Assignment: Expression of Genes in AML and ALL type Leukemia

The dataset available here contains expression level data for over 7000 genes in individuals with either AML or ALL type Leukemia.

Q1) Build a classification model to predict the type of Leukemia as AML or ALL. Your classifier should be intepretable, I.e., one can look at the model and figure out how it identifies the classes, e.g., rule-based classifiers, logistic regression, or decision trees. You may use more complex models but then you will have to apply explainable AI methods as a separate step to interpret results. Tune model parameters to achieve the best possible prediction results but ***do not*** use the test set for this tuning.

The dataset has already been divided into training and testing sets. Please use them for your training and testing purposes, respectively. Also, note that the data has already been normalized.

Q2) Using the model you developed in Q1, identify any patterns in gene expression that help distinguish between the types AML and ALL.

Q3) Without using the class labels in the data, apply an unsupervised technique such as clustering to group the data into two clusters. Report the purity of the best clustering you got (you will need to use the class labels to compute the purity though). Explain the use of this kind of unsupervised clustering in the case of analyzing gene expression data such as this.