Project 2: Heart Disease Prediction

Project Description

In this project, you will work on a Kaggle dataset that describes past records of patients having heart disease. Various attributes are included in the dataset; as such, it is a good candidate to perform all tasks related to data visualization and exploration, statistical analysis, as well as prediction based on machine learning. Please refer to the "Aboud Data" section of this dataset for further information, especially regarding the meaning of those attributes, which will be used as input for your objective in this project.

This dataset can be reached from here. Please refer to the original link for further information regarding this dataset.

Students are also expected to extract features from additional datasets and combine them with the original one to perform data aggregation. Some extra datasets can be found here:

https://www.kaggle.com/fedesoriano/heart-failure-prediction

https://www.kaggle.com/itachi9604/disease-symptom-description-dataset

https://www.kaggle.com/johnsmith88/heart-disease-dataset

https://www.kaggle.com/mazharkarimi/heart-disease-and-stroke-prevention

Expected task can be summarized as follows:

Exploratory Data Analysis

- Visualize each attribute separately with respect to AHD (0: no disease, 1: disease)
- Visualize the relationship between some selected attributes and the final disease state (AHD: disease, no disease). For instance, you may choose [age, sex, Chol] as one attribute group and visualize against whether AHD is 0 or 1 (no disease/disease). Try to do this for several different attribute groups
- Analyze Chol, Fbs and ExAng values with respect to age and sex
- Analyze & visualize how each attribute changes with respect to age and sex
- Analyze & visualize the relationship between each attribute (without AHD) and chest pain
- Analyze & visualize the relationship between two selected attributes, as well as whether they would lead to disease or not. That is, divide all the attributes into groups of two, and for each group, analyze whether that combination leads to disease or not. Consider all possible combinations.
- Introduction of at least four new features from additional datasets and visual explanations of them
- For both existing and new features, and usage of spatial visualization examples (as much as possible) are expected.

Statistical Analysis & Hypothesis Testing

- Statistical tests to check whether the values of chest pain, RestBP and Chol contribute to heart disease or not.
- Statistical tests to check whether age and sex contribute to heart disease or not

- Statistical tests to check if significant differences exist between age groups and sex that suffer from heart disease
- Statistical tests to check how all those attributes contribute to heart disease
- Utilizing at least four new features using extra datasets in hypothesis testing

Machine Learning

- Prediction of heart disease given values of age, sex, chest pain and RestBP
- Perform hyper-parameter tuning to increase model performance
- Students are expected to introduce new approaches to their selected model in A to investigate their effects on model performance
- Students are expected to utilize at least two different machine learning models for prediction and compare their performance
- Build a system that can predict whether a person could suffer from heart disease when provided with values of those attributes.
- Students should clearly present their reason for selecting that particular method
- Students are also expected to handle with scenarios of missing attributes. That is, when predicting, the provided personal data may have some missing attributes compared to the original training data
- Utilize at least four new features using additional datasets in machine learning models