Machine Learning Course Project in Biological Sciences

Objective: The project aims to utilize the learned data from the course in a real informatics problem.

Project Description:

The goal is to examine the genetic factors related to Non-Alcoholic Fatty Liver Disease (NAFLD) using RNA-Seq data. NAFLD is a prevalent liver disease in industrialized societies and can lead to advanced liver conditions such as cirrhosis if not diagnosed early. Early diagnosis is crucial for effective treatment, and while standard methods are used for diagnosis, using gene markers and panels designed for a better understanding of the liver conditions can be beneficial.

Data:

The study involves transcriptomic data from 192 liver tissue samples, including healthy individuals, those with non-advanced fibrosis, and those with advanced fibrosis. The first step involves downloading the data using SRA tools, followed by processing the data using tools like Hisat2, Stringtie, and limma-voom.

Project Tasks:

1. Classification and Feature Selection:

- Design classifiers to categorize samples into Normal, Non-Advanced, and Advanced Fibrosis categories.
- Use methods for dimension reduction and feature selection.
- Evaluate the models using Precision and Recall metrics.

2. Regression Analysis:

- Analyze the impact of variables such as gender, Body Mass Index (BMI), etc., on the classification.
- Separate models for males and females if it improves accuracy.

3. Advanced Tasks:

- Use Lasso regression to predict age and compare performance by separating samples into male and female groups.

Requirements:

- Students can work in groups of one or two.
- Each group must submit a written report and a presentation.
- The final presentation will be conducted virtually via Google Meet.

Submission:

- Email the report and code to kkavousi@yahoo.com by August 20-23.
- Include group members' names and student IDs in the email.