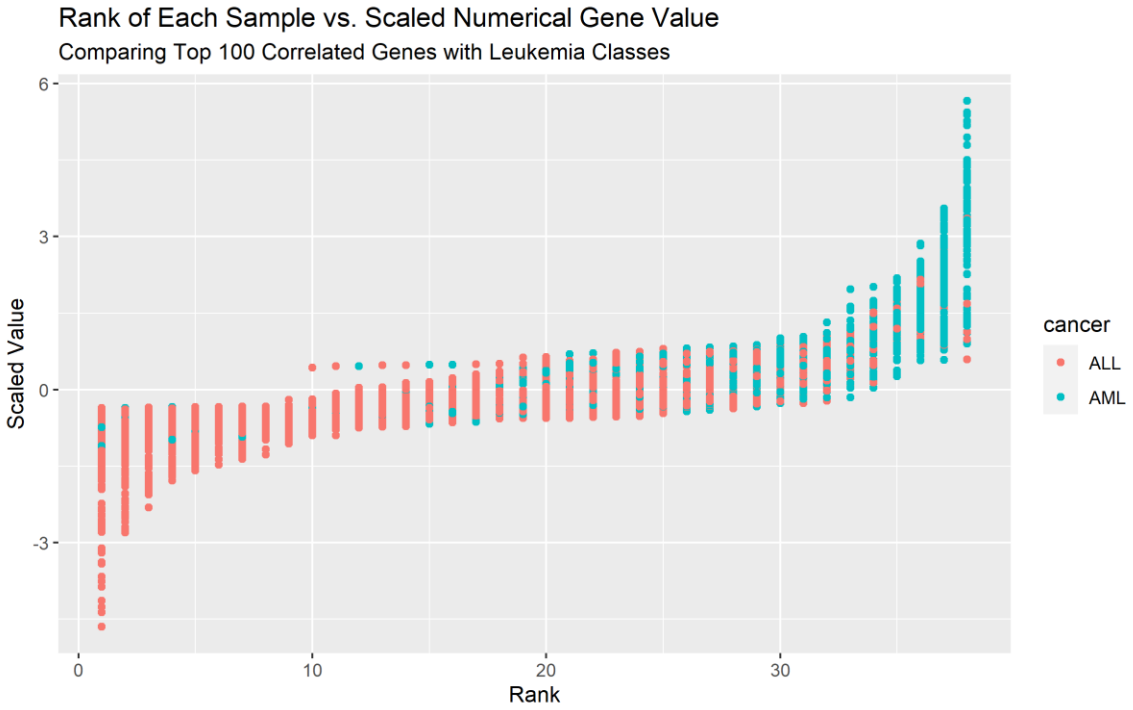


Figure 5: Highly Correlated Genes Numerical Data Visualized Against Class Separation

Since the categorical data can not be scaled, the percent count of the categorical data and class were graphed for each gene. To capture the difference between negatively and positively correlated genes, genes with positively correlated genes were graphed on bottom and negatively correlated genes on top. This allowed the distribution of A (absent), M (marginal), and P (present) categorical variables to be observed for each gene and how it correlates with the true class.

As assumed, for negatively correlated variables, the gene is typically A - absent for ALL while P – present for AML. For positively correlated genes, it was P – present for ALL while A – absent for AML.

This separation observed is bright news for the modeling using this categorical data of highly correlated genes.

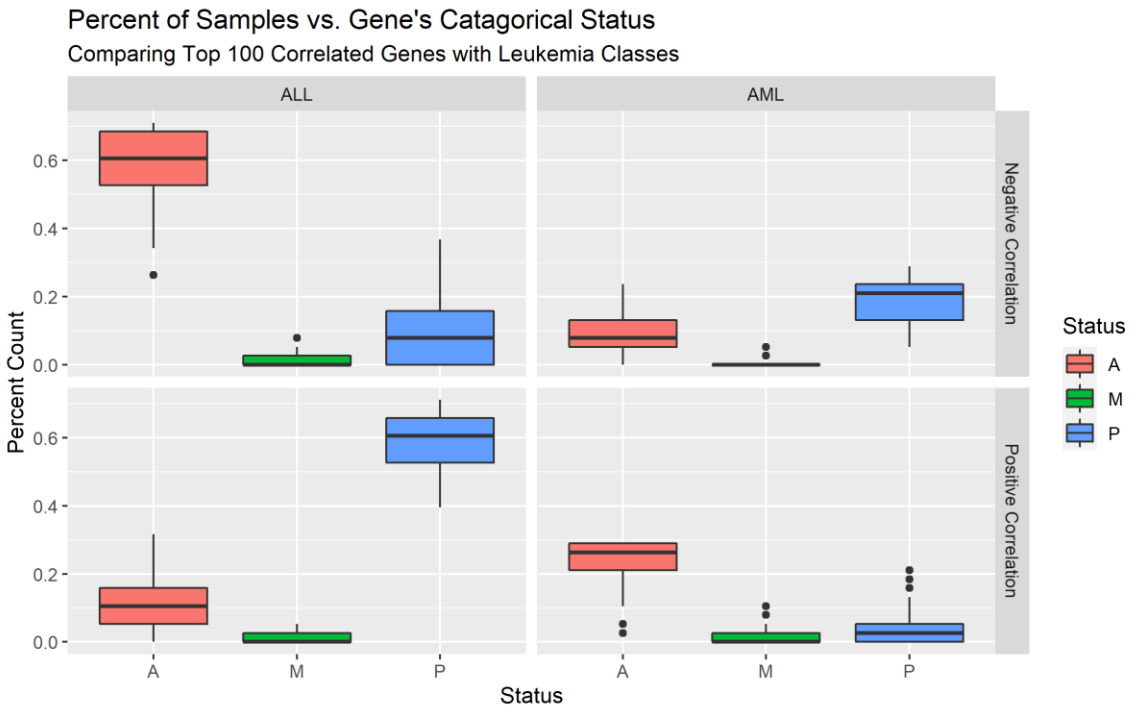


Figure 6: Highly Correlated Genes Categorical Data Visualized Against Class Separation and Correlation Direction

IV. Data Mining Techniques and Implementation

Given a dataset of gene expressions from a specified patient – the goal is to successfully classify if they have a specific condition. In the case of the current leukemia dataset – the patients have either Acute lymphocytic leukemia (ALL) or Acute myeloid leukemia (AML). The goal is to be able to separate the two classes using supervised learning classification models.

The complete the classification model various steps were taken. To start, the problem has to be defined and data sourced and explored. Next, the data was preprocessed. In the case of this problem given the nature of the data sourced, there was categorical and numerical data which are directly related by the same measurement / assay. Therefore, those were split into two categories and used for the next steps of the process.

Next, both numerical and categorical data needed to be reduced. Given the large amount of gene and relatively small number of samples. Therefore, two methods were used, PCA and a basic correlation matrix were used. As seen above, both methods saw some degree of separation. For these steps, only the training data was used and then applied to the test data from the findings as to not leak data into the test data.

Lastly various models were used and implemented. These included random forest, KNN, neural network and more. Once implemented they were analyzed for performance and compared against one another. As seen below, the flowchart outlines the steps of the process performed.

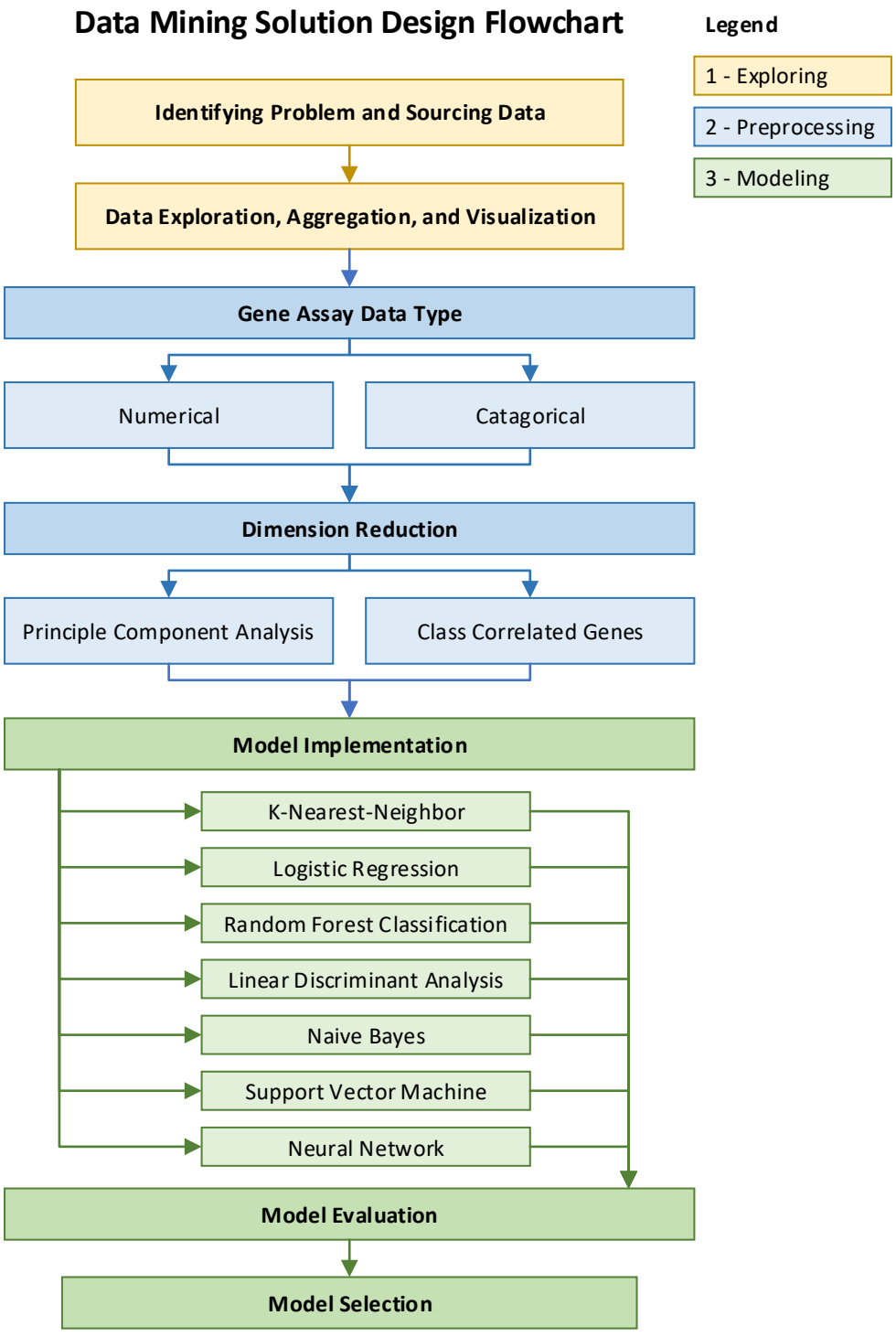


Figure 7: Data Mining Implementation Workflow Flowchart

V. Performance Evaluation

For each of the models, to obtain a training accuracy – a cross fold validation of 3 sets of 5 folds were preformed to determine any tuning of parameters. The metric used was accuracy of classification and the testing score was derived from the model using the sub-set testing data.

In the figure below, is an example of the visualization produced for each method to indicate the probability of each test sample given the reduction method and type of data. These values correlate with the table above given the test accuracy of a cutoff of 50%. Also, ROC curves were produced for each method and type of data / reduction method, but due to space were not included. Please refer to the code for full details.

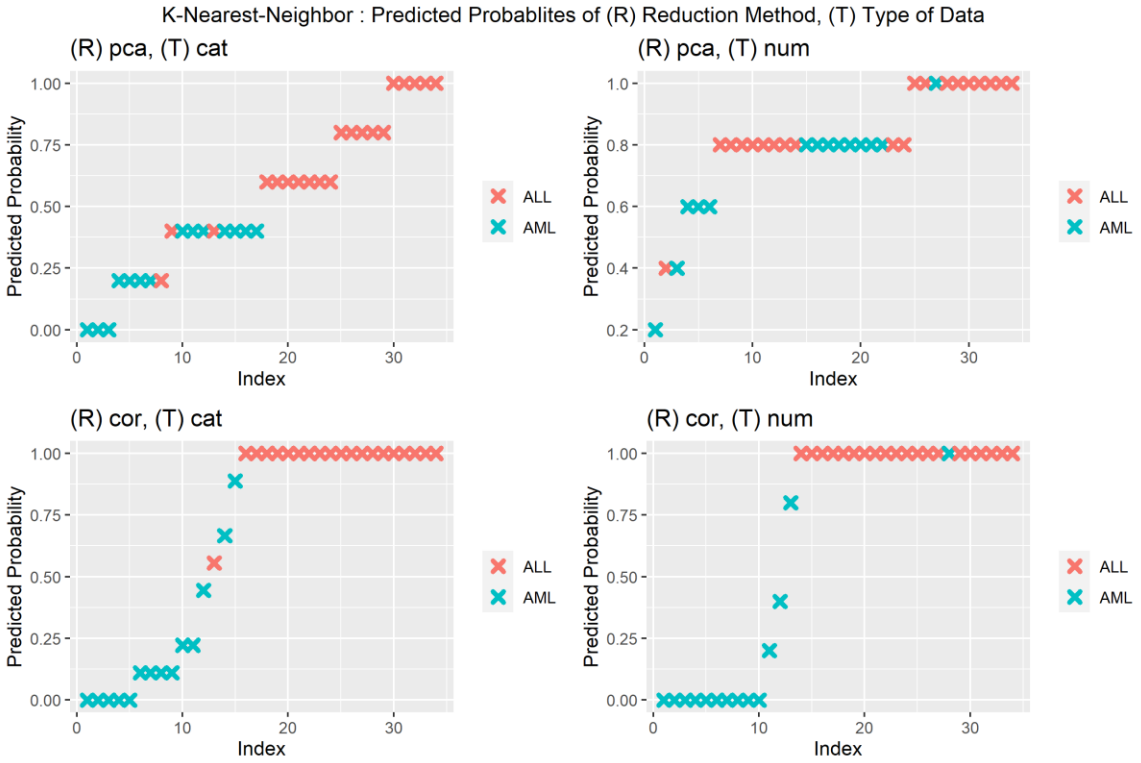


Figure 8: Example of Figure Showing Predicted Probabilities for KNN (produced for all methods but only included due to space)

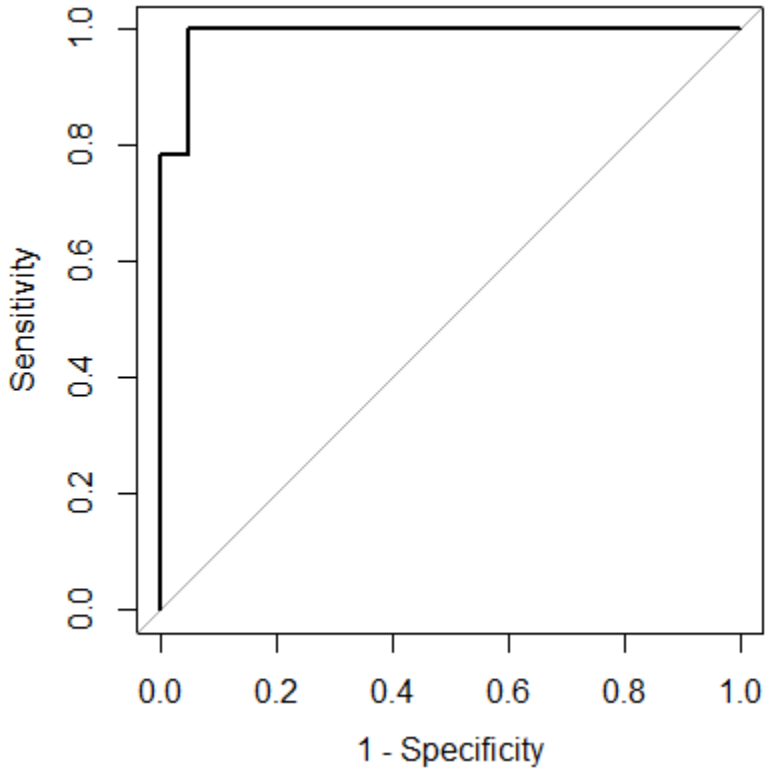


Figure 9: Example ROC Curve Produced per Model

For Full details of each model's predicted probabilities graph, refer to Appendix A. However, for a summary of the results view the table below, displaying the test accuracy, sensitivity, specificity and F1 score for each data set and model.

Method	Reduction	Type	Accuracy	Sensitivity	Specificity	F1
Log Regression	PCA	Cat	**0.91**	1.00	0.79	0.93
Log Regression	PCA	Num	**0.76**	0.95	0.50	0.83
Log Regression	Correlation	Cat	**0.88**	0.95	0.79	0.90
Log Regression	Correlation	Num	**0.94**	1.00	0.86	0.95
KNN	PCA	Cat	**0.91**	0.85	1.00	0.92
KNN	PCA	Num	**0.62**	0.95	0.14	0.75
KNN	Correlation	Cat	**0.94**	1.00	0.86	0.95
KNN	Correlation	Num	**0.94**	1.00	0.86	0.95
LDA	PCA	Cat	**0.91**	1.00	0.79	0.93
LDA	PCA	Num	**0.82**	1.00	0.57	0.87
LDA	Correlation	Cat	**0.88**	0.95	0.79	0.90

Method	Reduction	Type	Accuracy	Sensitivity	Specificity	F1
LDA	Correlation	Num	**0.88**	0.90	0.86	0.90
Naive Bayes	PCA	Cat	**0.88**	0.90	0.86	0.90
Naive Bayes	PCA	Num	**0.59**	0.95	0.07	0.73
Naive Bayes	Correlation	Cat	**0.94**	1.00	0.86	0.95
Naive Bayes	Correlation	Num	**0.91**	0.95	0.86	0.93
Neural Network	PCA	Cat	**0.97**	0.95	1.00	0.97
Neural Network	PCA	Num	**0.88**	0.80	1.00	0.89
Neural Network	Correlation	Cat	**0.97**	0.95	1.00	0.97
Neural Network	Correlation	Num	**0.97**	1.00	0.93	0.98
Random Forest	PCA	Cat	**0.68**	1.00	0.21	0.78
Random Forest	PCA	Num	**0.71**	1.00	0.29	0.80
Random Forest	Correlation	Cat	**0.88**	0.95	0.79	0.90
Random Forest	Correlation	Num	**0.97**	1.00	0.93	0.98
SVM	PCA	Cat	**0.91**	1.00	0.79	0.93
SVM	PCA	Num	**0.79**	1.00	0.50	0.85
SVM	Correlation	Cat	**0.91**	0.95	0.86	0.93
SVM	Correlation	Num	**0.97**	1.00	0.93	0.98

VI. Discussion and Recommendation

To summarize the results of each, the figure below shows the accuracy of each model and dataset. As observed, numerical + PCA dataset had not performed well against any of their counterparts. Comparatively, the correlated dataset performed well with both the numerical and categorical datasets. For the categorical data using PCA , the results are slightly worse than the correlated dataset. Overall, it seems like all methods preformed fairly similarly, achieving results of testing accuracies of + 85% using the correlated data across all models.

Figure 10: Comparing Model Performances with Various Datasets

For each of the results above, twenty components were used for PCA and 20 of the top correlated genes were used for the correlated dataset. To further dissect which genes are important to predicting, varying the number of genes chosen can be performed. This would need validation and training sets for each component amount.

VII. Summary

Overall, it was observed that mapping correlation associated with each of the genes for numerical and categorical data was greatly effective at predicting the correct class of Leukemia. Comparatively, it was observed performing PCA on the numerical dataset had performed poorly across all models. Between the numerical and categorical datasets in which correlated genes were chosen, the results are within variance, but numerical dataset performed slightly better.

Given the results, it is recommended that the top twenty genes used for the tests can accurately determine the class of Leukemia. The dataset best suited is the numerical dataset and implementing either Neural Network, SVM or random forest, however other models like Log Regression and Naive Bayes work well.

References

Dwivedi, A. (2018). Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Computing and Applications, 29(12), 1545-1554.

Golub, T., Slonim, D., Tamayo, P., & Huard, C. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286(5439), 531-7.

Farber, S., Diamond, L., Mercer, R., Sylvester, R., et al. (1948). Temporary Remissions in Acute Leukemia in Children Produced by Folic Acid Antagonist, 4-Aminopteroyl-Glutamic Acid (Aminopterin). N. Engl. J. Med. 238, 787-793.

Frei, E., Freireich, E., Gehan, E., Pinkel, D., Holland, J., Selawry, O., & Taylor, R. (1961). Studies of Sequential and Combination Antimetabolite Therapy in Acute Leukemia: 6-Mercaptopurine and Methotrexate. Blood, 18(4), 431-454.

Quaglino, D., & Hayhoe, F. (1959). Observations on the periodic acid-Schiff reaction in lymphoproliferative diseases. The Journal of Pathology and Bacteriology, 78(2).

Bennett, J., & Dutcher, T. (1969). The cytochemistry of acute leukemia: Observations on glycogen and neutral fat in bone marrow aspirates. Blood, 33(2), 341-347.

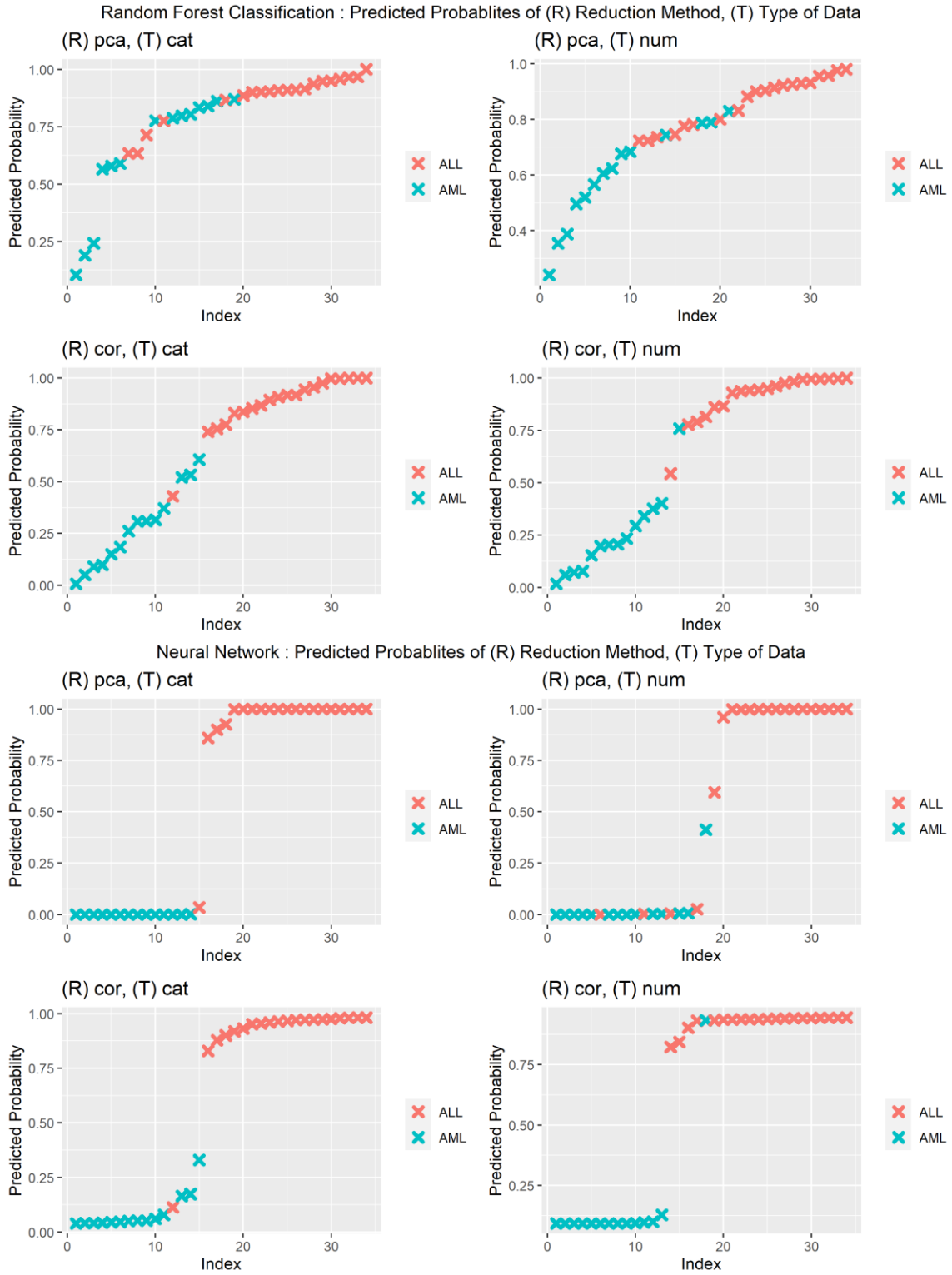
Graham, R., Lundholm, U., & Karnovsky, M. (1965). CYTOCHEMICAL DEMONSTRATION OF PEROXIDASE ACTIVITY WITH 3-AMINO-9-ETHYLCARBAZOLE. The Journal of Histochemistry and Cytochemistry: Official Journal of the Histochemistry Society, 13, 150-2.

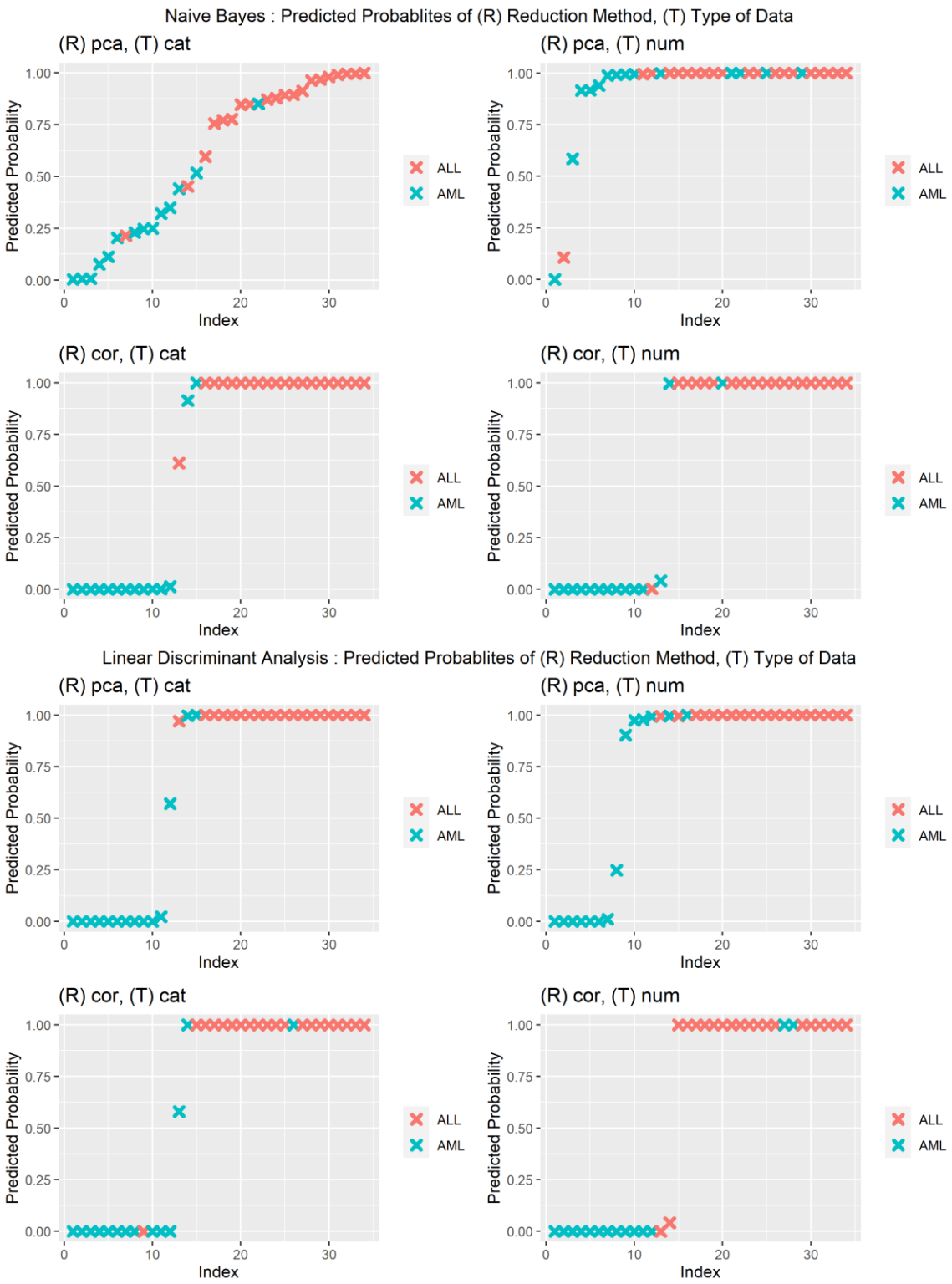
Tsukimoto, I., Wong, K. & Lampkin, B. (1976). Surface Markers and Prognostic Factors in Acute Lymphoblastic Leukemia. N. Engl. J. Med. 294, 245-248.

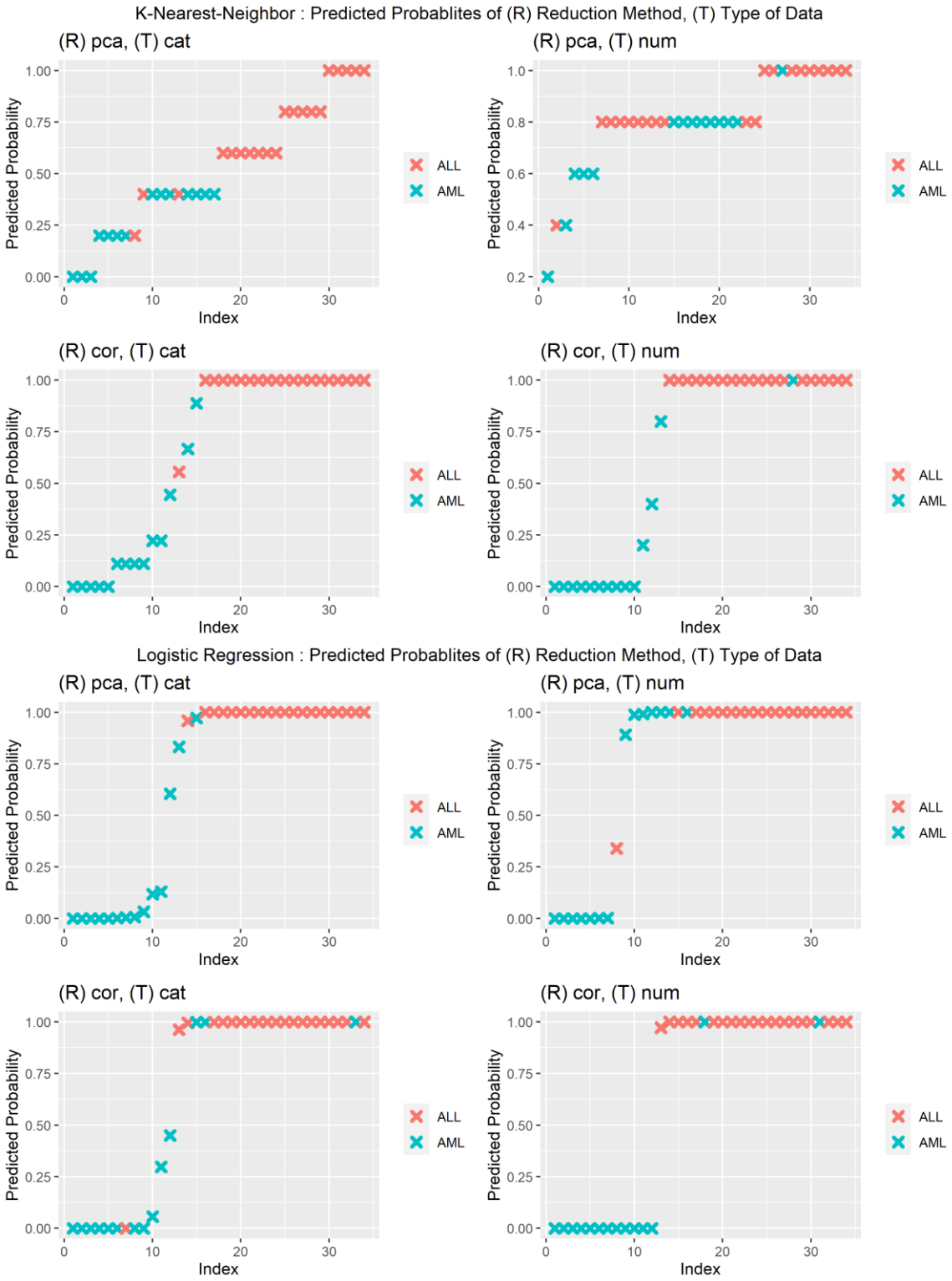
Schlossman, S., Chess, L., Humphreys, R., & Strominger, J. (1976). Distribution of Ia-like molecules on the surface of normal and leukemic human cells. Proceedings of the National Academy of Sciences of the United States of America, 73(4), 1288-1292.

Barker, G., Golub, T., Gilliland, D., Bohlander, S., Rowley, J., Heibert, S., Morgan, E. (1995). Fusion of the TEL gene on 12p13 to the AML1 gene on 21q22 in acute lymphoblastic leukemia. Proceedings of the National Academy of Sciences of the United States of America, 92(11), 4917-4921.

Appendix A: Results of Predicted Probabilities per Model




Naive Bayes : Predicted Probablites of (R) Reduction Method, (T) Type of Data
(R) pca, (T) cat
(R) pca, (T) num
(R) cor, (T) cat
(R) cor, (T) num
Predicted Probability
Index
ALL
AML
Linear Discriminant Analysis : Predicted Probablites of (R) Reduction Method, (T) Type of Data
(R) pca, (T) cat
(R) pca, (T) num
(R) cor, (T) cat
(R) cor, (T) num


K-Nearest-Neighbor : Predicted Probablites of (R) Reduction Method, (T) Type of Data
(R) pca, (T) cat
(R) pca, (T) num
(R) cor, (T) cat
(R) cor, (T) num
Logistic Regression : Predicted Probablites of (R) Reduction Method, (T) Type of Data
(R) pca, (T) cat
(R) pca, (T) num
(R) cor, (T) cat
(R) cor, (T) num
Predicted Probability
Index
ALL
AML