ECE 539 Proposal: Analysis and Prediction of Heart Disease

Xiaofei Qu

xqu29@wisc.edu

Jack Yuan

yuan88@wisc.edu

1. Overview

This project is to build a model to analyze the data about heart attack and make predictions based on the given data from Kaggle [1].

2. Background

Heart disease is one of the main causes of deaths each year in the United States [2]. Heart attack is one of the most common types of heart disease. There are around 805,000 people in the US who have a heart attack. In addition, there are one fifth heart attacks that are silent, which means that the person may be not even aware of it when damages to the heart have already occurred [2]. Therefore, it is important to be able to predict the occurrence of a heart attack and try to avoid it as early as possible.

There are currently different ways to predict heart disease, including a heart disease prediction system built based on data mining techniques, and a prediction system that applies data processing techniques. For the first effective heart disease prediction system (EHDPS), multilayer perceptron neural network (MLPNN) and backpropagation (BP) were applied to the training data set with a partition of 60% testing data and 40% training data[3]. Whereas the second system first checks the skewness, outliers, and distribution of the data, then applies the Lasso algorithm to the processed data [4].

3. Statement of Work

3.1. Datasets

The data the team is going to use comes from Kaggle. There is a total of 14 columns including age, sex, chest pain type, resting blood pressure, cholestoral level, fasting blood sugar, resting electrocardiographic result, maximum heart rate, exercise induced angina, previous peak, slope, number of major vessels, thal rate, and target output (1 if this person has heart attack and 0 if this person does not have heart attack). The main features that may be used are the age, sex, resting blood pressure, resting electrocardiographic result, and maximum heart rate because these features are easier to be obtained for a person. All the 14 columns of information are provided for each of the 303 people. The data for 303 people may not be rough for us to build an accurate prediction model, however, that's all the data Kaggle provides..

Based on the data given, the preliminary inspection is that there may be outliers that we need either not consider or include in the analysis. However, the data gives enough information that we can use to build the prediction model. From what others commented, the data is reliable and able to give accurate predictions even though the data size is small.

3.2. Method

We are planning on applying a logistic regression algorithm to the data to get a probability of getting a heart attack, and then making a decision tree or performing k-nearest neighbors (k-NN) algorithm or support vector machines (SVM) based on the different categories in the dataset. In addition, a confusion matrix will be made and a receiver operating characteristic (ROC) curve will be graphed to evaluate the accuracy of the predictions. Moreover, since only performing a single trail may not provide an accurate prediction, several trials may be needed with a partition of the data (70% testing and 30% training). Then, the average and standard deviation may be used to further analyze the data.

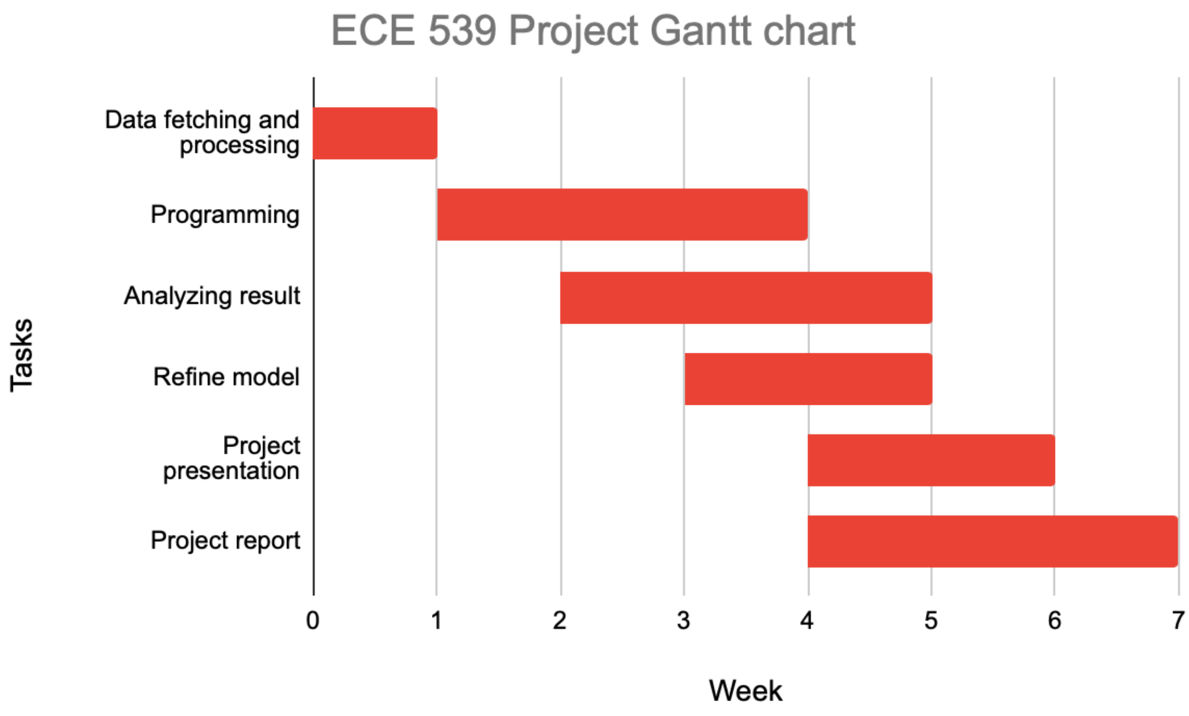
As mentioned in the background section, there are different methods to perform the heart disease prediction. For the EHDPS that uses data mining technique and applies a multi-layer perceptron network, it was proved that the accuracy of predicting heart disease is around 100% [3]. For the prediction system that applies data processing techniques, it is shown that KNeighbors classifier in machine learning has an accuracy of 84.8% [4].

Although there are different methods performed on heart attack prediction, the data that was used is either too old (in 1988) or unclear about the credibility. The data we are planning on using is updated in 2021, and it may provide an even more accurate prediction model on people nowadays since the environment people live in may change a lot compared to 1988's.

3.3. Outcome and Performance evaluation

As heart disease is a serious disease, it is crucial to not let any positive cases be classified as negative ones. While the classification rate of a classifier is an important factor, our main goal of this project is to maximize the true positive rate in the confusion matrix while maintaining a high classification rate (≥ 95%). In addition, we also want to make sure the ROC curve is as close to the top-left corner as possible. With such a model, doctors can safely determine whether a patient is in need of medical treatments.

4. Project Plan



GitHub repository link: https://github.com/yongleyuan/ece539-project

5. References

[1] R. Rahman, "Heart attack analysis & prediction dataset," *Kaggle*, 22-Mar-2021. [Online]. Available: https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset. [Accessed: 09-Mar-2022].

[2] "Heart disease facts," *Centers for Disease Control and Prevention*, 07-Feb-2022. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm. [Accessed: 09-Mar-2022].

[3] P. Singh, S. Singh, and G. S. Pandi-Jain, "Effective heart disease prediction system using data mining techniques," *International journal of nanomedicine*, 15-Mar-2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/. [Accessed: 09-Mar-2022].

[4] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, “Prediction of heart disease using a combination of machine learning and Deep Learning,” *Computational Intelligence and Neuroscience*, 01-Jul-2021. [Online]. Available: https://www.hindawi.com/journals/cin/2021/8387680/#methodology. [Accessed: 09-Mar-2022].