Identification of Signature Proteins using various Machine Learning Tools

We have two cohorts of samples, diseased and healthy, each cohort having 6 samples. Each sample has approximately 3000 features or proteins. We have to train our ML model such that it can predict diseased vs healthy samples/individuals.

Libraries Used:
We will be using sci-kit learn, NumPy, Pandas, and Matplotlib libraries for our project.

Plan:-
BIG DATA ANALYSIS PIPELINE:

Data generated is biological in nature and such large dataset needs an unbiased analysis for the identification of signature protein which can eventually classify the two cohorts irrespective of the biological technique used for data generation.Being biological in nature, there can be two reasons for the incorporation of missing values:-instrumental error or biological reason and during our analysis, we need to closely take into consideration both the aspects.
We would initially analyse the data in order to understand its distribution followed by correlation analysis among the samples within the cohorts for identifying the outlier samples. After preliminary data analysis of the two cohorts separately, we would take forward the dataset for missing value imputation as per our preliminary analysis followed by transformation, normalisation, scaling or standardisation as per the need of the data.Further we will be also using visualisation of the data at every step. We would the use various statistical tools such as f-test, t-test, fold chain analysis for identifying the significant proteins. After that we will be using PCA(principal component analysis) for dimensionality reduction followed by application of different classifiers and clustering tools for predicting the signature proteins and comparing the accuracies of different models developed.

Timeline:
1-2 days:- Preliminary Data Analysis, Missing Value Imputation
3-4 days :- Using different visualisation techniques for visualizing the data
1-2 days :- Various Statistical Analysis
1-2 days:- To Implement PCA on the data
4-5 days:- To implement different clustering and classifiers on the data

Other things:-
We will learn the basics of ML and how to implement ML techniques to real world problems.

Motivation:-
To understand the basic concepts of Machine Learning and various visualisation techniques.

Reference:
Machine Learning In Action by Peter Harrington