Final Project

The number of samples I got in this data is 359 sample for both AML patients and normal people (figure 1) by flow cytometry, there are 43 AML cases and 316 normal cases in each class (figure 2), where the number of features is 128, there is no missing data.

Since the number of features is big with small number of samples it will affect my analysis as it is high dimensional data, it may cause overfitting. And, the number of normal people is a lot more than AML people which may not produce a reliable model.

To overcome this problem, we may need to feature selection reduction we may apply techniques as PCA.

Since there is no missing data, this is good as there is no preprocessing.

A screenshot of a computer

Description automatically generated

Figure 1: data table

A screenshot of a computer

Description automatically generated

Figure 2: distribution

I picked four clustering methods to my data, I started with K-mean clustering where according to silhouette score (figure 3), since all are near to zero, so the clusters are not well separated, not well distinct, the best score was of 3 clusters with score 0.177 indicates slightly better separation but still indicates a significant amount of overlap and the second best was for 2 clusters with score 0.157 This score shows some level of distinction between the two clusters, but it's weak separation. This value indicates there is still overlap between the groups. 4 Clusters with score 0.075, this is a significant drop from three clusters, means that adding another cluster does not help to better separate the data. 6 Clusters scoring 0.074, a slight improvement over five clusters but still low. 7 Clusters score is 0.073 Similar to six clusters, the score is still low showing that further increasing the number of clusters continues to cause poor separation. 8 clusters score is 0.070, the increase in the number of clusters to eight does not provide better clarity or separation. The scatter plot for the K-means (figure 4) is not doing a great job separating the clusters but I think it is better than the others where there is some separation between classes. The second clustering method is Hierarchical Clustering with single linkage (Figure 5) doesn’t do well. We see that almost all cell types are clustered in C2 Failing to identify data points in other clusters. since all markers (cell types) are shown to belong to C2, which is misleading, the third method is Hierarchical Clustering with complete linkage (Figure 6) this also is not doing well as we can see there are a few points that is blue it is also favoring cluster 2 but it doing a little better than single linkage, the fourth method is Hierarchical Clustering with ward linkage (Figure 7) this is the best method of all three of the Hierarchical Clustering we can see some separation , Classification should be possible with this data, as there is some degree of separability between clusters. The performance of each method could vary based on how well it can handle the overlap and complexity shown in the plots. To enhance classification, we should do feature selection (dimensionality reduction) and regularization.

A screenshot of a computer

Description automatically generated

Figure 3: Silhouette score

A screenshot of a computer

Description automatically generated

Figure 4: K-means scatter plot.

A screenshot of a computer

Description automatically generated

Figure 5: Hierarchical Clustering with single linkage scatter plot

A screenshot of a computer

Description automatically generated

Figure 6: Hierarchical Clustering with complete linkage scatter plot

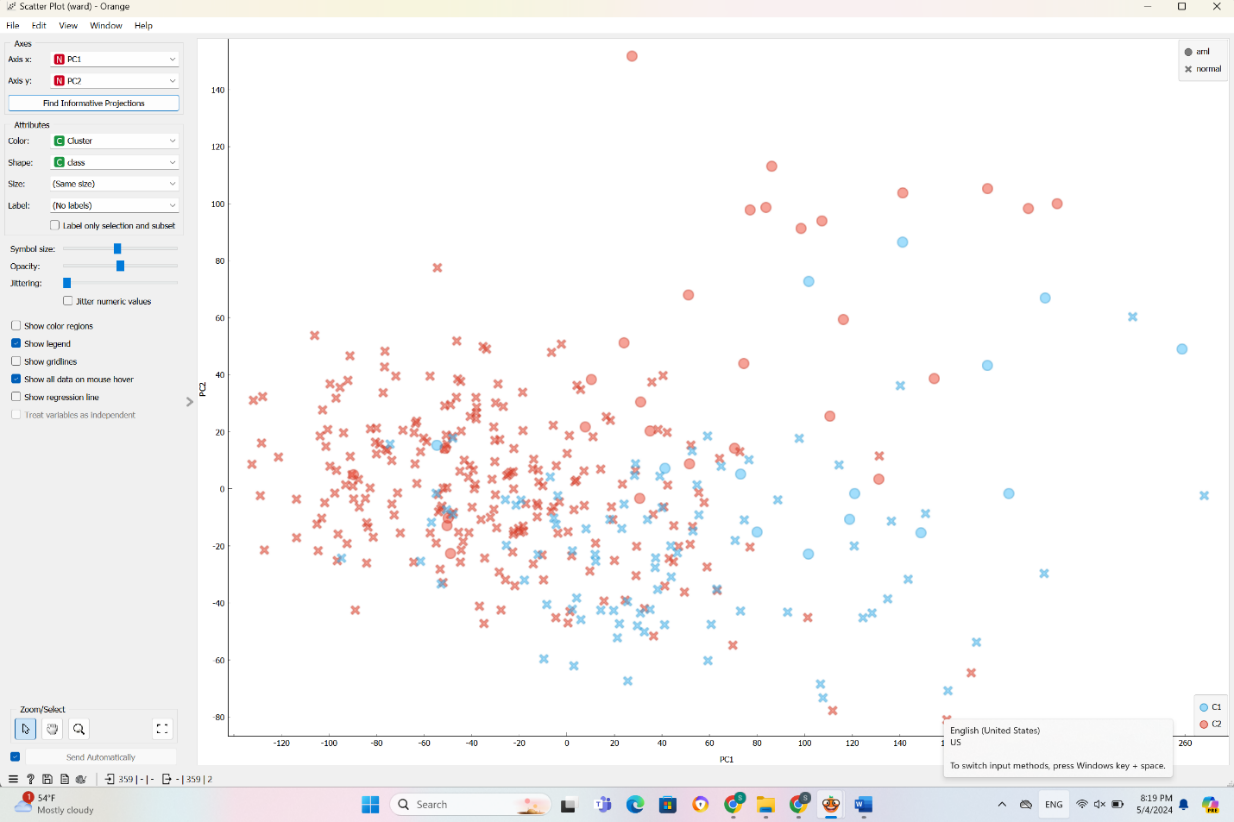
****

Figure 7: Hierarchical Clustering with ward linkage scatter plot

For data classification I picked four different classification methods, Naïve Bayes, Random Forest, Tree, and SVN.  
The first one is Naïve Bayes, I applied a preprocessing step, I changed number of intervals with equal width bins from 5 as a default to 2 I got the best test scores (figure 8) and confusion matrix (figure 9) with these parameters the model correctly predicted 29 AML cases, correctly identified 214 normal cases, incorrectly predicted 2 normal cases as AML, and failed to identify 7 cases of AML. The second method I picked is Random Forest the default is set number of trees to 50 trees and limit depth of individual tree is 3, and ensured no subsets smaller than 5 are split, when I changed the number of don’t split subset smaller than from 5 to 7 has best test score and confusion matrix (figure 10) The model correctly predicted 28 AML cases, correctly identified 219 normal cases, incorrectly predicted 3 normal cases as AML, and failed to identify 2 cases of AML.  
The third classification method I picked is tree the default used for the minimum number of instances in leaves was 2, I changed it a to 8 where it gives the strongest model, the best confusion matrix for this method with these parameters (figure 11) The model correctly predicted 22 AML cases, correctly identified 214 normal cases, incorrectly predicted 9 normal cases as AML, and failed to identify 7 cases of AML.  
The fourth and the best model that performed is SVM. This is the default Set C=1 to balance the margin maximization and training error minimization. Used a linear kernel to manage the simplicity and effectiveness of the model. When changing the parameters didn’t see shift in the confusion matrix (figure 12), the model correctly predicted 29 AML cases, correctly identified 221 normal cases, incorrectly predicted 2 normal cases as AML, and the model did not miss any AML cases.  
SVM shows the strongest performance among the models where it has the highest number of True Positives and True Negatives with the lowest number of False Positives and no False Negatives. This shows a high level of accuracy and a strong ability to correctly classify both AML and normal cases without missing any AML cases.  
Random Forest also performs well especially in minimizing False Negatives showing it is reliable at detecting AML cases.  
Naive Bayes and the Decision Tree shows a higher numbers of False Negatives compared to SVM and Random Forest making them less reliable, where missing a positive diagnosis has serious consequences, and also Decision Trees have the highest number of False Positives, which could lead to unnecessary anxiety and medical procedures.  
The goal of my model is to detect who is AML positive and who is normal where failure in detecting a disease like AML could be life-threatening, a model like SVM or Random Forest where it minimizes False Negatives is better It is better to get a false positive than false negative, we want to limit both but specially false negative the disease is dangerous if left untreated.

A screenshot of a computer

Description automatically generated

Figure 8: Classification test scores

A screenshot of a computer

Description automatically generated

Figure 9: Confusion matrix for Naïve Bayes

A screenshot of a computer

Description automatically generated

Figure 10: Confusion matrix for Random Forest

A screenshot of a computer

Description automatically generated

Figure 11: Confusion matrix for Tree

A screenshot of a computer

Description automatically generated

Figure 12: Confusion matrix for SVM

The subset of fitting data is used to fit the model (70% of data) where the models learn the relationships between the features and the outcome, used for training to balance between having enough data for learning and enough data for evaluation.  
Final evaluation data (testing) 30% of data for evaluating the models after they have been trained to estimate how well the model is expected to perform when making predictions on data it hasn’t seen during training, and validate the model’s ability to generalize This data is never used during the training to ensure that the performance metrics calculated from this set are not biased by any previous exposure to the same data. This avoids Overfitting by using separate data for training and testing to detect overfitting, where a model might perform exceptionally well on training data but poorly on any new data.

I used lasso regression to reduce the number of features without significantly degrading the model performance it helped in feature selection by shrinking less important feature coefficients to zero.  
Lasso regression is extremely close to SVM, I found that it didn’t make a different where the test score is almost the same at c=1, it gives the same exact test and score numbers (figure 13), it means the model is not degrading when adding the logistic regression. When setting the c = 0.18 there are 11 features that are not zero coefficient, the SVM score is better (figure 14), it means that Lasso did not remove any important information despite reducing the number of features. When changing from cross validation to test on test data is doing better it gives a better number (figure 15), This indicates that despite the reduction of features, no critical information was lost, and the model remains robust across different subsets of data.

A screenshot of a computer

Description automatically generated  
Figure 14: Test and score after Lasso regularization C = 1 (cross validation)

A screenshot of a computer

Description automatically generated

Figure 14: Test and score after Lasso regularization C = 0.18 (cross validation)

A screenshot of a computer

Description automatically generated

Figure 14: Test and score after Lasso regularization C = 0.18 (test on test data)

CD16 ad CD 56 are proteins that are important features for distinguishing normal and AML cells where they are showing more than once, it means they significantly contribute to distinguishing between normal and AML cells.  
CD16 is a type I transmembrane receptor with two external immunoglobulin-like domains. It serves as a low-affinity receptor for IgG, enabling natural killer (NK) cells to carry out antibody-dependent cellular cytotoxicity (ADCC), it also can identify certain unknown tumor antigens. This receptor initiates signaling through its connection with the adaptor proteins CD3ζ and FcεRγ, both containing the immunoreceptor tyrosine-based activation motif (ITAM), it is a highly potent activating receptor on newly isolated human NK cells, triggering significant cytotoxic activity and cytokine release (1). There is a relationship between CD16 surface expression on NK cells and its prognostic significance in acute myeloid leukemia (AML), where the downregulation of CD16 on NK cells is associated with a favorable prognosis in AML. This suggests that CD16 expression levels might serve as a potential biomarker for predicting outcomes in AML patients and could be important in the development of targeted immunotherapies (2).  
CD56 plays an important role in the innate immune response by enabling natural killer cells to produce significant levels of key immunoregulatory cytokines in response to monokine stimulation. These cytokines are essential for the body's early defense mechanisms against infections. The specific monokines present affect how much CD56 helps natural killer cells produce important immune chemicals. This makes CD56 a critical factor in the immune system's ability to respond effectively to pathogens (3).  
In acute myeloid leukemia (AML), CD56 expression is associated with a negative impact on treatment outcomes. It correlates with lower rates of complete remission and shorter overall survival. This adverse effect is due to the co-expression of P-glycoprotein (PGP), which facilitates drug efflux, reducing the efficacy of chemotherapy. This shows CD56 as an independent negative prognostic factor in AML (4).

References:

1. Ahmed, Md Selim, Moizur Rahman, Md Abdul Matin, Muhammad Jahangir Hossen, and Mahmudul Hasan Sikder. "Role of macrophages in systemic inflammation: wound healing." In *Recent Advancements in Microbial Diversity*, pp. 335-360. Academic Press, 2022.
2. Wang, Huaiyu, Xiaojuan Zhang, David Wald, Sha Gong, Wenjuan Wang, and Jincheng Wang. "Downregulation of CD16 surface expression on NK cells is a favorable prognostic factor in acute myeloid leukemia." *Blood* 140, no. Supplement 1 (2022): 11834-11835.
3. Cooper MA, Fehniger TA, Turner SC, Chen KS, Ghaheri BA, Ghayur T, Carson WE, Caligiuri MA. Human natural killer cells: a unique innate immunoregulatory role for the CD56(bright) subset. Blood. 2001 May 15;97(10):3146-51. doi: 10.1182/blood.v97.10.3146. PMID: 11342442.
4. Raspadori, Donatella, Daniela Damiani, Mariagrazia Michieli, Raffaella Stocchi, Sara Gentili, Alessandro Gozzetti, Paola Masolini et al. "CD56 and PGP expression in acute myeloid leukemia: impact on clinical outcome." haematologica 87, no. 11 (2002): 1135-1140.