**PROJECT**

**Molecular Classification of Cancer by Gene Expression Monitoring**

**Dataset:**

Gene expression dataset (Golub et al.,1999).

There are three datasets containing the initial (training, 38 samples) and independent (test, 34 samples) datasets used in the paper. The labels of each patient are separately provided (actual,72 samples)

Intensity values have been re-scaled such that overall intensities for each chip are equivalent.

These datasets have been converted to a comma separated value files (CSV). Each row represents a different gene Columns 1 and 2 are descriptions about that gene. Each numbered column is a patient.

Each patient has 7129 gene expression values - i.e., each patient has one value for each gene. The training data contain gene expression values for patients 1 through 38. The test data contain gene expression values for patients 39 through 72

**Problem Statement:**

This dataset comes from a proof-of-concept study published in 1999 by Golub et al. These datasets contain measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood. It showed how new cases of cancer could be classified by gene expression monitoring (via DNA microarray) and thereby provided a general approach for identifying new cancer classes and assigning tumors to known classes.

The task is to classify patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

**Evaluation Metrics:**

The evaluation metrics employed for this project are accuracy score and confusion matrix

**Approach:**

1. Loading the separate datasets.
2. Data cleaning and pre-processing of train and test data.
3. Removing columns corresponding to the probes used using list comprehension.
4. Transposing the dataframe so that each row is a patient and each column corresponds to the expression level of a gene.
5. Cleaning up the column names for training data-Removing Gene Description
6. Translating index types of training and testing datasets from strings to respective numeric type values for sorting.
7. Removing the columns with endogenous controls, which are not necessary, using regular expressions.
8. Importing labels (for the training and testing datasets) from the label csv file and translating index types to numeric.
9. Concatenating the labels to the test and train dataframes.
10. Exploratory Data Analysis
11. Setting X and y variables and Dimensionality Reduction with PCA-7129 columns reduced to 35 Principal Components.
12. Train\_Test\_Split using X\_pca and cancer type
13. Model Building and Comparison- The following models were used:
14. K means Clustering
15. K Nearest Neighbours
16. Naïve-Bayes
17. Support Vector Classifier
18. XGBoost Classifier
19. Logistic Regressor
20. Ridge Classifier
21. Random Forest Classifier
22. GridSearch CV was used for hyperparameter tuning.
23. Accuracy scores and Confusion matrices were used to visualize the models

**RESULT:**

1. The best Accuracy score was from K Nearest Neighbours (100%) and the second was by SVM, Logistic Regression and Ridge Classifier (92.37%).
2. Models were generated and Prediction done with tuned SVM, Logistic Regression and Ridge Classifiers with test data.