1. Topic:

Using Machine Learning to Predict Disease Outcomes and WT1 SNP Status in CN-AML Patients Based on Gene Expression Data

2. The ONE method you plan to investigate;

We will focus on Support Vector Machines (SVM) as the primary method for binary classification. SVM is well-suited for high-dimensional data, such as gene expression profiles, due to its ability to handle non-linear relationships and provide robust classification boundaries.

3. Your plan to conduct simulations, analyze real data, or develop a theoretical derivation of a new method. If real data analysis is included, specify the dataset to be used.

Data Source:

https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-1425.

There are 408,273 genotype SNPs and identified expression quantitative trait loci from measurements of 54,675 transcripts representing 20,599 genes in Epstein-Barr virus￼ Transformed lymphoblastoid cell lines. We executed genome-wide association scans for these traits.

Analysis Plan:

Simulations:

Preprocess the E-MTAB-1425 dataset (normalize expression levels, impute missing values, and select relevant features).

Generate synthetic datasets to understand SVM's sensitivity to hyperparameters (kernel type, C, gamma).

Evaluate its performance on data with varying levels of noise, sparsity, and dimensionality.

Apply SVM to predict:

1. Presence of WT1 SNP (rs16754).

2. Disease characteristics.

3. Use cross-validation to evaluate classification performance such as accuracy, precision, recall, and ROC-AUC.

Model Interpretation:

Utilize feature importance such as coefficients in linear SVM or SHAP values to identify key genes driving predictions.

Outcome:

Provide insights into the utility of SVM for gene-expression-based predictions.

Highlight biomarkers associated with WT1 SNP and disease outcomes.

Assess SVM's scalability and robustness for high-dimensional datasets.