Hassan Shah

# Leukemia Multi-Class Classification Based on Various Gene Expression Levels

## Introduction:

Leukemia refers to a broad category of cancers affecting blood cells and bone marrow. One of the most common forms of leukemia is acute myeloid leukemia (AML), which affects the myeloid cells, a type of white blood cell that helps to fight infection. AML is often treated with chemotherapy and a bone marrow transplant. However, the emergence of preventative medicine and machine learning algorithms has allowed researchers to develop predictive methods for AML by using transcriptomic and genomic data (Eckardt et al., 2020). In order to test the prediction capabilities of the current models being examined by researchers, this project aims to fit various statistical learning models to genomic data associated with AML.

The dataset used in this analysis is a leukemia microarray dataset from the Curated Microarray Database (CuMiDa), a repository containing 78 cancer microarray datasets designed to test various machine learning methods (Feltes et al., 2019). The primary goal of this analysis is to compare the prediction capabilities of different machine learning models to determine their efficacy in predicting leukemia types. Additionally, since there are five different classes in the output column of this dataset, various multi-class classification tools will be used to deal with a multi-class output. Lastly, in the three tree-based models, one can examine the importance each model assigns to different genes to determine which genes are weighted more significantly than others. This process can help evaluate whether only a few genes play a significant role or whether each model uses vastly different features.

**Methods:**

Before performing any analysis or fitting models, the dataset needs to be examined for any necessary preprocessing. Based on the fact that there are no null values, no data imputation techniques are required. Additionally, by using the 'describe' and 'corr' functions, the script returns the distributions and correlations, respectively, of the features in the dataset. Seeing as there are no significant issues with distribution or correlation, the next step is proceeding to the analysis of the feature set.

The dataset contains 64 samples of 22,284 gene expression levels and the associated leukemia class, which is the output. There are five different classes of leukemia in the output column, of which AML encompasses 40% while the other four classes span between 10% and 20%. The distribution of the output column is displayed below.

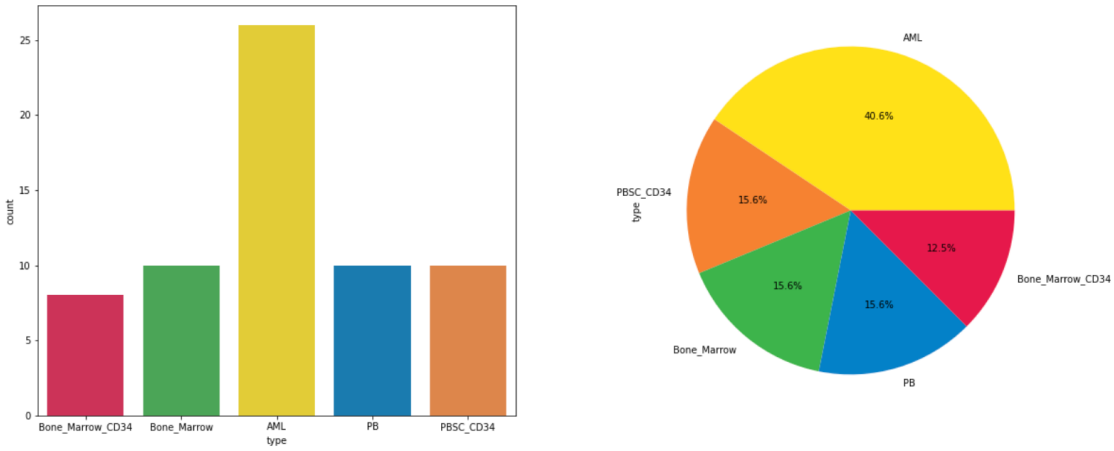


**Figure 1: Bar and Pie Plots of the Distribution of Output Column (Leukemia Sample)**

After separating the features and output and splitting the dataset into train (48 samples) and test sets (16 samples), it is now possible to begin working to fit learning methods to the data. However, before beginning, it can be valuable to test the 'DummyClassifier,' which acts as a baseline for accuracy by estimating accuracy if the model were always to guess the majority class. In this instance, guessing the majority class leads to an accuracy of 25%, so all models

should ideally predict with greater accuracy than 25% accuracy. As previously mentioned, the output in this dataset contains five different classes, which requires using various preprocessing methods to deal with multi-class classification. Specifically, this analysis will examine one vs. rest classification, one-hot encoding, and label encoding. Different learning methods require different multi-class processes, and a couple of different learning models will be used for each method.

The first multi-class classification method in this study is one vs. rest classification. One vs. rest (OvR) classification is a multi-class classification method in which a separate binary classifier is trained for each class in the dataset. Then, the binary classifier is trained to predict whether an example belongs to the target class or not, and the class with the highest predicted probability is selected as the final prediction. In this analysis, two models use OvR: logistic regression and support vector classification. As both these methods are linear models, it is generally recommended to standardize features.

Next, one-hot encoding is performed. One-hot encoding is a method used to convert multi-class categorical data into a numerical representation that machine learning algorithms can use. In one-hot encoding, each category is represented by multiple columns, with a “1” in the column corresponding to the category and a “0” in all other columns. Two models are fit using one-hot encoding: k-nearest neighbors and multinomial naive Bayes. Additionally, the features are scaled for one-hot encoding. Scaling is not necessary in this case but is helpful as these models are sensitive to scale.

The final method tested here is label encoding. Label encoding is a method for converting categorical data, represented as a string or an integer, to a numerical representation that machine learning algorithms can use. This conversion is achieved by assigning a unique integer to each

category or class in the data so that a single integer value can represent each category or class. In this case, decision trees, random forest, and extreme gradient boosting (XGBoost), three tree-based methods, will be fit to the data with label encoding. Additionally, since tree-based methods allow for the assessment of feature importance by determining which features create the most significant splits in trees, the top ten features of each tree model are recorded to assess significance.

**Results:**

All models were fit to the training data and assessed for prediction accuracy with the test data. For the one vs. rest classifiers, the logistic regression model provided a test accuracy of 93.75%, while the support vector classifier yielded a prediction accuracy of 87.5%. Next were the one-hot encoded models. K-nearest neighbors had a prediction accuracy of 93.75%, which was also the prediction accuracy for naïve Bayes. Finally, for the label-encoded tree models, the decision tree, random forest, and gradient boosting models had predictive accuracies of 93.75%, 100%, and 93.75%, respectively. A comparison of predictive accuracy amongst all learning models fit in the analysis is provided below.

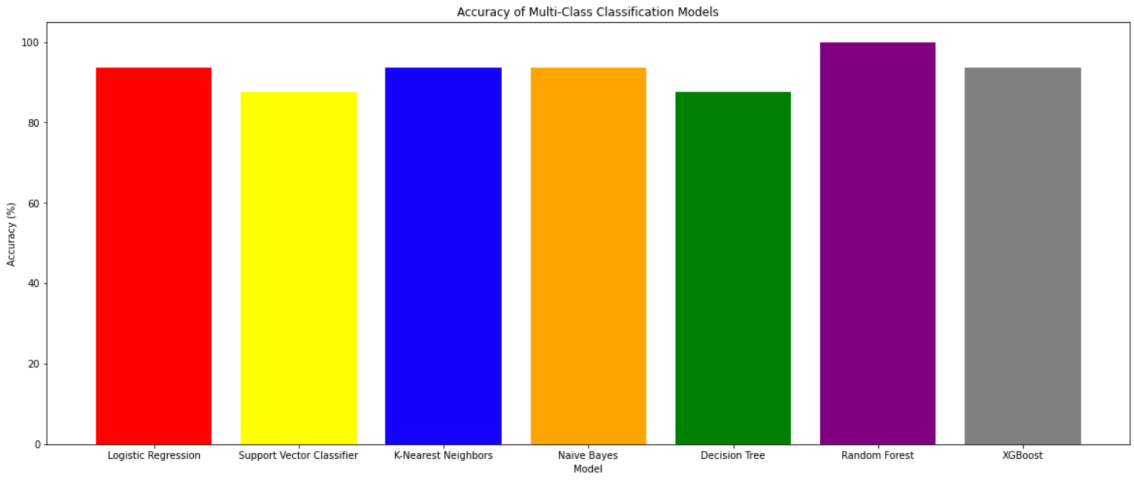


**Figure 2: Predictive Accuracy of Multi-Class Classification Methods**

Feature importance was determined through analysis of the top ten features in each tree-based model. In all three models, the '221268_s_at' gene levels acted as a significant gene. In fact, '221268_s_at' was the first or second most important feature in all three models.

**Discussion:**

Random forest with label encoding provided the highest accuracy of any model in this analysis. This finding makes sense, given that many studies have found random forest models to be some of the most adept algorithms at predicting AML. Since the number of samples in this dataset was limited, a prediction accuracy of 100% is likely not achievable with more samples and, subsequently, more irreducible error. However, the relative performance of random forest compared to other models employed in this project suggests that it may provide the best chance of being used in preventative AML care. In fact, other researchers with far larger datasets and samples have achieved predictive accuracies of around 93% using random forest models to predict AML (Dasariraju et al., 2020). Given the success of random forests in predicting leukemia, a future iteration of this analysis could be to use each of the various multi-class labeling techniques to determine if random forests can further be fine-tuned to predict well on larger datasets. However, any future analysis would require more data as this study's limited number of samples prevented further exploration, as the test accuracy for the random forest model was already 100%.

Apart from prediction accuracy, the other facets being examined were the different multi-class labeling methods and the feature importance of the tree models. Regarding labeling methods, none of one vs. rest classification, one hot-encoding, or label encoding resulted in significant differences in predictive accuracy. All models achieved at least 87.5% predictive accuracy, and the averages of each method amongst all models were not significantly dissimilar.

This lack of difference further suggests that each method is valid so long as it is used with compatible models and the necessary preprocessing is performed.

Finally, the tree-based models highlighted the importance of the '221268_s_at' feature. '221268_s_at' was the only feature to appear in each of the three tree models, and even more significantly, it was the first or second most important feature in each model. With '221268_s_at' seemingly playing an essential role in leukemia classification, further research was conducted on its biological function. Not much research is available on '221268_s_at', but it is known to be involved in producing SGPP1, a sphingolipid metabolite that regulates different biologic processes. A previous study from 2015 found that SGPP1 could be a prognostic biomarker for advanced gastric cancers (Gao et al., 2015). Although there seems to be relatively little current research on SGPP1's status in leukemia, it may very well be worth researching, given that the gene levels for '221268_s_at' appear to be highly relevant in determining leukemia class.

**Conclusion:**

This study aimed to evaluate the efficacy of different machine learning models in predicting leukemia class using the expression levels of over 20,000 genes from the Curated Microarray Database (CuMiDa). One vs. rest classification, one-hot encoding, and label encoding were the different multi-class classification techniques used to support these models in dealing with multiple classes in the output. The results of this study showed that the random forest model predicted the leukemia class of test samples with 100% accuracy. However, this number would likely decrease with more data as there were only 16 test samples. Additionally, the three tree-based models evaluated the gene expression levels of the '221268_s_at' gene to be most significant in determining leukemia class. Although not much is known about

‘221268_s_at’, it is known to produce SGPP1, a sphingolipid that has been touted as a potential biomarker for advanced gastric cancer. Further studies should be done to determine whether SGPP1 may also be a potential biomarker for AML.

**References:**

Dasariraju, S., Huo, M., & McCalla, S. (2020). Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random forest algorithm. Bioengineering, 7(4), 120.

Eckardt, J. N., Bornhäuser, M., Wendt, K., & Middeke, J. M. (2020). Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects. Blood Advances, 4(23), 6077-6085.

Feltes, B. C., Chandelier, E. B., Grisci, B. I., & Dorn, M. (2019). Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. Journal of Computational Biology, 26(4), 376-386.

Gao, X. Y., Li, L., Wang, X. H., Wen, X. Z., Ji, K., Ye, L., ... & Ji, J. F. (2015). Inhibition of sphingosine-1-phosphate phosphatase 1 promotes cancer cells migration in gastric cancer: Clinical implications Corrigendum in/10.3892/or. 2018.6269. Oncology reports, 34(4), 1977-1987.