Heart Disease Prediction

Project Proposal

In this project, we will use the Kaggle dataset (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset) and Machine Learning to make predictions about whether a person is suffering from Heart Disease or not. To decide the best machine learning model, we will first execute the dataset to identify the model with high accuracy and F1 Score between Logistic Regression, Random Forest Classifier, Support Vector Machine, K Nearest Neighbours, and XG Boost. Then, from there we will create an ML model to make accurate predictions on whether a person is suffering from Heart Disease or not.

The dataset has 76 attributes, but all published experiments will refer to using a subset of 14 of them, as shown below:

- age: age of the patient
- sex: sex of the patient
- cp: chest pain type (4 values)
- trestbps: resting blood pressure
- chol: serum cholesterol in mg/dl
- fbs: fasting blood sugar > 120 mg/dl
- restecg: resting electrocardiographic results (values 0,1,2)
- thalach: maximum heart rate achieved
- exang: exercise-induced angina
- oldpeak: ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment
- ca: number of major vessels (0–3) colored by fluoroscopy

- thal: 0 = normal; 1 = fixed defect; 2 = reversible defect
- target: the presence of heart disease in the patient

Project Questions

To ensure that the final model is very accurate, this project will have several questions to address:

- What factors influence the likelihood of heart disease in patients according to the dataset?
- What is the distribution of cholesterol levels among patients with heart disease, and how does it compare to those without heart disease? Please visualize this using a histogram or a density plot.
- Is there a relationship between the number of chest pain types experienced by patients and the likelihood of heart disease? Plot a bar chart to illustrate.
- Can you use a heatmap to depict the correlation between different variables in the dataset, such as age, cholesterol levels, and maximum heart rate achieved?
- How do the various machine learning models, including Logistic Regression, Random Forest Classifier, Support Vector Machine, K Nearest Neighbours, and XG Boost, compare in their ability to predict heart disease based on the provided dataset?
- Could you develop a predictive model to predict the likelihood of heart disease in patients based on their medical attributes and demographic information, using the best-performing machine learning model identified in the analysis?

Proposed Project execution process

In this proposed project, we aim to conduct exploratory data analysis (EDA), model development, and model validation to predict heart disease based on a comprehensive dataset. The initial phase of EDA involves data collection, preprocessing, and visualization to understand the dataset's characteristics and relationships between variables. We will use

Python libraries such as Matplotlib and Seaborn to create visualizations and descriptive statistics to understand the data better.

More so, we will proceed with model development, where we will perform feature engineering, model selection, training, hyperparameter tuning, and evaluation using various machine learning algorithms suitable for classification tasks. These algorithms include Logistic Regression, Random Forest Classifier, Support Vector Machine, K Nearest Neighbours, and XG Boost. We will select the best-performing model based on evaluation metrics like accuracy, precision, recall, and F1-score.

Also, we will validate the selected model using cross-validation techniques to ensure its generalizability and interpret its predictions and feature importance. Visualization of the model's performance metrics will be provided to facilitate a clear understanding and documentation of the entire process for reproducibility and future reference. Through this project, we aim to develop a reliable predictive model for heart disease, which could potentially aid in early diagnosis and intervention.