ROBERT H. SMITH
SCHOOL OF BUSINESS

BUDT704: DATA PROCESSING AND ANALYSIS IN PYTHON

“ HEART DISEASE : Prediction Using Machine Learning ”

SAI VINAY BHIMANA | PREM MALDE | RASIKA PANDE | DEVIKA RAHEJA

GROUP: 507-E

Mentor: Prof. Peng Huang

INTRODUCTION

BACKGROUND

Heart disease is considered to be the leading cause of death worldwide. It contributes to approximately 17.9 million deaths annually with 32% accounting to Cardiovascular Diseases with a large proportion of these deaths happening prematurity - below the age of 70 years. This alarming finding has made it necessary to understand the variables that contribute to heart disease and develop effective methods for its timely diagnosis and long term prevention.

The causes of heart disease are multifaceted and often interrelated: bad diet, lack of exercise, stress, smoking, and a high alcohol intake in addition to physiological indicators such as high cholesterol, high blood pressure, diabetes, and obesity. Heart Disease can also be aggravated by socio-economic circumstances such as genetic propensity, and environmental factors. These are important characteristics to recognize, as the chances of cardiovascular events can be reduced by early detection and action. By encouraging healthier lifestyles, healthcare providers and policy makers can reduce the effects of heart disease on individuals and healthcare systems.

Our project aims to fill the gap between raw data and actionable insights to enable potential users such as healthcare professionals, insurance companies and individuals to make informed decisions. The focus of "Heart Disease Prediction" is to find predictors for heart disease using a rich set of clinical features to contribute toward the identification of key trends and risk factors for better targeted prevention and early detection. With the use of different machine learning models, such as Logistic Regression and KNN, we hope to provide healthcare professionals with predictive tools that could save lives. These results may further open a door to new healthcare solutions, personalized interventions, and improved patient outcomes, contributing to fighting heart disease globally.

RESEARCH METHOD SUMMARY

Using exploratory data analysis and predictive modeling, we plan to churn meaningful insights from the datasets. We will pursue the following 5 step process: Our methodology follows a five-step process:

- *Data Wrangling Journey*
- *Preliminary Statistical Analysis and Hypothesis Building:*
- *Plotting Trends and Identifying Outliers*
- *Model Preparation and Predictive Modeling*
- *Summarizing Recommendations*

VARIABLES IN THE DATASET

The combination of demographic, clinical, and behavioral variables in this dataset makes it ideal for exploratory data analysis, identification of trends, and predictive modeling in an attempt to comprehend and deal with heart disease risk

- **Age**: Indicates the individual's age in years, which is an important demographic element determining heart disease risk.
- **Gender**: Indicates whether the individual is male or female, allowing for gender-specific variations in heart disease prevalence.
- **Cholesterol**: The individual's cholesterol level measured in mg/dL, an essential marker for cardiovascular health.

- **Blood Pressure**: Systolic blood pressure in mmHg, highlighting potential hypertension issues linked to heart disease.
- **Heart Rate**: The heart rate in beats per minute, reflecting the individual's cardiovascular performance.
- **Smoking**: Smoking status categorized as Never, Former, or Current, a known lifestyle risk factor for heart disease.
- **Alcohol Intake**: Frequency of alcohol consumption classified as None, Moderate, or Heavy
- **Exercise Hours**: The number of hours of physical exercise per week, providing insights into the individual's fitness level.
- **Family History**: Indicates whether there is a family history of heart disease (Yes/No), an important genetic risk factor.
- **Diabetes**: Identifies if the individual has diabetes (Yes/No)
- **Obesity:** Indicates obesity status (Yes/No), reflecting the individual's body weight and potential health risks.
- **Stress Level**: Stress level on a scale from 1 to 10
- **Blood Sugar**: Fasting blood sugar level measured in mg/dL
- **Exercise-Induced Angina**: Indicates the presence of angina triggered by exercise (Yes/No), a symptom of reduced blood flow to the heart.
- **Chest Pain Type**: Categorized as Typical Angina, Atypical Angina, Non-Anginal Pain, or Asymptomatic, providing diagnostic insights.
- **Heart Disease**: The target variable that signifies the absence (0: No) or presence (1: Yes) of heart disease, enabling predictive modeling.

EXPLORATORY DATA ANALYSIS

We used shape to check if all the data was correctly read into python. 1000 rows and 16 columns were found using the df.shape function. There are 8 integer data types and 8 object data types in the initial data. We also found 340 null values in the 'Alcohol Intake' column using the df.info() function below.

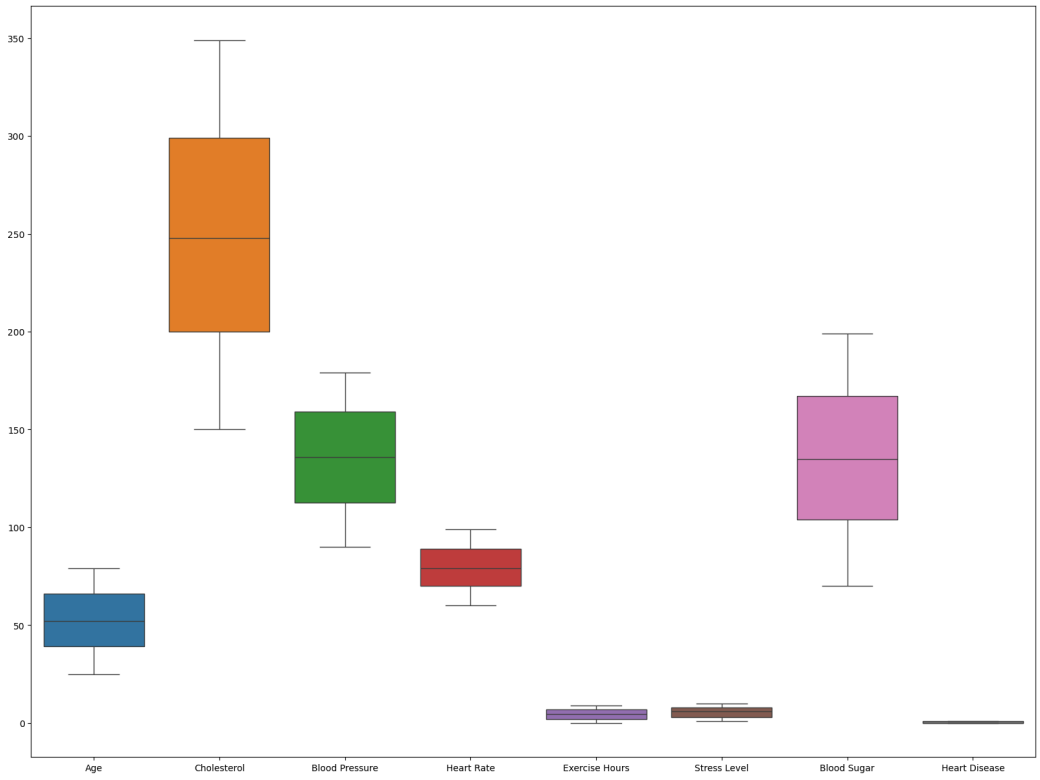
Based on the statistics summary below using the describe function, we are able to see that the 'Age' ranges from 25 - 79 with an average age of ~52 in the dataset. The average 'Stress Level' and 'Exercise Hours' indicated by people in the dataset is approximately 5.64 and 4.52 respectively. The 'Cholesterol' variable has a large range of 199 in the dataset. These data points can be visualized using the box plot below.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 16 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Age                      1000 non-null   int64
 1   Gender                   1000 non-null   object
 2   Cholesterol              1000 non-null   int64
 3   Blood Pressure           1000 non-null   int64
 4   Heart Rate               1000 non-null   int64
 5   Smoking                  1000 non-null   object
 6   Alcohol Intake           660 non-null    object
 7   Exercise Hours           1000 non-null   int64
 8   Family History           1000 non-null   object
 9   Diabetes                 1000 non-null   object
 10  Obesity                  1000 non-null   object
 11  Stress Level             1000 non-null   int64
 12  Blood Sugar              1000 non-null   int64
 13  Exercise Induced Angina  1000 non-null   object
 14  Chest Pain Type          1000 non-null   object
 15  Heart Disease            1000 non-null   int64
dtypes: int64(8), object(8)
memory usage: 125.1+ KB
```

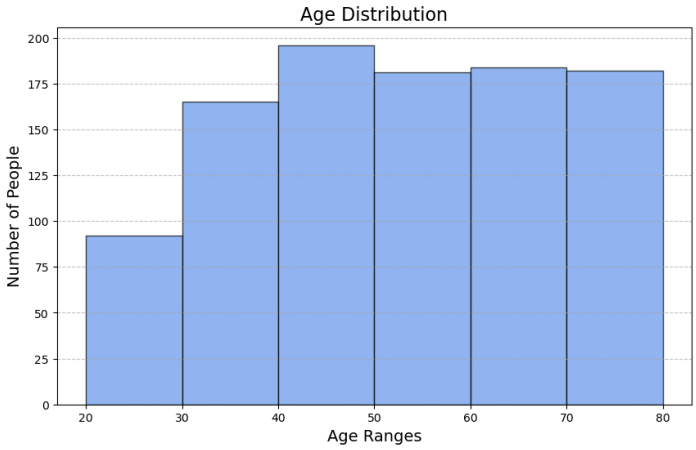
```
df.describe()
```

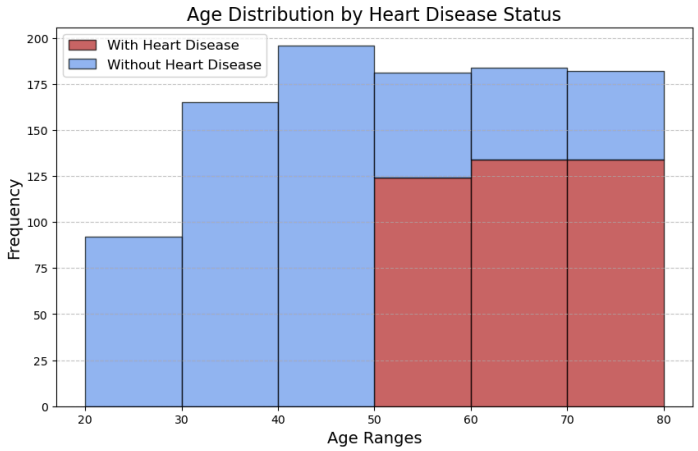
	Age	Cholesterol	Blood Pressure	Heart Rate	Exercise Hours	Stress Level	Blood Sugar	Heart Disease
count	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	52.293000	249.939000	135.2810	79.204000	4.529000	5.646000	134.941000	0.392000
std	15.727126	57.914673	26.3883	11.486092	2.934241	2.831024	36.699624	0.488441
min	25.000000	150.000000	90.0000	60.000000	0.000000	1.000000	70.000000	0.000000
25%	39.000000	200.000000	112.7500	70.000000	2.000000	3.000000	104.000000	0.000000
50%	52.000000	248.000000	136.0000	79.000000	4.500000	6.000000	135.000000	0.000000
75%	66.000000	299.000000	159.0000	89.000000	7.000000	8.000000	167.000000	1.000000
max	79.000000	349.000000	179.0000	99.000000	9.000000	10.000000	199.000000	1.000000



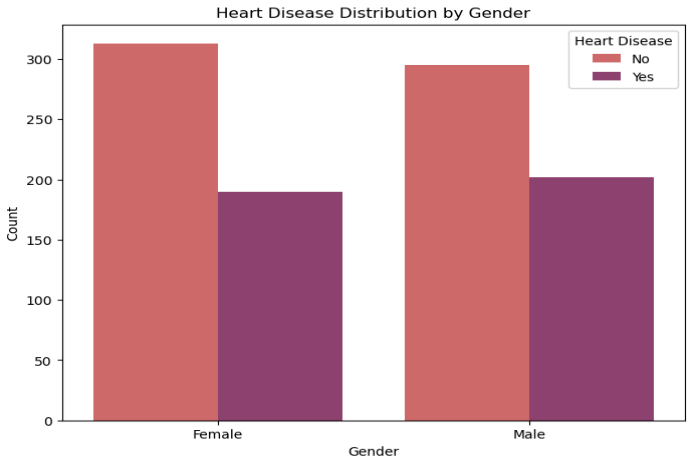
We ran multiple preliminary descriptive statistical functions and performed univariate, bivariate and multivariate analysis on the data to get a better understanding of how certain variables impact our dependent variable 'Heart Disease'. Some of our findings & insights are discussed below –

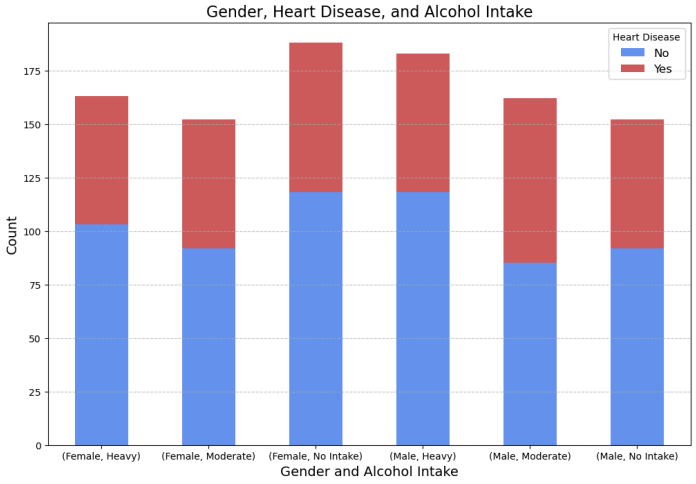
1. Age : There are almost twice as many people who are above the age of 50 in the dataset as compared to people between the ages of 20-40. It is also evident from the 'Age Distribution by Heart Disease Status' graph that people above the age of 50 are primarily impacted by heart disease as compared to people under the age of 50.



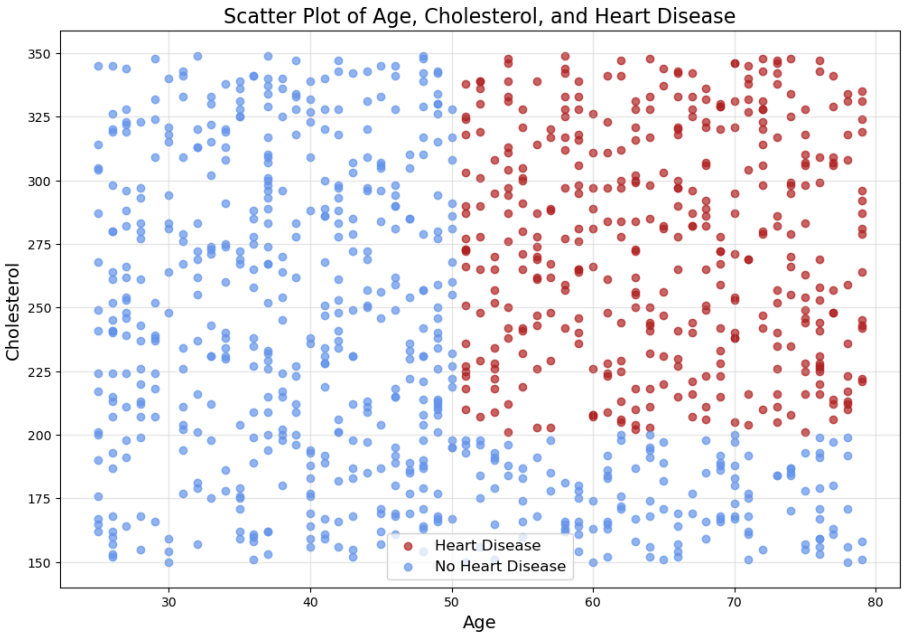


2. Gender: The dataset is almost equally balanced between males and females. Based on our visualization of 'Gender' vs 'Heart Disease', it is evident that gender might not play as strong a role in the prevalence of heart disease. However, when comparing the impact of 'alcohol intake' by gender on heart disease, it becomes evident that almost 50% of males who indicated a 'moderate intake' of alcohol were more likely to have heart disease.

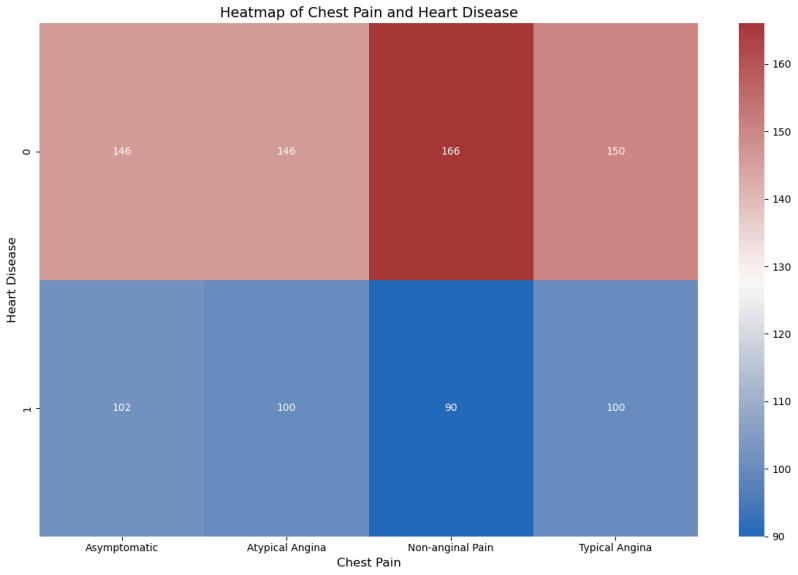




3. Cholesterol: Based on our analysis of cholesterol, we found that individuals who have a cholesterol level above 200 are more likely to have heart disease. Comparing that with our earlier analysis of age, we found that there is a significant correlation between age, cholesterol and heart disease. All individuals in the dataset who are above the age of 50 and have cholesterol level above 200 indicated heart disease (scatterplot below)



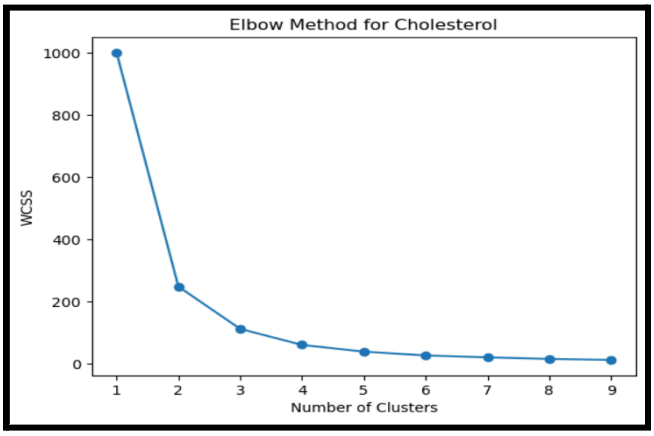
4. Chest Pain Type: It was expected that chest-pain types, specifically Typical angina which is a classic symptom of heart disease, would have an impact on heart disease. However it does not have the strongest association with heart disease in this dataset, suggesting potential underdiagnosis in cases with other pain types. This insight is further driven by 'Asymptomatic' and 'Atypical angina' having relatively balanced distributions of heart disease cases, which may warrant deeper analysis into other cofactors.



CLUSTERING

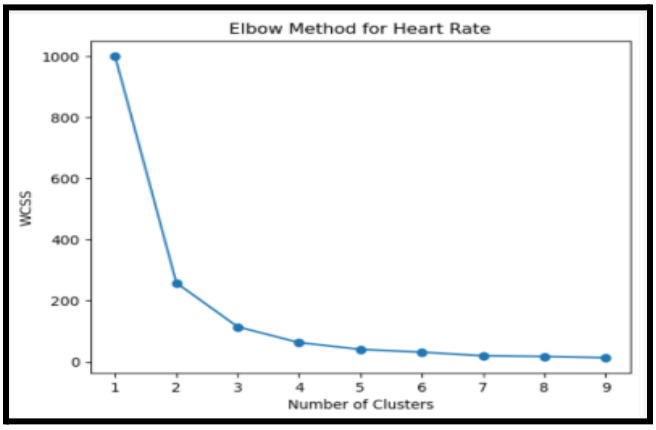
Cluster analysis was performed to further explore the data and identify the groups within the data for several factors for heart disease patients. Out of multiple clustering forms, we opted for the Elbow Method to help determine the number of clusters since this is more accurate as compared to other

methods. This curve recommended 3 clusters for all the variables affecting. Some of them have been discussed below:



Cholesterol Levels	Heart disease patient count
150-218	47
219-285	171
286-349	174

As seen in the elbow curve, it suggests forming 3 clusters. Following table shows the clusters of cholesterol ranges and number of heart patients within. This gives more clarity for prediction as compared to the scatter plot where the cholesterol level was 200+. Here, the difference between the first and second range is large enough to give clarity for estimates.



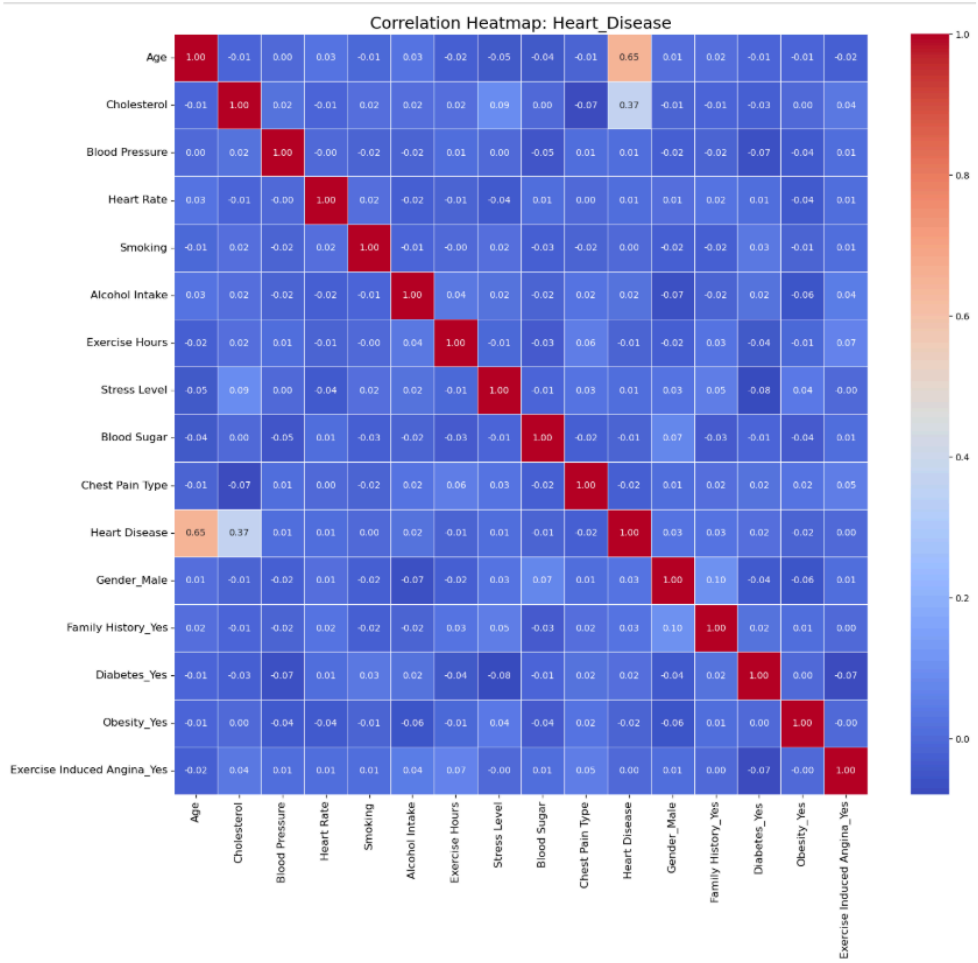
Heart Rate Ranges	Heart disease patient count
60-73	138
74-87	141
88-99	113

For heart rate too, the Elbow curve recommends forming 3 clusters. However, the number of heart patients in each of these clusters are similar in count. Also, the medically defined ranges for heart disease probabilities are different from the above. Hence, to get more clarity and a relatively accurate forecasting, we proceeded with predictive modelling.

PREDICTIVE MODELLING

Checking for Correlation & Multicollinearity

Before we construct our models, we considered it essential to assess multicollinearity. To achieve this, we examined the correlation table matrix, seeking noteworthy patterns of strong positive or negative correlations that could offer valuable insights



Shades of red indicate stronger correlations and shades closer to blue indicate no correlation to weak correlation.

VIF Analysis:

Another measure used to determine whether multicollinearity existed among the dataset's features was the Variance Inflation Factor (VIF) analysis. The findings showed that there was very little multicollinearity, with all VIF values being near 1. This implies that each feature contributes to the model in a unique way, free of duplicate or overlapping information, because the characteristics are essentially independent of one another.

These results confirm that the dataset is stable and reliable enough to be used with Random Forest and other machine learning methods. It also shows how the feature set is appropriate for other modelling techniques, such linear or logistic regression, where multicollinearity could otherwise risk the interpretation of results, even though Random Forest is naturally robust to multicollinearity. Based on this analysis, no features need to be changed or eliminated.

Train-Test Split

Utilizing the train_test_split function from the sklearn.model_selection module, we partitioned the data into training and testing sets. The predictor variables were separated from the target variable, with the split ratio configured at 70% for training and 30% for testing. The random state was set to 42.

Logistic Regression

Logistic Regression was chosen as the modeling approach due to its suitability for binary classification tasks, precisely the prediction of heart disease presence or absence in this context. We got an accuracy score of 83.33% indicating that the model accurately predicted the outcome in the test set for 83.33% of the cases. This high accuracy suggests that the model effectively captured the underlying patterns in the dataset, demonstrating its ability to discern between individuals with and without heart disease based on the selected features.

Precision measures how accurately the model predicts 'heart disease,' while recall reflects how well it identifies all individuals who truly have the condition. In heart disease prediction, having a high recall is essential. This ensures that the model detects as many actual patients as possible. A missed true case might result in a person not receiving the necessary care, which could lead to serious health consequences. While precision is crucial to minimize false positives, more emphasis is placed on recall to reduce the possibility of missing heart disease cases.

Class	Precision	Recall	F1-score	Support
No Heart Disease	0.83	0.88	0.86	171
Heart Disease	0.83	0.77	0.70	129

K-Nearest Neighbors

In order to increase our accuracy and precision value we tried a few more models. One of them is the KNN model. As a non-parametric algorithm, KNN employs lazy learning, meaning it does not make strong assumptions about the underlying distribution of the data. We got an accuracy score of 87.3%. KNN outperformed Logistic Regression due to its ability to handle non-linear decision boundaries.

Class	Precision	Recall	F1-score	Support
No Heart Disease	0.87	0.92	0.89	171
Heart Disease	0.88	0.81	0.85	129

Random Forest

Model Performance Analysis

The Random Forest model was created and its ability to predict heart disease was assessed. Using learning curves and a confusion matrix, the model's performance was examined.

Confusion Matrix Insights: With only one false negative, the model accurately predicted 171 cases of "No Heart Disease" and 128 cases of "Heart Disease," demonstrating outstanding performance. This shows how well the model can classify the data with a low error rate.

Accuracy of Cross-Validation: The model's cross-validation accuracy of 99.8% indicates that it consistently performs well across various data subsets, showing its robustness.

Overfitting Analysis

To make sure the training data was not being overfitted by the model we split the accuracies:

Training Accuracy: The model's perfect fit to the training data was demonstrated by the training accuracy, which was constantly 100%.
Validation Accuracy: Across a range of training set sizes, validation accuracy stayed relatively similar to training accuracy, suggesting that the model is not overfitting and that it generalises well to unknown data.

Feature Importance Analysis
The feature importance analysis provided valuable insights into the predictors driving the model's performance:
Important Predictors: It was found that age and cholesterol had a greater impact on heart disease prediction than any other characteristic.
Less Influential Elements: It was discovered that lifestyle and demographic variables including gender, stress level, and smoking had little impact on the model's predictions.
The Random Forest algorithm's outstanding accuracy and dependability were a result of its capacity to prioritise these important features while carefully handling less important ones.

RECOMMENDATIONS

- To increase sample size of data to improve representation of people less than the age of 40
- Healthcare providers should evaluate atypical/ typical angina and other chest pain types to prevent misdiagnosis since accuracy is the most critical in medical.
- Insurance companies can develop a better coverage plan for people above the age of 50 by understanding the most impactful factors and accordingly pricing the products.
- Individuals should understand the risks of cholesterol, stress, intake of alcohol and other factors and accordingly strive for a fit and healthy lifestyle to reduce the chances of heart diseases.
- Startup's can leverage this data to build and leverage products that track cholesterol, heart rate, stress levels and much more such as smart watches, smart rings, weighing scales etc.

LIMITATIONS:

Using cluster analysis with the Elbow method, clusters have been formed for heart rate, cholesterol, blood pressure group, and all other factors impacting. However, except for cholesterol, all factors showed a similar number of patients with a heart disease which limits accuracy to predict chances of heart disease. In such scenarios, domain knowledge, general consensus and alternative prediction models provided better accuracy.

REFERENCES:

Mammadov, R. (2024). *Heart Disease prediction*. Kaggle.com. https://www.kaggle.com/datasets/rashadrmammadov/heart-disease-prediction?resource=download

WHO. (2021, June 11). *Cardiovascular Diseases (CVDs)*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)