Reichman University

Efi Arazi School of Computer Science
Machine Learning and Data Science M.Sc. Program

Final Project Proposal

Title: **Feature selection in Gene Expression (high dimensional data)**
Date: Mar 2023

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of the model. The data features that are used to train machine learning models may dramatically influence the performance of the model. For example, irrelevant or partially relevant features can negatively impact model performance.
Feature Selection is the process where we automatically or manually select those features which contribute most to the predicted variable or to the output of interest.

Genes encode proteins and proteins dictate cell function. Therefore, the thousands of genes expressed in a particular cell determine the function of that cell. Moreover, each step in the flow of information from DNA to RNA to protein provides the cell with a potential control point for self-regulating its functions by adjusting the amount and type of proteins it manufactures.
Gene expression datasets typically span several thousands of genes, which is very high dimension from a machine learning perspective.

Feature selection in high-dimensional data is a challenging task for several reasons. First, from the computational perspective – some ML algorithms and feature selection techniques are impractical in this environment. Second, even if you can run the process, taking into consideration all the relations between the features, is not an easy task. For example, removing all correlated features may be impractical; therefore, the correlations can influence the results.

In this project we will aim to take a gene expression data as input. The partitioning of the data will be driven by existing clinical classification. You will implement several feature selection techniques to select the best genes for predicting the clinical class.
We provide the first dataset – Breast cancer gene expression.
The goal will be a complete ML pipeline – but the focus will be on the feature selection part.

This work will have several tasks:

1. Implement feature selection pipeline for high dimensional data.
2. Compare different feature selection technique results.
3. Build infrastructure that will expedite the process (parallelized feature selection, etc...)
4. Test your pipeline on a different dataset (not necessarily gene expression) – you need to find one and run your process on it.

The end result of the project will be software (or a working pipeline) that can be run on an input dataset. The students will also write a final report to describe the project results, which will be assessed for its academic quality.