Heart Disease Prediction

DATA1030 Midterm Report – Yash Bafna

Supervised by Prof. Andras Zsom

Section 1. INTRODUCTION

Heart diseases have become extremely common these days due to changing lifestyle. High blood pressure, diabetes, hyper tension and cholesterol are some of the major factors that lead to heart attacks. With the advancement of technology, we can now predict using some important factors how likely a person is to get a heart disease such as coronary-artery disease, cardio-vascular, stroke, heart failure and much more, thereby helping the person to prevent it and save their life. Hence, in my opinion, this is a very important and interesting topic to work upon.

The dataset I will be working on is from UCI's Machine Learning repository. This dataset consists of thirteen columns representing different parameters or features which will be used to predict if the person is likely to have a heart related disease or not and a target column titled 'target' which informs us whether the person has a heart disease or not. The target variable only holds values zero and one, where zero indicates no presence of heart disease and one indicates that the person has a heart disease. This dataset has three hundred and three datapoints. It is a classification-based problem where we can classify a given person into one of the two options, either has a disease or doesn't have one. The fourteen columns can be classified into two types, continuous columns and categorical columns.

Following are categorical columns and the values they can take:
sex - sex (1 = male; 0 = female)
cp - chest pain; 1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
fbs - fasting blood sugar > 120 mg/dl: 1 = true; 0 = false
restecg - resting electrocardiographic results; 0: normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
exang - exercise induced angina; 1 = yes; 0 = no
slope - the slope of the peak exercise ST segment; 1: upsloping 2: flat 3: downsloping
ca - number of major vessels (0-3) colored by flourosopy
thal - 1 = normal; 2 = fixed defect; 3 = reversable defect

Following are categorical columns and the values they can take:
age - age in years
trestbps - resting blood pressure (in mm Hg on admission to the hospital)
chol - serum cholestoral in mg/dl
thalach - maximum heart rate achieved

oldpeak - ST depression induced by exercise relative to rest

Literature Survey:

1. Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization - Muhammad Saqib Nawaz, Bilal Shoaib, and Muhammad Adeel Ashraf [1] https://doi.org/10.1016/j.heliyon.2021.e06948

This paper talks about the simulation results of Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization, NB, SVM, K-NN, RF, and ANN classifiers and concluded that Intelligent CVD Prediction Empowered with GDO achieved maximum accuracy (98.54%) which they could achieve through optimization algorithms.

2. Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction - Emrana Kabir Hashi and Md. Shahid Uz Zaman [2] https://doi.org/10.33736/jaspe.2639.2020

This paper talks about the use of data mining algorithms like NB, DT, and Random-Forest on the same dataset. These performance techniques showed that the Random Forest Algorithm produced the highest accuracy for heart disease prediction, which is 90.16%

Section 2. EXPLORATORY DATA ANALYSIS

This section contains some of the graphs that were created during exploratory data analysis.

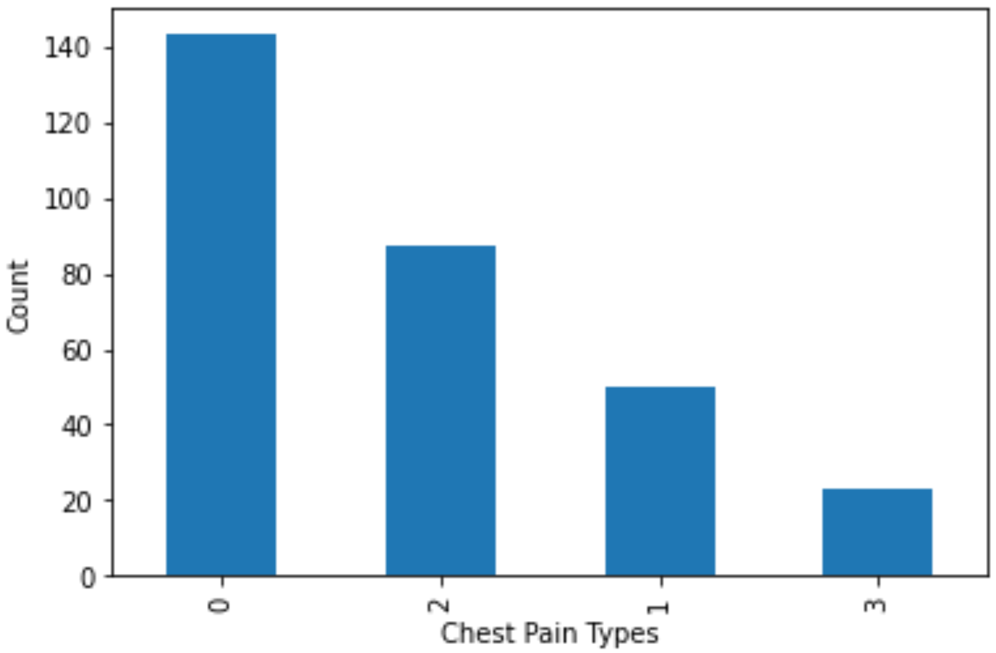

Figure 1: The above bar graph shows the total count for each of the four types of chest pains. As mentioned earlier, there are four types of chest pains: 0: no pain 1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic. As we can see, most people do not have any chest pain.

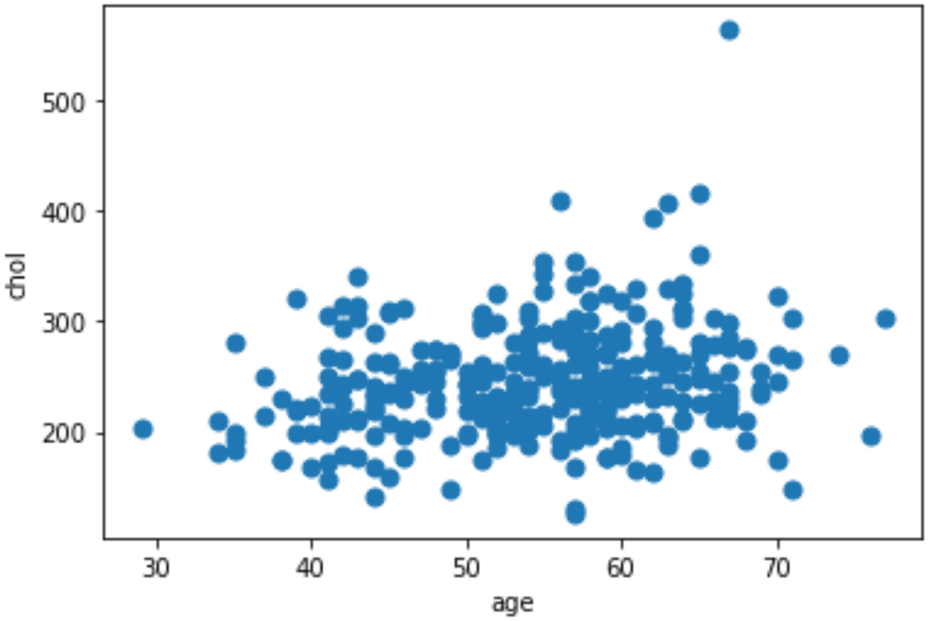

Figure 2: This scatter plot shows the relation between the age and cholesterol levels. Both the columns are continuous in this case. If we were to draw conclusions from this scatterplot, we could conclude that with increasing age, there is an overall slight increase in cholesterol levels.

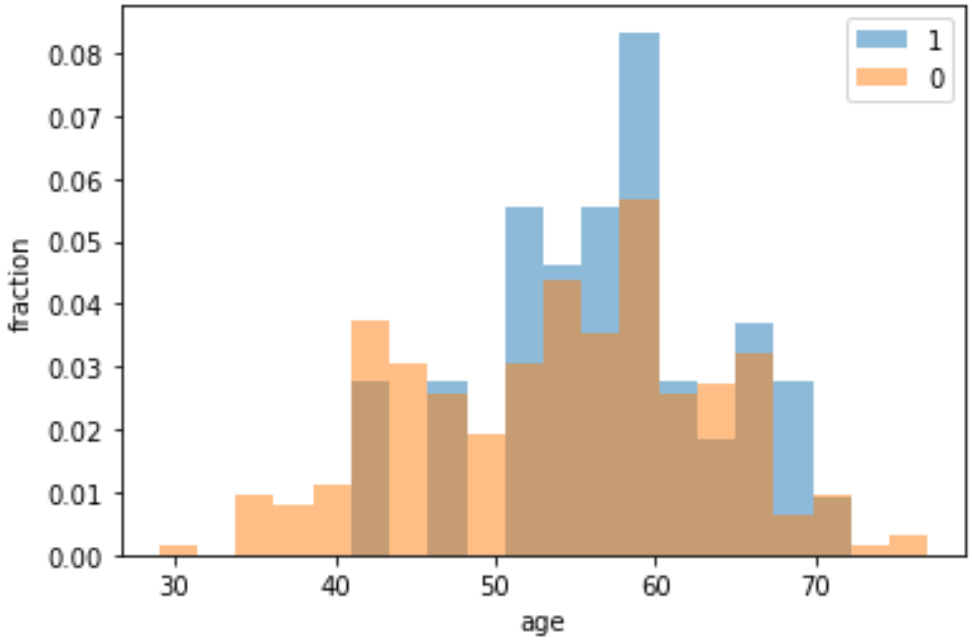

Figure 3: This is category specific histogram graph which plots the columns age, which is continuous and fasting blood sugar, which is categorical. Orange ones indicate people with fasting blood sugar greater than 120 mg/dl and blue ones indicate people with fasting blood sugar less than 120 mg/dl.

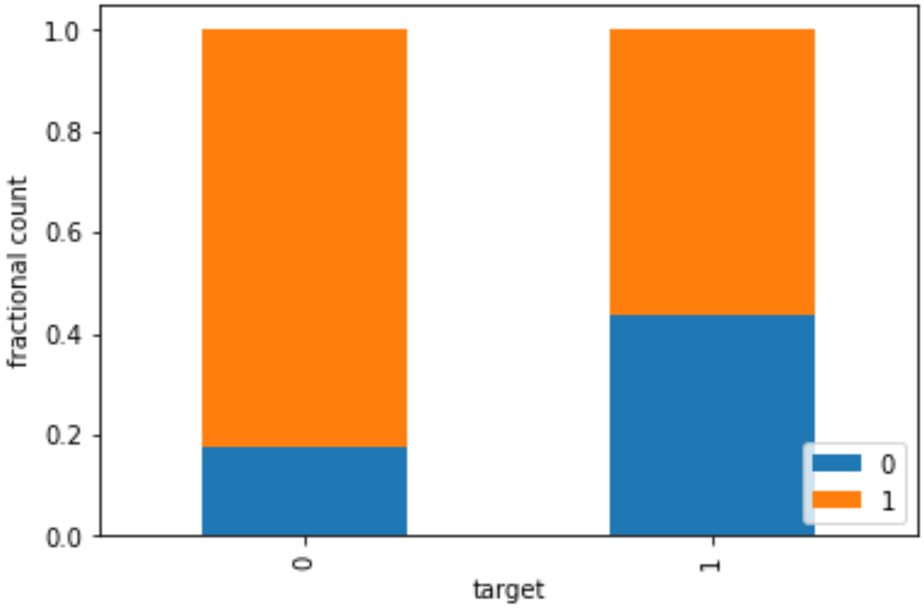

Figure 4: This stacked bar graph plots two categorical columns which are sex and target columns. 1 indicates males and 0 indicates females for sex column. For target column, which is plotted across the x axis, 0 indicates that the person does not have disease whereas 1 indicates that the person has disease.

Section 3. METHODS

Section 3.1. DATA SPLITTING & PREPROCESSING

The dataset for this project is independent and identically distributed (IID) dataset. We used the train_test_split function to split the data into training, testing and validation datasets. 60% of the dataset was assigned to training stage, and 20% for testing and 20% for validation stage. Since the dataset is not very large, I decided to go for this splitting configuration as this way, there is enough data to train, test and validate. This dataset does not follow group structure and is not a time-series data. I applied OneHotEncoder on categorical features since I am working with unordered categorical data and MinMaxEncoder on continuous features since continuous feature values for all the columns are reasonably bounded. List of categorical and continuous columns are given in Section 1. The number of features after preprocessing is 13.

Section 4. REFERNCES

1. Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization - Muhammad Saqib Nawaz, Bilal Shoaib, and Muhammad Adeel Ashraf
https://doi.org/10.1016/j.heliyon.2021.e06948 [1]
2. Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction - Emrana Kabir Hashi and Md. Shahid Uz Zaman
https://doi.org/10.33736/jaspe.2639.2020 [2]
3. Heart Disease Detection Using Machine Learning - Chithambaram T, Logesh Kannan N and Gowsalya M
https://doi.org/10.21203/rs.3.rs-97004/v1