Data mining techniques to classify cancer types based on Gene Expression Data

Project Description:

This project focuses on the application of Data Mining techniques to classify cancer types based on gene expression data, using insights derived from the foundational study "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" by Golub et al. The original study demonstrated how DNA microarray data could systematically differentiate between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), pioneering a shift from morphology-based cancer diagnosis to molecular profiling.

Building on this research, the team will work with real-world gene expression datasets to replicate and expand upon methods like clustering (class discovery) and supervised classification (class prediction). They will implement algorithms such as K-Means clustering, Logistic Regression, Support Vector Machines, Random Forests, XGBoost, and Neural Networks, while exploring the effects of dimensionality reduction (e.g., PCA) and model tuning. The work will include rigorous model evaluation using cross-validation, confusion matrices, ROC analysis, and feature importance studies. Advanced components like SHAP value interpretation and t-SNE visualization are also integrated to enhance model transparency and data understanding.

This project not only provides practical experience in bioinformatics and Data Mining, but also echoes the broader goals of personalized medicine—leveraging molecular data to guide diagnosis and treatment. The outcome will be a complete, well-documented classification pipeline along with an analytical report, demonstrating the feasibility and power of gene expression monitoring for cancer classification.

Read the paper here - https://drive.google.com/file/d/1Q2CiHQdFsntX8DX-LVpfg_kr3KRGt4WA/view?usp=drive_link (though the paper is old, it is a landmark paper in cancer prediction and it is a gold standard dataset even today. They did not have machine learning during those days to analyse the data. Thus, this paper provides good data that is suitable for today's machine learning):

Download the data here - https://drive.google.com/drive/folders/1ah-PGT90UCI6JH0vY5mh5pIOCwxgs4EJ?usp=sharing

In the data, target 0 denotes ALL (Acute Lymphoblastic Leukemia) and 1 denotes AML (Acute Myeloid Leukemia)

Task
Install environment + setup all libraries
Read and fully understand the dataset research paper
Load the dataset, check basic structure, missing values, class balance
Perform detailed EDA: histograms, correlation heatmaps, boxplots
Preprocessing: scale features, PCA (retain 80-90% variance)
Build Baseline Classifier (DummyClassifier, simple Logistic Regression)
Implement K-Means Clustering; visualize clusters before/after PCA
Evaluate Clustering: Silhouette score, Elbow method, Inertia plots
Implement t-SNE visualization of data (2D scatter)
Preprocess data variants: with/without PCA for future models
Implement Naive Bayes Classifier
Implement 5-fold Cross-validation for Naive Bayes
Implement Logistic Regression with hyperparameter tuning (regularization)
Implement 5-fold Cross-validation for Logistic Regression
Compare Naive Bayes vs Logistic Regression: Accuracy, F1, ROC curves
Implement Support Vector Machine (SVM) with linear kernel
Implement SVM with RBF kernel; tune C and gamma
Implement Random Forest Classifier
Tune Random Forest: n_estimators, max_depth, min_samples_split
Plot Feature Importance for Random Forest
Implement XGBoost Classifier + PCA preprocessing
Implement XGBoost + GridSearchCV for hyperparameter tuning
Implement XGBoost without PCA and compare models
Plot XGBoost Feature Importance; try early stopping

(Optional Advanced) Implement SHAP values for XGBoost interpretability
Build a simple Neural Network (MLP) with 2 hidden layers
Tune Neural Network: number of neurons, activations, dropout
Compare XGBoost vs Neural Network on test metrics
Prepare graphs: Confusion Matrices, ROC curves, Accuracy/F1 tables
Buffer + Write Final Project Report (2–3 pages)

Evaluation

Tasks	Key Goals	Skills Evaluated
Install environment, Read dataset paper, Load and explore data, EDA (plots), Preprocessing (scaling, PCA)	Dataset loaded and explored, EDA completed, PCA applied (retain 80-90% variance)	Python setup, Data exploration, Preprocessing techniques
Build baseline classifier, Implement K-Means clustering, Evaluate clusters, Visualize with t-SNE	Baseline model ready, Clustering implemented and evaluated, t-SNE visualization plotted	Model building, Clustering analysis, Visualization of high-dimensional data
Implement Naive Bayes and Logistic Regression, Tune hyperparameters, Cross-validation, ROC analysis	Models cross- validated, Hyperparameters tuned, ROC curves plotted	Supervised model evaluation, Cross-validation skills, Metric interpretation
Implement SVM (linear and RBF), Tune SVM, Implement Random Forest, Plot feature importance	SVM and RF models tuned, Feature importance visualized	Advanced model optimization, Ensemble methods, Feature analysis

Implement XGBoost (with/without PCA), Hyperparameter tuning (GridSearchCV), Optional SHAP values	XGBoost models compared (PCA vs no PCA), SHAP analysis attempted (optional)	Boosting expertise, Hyperparameter search, Explainable ML (optional)
Build Neural Network, Tune architecture (layers, neurons, dropout), Compare NN vs XGBoost, Final report writing	Neural Network tuned, Models compared, Final report prepared	Deep learning basics, Model comparison, Final reporting and documentation

****