Machine Learning Course Project in Biological Sciences

Project Goal:

The goal of this project is to use the knowledge learned in the course to address a real bioinformatics problem using data from a study on Non-Alcoholic Fatty Liver Disease (NAFLD) through RNA-Seq transcriptomics.

NAFLD is a common liver disease in industrialized societies, characterized by fat accumulation in the liver without alcohol consumption. It can progress to inflammation and liver fibrosis, which can further develop into cirrhosis and liver cancer. Early diagnosis and treatment are essential as the disease is often reversible with timely intervention. Standard diagnostic methods, although useful, have limitations, and genetic markers could improve the understanding and diagnosis of NAFLD.

Data:

The data set includes RNA-Seq transcriptomics data from liver tissues of 192 individuals, including healthy individuals and patients with advanced fibrosis. The data is provided as run_accession_list.txt, which includes SRA run identifiers. These data need to be downloaded and processed using tools such as Hisat2, Stringtie, and limma-voom.

Step 1: Download and Process Data

The RNA-Seq data should be downloaded, processed, and normalized using advanced methods. The processed data should be aligned to the GRCH38 reference genome.

Step 2: Feature Selection and Dimensionality Reduction

Using various methods for feature selection and dimensionality reduction, identify important genetic features. Consider techniques like precision and recall in classification tasks to differentiate between normal, advanced fibrosis, and non-advanced fibrosis samples.

Step 3: Classifier Design

Design classifiers to distinguish between different classes (normal, advanced fibrosis, and non-advanced fibrosis) using selected features. Evaluate the minimum number of features required for effective classification and the parameters that improve classifier performance.

Step 4: Effect of Covariates

Investigate the impact of covariates such as Body Mass Index (BMI), gender, and age on the classification performance. Assess whether including these covariates improves the accuracy of the machine learning models.

Step 5: Age Prediction

Design a regression model to predict age using genetic features from the meta_data.csv file. Use methods like LASSO regression to identify the minimum number of features required for accurate age prediction.

Important Points:

- Projects can be done individually or in pairs.

- A written report including the methodology, answers to the questions, and relevant code must be submitted by the oral presentation date.

- Presentations will be held virtually via Google Meet between August 20 and 23.

- Each group must designate a representative to email the names, surnames, and student numbers of group members to kkavousi@yahoo.com.