Heart Disease Prediction

Johanna Kasenurm, Elisabet Hein

Github Link: https://github.com/Elisabethein/DataScienceProject

Task 2

Background:

According to the World Health Organization (WHO) heart diseases, also known as cardiovascular diseases (CVDs) are causing the most deaths around the world. CVDs include different heart and blood vessel-related diseases, such as coronary heart disease (CHD), cerebrovascular disease, and rheumatic heart disease (RHD). It is argued that CVDs are among the deadliest diseases in both developed and developing countries. According to WHO CVDs take an estimated 17.9 million lives each year which is about 32% of all deaths globally. However, the chance of survival is higher if the diagnosis is made early enough. Therefore machine learning has become one of the most important tools in solving this problem by helping to predict which people are at risk of heart diseases.

Business goals

Our primary objectives involve exploring diverse machine learning techniques to identify the most accurate predictive model for assessing the risk of heart disease. Another goal is to look deeper into the provided features and see which factors influence or boost the development of the diseases the most. Successfully achieving these goals would empower doctors to anticipate their patients' health risks at an early stage.

Business success criteria

The project will be deemed successful if the constructed model accurately predicts heart disease risk on test data with a precision of 92%. This achievement would validate the model's applicability to other datasets..

Inventory of resources

We have a dataset from the Kaggle platform called Heart Disease Dataset (Comprehensive) which can be found here. (https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final/data) We have two people working on this project and have access to the Deepnote environment, which allows us to work simultaneously on our project.

Requirements, assumptions, and constraints

The project must be finished by 11.12.2023 and must be available in our GitHub repository. We also need to present a poster of our project.

Risks and contingencies

The main risk could be not finding the right model for a problem like ours. It is very important to find the most optimal technique that works well on other datasets too. According to WHO, selecting a suitable algorithm for a specific dataset is a big challenge in bioinformatics. Consequently, selecting good feature selection or classification algorithms is also a big challenge in this field.
As a solution, we will try out the techniques and models we know and select the best one of them. We will also investigate the feature importance of each model.

Terminology

We intend to use the following terms in our project and dataset:

Chest Pain Type - The type of chest pain is categorized into 1 typical, 2 typical angina, 3 non-anginal pain, and 4 asymptomatic.
Typical chest pain is usually felt in the centre of the chest, behind the breastbone.
Typical Angina is generally characterized by a feeling of pressure or squeezing in the chest. It may also be accompanied by pain radiating to the arms, neck, jaw, shoulder, or back.
Non-anginal pain refers to chest discomfort that doesn't originate from the heart. It may be musculoskeletal (connective tissues in the human body) or related to other organs in the chest.
Asymptomatic pain means the absence of noticeable symptoms or chest pain related to heart issues. People may not experience chest pain or discomfort despite potentially having risk factors for heart disease
Cholesterol is a fatty substance, and elevated levels can contribute to the buildup of plaques in arteries, increasing the risk of heart disease.
Exercise Angina is the pain experienced in the chest during exercise.
The ST segment is a portion of the electrocardiogram (ECG or EKG) that represents the time between ventricular depolarization (when the heart's lower chambers contract) and repolarization (when the chambers relax).
The ST slope refers to the direction or inclination of the ST segment concerning the baseline of the ECG (electrocardio diagram).
Oldpeak refers to the amount of depression (or downward displacement) of the ST segment on the ECG during or after exercise compared to the baseline ST segment at rest.
Serum is an amber-coloured, protein-rich liquid that separates out when blood coagulates.

Costs and benefits

The approximate time of this project is 60 hours. If the project is successful, there could be a new way to prevent heart disease around the world.

Data-mining goals

Our goal is to build different models like RandomForest, DesicionTree, KNN and NN using our data. We also want to create a repository and a poster for our project that concludes our work.

Data-mining success criteria

We consider our project to be successful if the model we have built predicts the heart disease risk on our test data 92% correctly.

Task 3

Outline data requirements

Our data must be in the form of a CSV file.

Verify data availability

We have found the necessary data from the Kaggle platform and it is available for us to use. The dataset we are planning to use can be found here.
(https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final/data)

Define selection criteria

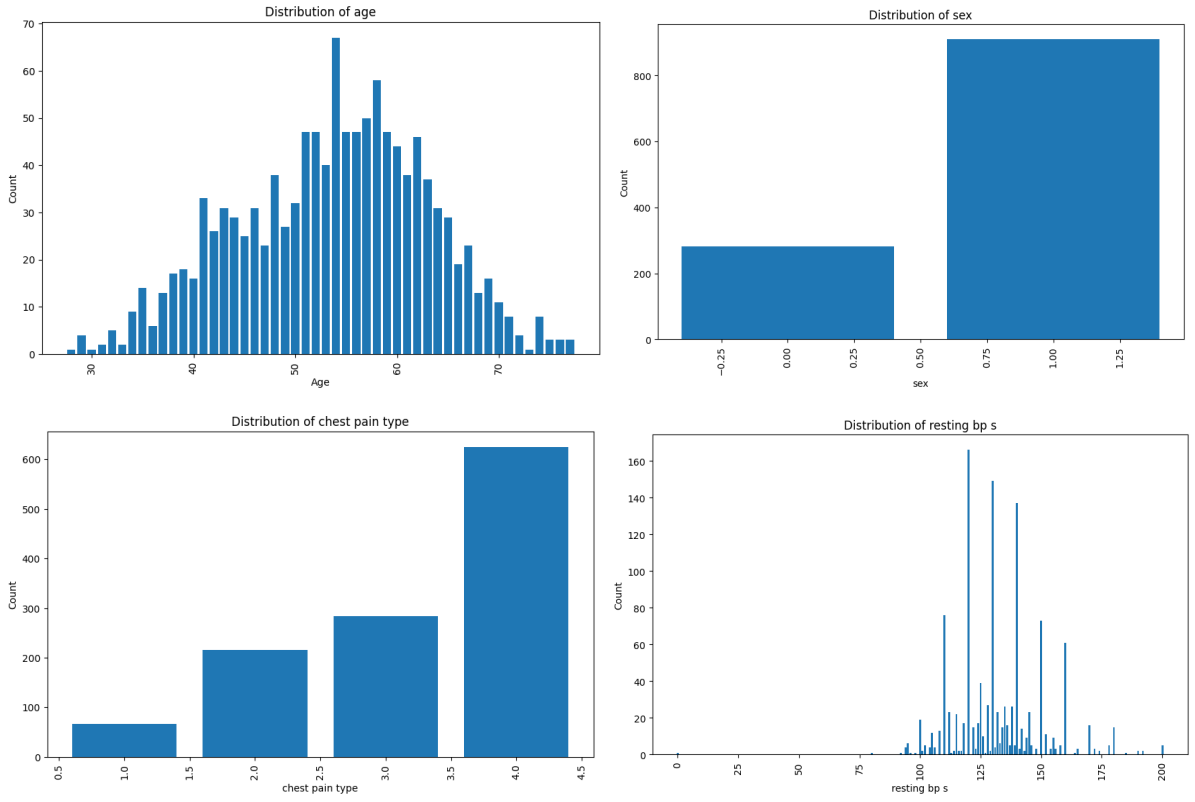
Link to our datafile: DataScienceProject/heart_statlog_cleveland_hungary_final.csv at main · Elisabethein/DataScienceProject (github.com)
Our initial plan is to use all the attributes within this file and perform an evaluation of feature importance to find which attributes affect our predictions.

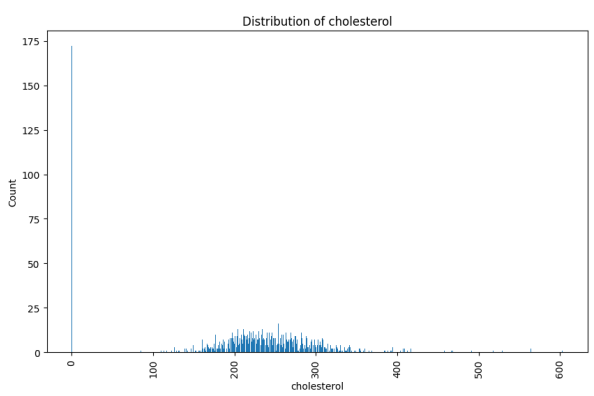
Describing data

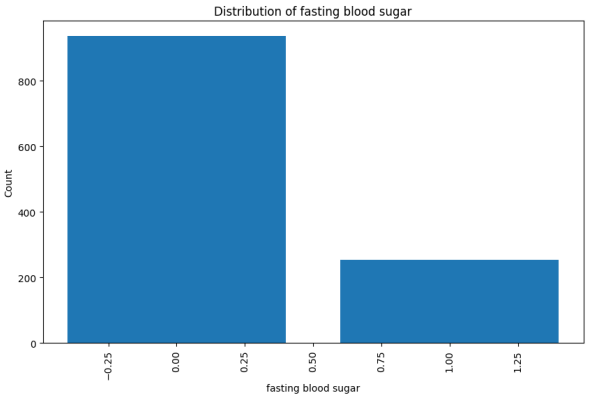
This database has data from different datasets, which have been combined over 11 common features. There is data from Cleveland with 303 instances, from Hungary with 294 instances, from Switzerland with 123 instances, from Long Beach VA with 200 instances and from another general dataset with 270 instances. The total number of instances in this dataset is 1190 and the number of features is 11.
The features are:

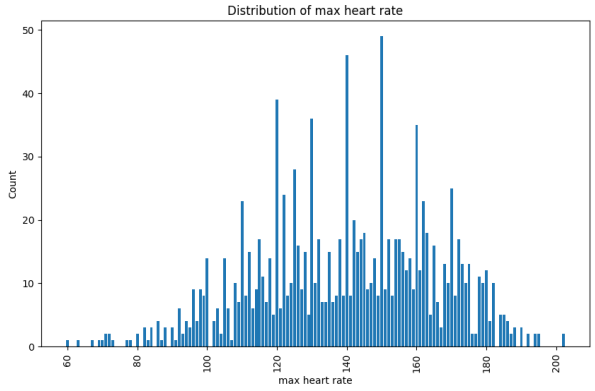
1) age (numeric). Patient's age in years.
2) sex (nominal). Patient's gender. Male is 1 and Female is 0.
3) chest pain type (categorical). 1 - typical; 2 - typical angina; 3 - non-angina; 4 - asymptomatic.
4) resting bs s (numeric). Level of blood pressure at resting mode in mm/GG.
5) cholesterol (numeric). Serum cholesterol in mg/dl.
6) fasting blood sugar (nominal). Blood sugar levels on fasting > 120 mg/dl represent 1 in case of true and 0 as false.
7) resting ecg. Results of the electrocardiogram while at rest are represented in 3 distinct values 0: Normal 1: Abnormality in ST-T wave 2: Left ventricular hypertrophy
8) max heart rate (numeric).
9) exercise angina (nominal). Angina induced by exercise 0: no and 1: yes
10) oldpeak (numeric). Exercise induced ST-depression in comparison with the state of rest.
11) ST slope. ST segment measured in terms of slope during peak exercise 0: Normal 1: Upsloping 2: Flat 3: Downsloping
12) target. Heart risk, 1: heart disease 0: normal

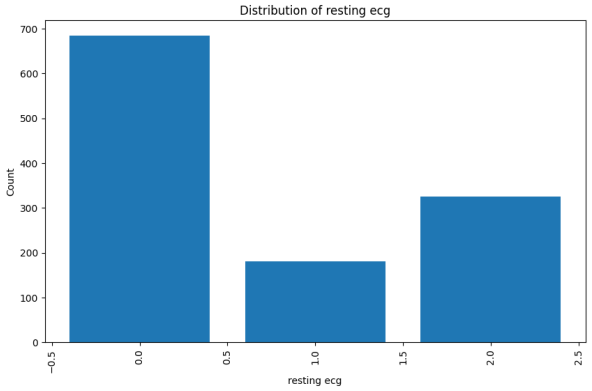
Exploring data

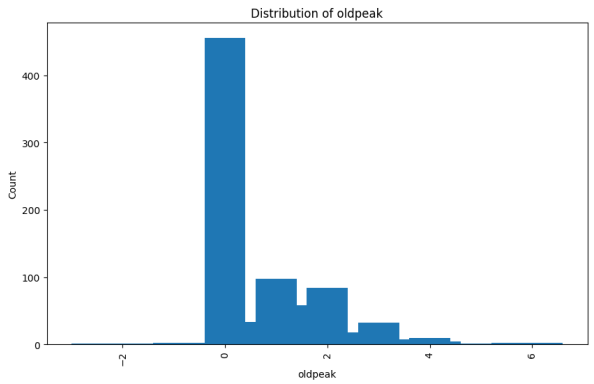


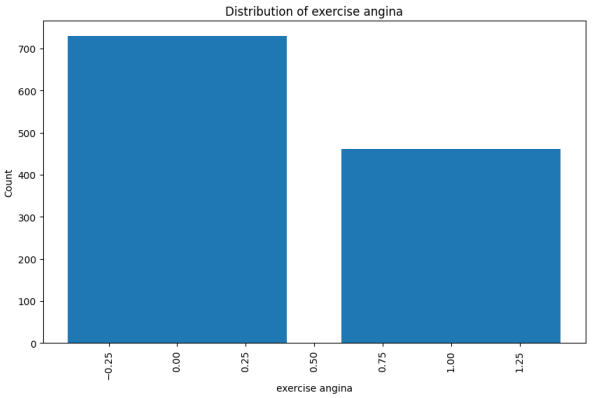


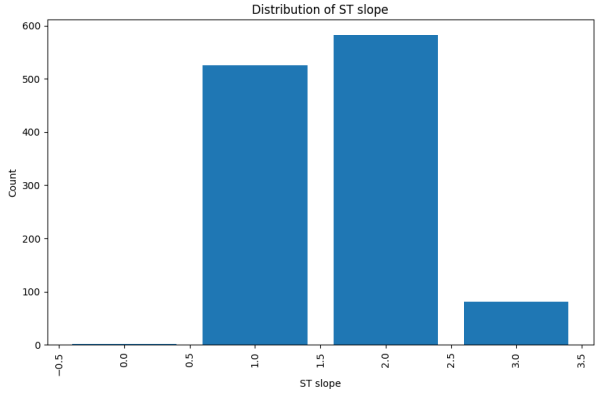


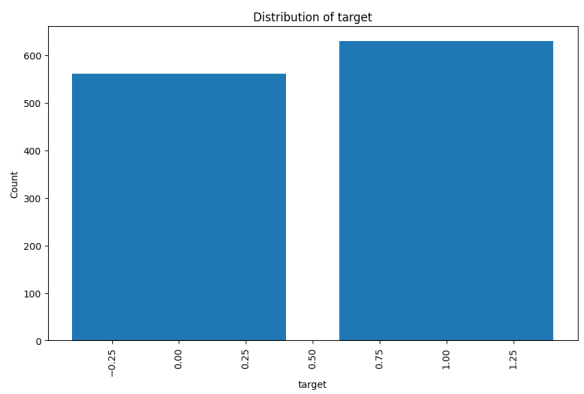












Looking at our data we noticed that there are no missing values and all the values are numeric, which means we don't need to clean the data and there is no need to make dummies in order to transform the column values to numeric. We performed data visualization in order to see if the features have a normal distribution and we found that most of them have, such as age or heart rate. We intend to explore why some values have larger

counts than others and what role they play in the result, such as cholesterol. It's like solving a puzzle to see which factors matter most. This will help us understand what's going on in our data and make our analysis more spot-on later on.

Verifying data quality

We have determined that our dataset is sufficiently rich in qualitative information to meet the requirements of our project. We identified the absence of any missing values, ensuring the completeness of our dataset. Nevertheless, we observed certain attributes exhibiting an imbalance, indicating variations in the distribution of values. Despite this, it's reassuring to note that our target value, the key variable of interest, demonstrates a relatively balanced distribution. This balance in the target value enhances the robustness of our dataset, allowing for more reliable and accurate analyses in the subsequent stages of our project.

Task 4

Make a detailed plan of your project with a list of tasks. There should be at least five tasks. Specify how many hours each team member will contribute to each task.

1) Get familiar with this topic and search for background information (3h)
2) Feature importance evaluation, selecting data that is important for our task (1h)
3) Visualizing data to understand it better (4h)
4) Build a KNN model (2h)
5) Build a Random Forest model (2h)
6) Build a Neural network model (2h)
7) Build a Decision tree model (2h)
8) Potentially add a user input function (3h)
9) Refactor our work (3h)
10) Write a conclusion of the project (3h)
11) Create a poster to present this project (3h)

List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

1) Deepnote for real-time collaboration
2) Sklearn library for classifiers, scalers, and data transformation
3) Matplotlib for data visualization
4) TensorFlow for neural network implementation
5) Pandas for data manipulation
6) NumPy for numerical operations