Expression of Genes in AML and ALL type Leukemia

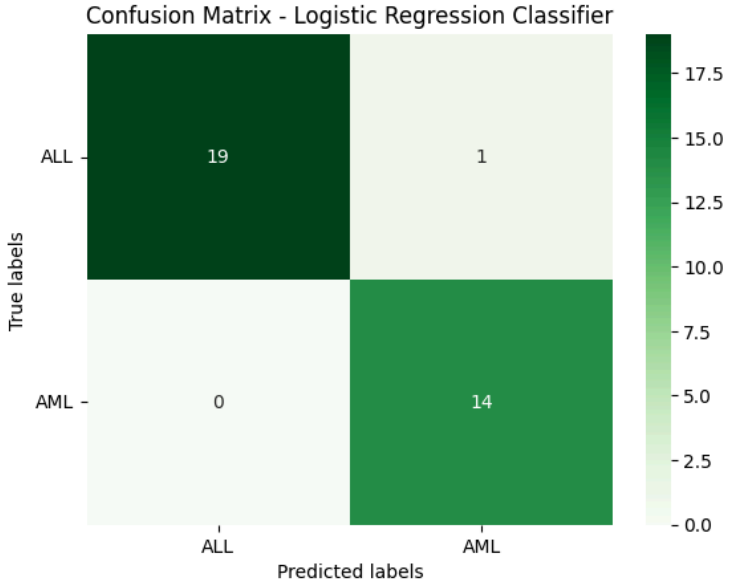
Dataset

There are two datasets containing the initial (training, 38 samples) and independent (test, 34 samples) datasets used in the paper "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" by Golub et al. These datasets contain measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood. The classes ALL and AML are encoded as 0 and 1 respectively.

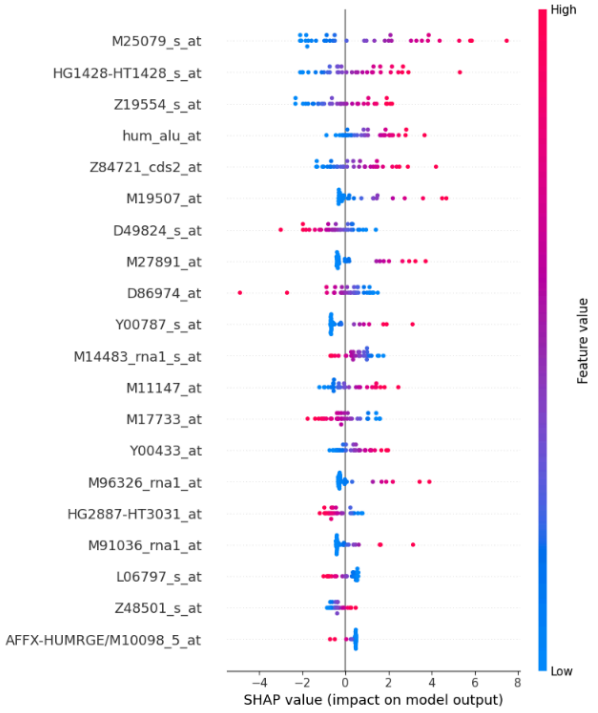
(1) Classifier

2 models were built and the logistic regression model was chosen as the best considering its higher accuracy.

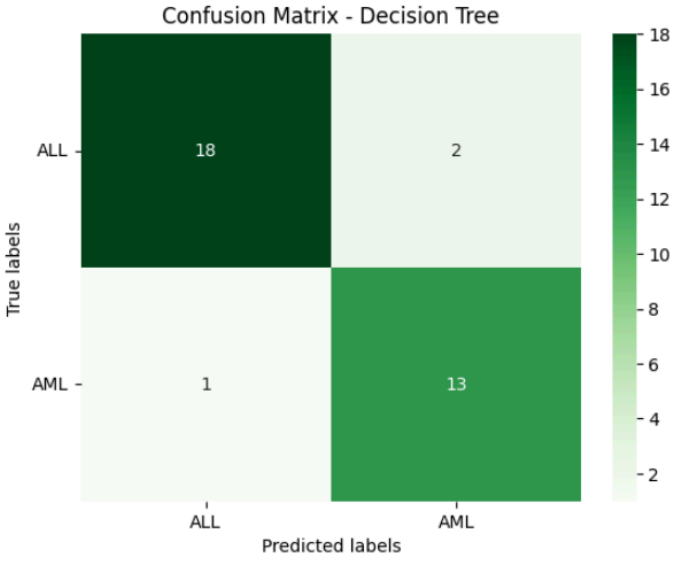
Model 1: Logistic Regression model
Accuracy = 97.06%
Confusion matrix:



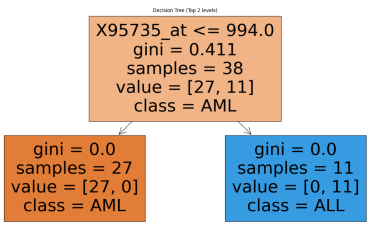
Summary Plot of Feature Importance using SHAP:



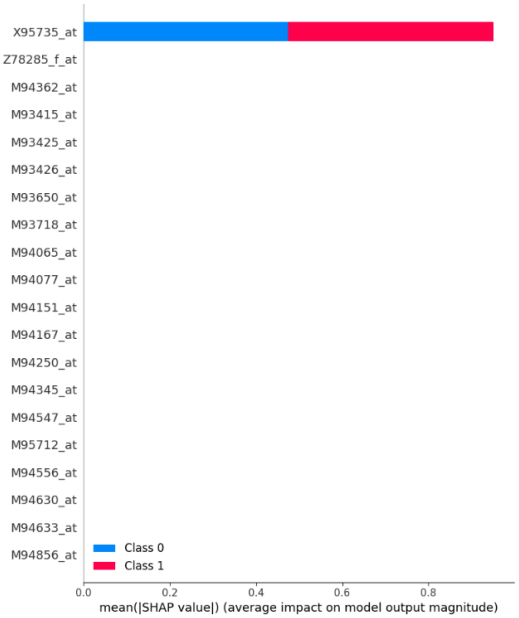
Model 2: Decision Tree
Accuracy = 91.18%
Confusion matrix



Self interpretability of the decision tree:



Summary Plot of Feature Importance using SHAP:



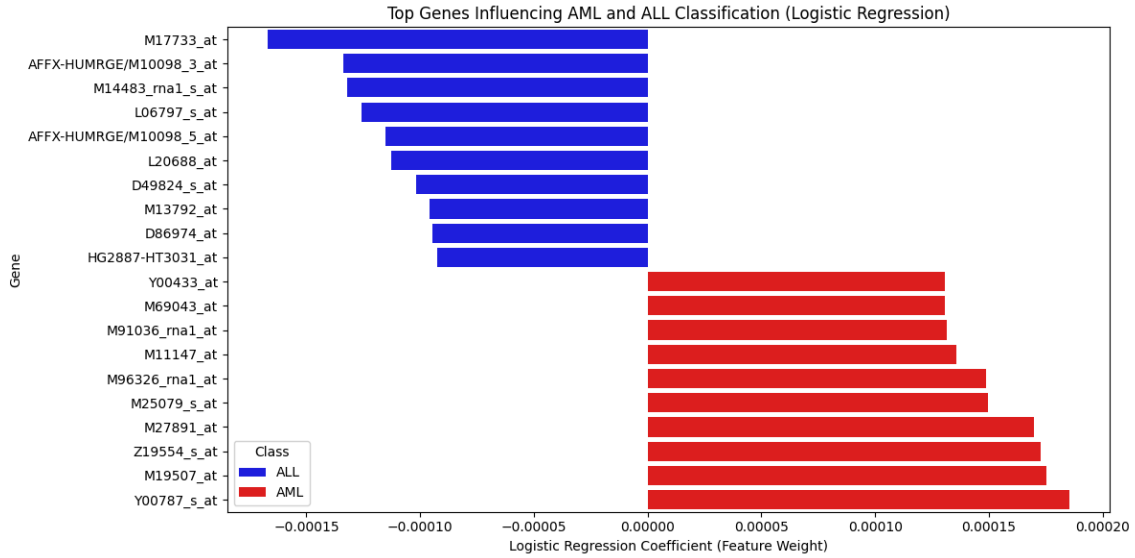
(2) Patterns in gene expression that help distinguish between the types AML and ALL

Top genes associated with AML (positive weights):

- Y00787_s_at: 0.0002
- M19507_at: 0.0002
- Z19554_s_at: 0.0002
- M27891_at: 0.0002
- M25079_s_at: 0.0001
- M96326_rna1_at: 0.0001
- M11147_at: 0.0001
- M91036_rna1_at: 0.0001
- M69043_at: 0.0001
- Y00433_at: 0.0001

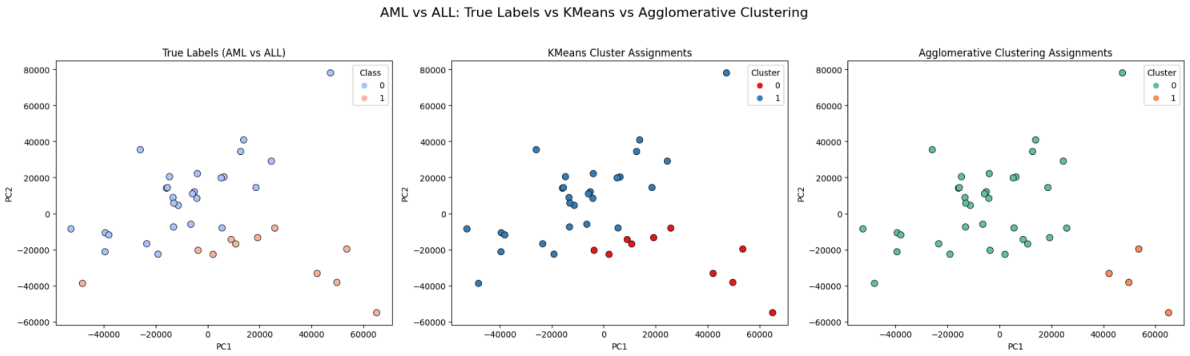
Top genes associated with ALL (negative weights):

- M17733_at: -0.0002
- AFFX-HUMRGE/M10098_3_at: -0.0001
- M14483_rna1_s_at: -0.0001
- L06797_s_at: -0.0001
- AFFX-HUMRGE/M10098_5_at: -0.0001
- L20688_at: -0.0001
- D49824_s_at: -0.0001
- M13792_at: -0.0001
- D86974_at: -0.0001
- HG2887-HT3031_at: -0.0001



(3) Clustering

Tried K-Means clustering and Agglomerative clustering. K-Means clustering showed a higher purity than the Agglomerative clustering.
Purity of the kmeans clustering: 0.9737
Purity of Agglomerative Clustering (linkage='ward'): 0.8158



Unsupervised clustering techniques like KMeans and Agglomerative Clustering are powerful tools for discovering hidden patterns in gene expression data without needing known class labels (e.g., disease types).

Due to the great dimensionality and complexity of the data, unsupervised clustering is especially useful in gene expression analysis. It aids in the discovery of natural groups by detecting similarities in gene expression patterns between samples. Clustering is also important in exploratory analysis before supervised classification because it allows researchers to validate the data's underlying structure, detect potential subtypes or outliers, and even suggest biomarkers based on genes that contribute to cluster separation.