Problem Statement:

The aim of this project alludes to the patient's having a heart disease with the help of analysing the following chosen attributes:

1. id: Each person's unique patient ID
2. age: The Person's age in years
3. sex: The person's sex
4. dataset: Location of the dataset
5. cp: Different types of chest pains experienced
6. trestbps: Resting Blood Pressure of the patients
7. chol: The person's cholesterol measurement
8. fbs: Fasting blood sugar of the patient
9. restecg: Resting ECG measurement of the Patient
10. thalch: Maximum heart rate achieved by the Patient
11. exang: exercise induced angina of the patient
12. oldpeak: ST depression induced by exercise relative to rest
13. slope: the slope of the peak exercise ST segment
14. thal: levels of thalassemia
15. num: Stages of heart diseases from 0 to 4

5 Questions about the dataset:

1. What is the distribution of the patients on the basis of their age and genders?
2. Which age group is more likely to have a heart disease of any stage?
3. Which type of chest pain is most common among the patients?
4. What is the relation between the cholesterol level and the Resting ECG measurement of the patients?
5. Considering all the parameters, what is the probability that a normal person has a heart disease or not?

Project Analysis:

- After reviewing the description of the dataset, we get to know that there are 920 observations in the dataset with 16 columns. To fix this, we remove the 'ca' column. After this, we remove all the null values from all the columns from our dataset. Finally, we get our new dataset with 371 observations and 15 values.
- Then, we checked the number of patients with no hear diseases (i.e., observation 0 in the 'num' column, patients with 1st stage heart diseases, patients with 2nd stage heart diseases, patients with 3rd stage heart diseases and patients with 4th stage heart diseases.
- Then we plotted the distributions of patients on the basis of their age and gender.
- We then plotted the distribution of age and gender of patients with no heart diseases, patients with stage 1 heart disease and patients with stage 2 heart disease.
- After that, we plotted the distribution of different types of chest pains in the patients.
- Then, we plotted the colourmap for resting ECG of patients with no heart diseases and different stages of heart diseases.
- Then, we plotted the colourmap for different types of chest pain in patients with no heart diseases and different stages of heart diseases.
- After this, we have plotted the scatterplot of the resting ECG and the cholesterol level of the patients to see the relation between them.
- Then, we have performed One Hot encoding for the use of Modelling.
- We split the dataset into Training and Testing after that, so that we can view the number of observations and columns in each of them.
- We have then used the Minmaxscaler to bring all the values a little closer to each other, so that the model becomes more efficient.
- Using Random Forrest for the prediction, which helps in fitting the training and testing dataset.

- After that, we plot the confusion matrix to view the accuracy of our model's predictions.
- The maximum accuracy of the model came out to be 58.6667%.

Possible Future Work and Ethical Quandaries:

- The data was originally 920; after the null values were removed, the data shape was around 371. So, I believe the accuracy which I got from the prediction of the model comes out to be in between 50% to 58.6667%, which is fairly good considering the paucity of data. In the future if there is more work to be done on this dataset, my suggestion would be to increase the number of rows and columns, i.e., take a bigger and more accurate dataset, so that we can get more accurate predictability.
- Working with the data without the knowledge of any patient's identify is another problem I'd like to fix with this model in the future. I think it's important for this information concerning people's heart conditions to be kept private because it's so delicate.