Final Project Plan (COMP 4448)

Complete this document as part of your project plan about your final project. Your answers need to be short and brief. About one to three sentences for each question is fine. Write your answers under each question below. Make sure to maintain the question numbering.

Maximum Grade Points = 50 points

1. What is purpose of your project? (This should include the algorithms you are comparing, the output variable and the input variables captured in a general way).

The purpose of this project is to create and compare different classification algorithms that can predict if a person has heart disease. The algorithms will be SGD, Random Forest, KNN, Linear SVC and SVC.

1. What is the significance of your project? That is, why is this project important? What potential positive impact does your project have?

Heart disease is a leading cause of death in the US and often goes unnoticed until a person suffers a heart attack. If it is possible to predict if a person has heart disease based on lifestyle factors, then medical intervention is possible before a heart attack potentially saving thousands of lives annually and millions in medical costs.

1. What is the research question? (This should include the algorithms you are comparing, the output variable and the input variables captured in a general way. A research question is the purpose of your project flipped into a question).  
   Can classification algorithms predict heart disease?  
   What classification algorithm is best for predicting if a person has heart disease?
2. What algorithms will you be comparing to answer your research question?

The 5 algorithms will be SGD, Random Forest, KNN, Linear SVC and general SVC.

1. What is your output variable?

The column name is HeartDiseaseorAttack. That is, a person is diagnosed with heart disease or had a heart attack.

1. What are the input variables for your analysis? You should list at least five input variables in the dataset to be used for analysis/modeling.

The column names are: HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, Diabetes, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income

2 bad models in a row and I reduced the variables to HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, Diabetes, and Age

1. What kinds of data preprocessing are you anticipating to perform on the dataset before modeling?

The dataset is clean already so scaling mainly. After EDA I will probably drop some columns.

1. What exploratory data analyses (on which variables) do you plan to do before modeling?

I will run a Pandas Profiling report to see what anomalies exist in the data that need to be handled.

1. What visualizations (on which variables) do you plan to do before modeling?

The Pandas Profiling report supplies comprehensive statistical visualizations for all variables.

1. What is the link to your data source? If your dataset is not from an internet data source, you will need to upload it to GitHub and share the link here. Avoid using common internet datasets such as Iris dataset, mushroom dataset, etc. Your dataset should be something interesting to you and uncommon). Feel free to use a dataset from this link: <https://vincentarelbundock.github.io/Rdatasets/datasets.html> or from some other source such as Kaggle competition but don’t present any copy and paste internet solutions for your final project. You will use the tools you have learned from this course to do your own project independently. Please, provide the link to your data source here:

Data:

<https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset>

Notebook explaining the cleaning process:

<https://www.kaggle.com/code/alexteboul/heart-disease-health-indicators-dataset-notebook/notebook>