Phase 1 - Business and Data Understanding:

Dikla Itah

1. Q: What are the business needs/problems you are trying to address

   A: There is not enough money for the state and the health system to support all heart patients who require treatment and financial support.

2. Q: What are the objectives of the Data Science project – use supervised learning (regression or classification) and define what is the target.

   A: The goal is to predict whether someone is at risk of having a heart disease. The target variable is categorical and the model will be a classification model.

3. Q: What kind of data is available.

   A:

   1. Age: Patients Age in years (Numeric)
   2. Sex: Gender (Male : 1; Female : 0) (Nominal)
   3. cp: Type of chest pain experienced by patient. This term categorized into 4 category.
      0 typical angina, 1 atypical angina, 2 non- anginal pain, 3 asymptomatic (Nominal)
   4. trestbps: patient's level of blood pressure at resting mode in mm/HG (Numerical)
   5. chol: Serum cholesterol in mg/dl (Numeric)
   6. fbs: Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal)
   7. restecg: Result of electrocardiogram while at rest are represented in 3 distinct values
      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 hypertrophyby Estes' criteria (Nominal)
   8. thalach: Maximum heart rate achieved (Numeric)
   9. exang: Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal)
   10. oldpeak: Exercise induced ST-depression in relative with the state of rest (Numeric)
   11. slope: ST segment measured in terms of slope during peak exercise
       0: up sloping; 1: flat; 2: down sloping(Nominal)
   12. ca: The number of major vessels (0–3)(nominal)
   13. thal: A blood disorder called thalassemia
       0: NULL 1: normal blood flow 2: fixed defect (no blood flow in some part of the heart) 3: reversible defect (a blood flow is observed but it is not normal(nominal)
   14. target: It is the target variable which we have to predict 1 means patient is suffering from heart disease and 0 means patient is normal.

4. Q: How large is your data.

   A: 303 rows and 14 columns

5. Q: Main features

   A: There is not a lot of features, all of them looks related to the problem.

6. Q: Can you find a new data set online that you could merge and increase your insights

A: Heart Attack Analysis & Prediction Dataset-kaggle.

304 rows and 11 columns that are very similar to the data we have.

the same columns.

We can also add a dataset on deaths related to heart disease by age and gender.

1.2 Insights plots:

Fig 1. This graph shows a big connection between age and the chance of getting heart disease. As you go up in age group, the chance increases

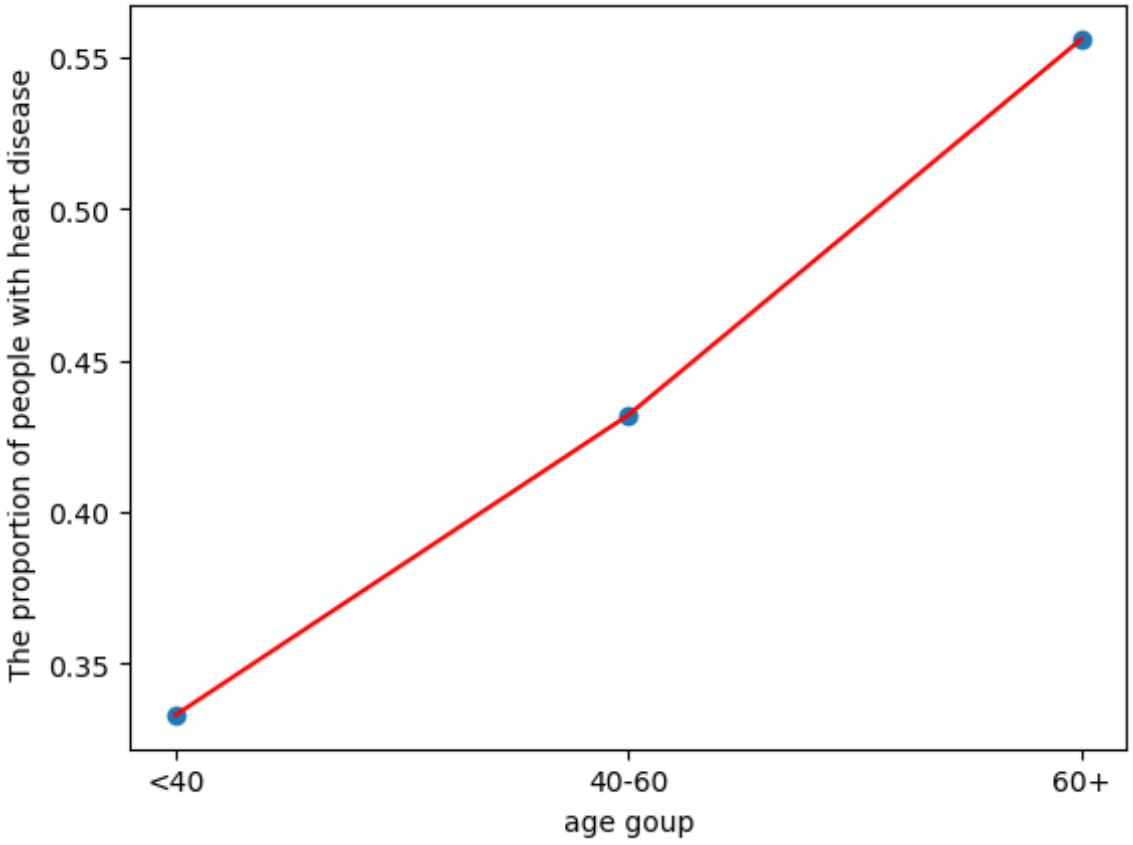


Fig 2. This graph shows that a significant amount of people have heart disease with no symptoms at all (cp=3, target=1)

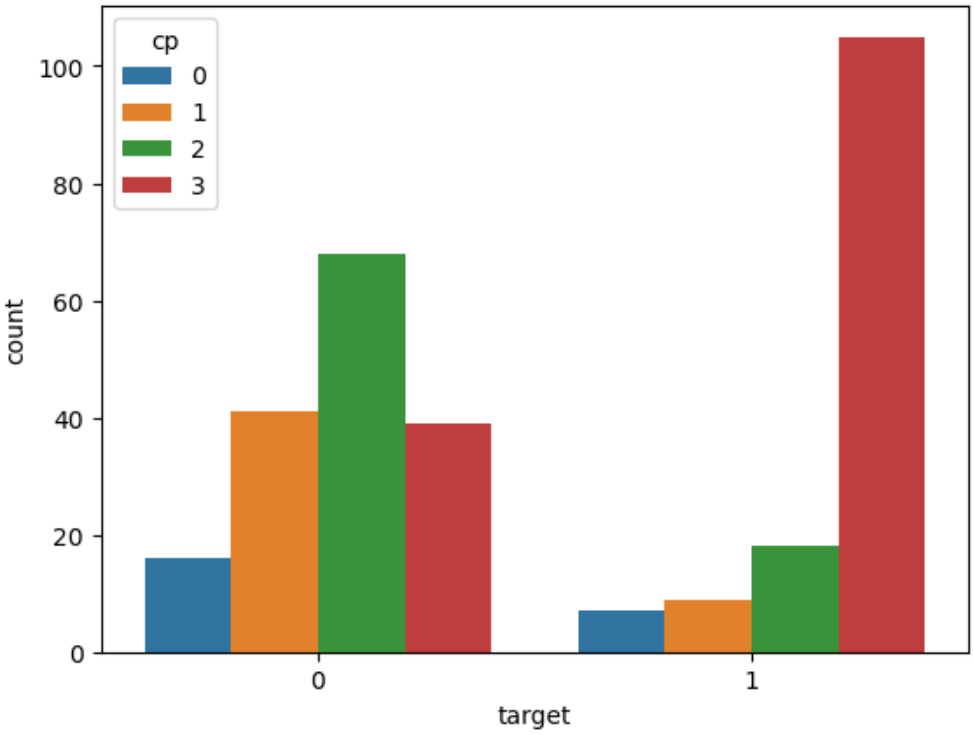
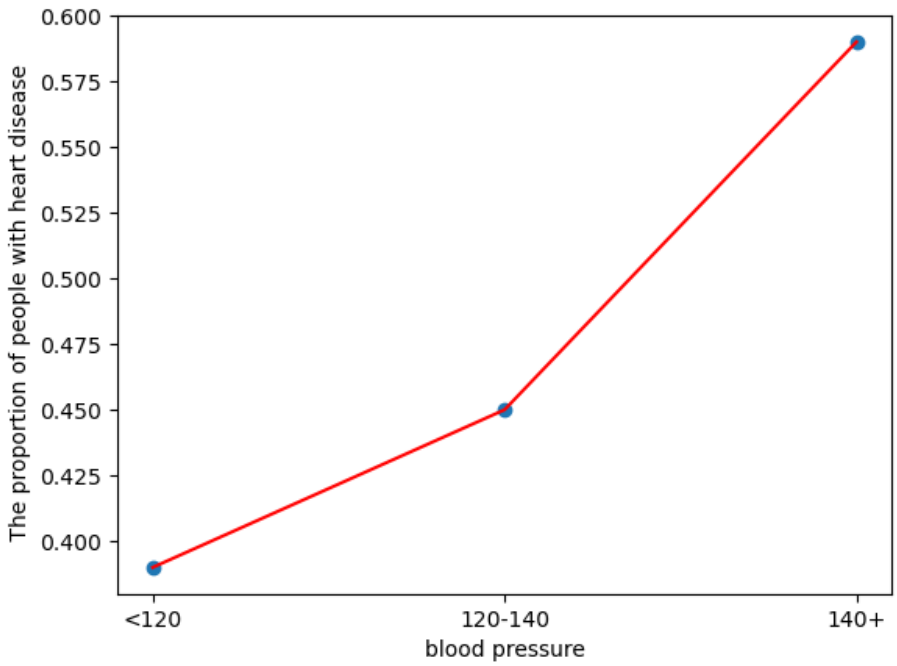


Fig 3. This graph shows a connection between blood pressure levels and the chance of getting heart disease. As you go up in blood pressure, the chance increases.



Phase 2 – Data preparation, model training and conclusions:

1. Data Preparation:

   I checked if the data is balanced and found out he is.

   I have searched for outliers columns and then replaced the outliers with the median of that feature.

   Then I used a standard scaler to normalize the data frame.

2. Model training:

   I tried the Decision Tree model, Random forest model, KNN model, SVM, and MLP model.

   I tried to remove features after using feature importance but there was no improvement.

   Then I tried to mix a couple of models with voting.

3. Conclusions:

   The best model is KNN model with an accuracy of 84%

   The model is a model that tries to predict heart disease and that's why the recall for target=1 is very important.

   We can improve the recall by lowering the threshold and remembering that some of the patients we diagnosed as heart disease patients are not sick and probably after a couple of tests we will figure it out.

   It's best that we diagnose more patients as sick and perform more tests to figure it out than tell them they are fine when they aren't.

   Lowering the threshold to 0.3 on the KNN model will improve the recall for the target=1 to 0.9 which means that we found 90% of the sick people (and some others).

   We need to be careful about lowering the threshold because we can get to the point where the model means nothing and sends everyone to take more tests.

   The data is very small, to get better accuracy there has to be more data.

   My next step would be to find more data and add it to the data frame.