Assignment: Expression of Genes in AML & ALL type Leukemia

Name: Chirath Hettiarachchi
Index No: 188098R

Question 1

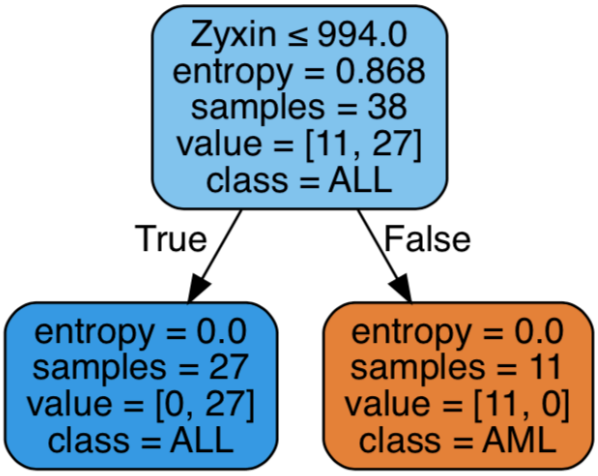
- Please find the **“classifier.py”** file attached.
- The train / test datasets were initially processed in order to clearly identify the features (genes) and the corresponding labels. (AML / ALL)
- The duplicate genes were removed.
- The number of genes were really large; hence a feature reduction was required. In the research paper related to the dataset, the researchers had selected 50 genes with the highest correlation. This was done based on a technique they developed. (“Neighborhood Analysis”)
- Instead, I used an ANOVA test to calculate the correlation between the target labels and the features. Based on the probability scores the best 50 genes were selected. (Lowest Probability values. As it is corresponding to the null hypothesis) [Note: There is a slight difference between the selected genes between the two methods.]
- Subsequently a binary classification was carried out using a Decision Tree classifier, in order to ensure the interpretability of the results.
- The model parameters were selected appropriately to reduce overfitting. (Ex: The maximum depth of the tree was limited to 5, etc) The hyperparameters were tuned, and stratified cross validation (5-skf) was used to select the best parameters.
- The test subjects (separate file) was then used to evaluate the accuracy of the model. All the subjected were classified **91.18%** accurately. (31/34 Test subjects)

Confusion Matrix		Predicted	
		AML	ALL
Actual	AML	13	1
	ALL	2	18

Question 02

Identify patterns in the gene expression that help distinguish between the types AML & ALL.

The rules identified by the Decision Tree Classifier was visualized in order to identify potential patterns.



From the above obtained result, it can be identified that **"Zyxin"** is a key gene to classify ALL and AML categories. And that if the gene expression is less than 994.0 it would classify ALL and if not would classify AML.

From the conducted ANOVA test also **"Zyxin"** achieved the second highest rank.

Rankings from the ANOVA test (only first 8 genes presented)

1. Leukotriene C4 synthase (LTC4S) gene
2. **Zyxin**
3. FAH Fumarylacetoacetate
4. LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
5. CD33 CD33 antigen (differentiation antigen)
6. Liver mRNA for interferon-gamma inducing factor(IGIF)
7. PRG1 Proteoglycan 1, secretory granule
8. GB DEF = Homeodomain protein HoxA9 mRNA

In order to validate the results, I obtained further, I checked with the results presented in the research paper. (The results from the research paper is presented below)

Discussion: Comparing Obtained Results & Research Paper Results.

1. The results here also show, that “Zyxin” is one of the top genes (2nd best) expressed in type AML. And that it is not expressed in type ALL. (please look at the gene expression in the AML section)
2. However, other main genes under AML is also present in ALL type. So, the classifier hasn’t selected those genes for the classification. (Ex: Fumarylacetoacetate)

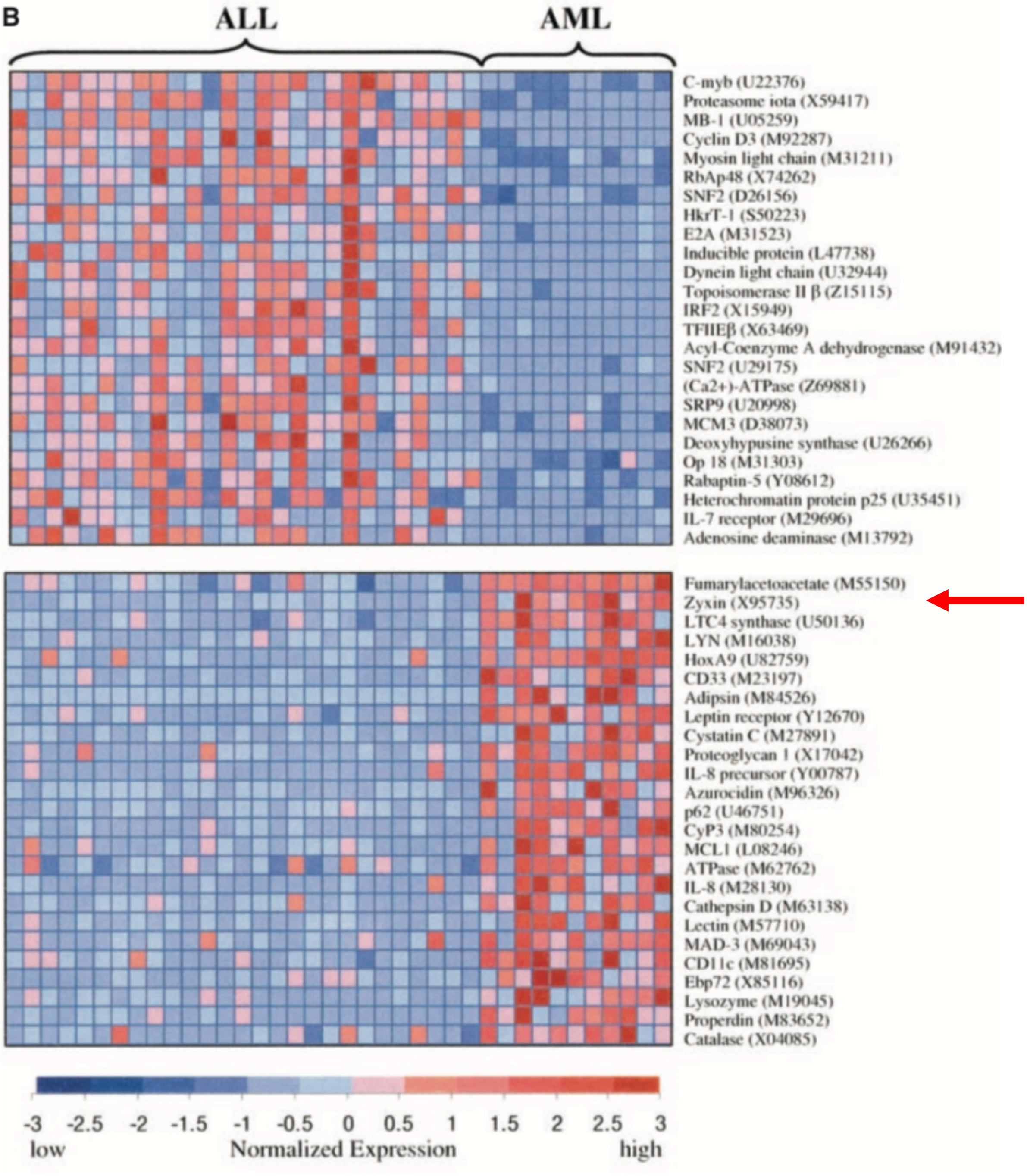


Figure 2. Research Paper Gene Expression Results

Question 03

Unsupervised Clustering. (Please fins the code also in the clustering.py file attached)

- The two datasets were combined (train & test) in order to increase the number of available data points.
- The above selected 50 genes were used as features for a k-means clustering with 2 classes.
- In order to calculate the purity of the unsupervised learning; a confusion matrix was derived.

Confusion Matrix		Unsupervised Predicted Label	
		AML	ALL
Actual Label	AML	20	5
	ALL	0	47

- The purity of the clustering is **93.06%** (67/72).

Explain the use of this kind of unsupervised clustering in the case of analyzing gene expression data such as this.

- The unsupervised clustering could be used to learn new types of cancers which were previously unknown, which is of utmost important for treatment. (In instances where we do not know the type of the cancer. i.e the Label)
- Clear identification of the type of cancer enables the targeting of specific therapies to pathogenetically distinct tumor types, to maximize efficacy and minimize toxicity.
- This kind of clustering is important as it can be done solely focusing on the gene expression data, independent of previous biological knowledge.