Predicting the Presence of Cardiovascular Disease in Adult Patients Using a Random Forest

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
    - According to the World Health Organization: “Cardiovascular diseases (CVDs) are the leading cause of death globally…..It is important to detect cardiovascular disease as early as possible so that management with counseling and medicines can begin” (*Cardiovascular Diseases (CVDs)*, 2021). Therefore, A model that can quickly predict the possible presence of cardiovascular disease would be an invaluable tool for medical professionals and patients in the early detection and diagnosis of CVD.
- What industry/realm/domain does this apply to?
    - Life Science/Healthcare
- What is the motivation behind your project?
    - My degree is in biochemistry and I have a background related to healthcare. I am passionate about science and healthcare and as mentioned earlier, cardiovascular diseases are a leading cause of death worldwide.

Data Understanding

- What data will you collect?
    - I will need health data that contains some basic information such as age and gender; but more importantly, data that contains features associated with the risk factors for certain cardiovascular diseases such as weight, blood pressure, cholesterol, smoker vs non-smoker, etc.
- Is there a plan for how to get the data (API request, direct download, etc.)?
    - I obtained data from The National Health and Nutrition Examination Survey (NHANES). The data is slightly difficult to work with and would require an extensive effort in cleaning and preparing it for modeling. Because of the foreseen difficulties that might occur, I decided it would be a good Idea to have a backup source of data and downloaded cardiovascular disease information from kaggle.
- Are the features that will be used described clearly?
    - Yes.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - I will need to convert the data into a usable categorical format.

- What are some of the cleaning/pre-processing challenges for this data?

    - If the NHANES data is used, It has a very large amount of data that will not be relevant to the project at hand. This will require a laborious effort to find and extract the necessary features, as well as joining the various tables. There are also a large number of null values.I will have to determine the best course of action in taking care of the null values.On top of that, all of the data fields are encoded requiring me to decode each field and record.

Modeling

- What modeling techniques are most appropriate for your problem?
    - Most of the standard modeling techniques will be appropriate such as train/test split, dummy variables for categorical data, and standardization of the data.
- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
    - The presence of Cardiovascular Disease.
- Is this a regression or classification problem?
    - Classification

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?
    - I will evaluate it and compare based on various metrics used for classification problems.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
    - I will probably start with a simple logistic regression but compare its performance to other algorithms. I am planning on ultimately using a random forest depending on how well it performs compared to the others.

References

Cardiovascular diseases (CVDs). (2021, June 11). World Health Organization (WHO). Retrieved December 8, 2022, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)