Yale University

Heart Disease Prediction

Jingyu (Kristal) Zhou

BIS 634

Supervisor: Robert McDougal

TABLE OF CONTENTS

1 Introduction

1.1 Background

Heart disease, also known as cardiovascular disease (CVD), is a major disorder of the heart and blood vessels that cause the death of approximately 17 million people worldwide each year [1]. One or combined risk factors including specific health conditions, age, lifestyle, and family history can lead to heart disease [2]. Particularly, about half of Americans (47%) have at least one of three major risk factors for heart disease: high blood pressure, high cholesterol, and smoking [2].

Due to the recent rapid advancement of techniques in data analytics, people with multiple risk factors could be benefited from early detection and management of the heart disease. One emerging prediction approach is machine learning, which uses a training dataset to create a model by detecting the patterns in the input dataset [3]. Machine learning algorithms, such as K-Nearest Neighbor (KNN) and Random Forest Classifier have been used to assists prediction and diagnosis of a disease [4,5]. This report will discuss and compare the classification task of heart disease prediction using two machine learning models KNN and Random Forest with the given dataset.

1.2 Data Resource and FAIRness

The dataset named “Heart Failure Prediction” under the “heart_failure_clinical_records” is from Kaggle, a public data Repositories. It was created by fedesoriano, who combined different datasets with over 11 common features [5]. The data was stored in .csv format. The metadata of the dataset is also available and licensed on Kaggle. The original five dataset used can be found from UCI Machine Learning Repository.

This dataset is following the **FAIRness** principle:

1. Findability: This dataset is public, which means that the data and metadata can be found easily by everyone by searching with key word “heart failure prediction” on Kaggle.

2. Accessibility: Since this dataset is open and free from a public repository, no permission is needed to request the data. Both the dataset and its metadata can be downloaded from Kaggle.

2. Interoperability: The dataset is stored in .csv format and its metadata stored in .data format. Description of the dataset includes qualified references and acknowledgements to the original 5 datasets.

2. Reusability: This data has a clear description of usage and licensure. The provenance is UCI Machine Learning Repository, also an open data source.

1.3 Analysis Questions

- Which age range/gender has the highest prevalence of heart disease?
- How are the three key biometric factors -- blood pressure/cholesterol level/blood sugar related to heart disease?
- Which attribute is most correlated with heart disease?
- Which machine learning model will have better performance in predicting heart disease using the given dataset?

The reason I chose the first question because I was interested to see the pattern of heart disease distribution in two different demographic groups. Since other demographic factors such as race/ethnicity, income level, education, etc. are not included in the dataset, I might have limited comparison. Still, it is worthy for analysis because gender and age are known to have different risk for heart disease.

I chose the second question due to my curiosity of confirming the relationship of those three key modifiable risk factors with heart disease. For the third question, I am interested in learning which risk factor for heart disease would be more concerning than others. And finally, I want to investigate the implication of machine learning in disease prediction.

2 Exploratory Data Analysis

2.1 Data Statistics and Preprocessing

2.1.1 Data Overview

Size and format: The data is stored in .csv format and with 918 entries, with a size of 36KB. There are 12 variables in the data, including:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: depression of ST segment in the EKG induced by exercise relative to rest: ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

2.1.2 Data Preprocessing

Since it is a very clean dataset, I did not preform much in this section.

- Missingness and Duplication

There are no missing or duplicated data. The author of the dataset already did the data cleaning. While combing the original five datasets, the author removed 272 entries from total 1190 entries [6].

• Data Manipulation

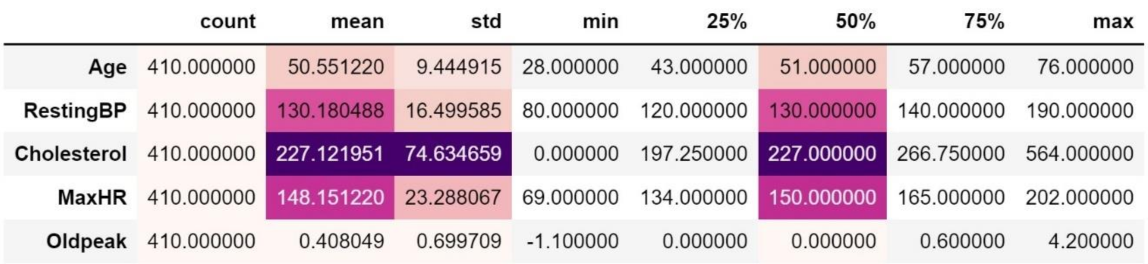
In order to present the data in a more meaningful way in exploratory analysis, I replaced 0 and 1 in "HeartDisease" with "No" and "Yes" as well as in "FastingBS" with ">120 mg/dl" and "Normal". I also replaced "Y" and "N" with "Yes/No" in "ExerciseAngina". Before double checking the description for each attribute, I wrongly thought that "FastingBS" was also a numerical attribute. However, it is in fact a categorical attribute with only two categories: 1: if FastingBS > 120 mg/dl, 0: normal.

2.2 Summary Statistics

This section will discuss some important statistics in the dataset, including checking for outliers.

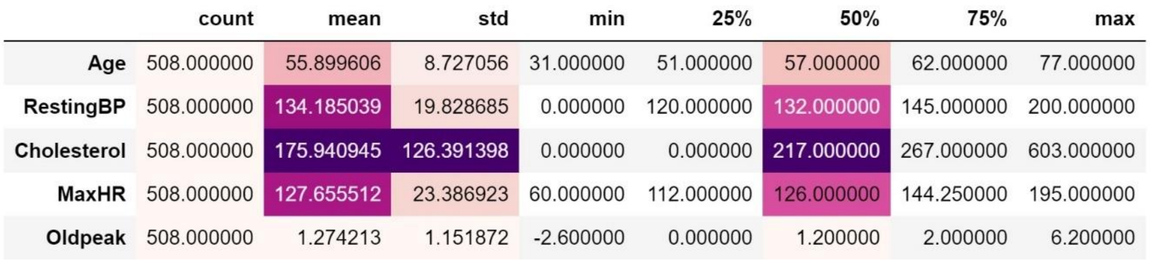
- Summary statistics of numerical values in heart disease vs no heart disease

In the below figures, we could see that the mean age for people who have heart disease is slightly higher than those who do not have heart disease. Mean resting blood pressures are similar in both outcome groups, while people with heart disease have slightly higher ranges of blood pressures. It is also not surprising to see that means of ST depressing induced by exercise (Old peak) is much higher among those who have heart disease. Generally, ST segment elevation or depression up to 0.1 mV is considered within normal limits [9]. However, means of cholesterol and heart rate levels are higher among those who do not have heart disease. This is contradictory with current findings since high cholesterol is linked with a higher risk of CVD [7]. Accelerated heart rate is also known to be clinically significant in the association with heart disease [8]. Nevertheless, individual risk factor might not be sufficient to contribute occurrence of heart disease.



	count	mean	std	min	25%	50%	75%	max
Age	410.000000	50.551220	9.444915	28.000000	43.000000	51.000000	57.000000	76.000000
RestingBP	410.000000	130.180488	16.499585	80.000000	120.000000	130.000000	140.000000	190.000000
Cholesterol	410.000000	227.121951	74.634659	0.000000	197.250000	227.000000	266.750000	564.000000
MaxHR	410.000000	148.151220	23.288067	69.000000	134.000000	150.000000	165.000000	202.000000
Oldpeak	410.000000	0.408049	0.699709	-1.100000	0.000000	0.000000	0.600000	4.200000

Figure 1: *Summary statistics for no heart disease*



	count	mean	std	min	25%	50%	75%	max
Age	508.000000	55.899606	8.727056	31.000000	51.000000	57.000000	62.000000	77.000000
RestingBP	508.000000	134.185039	19.828685	0.000000	120.000000	132.000000	145.000000	200.000000
Cholesterol	508.000000	175.940945	126.391398	0.000000	0.000000	217.000000	267.000000	603.000000
MaxHR	508.000000	127.655512	23.386923	60.000000	112.000000	126.000000	144.250000	195.000000
Oldpeak	508.000000	1.274213	1.151872	-2.600000	0.000000	1.200000	2.000000	6.200000

Figure 2: *Summary statistics for heart disease*

This prompted me to investigate more about each feature as well as their correlation with each other and heart disease. I also wonder if investigators had handled all the data entries properly. For example, the dataset includes large amounts of participants with zero cholesterol, which made me suspicious of correctness of the data (Figure 3). It might be possible that people have a very low total cholesterols and have little risk in having other diseases such as cancer and stroke [10]. Still, zero cholesterol seems quite impossible. However, information beyond what I could access from Kaggle remained unknown.

- Checking and treating outliers

  I checked the outliers by filtering those are not inside the interquartile range. There are outliers in the some of the features in the dataset:

  - No outliers in Age
  - There are 28 outliers in RestingBP
  - There are 183 outliers in Cholesterol
  - There are 2 outliers in MaxHR
  - There are 16 outliers in Oldpeak

  For this analysis, I would only remove outliers that would skew the data and affect prediction of outcome. The total observations after removing outliers are 702.

2.3 Univariate Analysis

2.3.1 Numerical Features

As I mentioned earlier, there are lots of outliers in Cholesterol with zero levels (Figure 3). After removing the outliers (Figure 4), we could still see that most participants in the dataset aged in 50 to 60, had cholesterol levels around 200, experienced tachycardia (Maximum heart rate > 140) and prehypertension (120-139 mmHg), and normal ST depression.

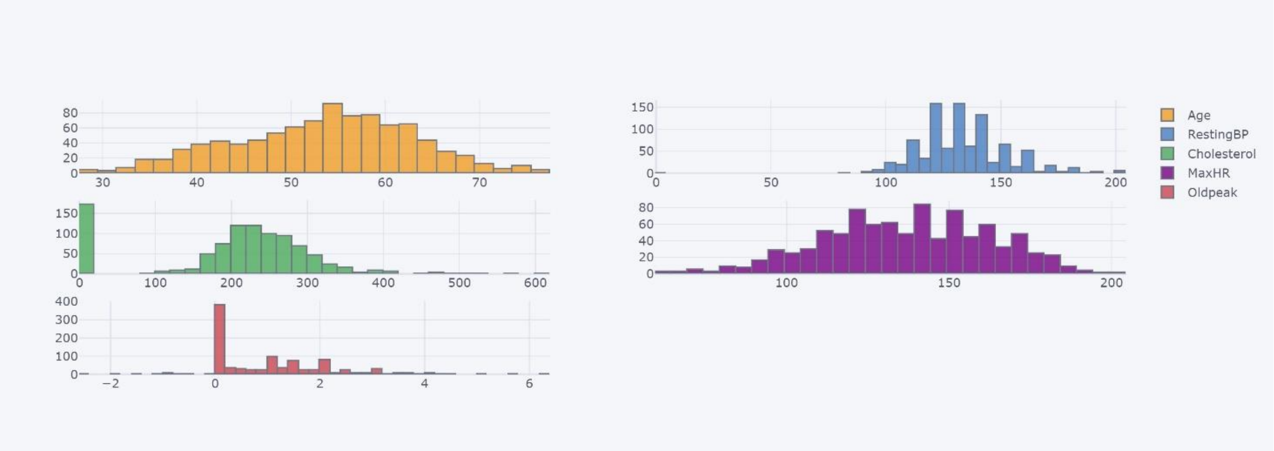


Figure 3. *Distribution of Numerical Features with Outliers*

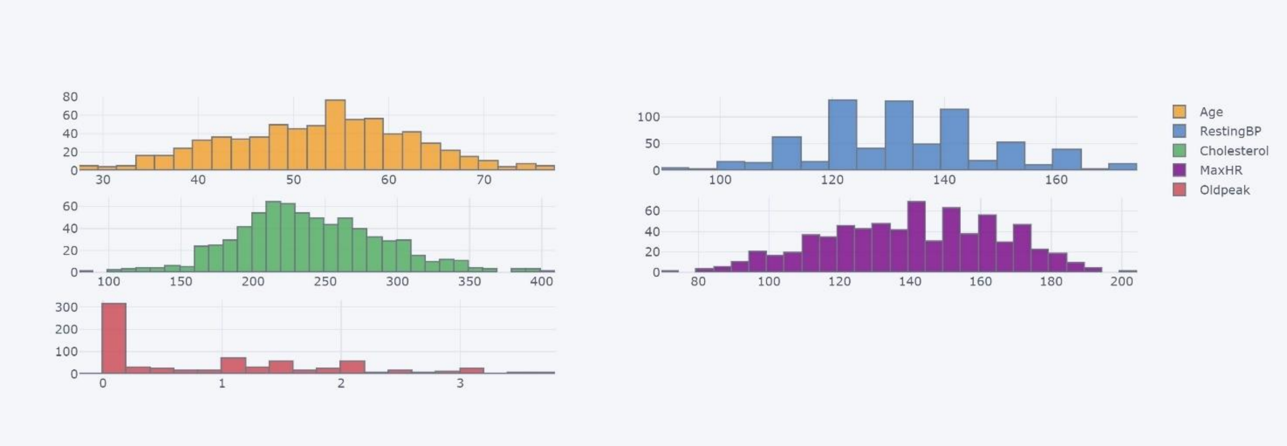


Figure 4. *Distribution of Numerical Features without Outliers*

1. Which age range has highest prevalence of heart disease?

From figure 5 we could see that people with heart disease had an older age range compared to those who did not suffer in heart disease. Most people with heart disease concentrated in around 50 to 65, with a highest age group from about 58 to 60.

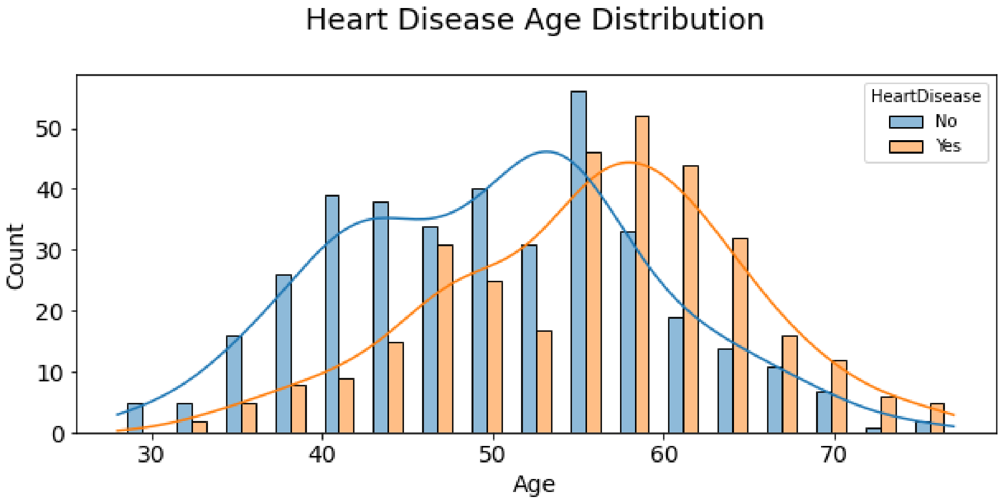


Figure 5. *Distribution of Age and Heart Disease*

2. How is blood pressure and cholesterol levels related to heart disease?

Not surprisingly, the distribution of high blood pressure and high cholesterol levels resembled that of heart disease. Most people with heart disease also had hypertension (>140 mmHg) and hyperlipidemia (>200 mg/dL), although not in extremely high levels. Although it was not clearly stated in the descriptions of dataset on Kaggle, I assumed that “RestingBP” referred to systolic blood pressure and “Cholesterol” as total cholesterol levels.

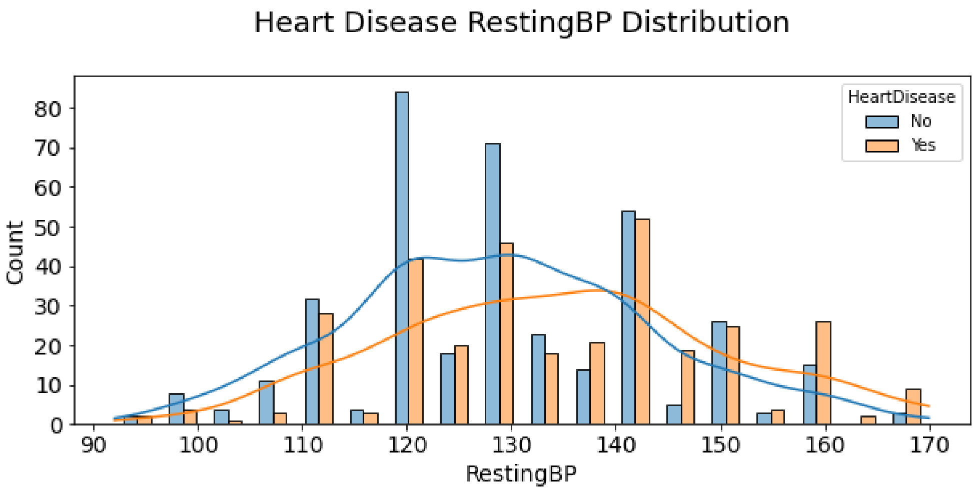


Figure 6. Distribution of Resting Blood Pressure and Heart Disease

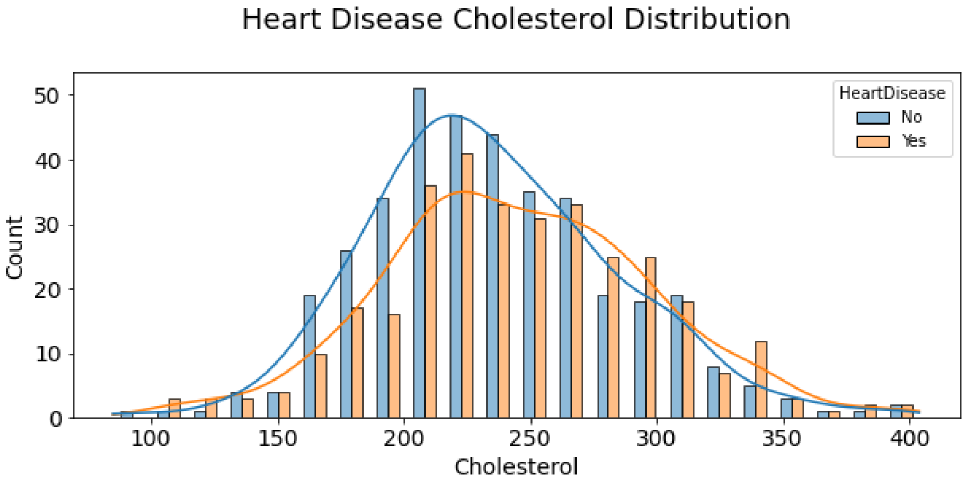


Figure 7. *Distribution of Cholesterol Levels and Heart Disease*

2.3.2 Categorical Features

1. Which gender has the highest prevalence of heart disease?

Although from figure 8 we might prompt to conclude that more males suffered in heart disease compared with females, we need to be cautious since males were three times as many as females in the dataset (figure 9).

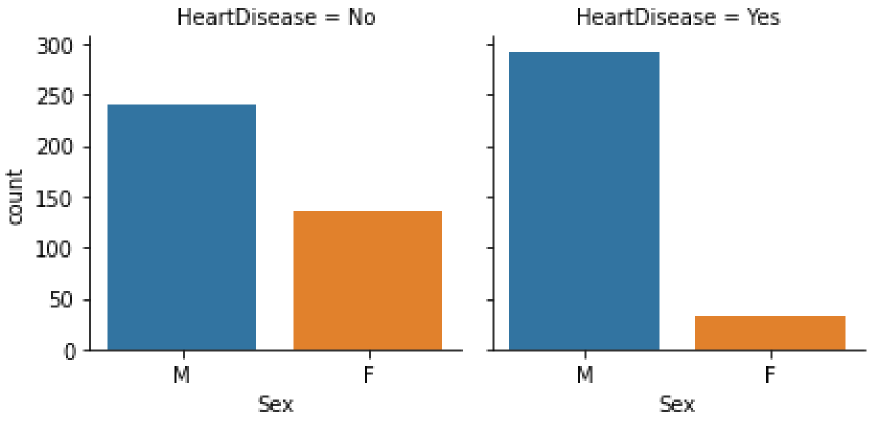


Figure 8. *Distribution of Gender and Heart Disease*

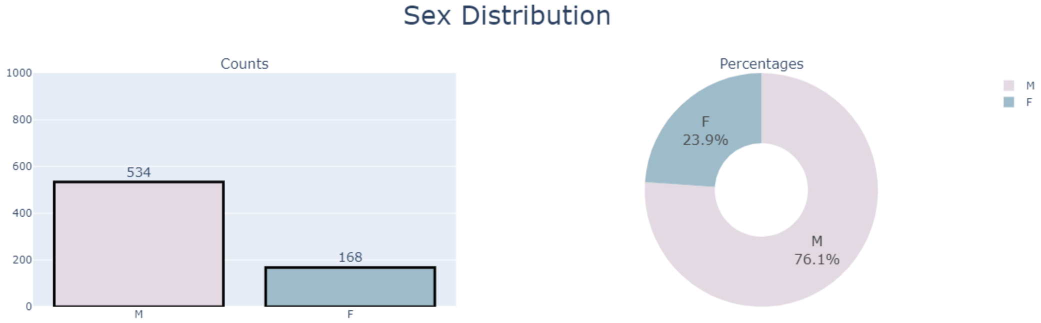


Figure 9. *Distribution of Gender in the Dataset*

2. How is blood sugar related to heart disease?

From figure 11 we could also observe an unbalanced distribution of normal and abnormal fasting blood sugar levels. Thus, figure 10 could be misleading since much fewer people with heart disease had high levels of fasting blood sugar.

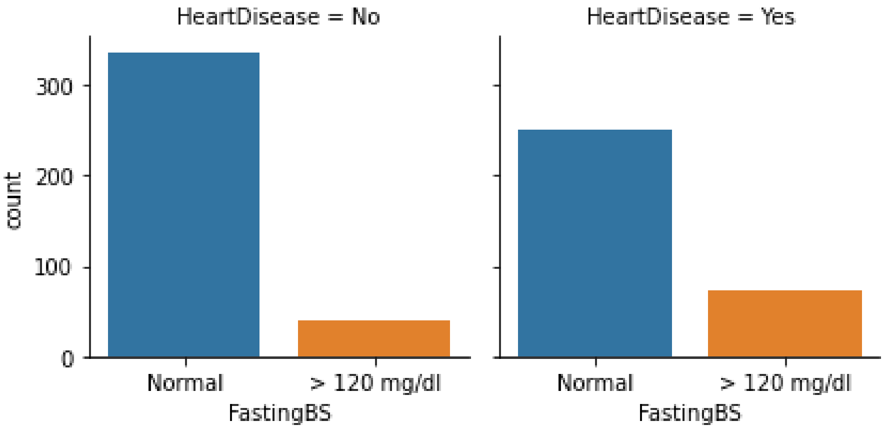


Figure 10. *Distribution of Fasting Blood Sugar and Heart Disease*

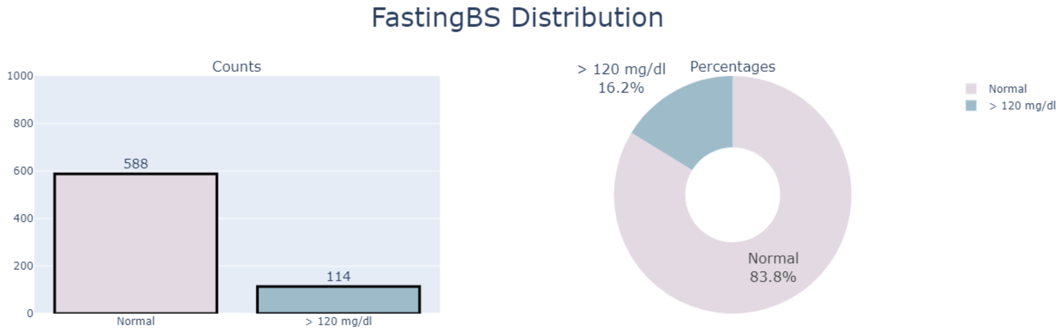


Figure 11. *Distribution of Fasting Blood Sugar in the Dataset*

In conclusion, univariate analysis alone could not provide enough information to understand the association of features and heart disease. We also need to conduct bivariate analysis and examine the correlation among the features in the dataset.

2.4 Bivariate Analysis

2.4.1 Relationships of risk factors in different feature groups

In this section, I will discuss the relationship between age and the three key risk factors (high blood pressure, high cholesterol, and high blood sugar) in different genders.

Using the facet plot method in Plotly, we could examine the relationship of risk factors with heart disease in different ways. It is interesting that older people (>50) with heart disease in either sex had higher cholesterol levels, especially if they had diabetes (figure 12). Further, people who had both high cholesterol and high blood sugar levels would be more likely to have heart disease.

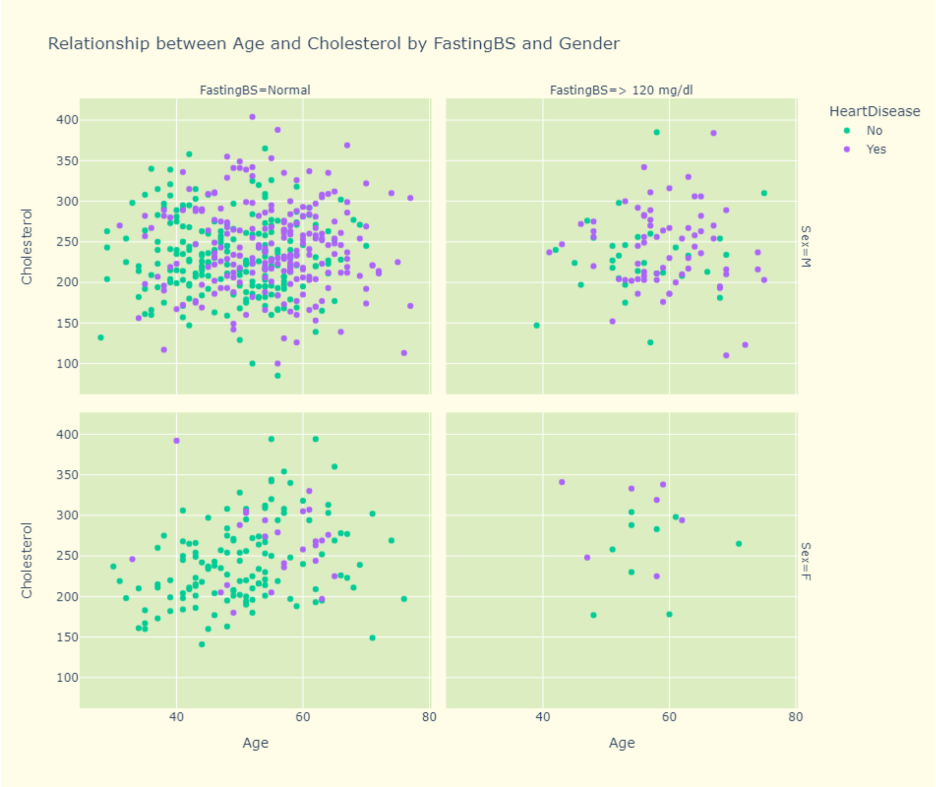


Figure 12. *Relationship between Age and Cholesterol in Blood Sugar and Gender groups*

Figure 13 provided a more interesting finding of the association of fasting blood sugar and blood pressure. Intuitively, I would tend to associate blood sugar and cholesterol or blood pressure with cholesterol. However, the results in figure 13 demonstrate that people with both high blood pressure and high blood sugar would have higher chance of getting heart disease. Those who had extremely high blood pressure and high blood sugar were more likely to have heart disease. Nevertheless, high blood pressure alone seems to be a stronger risk factor in either sex for getting heart disease.

Moreover, both figures 12 and 13 reveal that people older than 40 in both sexes have higher probability of having two out of three combined risk factors.

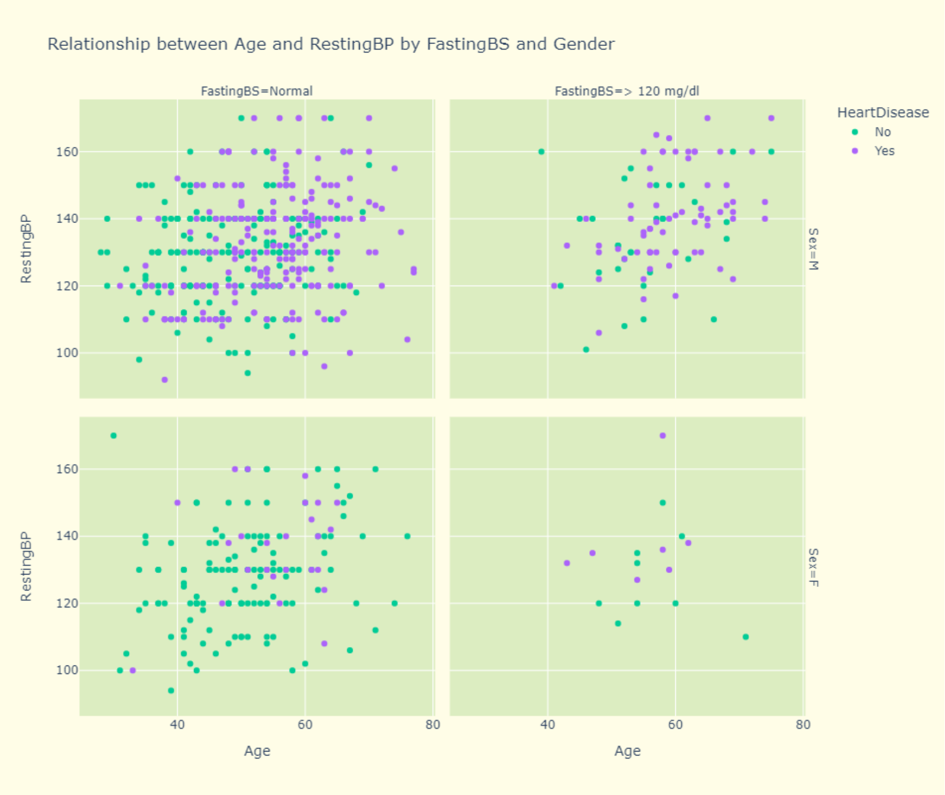


Figure 12. *Relationship between Age and Resting BP in Blood Sugar and Gender groups*

2.4.2 Feature Correlation

The correlation matrix heatmap is good to investigate the strongest risk factors for heart disease and to check whether there are highly correlated features amongst the independent variables as well. Figure 13 shows that heart disease is correlated with most features. Surprisingly, the three features ST slope, maximum heart rate and chest pain types that could be clinical indications of heart disease are moderately to strongly not correlated with heart disease in this dataset. Additionally, fasting blood sugar is slightly negative correlated with heart disease. The strongest unmodifiable risk factors are age and sex, while exercise angina and ST depression being the strongest modifiable risk factors. The three risk factors discussed earlier -- blood pressure, cholesterol levels and fasting blood sugar – do not have strong collinearity with each other. Generally, there are no strong correlated features in the dataset.

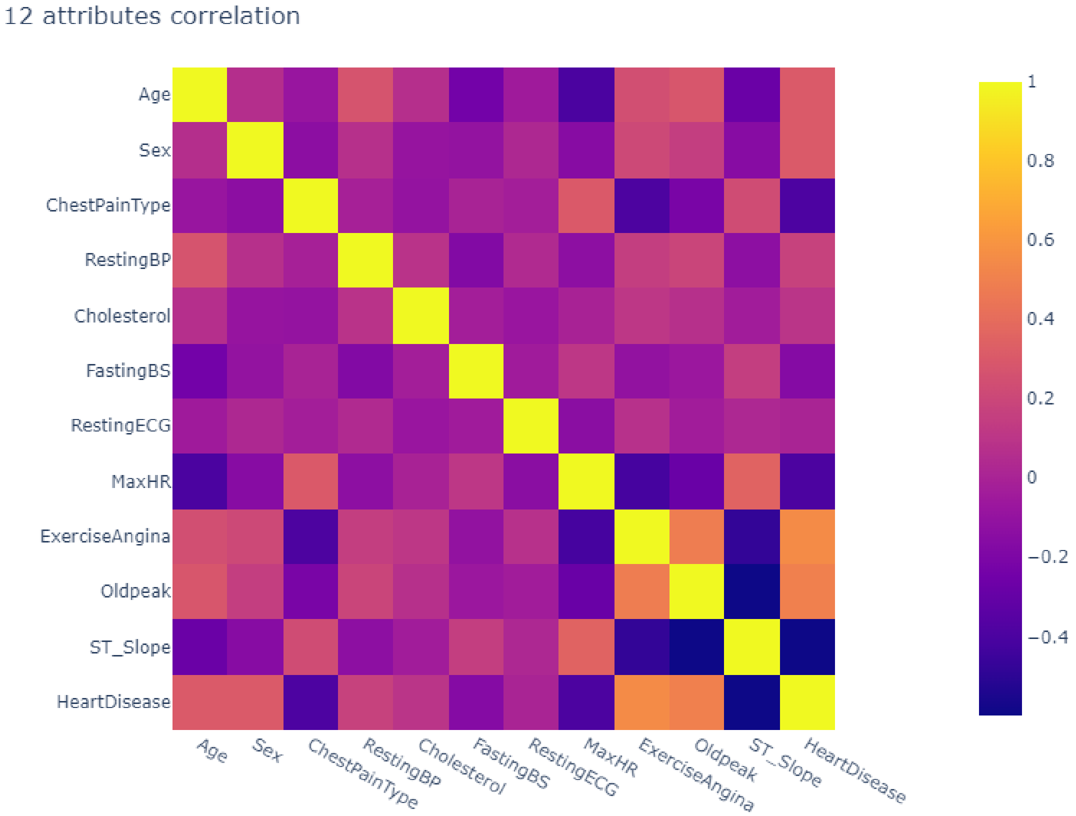


Figure 13. *Feature Correlation*

3 Machine Learning Model Prediction

3.1 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is one of the simplest supervised machine learning algorithms and widely used for classification problems [11]. This algorithm assumes the similarity between the proximity of new data and available data, then classify the new data based on its distance with the available data. Since the most critical step in KNN is to calculate distance between data points, I needed to convert the categorical variables to numerical variables by assigning dummy values (e.g., 0 for "No" and 1 for "Yes). Another critical consideration in KNN is to choose the right value K. The comparison of differences between the predicted outcome and the actual outcome (error rate) using a wide range of k (1-50) is shown in figure 14. I observed that k = 5 had the lowest error rate of 14.22%. As we could see below, a small k value such as 1 or 2 can lead to very unstable prediction, especially when we only have two categories of outcome in the data. The prediction becomes more stable as k increases to over 20. Nevertheless, we still have quite high error rates in around 30 and 50. Thus I would choose k=5 to predict heart disease for my dataset.

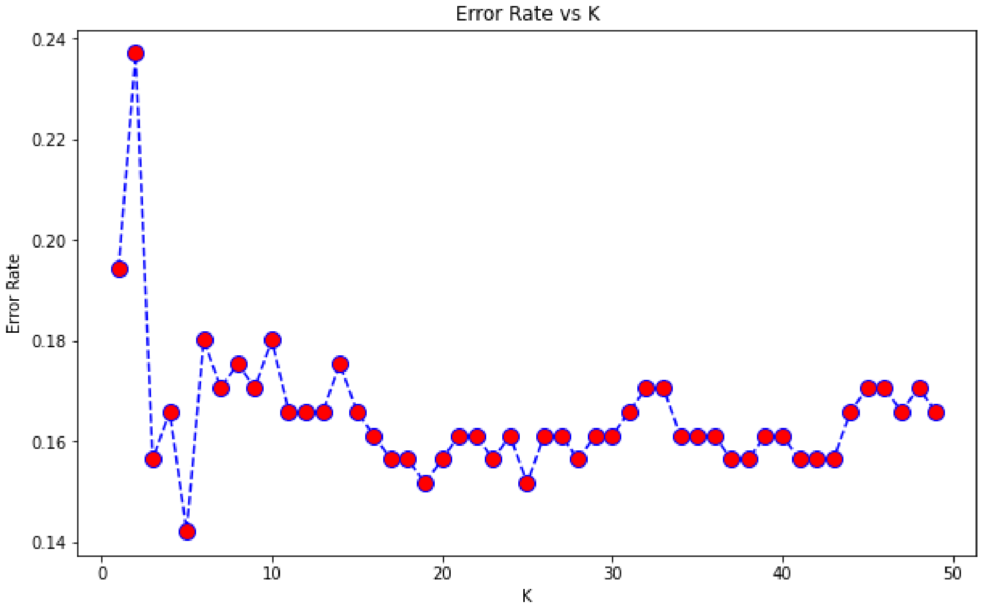


Figure 14. *Error Rate vs K*

3.2 Random Forest Classifier

Random Forest is also a supervised machine learning algorithm and used to solve classification problems [12]. Classification in random forests apply decision tree method to predict the outcome, using the training data to feed different decision trees [12]. Every decision trees provide a specific outcome for a voting system to select the majority of outcome, which becomes the final output of the forest [12]. An advantage of random forest over KNN is that it can efficiently predict outcomes with large datasets, while KNN algorithm could perform significantly slower as features or predictor variables increase [11]. I assumed that random forest performed better than KNN since we have 12 features in the dataset.

I also evaluated the error rates for different n estimators in Random Forest Classifier model. As shown in figure 14, this model has poorer performance in very low value of n estimators but improves significantly as n becomes larger. The error rates are closed to convergence as n goes over 150. In my observation, n=19, 27 and 28 had the lowest error rates with 13.27%, slightly lower than the best performance in KNN model.

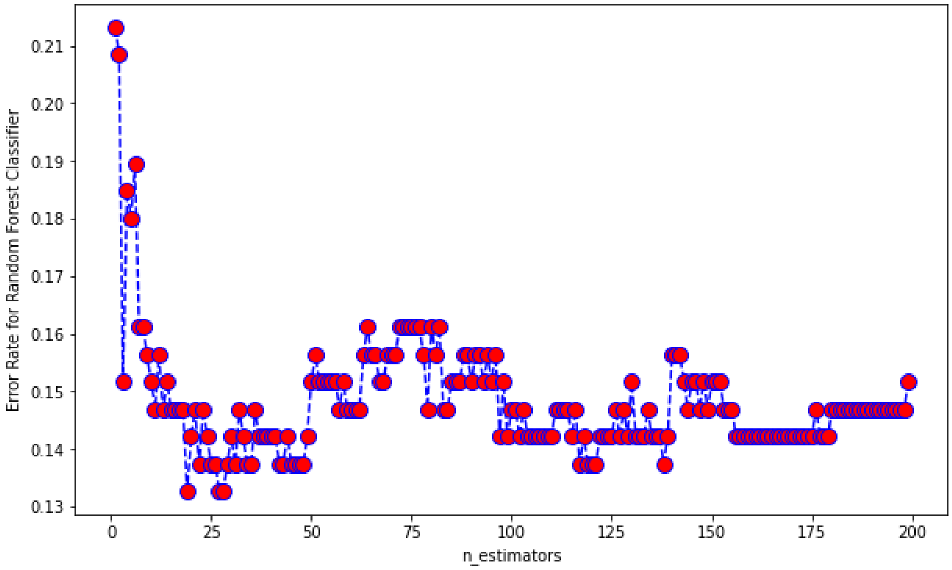


Figure 14. *Error Rate vs n Estimators*

Before implementing with either model, I encoded the categorical data (e.g., gender, chest pain type, and heart disease) using “Label Encoder” from the scikit-learn preprocessing module; standardized all predictors for the outcome heart disease using "Standard Scaler" function; and split the data into 30% test dataset and 70% test dataset using scikit-learn library.

3.3 Model Performance

In comparison, Random Forest Classifier model has a better and balanced average performance than KNN model. If we look at the precision and recall in Table 1 and 2, we could see that prediction for heart disease has higher precision for KNN model, whereas similar precision in Random Forest Classifier model. On the other hand, the fraction of positives that were correctly identified (recall) in KNN model is lower. Not surprisingly, the weighted harmonic mean of precision and recall is higher in Random Forest Classifier model. This confirmed to my early estimate that Random Forest Classifier model beat KNN for the given dataset. In terms of accuracy, however, both models are not ideal for the prediction task.

KNN Model	**Precision**	**Recall**	**F1-score**	**Support**
No heart disease	0.79	0.91	0.85	105
Heart disease	0.90	0.75	0.82	106
Accuracy			0.83	211
Macro average	0.84	0.83	0.83	211
Weight average	0.84	0.83	0.83	211

Table 1: Error rates for KNN Model

RFC Model	**Precision**	**Recall**	**F1-score**	**Support**
No heart disease	0.83	0.87	0.85	105
Heart disease	0.86	0.83	0.85	106
Accuracy			0.85	211
Macro average	0.85	0.85	0.85	211
Weight average	0.85	0.85	0.85	211

Table 2: Error rates for RFC Mode

4 API and web front-end

This section describes the API back-end and web front-end.

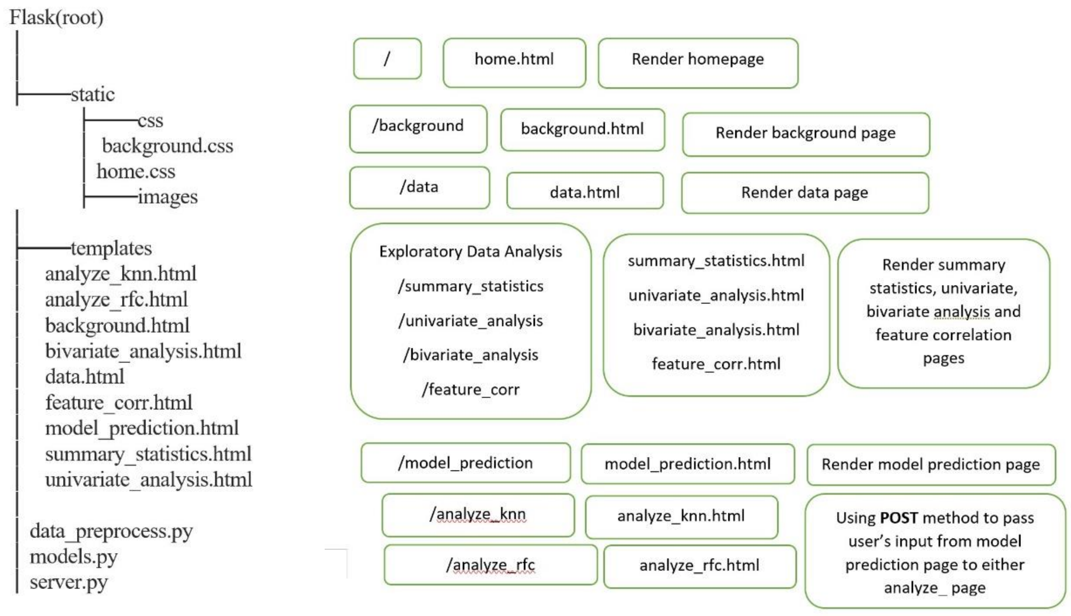
4.1 Backend API server

The file directory of the Flask API is shown in figure 15 (a).

I created a Flask app inside server.py, which including all the API routes that correspond with the html files inside the templates folder. The function of each route and its according html pages is displayed in figure 15(b). I imported data_preprocessing.py and models.py in the server.py for the following purposes: 1) data_preprocessing.py : remove outliers and preprocess categorical features and 2): models.py run the machine learning models function.

Among all the routes, the /model_prediction takes in users' input value and redirect to either analyze_ page, according to the model users choose. The /analyze_knn and /analyze_rfc request the users' input generated by /model_prediction, then use **POST** method to show the prediction results.

Simply running the app in server.py app.run in local host will launch the website.



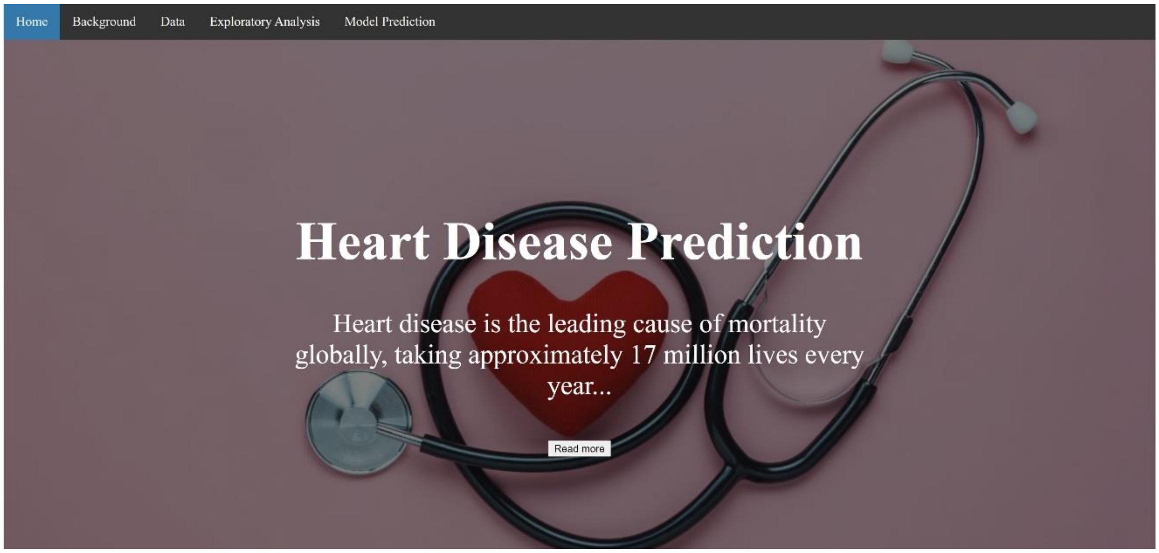
(a) *Fire* Directory **(b)** *API framework*

Figure 15. *Back-end API*

4.2 Web Front-end

There are eight pages in the web front-end:

1. The **Homepage** (figure 16) and **Background** page is a navigation and introduction of the project, clicking either “**Read more**” or “**Background**” will navigate users to the introduction of the topic -- heart disease and its prediction by machine learning.



Heart Disease Prediction

Heart disease is the leading cause of mortality globally, taking approximately 17 million lives every year...

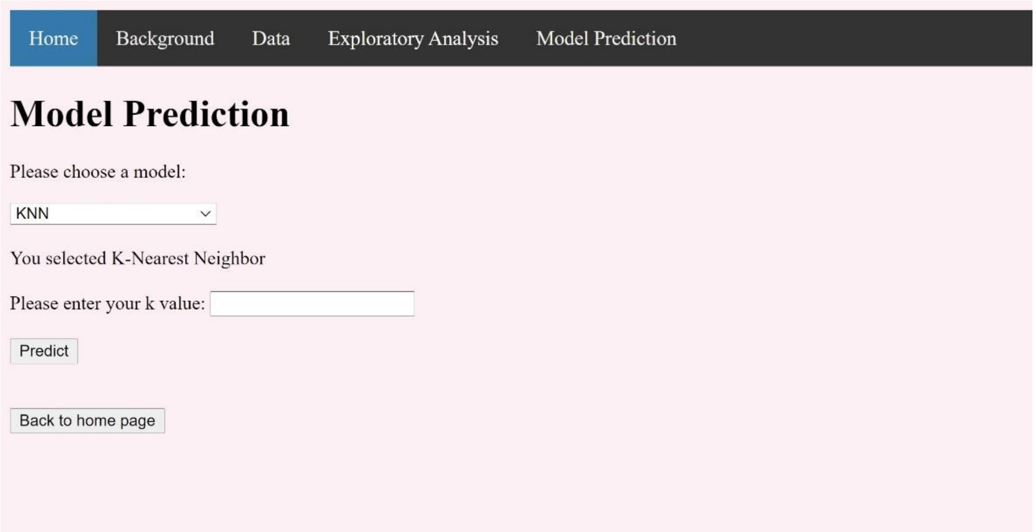
Figure 16. *Web Front-end Homepage*

2. The **Data** page contains the table of the dataset, along with the dataset’s origins and each feature, FAIRness principles, and data preprocessing. Users can move around the table in arbitrary order they like or enlarging a specific column to closely examine it.
3. Under the **Exploratory Analysis** tab, there are four pages:
    a. The **Summary Statistics** page describes descriptive statistics for numerical features, examination of outliers, and distribution of numerical features with vs without outliers.
    b. Next, the **Univariate Analysis** page shows four histograms that display distribution of four critical risk factors of heart disease (age, gender, blood pressure, and cholesterol level) along with the distribution of heart disease in the dataset.

c. The **Bivariate Analysis** simply display the relationship between age and resting blood pressure in different gender and fasting blood sugar groups, as well as relationship between age and Cholesterol in Blood Sugar and Gender groups. Hovering on the plot, users will be able to see the data description of specific points (age, gender, fasting blood sugar, and resting blood pressure/cholesterol).

d. **Feature Correlation** page shows the correlation heatmap between heart disease and the rest of feature, as well as the correlation among those features. Again, user can hover on the heatmap, user can view the specific values of correlation between two features.

4. The **Model Prediction** page allows the users to predict heart disease using two different machine learning models and experiment with different k values or n estimators. Figure 17 is an example of how to the model prediction: After clicking "KNN", the page will ask users to enter a k value.



Model Prediction

Please choose a model:

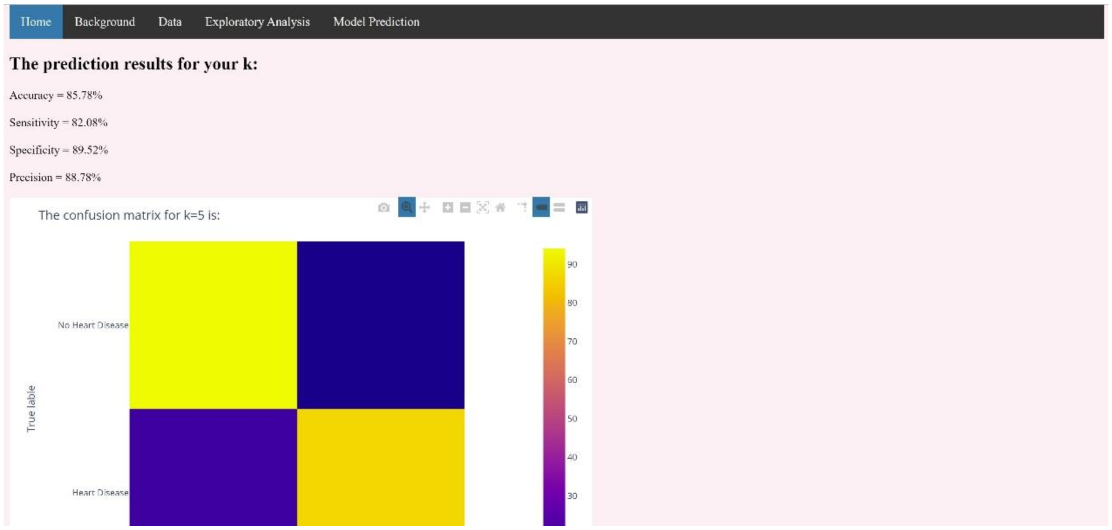
KNN

You selected K-Nearest Neighbor

Please enter your k value:

Figure 17. *Model Prediction Page*

Then, it will navigate users to the result page as shown in Figure 18, which is an example of the predicting result using k=5. Users can hover on the confusion matrix to view corresponding values. At the bottom of the result page, there is a "**Back**" bottom that can navigate the users back to **Model Prediction** page.



The prediction results for your k:

Accuracy = 85.78%

Sensitivity = 82.08%

Specificity = 89.52%

Precision = 88.78%

The confusion matrix for k=5 is:

Figure 18. *Example Result Page*

5 Discussion

5.1 Findings

• People aged around 50-65 have higher probability of getting heart disease. Males seem to be more affected by heart disease. However, since distribution of males and females are imbalanced in the dataset, the results might have limited indication.

• People with two combined factors out of high blood pressure, high cholesterol, and high blood sugar have greater chances of having heart disease. Each factor alone has no dominating risk for heart disease.

• The strongest unmodifiable risk factor are age and sex, while exercise angina and ST depression being the strongest modifiable risk factors. No strong correlated factors for heart disease observed in the dataset.

• Both models are not ideal for predicting heart disease. However, Random Forest Estimators have better and more stable performance than KNN model using the given dataset.

5.2 Difficulties

• There are some difficulties in interpreting the measurement of features in the dataset. For example, it is not clarified that whether "Cholesterol" in the dataset means total, low-density lipoprotein (LDL) or high-density lipoprotein (HDL).

• There are too many outliers in one feature with unreal values (Cholesterol). Removing the outliers can shrink the data. It is possible that there is not enough training data for the machine learning models.

• It is challenging to choose a more appropriate visualization technique to show the relationship among two or three features. In the future, I might consider using multilinear regressing model to predict the outcome using all features.

• Developing an efficient API and web frontend is also challenging. In the future, I will consider improve the API and html to allow users to navigate to specific pages with clicking fewer tabs

6 References

[1] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* **20,** 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

[2] Centers for Disease Control and Prevention. Heart Disease, 2021. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/risk_factors.htm

[3] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN COMPUT. SCI.* **1,** 345 (2020). https://doi.org/10.1007/s42979-020-00365-y

[4] Deepthi Y., Kalyan K.P., Vyas M., Radhika K., Babu D.K., Krishna Rao N.V. (2020) Disease Prediction Based on Symptoms Using Machine Learning. In: Sikander A., Acharjee D., Chanda C., Mondal P., Verma P. (eds) Energy Systems, Drives and Automations. Lecture Notes in Electrical Engineering, vol 664. Springer, Singapore. https://doi.org/10.1007/978-981-15-5089-8_55

[5] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[6] fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[7] Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 pdf icon[PDF-494K]. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Accessed May 9, 2019.

[8] Perret-Guillaume C, Joly L, Benetos A. Heart rate as a risk factor for cardiovascular disease. Prog Cardiovasc Dis. 2009 Jul-Aug;52(1):6-10. doi: 10.1016/j.pcad.2009.05.003. PMID: 19615487.

[9] Kashou AH, Basit H, Malik A. ST Segment. [Updated 2021 Aug 11]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459364/

[10] Lopez-Jimenez F (expert opinion). Mayo Clinic. Rochester, Minn. Nov. 25, 2020. https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/expert-answers/cholesterol-level/faq-20057952

[11] Harrison, O. Machine Learning Basics with the K-Nearest Neighbors Algorithm. *Towards Data Science*. Sep. 10, 2018. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[12] Mbaabu, O. Introduction to Random Forest in Machine Learning. *Section*. Dec.11, 2020. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/