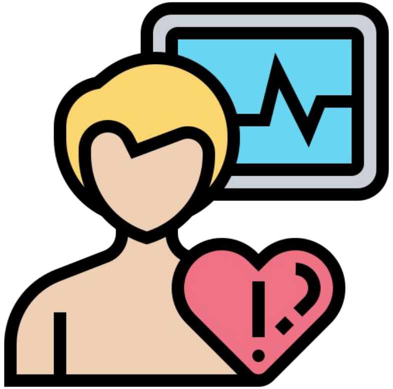
Classifying Cardiovascular Risk:
A Machine Learning Approach using the UCI Heart Disease Dataset



A project for the course

AT82.01 Computer Programming for Data Science and Artificial Intelligence

Group - 10

Tanzil Al Sabah - st123845

Md Shafi Ud Doula - st124047

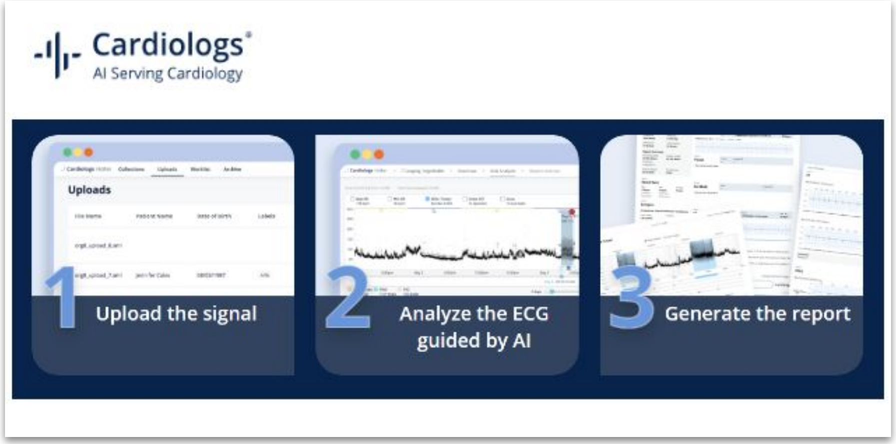
Introduction(1/1)

- **Background:** Heart disease is a global health concern and a leading cause of mortality.
- **Objective:** This project aims to develop a heart disease classification model to aid detection and improve patient outcomes.
- **Relevance:** Such a model can significantly assist healthcare professionals in diagnosing heart conditions promptly, ultimately saving lives. That aligns with the industry's overarching trend of leveraging machine learning for predictive analytics and personalized healthcare.

Related Work (1/2)

Health Monitoring Wearables: Devices like the Apple Watch and Fitbit have incorporated heart rate monitoring, providing insights into heart health. However, their classification algorithms often focus on wellness over medical validation.

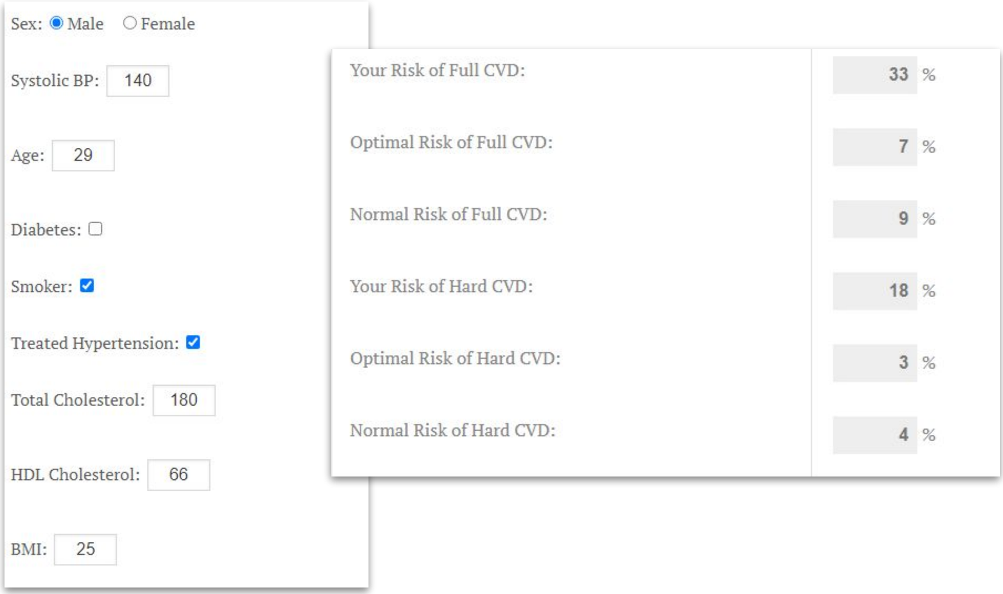
AI-driven Diagnostic Platforms: AI-driven platforms like Zebra Medical Vision and PathAI offer diagnostic insights based on imaging and other clinical data. Yet, their broad focus on multiple diseases sometimes leaves room for more specialized, heart-focused solutions.



Related Work(2/2)

Framingham Heart Study: This study introduced the concept of "risk scores," categorizing individuals based on their likelihood of developing CVDs.

However, its age and demographic-specific data might not resonate with today's diverse and global population.



Sex: ● Male ○ Female

Systolic BP: 140

Age: 29

Diabetes: □

Smoker: ☑

Treated Hypertension: ☑

Total Cholesterol: 180

HDL Cholesterol: 66

BMI: 25

Your Risk of Full CVD:	33 %
Optimal Risk of Full CVD:	7 %
Normal Risk of Full CVD:	9 %
Your Risk of Hard CVD:	18 %
Optimal Risk of Hard CVD:	3 %
Normal Risk of Hard CVD:	4 %

**The given data is just for demonstration purposes.*

Problem Statement(1/1)

- Current diagnostic methodologies for cardiovascular diseases rely heavily on invasive procedures, leading to limitations in early detection and intervention.

- Existing classification tools do not offer dedicated, medically-validated solutions for cardiovascular risk assessment.

- Legacy datasets may not comprehensively represent the diverse risk factors and manifestations of modern heart diseases.

Solution Approach and Dataset Selection(1/2)

By focusing on **classification**, this project seeks to address these gaps, leveraging the UCI dataset to create a specialized tool that categorizes individuals based on their cardiovascular risk, facilitating timely and appropriate medical interventions.

Advantages of the UCI Heart Disease Dataset:

1. **Rich Feature Set:** Offers a wide array of clinical parameters, aiding in comprehensive risk assessment.
2. **Historical Data:** Provides past patient records, essential for training accurate predictive models.
3. **Diverse Data Points:** Ensures the developed model is broad in its applicability, catering to various demographics.
4. **Feature Importance Analysis:** Allows for determining key clinical parameters that heavily influence heart disease risk.
5. **Benchmarking & Comparison:** Being a widely-recognized dataset, it enables comparison with existing research, highlighting the project's significance.
6. **Accelerated Prototyping:** Being pre-curated, it reduces preprocessing time, expediting model development.

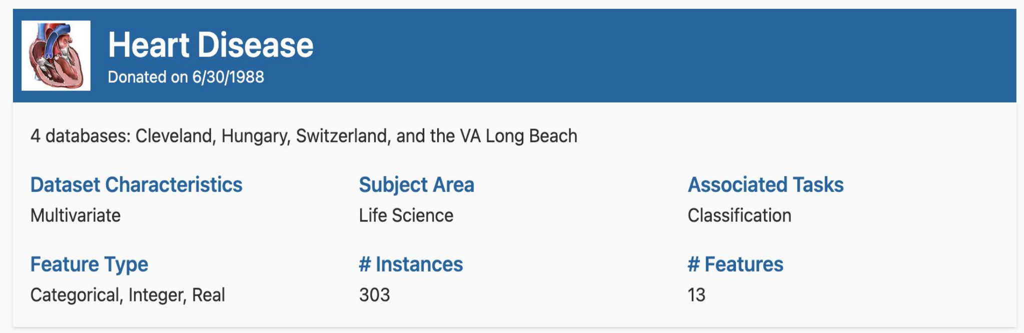
Solution Approach and Dataset Selection (1/2)

Data Source: Heart Disease UCI Dataset

Data Characteristics:

- The dataset comprises 76 attributes, with a significant focus on a subset of 14 key attributes.
- The dataset includes information such as age, sex, cholesterol levels, chest pain type, and other relevant patient attributes.
- The target attribute of interest is "goal," which indicates the presence of heart disease, represented as an integer ranging from 0 (no presence) to 4.

Training and Evaluation: The dataset forms the foundation for training and evaluating the machine learning model.



Heart Disease

Donated on 6/30/1988

4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Life Science	Classification
Feature Type	**# Instances**	**# Features**
Categorical, Integer, Real	303	13

Figure-: UCI Dataset Summary

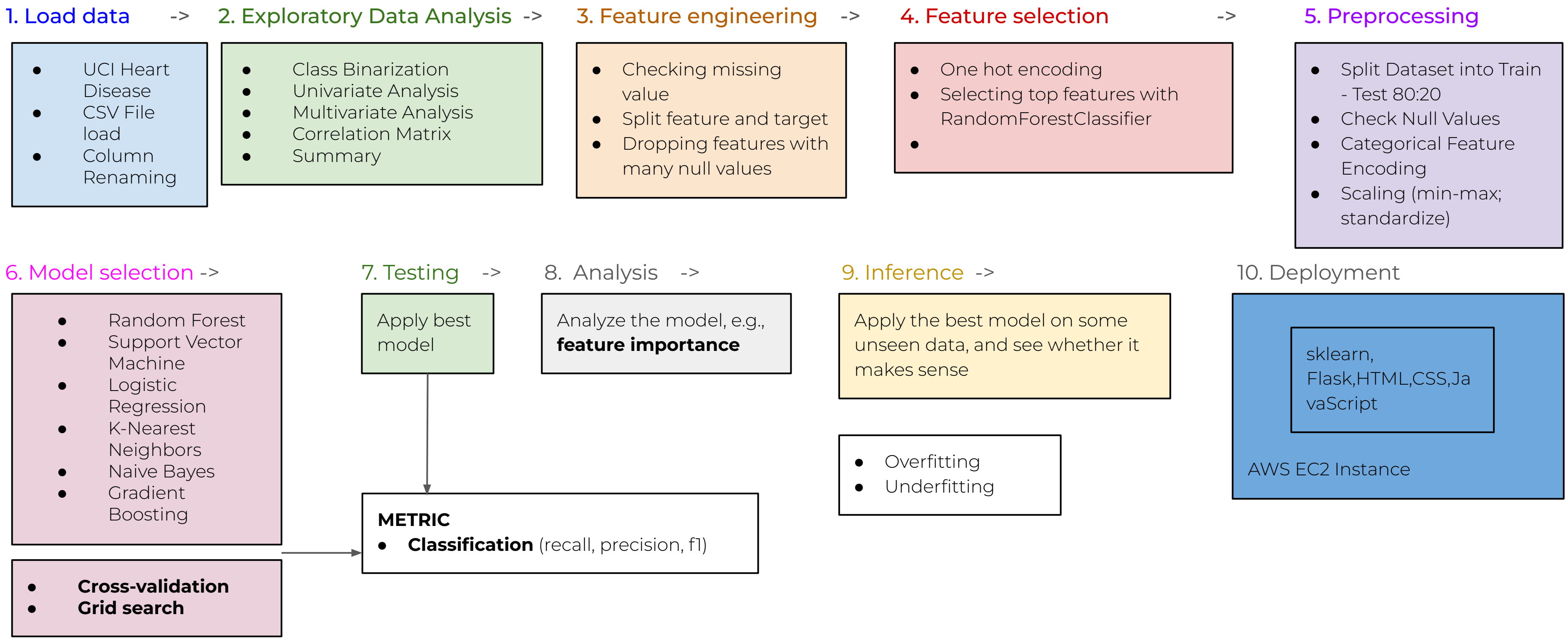
Dataset URL: https://archive.ics.uci.edu/dataset/45/heart+disease

Methodology (1/10)

The project will follow a structured methodology, including:

- **Exploratory Data Analysis (EDA):** This phase involves understanding the dataset, visualizing key features, and identifying patterns and outliers.
- **Feature Selection:** Selecting the most informative features for classification.
- **Data Pre-processing:** Data cleaning and feature engineering will be performed to ensure data quality and relevance.
- **Model Development:** Creating machine learning models, including decision trees, random forests, and deep learning models.
- **Hyperparameter Tuning:** Optimizing model parameters for improved accuracy.
- **Cross-validation:** Ensuring model robustness and generalization.
- **Model Evaluation:** Evaluating model performance using metrics like accuracy, precision, recall, and F1-score.

Methodology(2/10)



1. Load data ->

- UCI Heart Disease
- CSV File load
- Column Renaming

2. Exploratory Data Analysis ->

- Class Binarization
- Univariate Analysis
- Multivariate Analysis
- Correlation Matrix
- Summary

3. Feature engineering ->

- Checking missing value
- Split feature and target
- Dropping features with many null values

4. Feature selection ->

- One hot encoding
- Selecting top features with RandomForestClassifier
-

5. Preprocessing

- Split Dataset into Train - Test 80:20
- Check Null Values
- Categorical Feature Encoding
- Scaling (min-max; standardize)

6. Model selection ->

- Random Forest
- Support Vector Machine
- Logistic Regression
- K-Nearest Neighbors
- Naive Bayes
- Gradient Boosting

- **Cross-validation**
- **Grid search**

7. Testing ->

Apply best model

METRIC

- **Classification** (recall, precision, f1)

8. Analysis ->

Analyze the model, e.g., **feature importance**

9. Inference ->

Apply the best model on some unseen data, and see whether it makes sense

- Overfitting
- Underfitting

10. Deployment

sklearn, Flask,HTML,CSS,JavaScript

AWS EC2 Instance

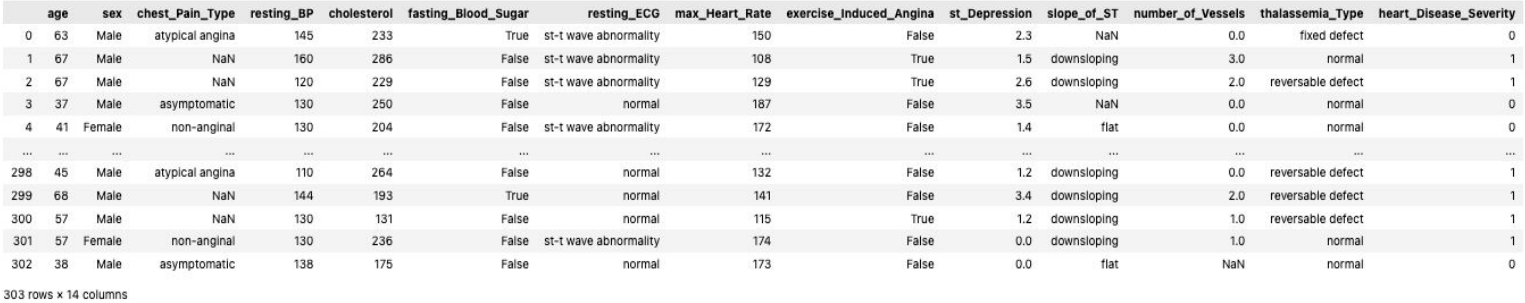
Process flow of Methodology

Methodology: Exploratory Data Analysis (3/10)

Dataset Columns:

* id (Unique id for each patient)
* age (Age of the patient in years)
* origin (place of study)
* sex (Male/Female)
* cp chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
* trestbps resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
* chol (serum cholesterol in mg/dl)
* fbs (if fasting blood sugar > 120 mg/dl)
* restecg (resting electrocardiographic results)
* Values: [normal, stt abnormality, lv hypertrophy]
* thalach: maximum heart rate achieved
* exang: exercise-induced angina (True/ False)
* oldpeak: ST depression induced by exercise relative to rest
* slope: the slope of the peak exercise ST segment
* ca: number of major vessels (0-3) colored by fluoroscopy
* thal: [normal; fixed defect; reversible defect]
* num: the predicted attribute

Methodology: Exploratory Data Analysis (4/10)

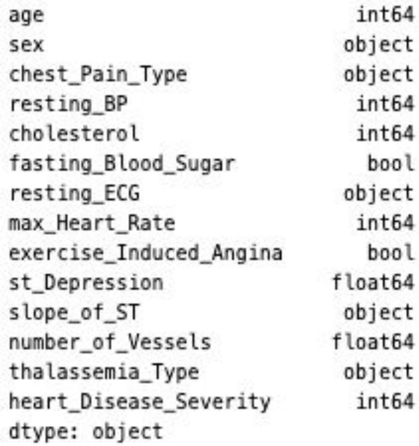


	age	sex	chest_Pain_Type	resting_BP	cholesterol	fasting_Blood_Sugar	resting_ECG	max_Heart_Rate	exercise_Induced_Angina	st_Depression	slope_of_ST	number_of_Vessels	thalassemia_Type	heart_Disease_Severity
0	63	Male	atypical angina	145	233	True	st-t wave abnormality	150	False	2.3	NaN	0.0	fixed defect	0
1	67	Male	NaN	160	286	False	st-t wave abnormality	108	True	1.5	downsloping	3.0	normal	1
2	67	Male	NaN	120	229	False	st-t wave abnormality	129	True	2.6	downsloping	2.0	reversable defect	1
3	37	Male	asymptomatic	130	250	False	normal	187	False	3.5	NaN	0.0	normal	0
4	41	Female	non-anginal	130	204	False	st-t wave abnormality	172	False	1.4	flat	0.0	normal	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	45	Male	atypical angina	110	264	False	normal	132	False	1.2	downsloping	0.0	reversable defect	1
299	68	Male	NaN	144	193	True	normal	141	False	3.4	downsloping	2.0	reversable defect	1
300	57	Male	NaN	130	131	False	normal	115	True	1.2	downsloping	1.0	reversable defect	1
301	57	Female	non-anginal	130	236	False	st-t wave abnormality	174	False	0.0	downsloping	1.0	normal	1
302	38	Male	asymptomatic	138	175	False	normal	173	False	0.0	flat	NaN	normal	0

303 rows × 14 columns

Figure-: Data in panda dataframe

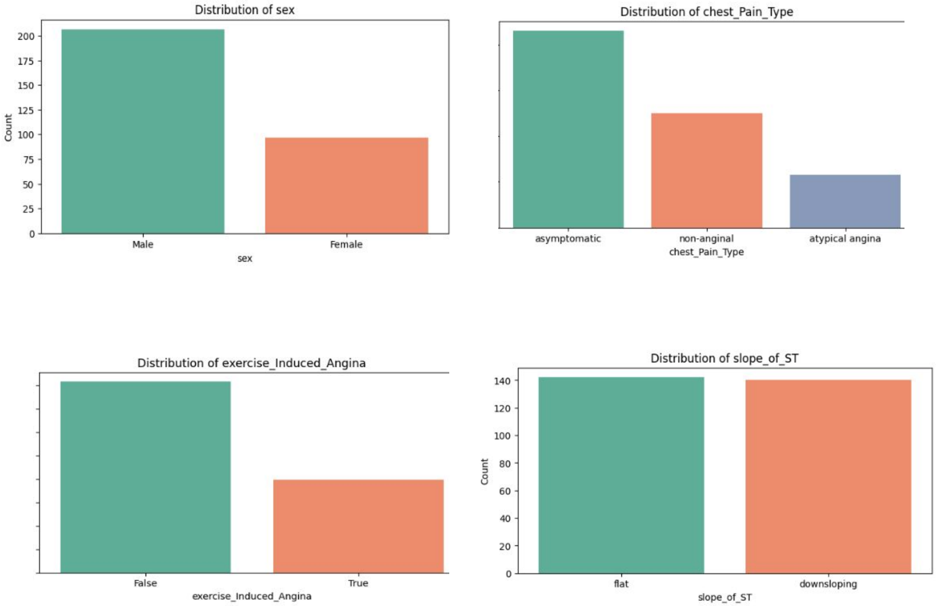
- **Number of Sample:** 303
- **Number of Columns:** 14
- **Number of Numerical Columns:** 7
- **Number of Boolean Columns:** 2
- **Number of Categorical Columns:** 5



```
age                          int64
sex                         object
chest_Pain_Type             object
resting_BP                   int64
cholesterol                  int64
fasting_Blood_Sugar           bool
resting_ECG                 object
max_Heart_Rate               int64
exercise_Induced_Angina       bool
st_Depression              float64
slope_of_ST                 object
number_of_Vessels          float64
thalassemia_Type            object
heart_Disease_Severity       int64
dtype: object
```

Figure-: Data types of each attributes

Methodology: Exploratory Data Analysis (5/11)



Sex: More males than females in the dataset.

Chest Pain Type: A significant number of patients report asymptomatic or silent(symptompless) chest pain.

Exercise-Induced Angina: A considerable number of individuals do not experience angina induced by exercise.

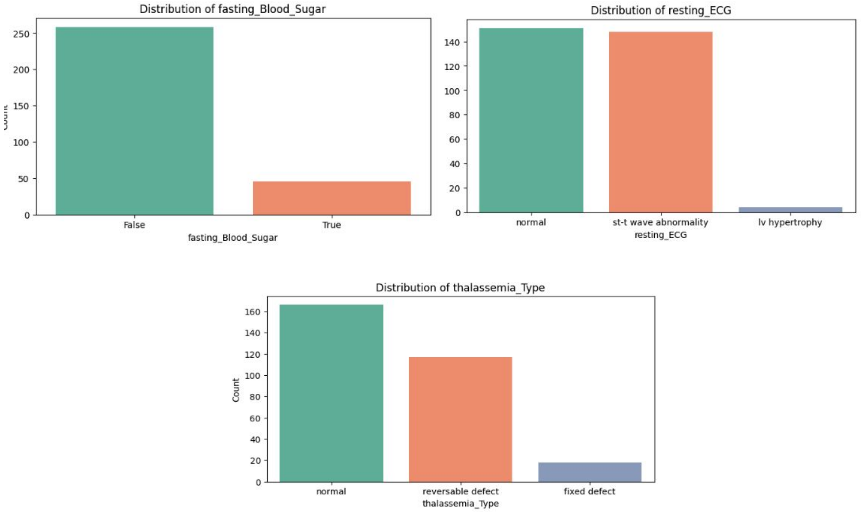
Univariate Analysis of some categorical features

For Details: preliminary_EDA_Model_Notebook.ipynb

Methodology: Exploratory Data Analysis (5/11)

Fasting Blood Sugar: True for most individuals. That means most individuals have fasting blood sugar below 120 mg/dL.

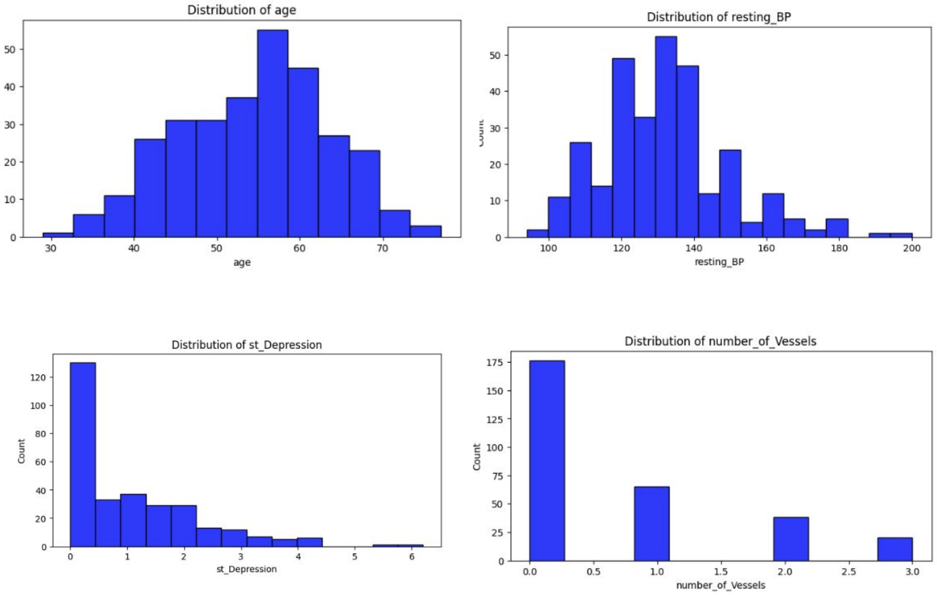
Resting ECG: The number of people reporting normal ECG is similar to the number of people reporting some abnormalities.



Univariate Analysis of some categorical features

For Details: preliminary_EDA_Model_Notebook.ipynb

Methodology: Exploratory Data Analysis (6/11)



Univariate Analysis of Numerical Features

Age: Shows a roughly bell-shaped distribution, indicating a wide range of ages in the dataset.

Resting Blood Pressure (trestbps): The distribution is centered around 120-140 mm Hg, which is within the typical range.

st_Depression : Right skewed - Most individuals in the dataset have little to no ST depression. This is typical for a general population without significant heart disease.

Number of vessels : There are non-zero values which might indicate a need for further cardiac evaluation and intervention among these individuals.

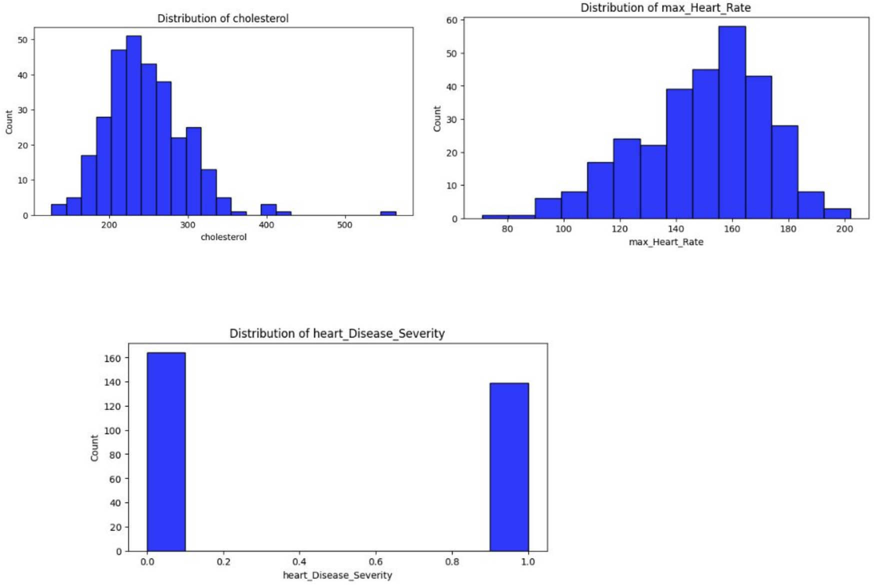
For Details: preliminary_EDA_Model_Notebook.ipynb

Methodology: Exploratory Data Analysis (6/11)

Cholesterol (chol): The data is somewhat right-skewed, indicating a few individuals with very high cholesterol levels.

Maximum Heart Rate (thalach): Exhibits a wide range, with most values clustering around 150-170 bpm.

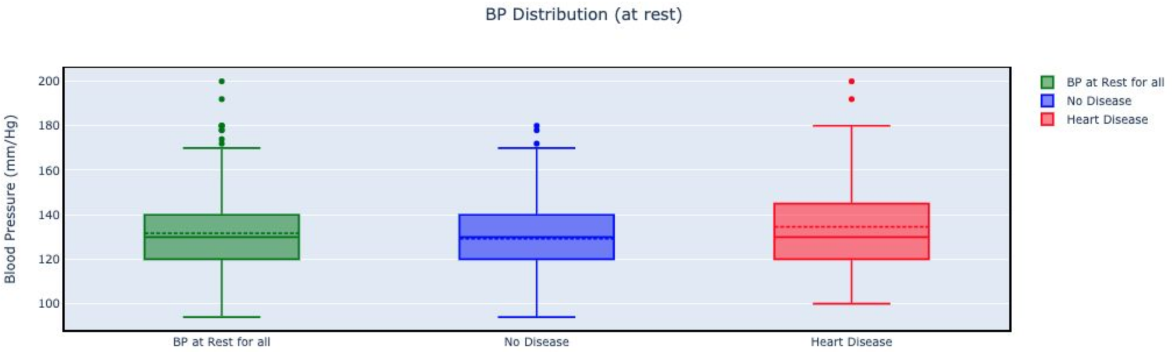
Heart Disease Severity (num): A large number of individuals have no heart disease, but there's also a substantial number with varying degrees of severity. (Here we have taken all the severities as 1)



Univariate Analysis of Numerical Features

For Details: preliminary_EDA_Model_Notebook.ipynb

Methodology: Exploratory Data Analysis (7/11)



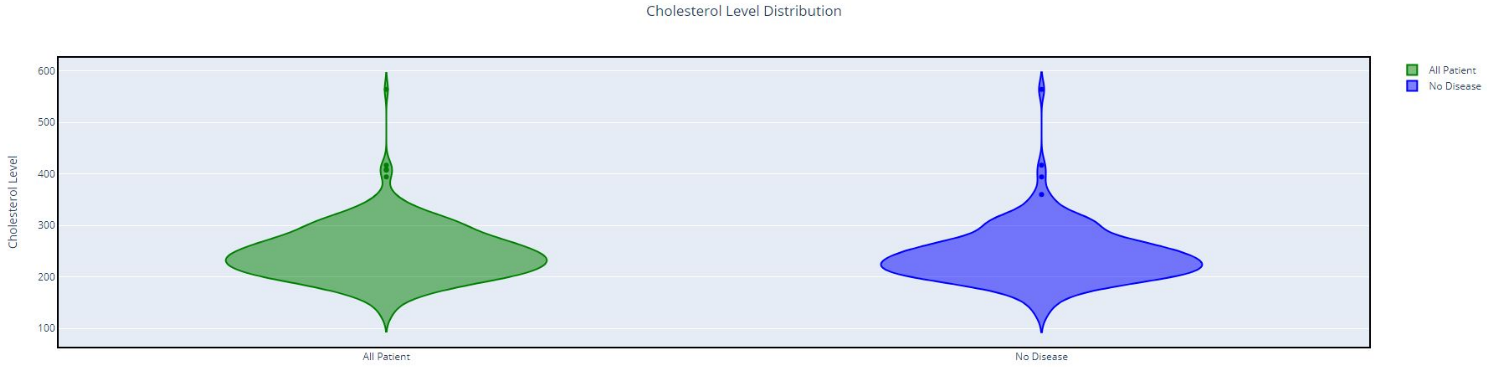
This boxplot comparison shows the distribution of resting blood pressure among all patients, those with no heart disease, and those with heart disease:

- **All Patients (Green Boxplot):** The median resting blood pressure for all patients is around 130 mmHg. The interquartile range (IQR) spans from approximately 120 to 140 mmHg, indicating that half of the patients' blood pressure falls within this range. There are a few outliers on the higher end, with the maximum blood pressure reaching 200 mmHg.
- **No Disease (Blue Boxplot):** For patients with no heart disease, the median is also around 130 mmHg, similar to the overall median. The IQR is slightly narrower than for all patients, indicating less variability among those without heart disease. The maximum without being an outlier is close to 170 mmHg, and there are a few outliers noted above this value.
- **Heart Disease (Red Boxplot):** Patients with heart disease show a median resting blood pressure a bit lower than the overall group, at around 130 mmHg. The IQR for this group is broader, with the lower quartile at 120 mmHg and the upper quartile at 145 mmHg. There are several extreme outliers, with the highest recorded blood pressure at 200 mmHg.

Some of the Multivariate Analysis of features with their distribution

For Details: preliminary_EDA_Model_Notebook.ipynb

Methodology: Exploratory Data Analysis (7/11)

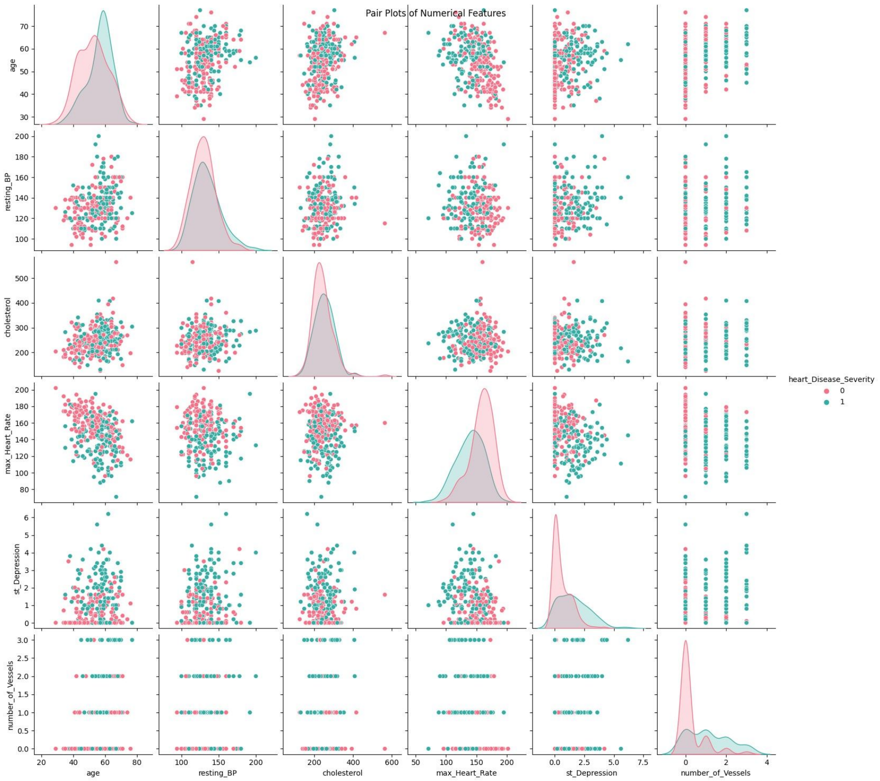


- **All Patients (Green Violin Plot):** This plot likely represents the cholesterol level distribution across the entire patient dataset. It shows a wide range of cholesterol levels with a median around 240 mg/dL, which is close to the borderline high level according to various health organizations. The plot also shows a long tail towards higher cholesterol levels, indicating that some patients have significantly high cholesterol.
- **Patients with No Disease (Blue Violin Plot):** The distribution for patients with no heart disease shows a median slightly lower than that of the general population, around 235 mg/dL. The distribution is less spread out than in the overall patient population, indicating less variability in cholesterol levels among patients without heart disease. The bulk of the distribution is between approximately 200 mg/dL and 300 mg/dL.

Some of the Multivariate Analysis of features with their distribution

For Details: preliminary_EDA_Model_Notebook.ipynb

Methodology: Exploratory Data Analysis (8/11)



Key Takeaways:

The left-skewed distribution for individuals with heart disease suggests that younger individuals are less frequently represented in the heart disease group. Conversely, older individuals are more commonly affected by heart disease, which aligns with medical understanding that the risk of heart disease generally increases with age.

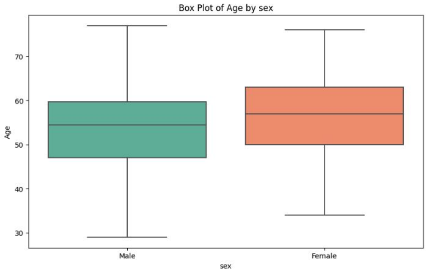
Cholesterol levels are seen to increase with age in the dataset, and higher cholesterol levels are associated with more heart disease reports among older individuals, this reflects a well-established risk factor for heart disease. Elevated cholesterol can lead to atherosclerosis, a condition that narrows and hardens arteries, potentially leading to heart disease.

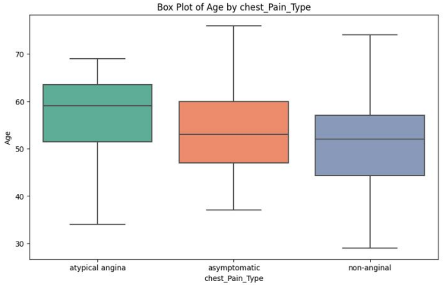
A significant maximum heart rate is generally a sign of better cardiovascular fitness. The observation that individuals with higher maximum heart rates tend to report less heart disease could suggest that better cardiovascular health, reflected by the ability to reach higher heart rates, is protective against heart disease. This could also be related to physical activity levels, as more active individuals often have a higher maximum heart rate.

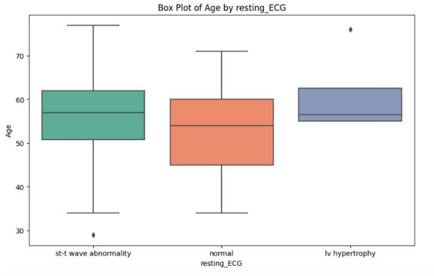
Pair plots were used to visualize the relationships between pairs of numerical features.

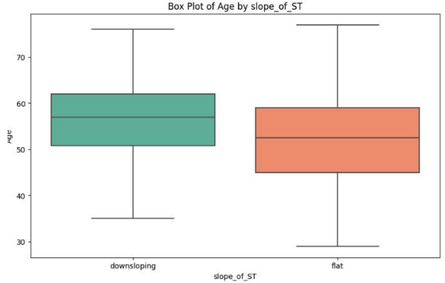
For Details: preliminary_EDA_Model_Notebook.ipynb

Methodology: Exploratory Data Analysis (9/11)









Age by Sex
The age distribution for males appears to have a slightly higher median age compared to females. Both distributions have a similar range of ages, as indicated by the length of the boxes and whiskers. Neither group shows extreme outliers, suggesting that the ages are fairly normally distributed within the biological sex categories.

Age by Slope of ST
Individuals with a flat slope tend to be older than those with a downsloping ST segment, as indicated by the higher median in the "flat" category. The age range is fairly similar across both categories.

Age by Resting ECG
Individuals with ST-T wave abnormality and normal ECG results have similar median ages, but the ST-T wave abnormality group has a tighter age range. Those categorized with LV hypertrophy appear to have a higher median age and the age range is narrower compared to the other two categories.

Age by Chest Pain Type
In terms of chest pain type, those with asymptomatic chest pain are older on average than those with atypical angina or non-anginal pain.

Multivariate Analysis (visualize the distribution of age with respect to different categorical features)

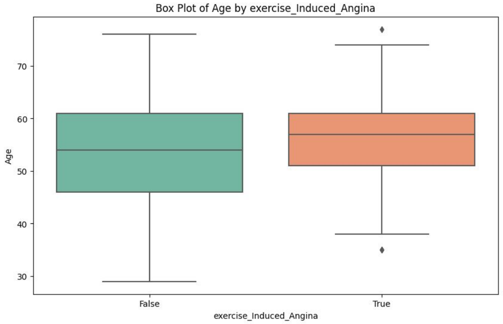
For Details: preliminary_EDA_Model_Notebook.ipynb

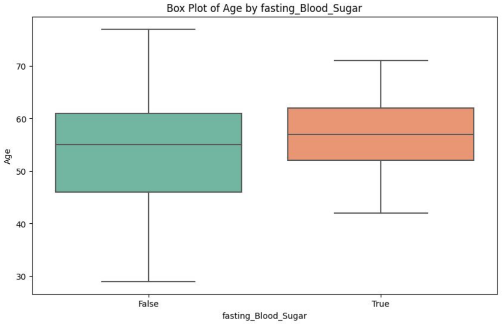
Methodology: Exploratory Data Analysis (9/11)

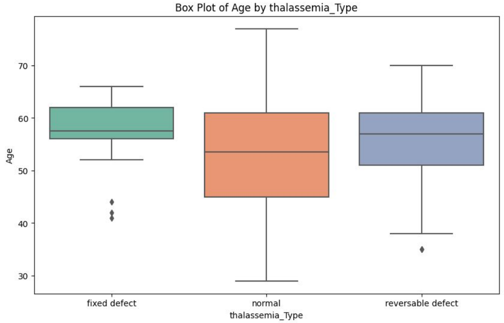
Age by Exercise Induced Angina
Those who experience exercise-induced angina appear to be older on average, as the median of the "True" category is higher. Both categories show a similar range of ages, but the "True" category has outliers on the lower age end

Age by Fasting Blood Sugar
Those with fasting blood sugar above the threshold tend to be older, as seen by the higher median age in the "True" category. There is a wider age range among those with higher fasting blood sugar, indicated by the taller box and longer whiskers.

Age by Thalassemia
For thalassemia, those with a fixed defect have a lower median age compared to the normal and reversible defect categories.



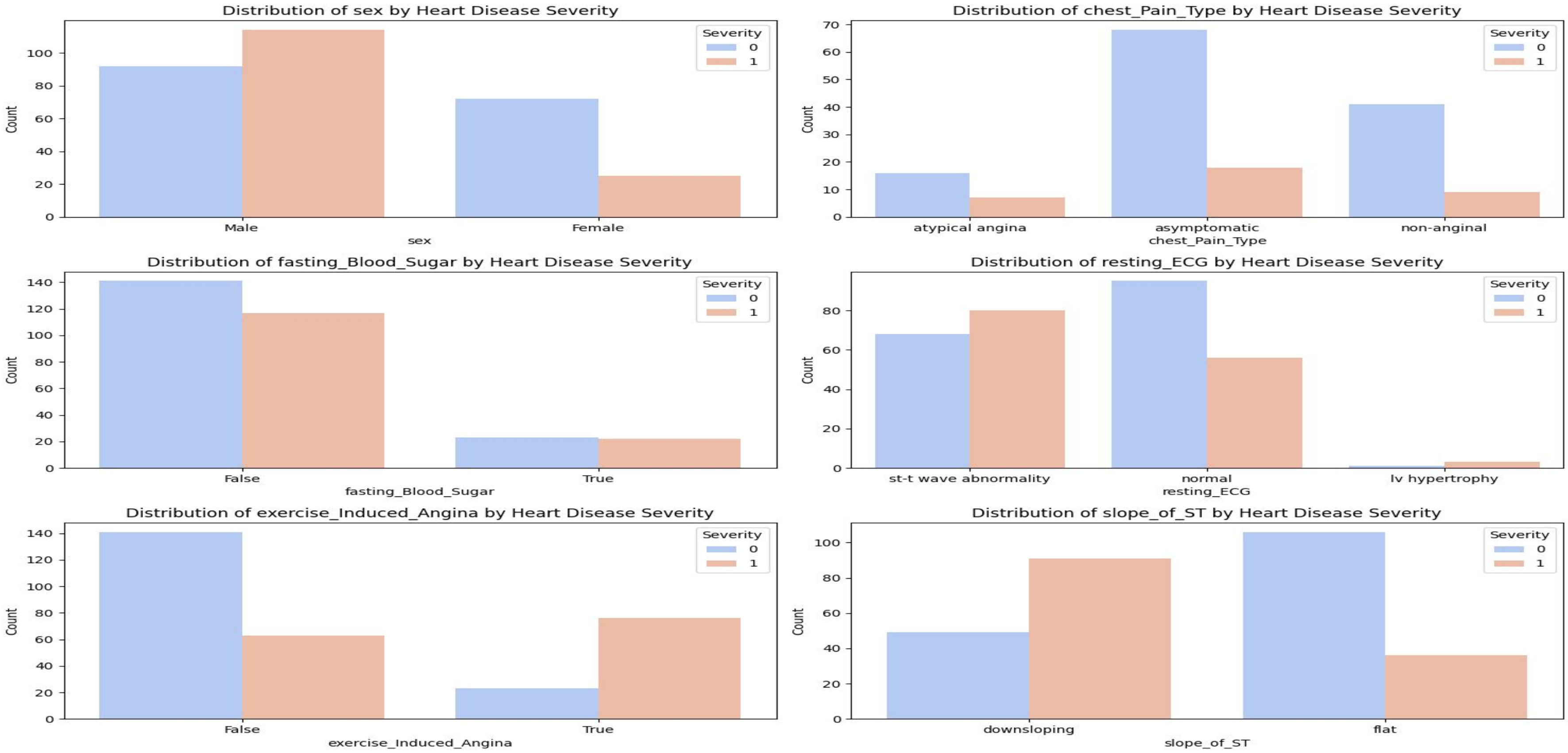




Multivariate Analysis (visualize the distribution of age with respect to different categorical features)

For Details: preliminary_EDA_Model_Notebook.ipynb

Methodology: Exploratory Data Analysis (10/11)



Multivariate Analysis (distribution of various features with Heart Disease severity, 0 means absence of heart disease, 1 means possibility to have heart disease.

For Details: preliminary_EDA_Model_Notebook.ipynb

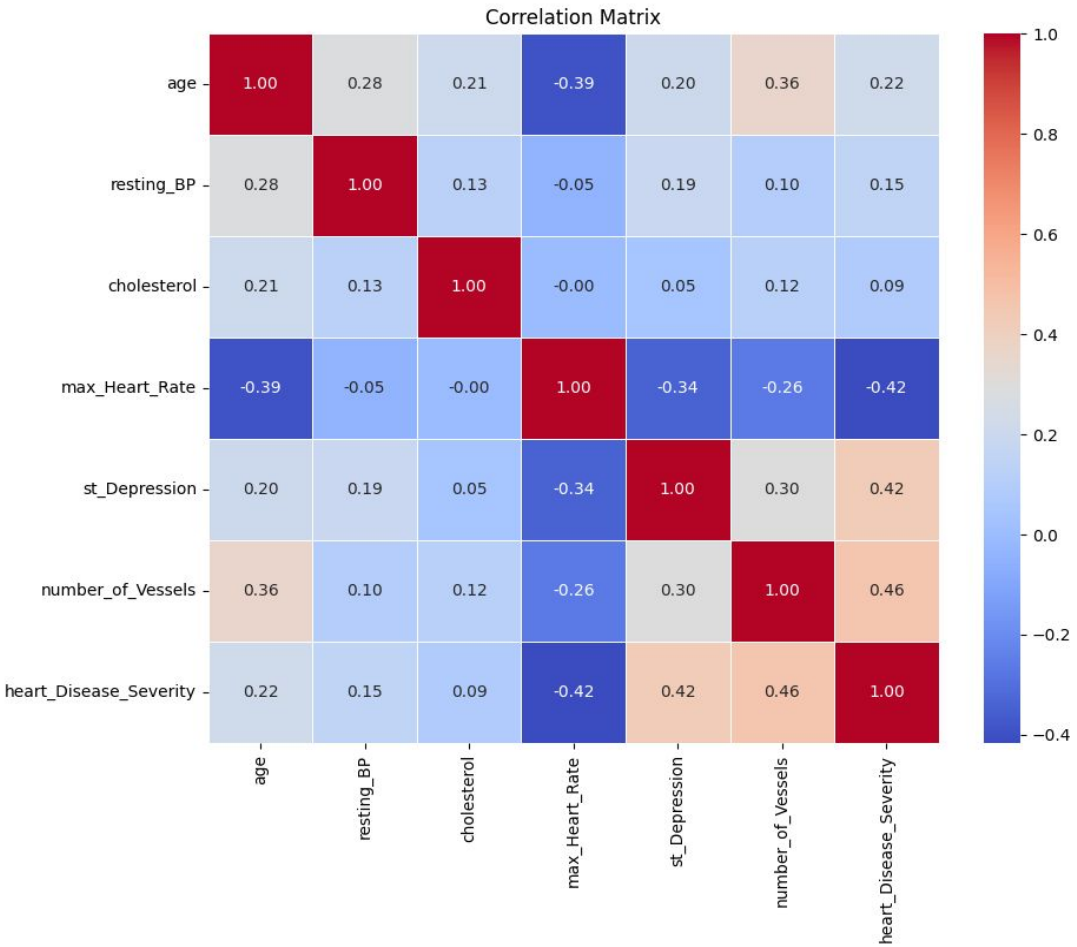
Methodology: Exploratory Data Analysis (10/11)

- In the dataset, males report heart disease more often than females, possibly due to genetic, lifestyle, and societal factors.
- Asymptomatic chest pain is common among both healthy individuals and those with heart disease, highlighting the need for in-depth diagnostics.
- Normal fasting blood sugar levels are prevalent in both groups, suggesting that high blood sugar may not be a direct heart disease indicator, possibly due to effective diabetes control.
- Resting ECG abnormalities are more common in heart disease patients, as these can signal conditions like ischemia.
- Individuals with exercise-induced angina, a symptom of coronary artery disease, frequently report heart disease.
- A downsloping ST segment is traditionally seen as normal but is linked to heart disease in this dataset, suggesting it may still pose a risk. Conversely, a flat ST segment is not solely indicative of heart disease.
- Lastly, fixed or reversible defect thalassemia types are more prevalent in heart disease cases, which may correlate with increased risk due to their impact on hemoglobin production and cardiac workload.

Multivariate Analysis (distribution of various features with Heart Disease severity, 0 means absence of heart disease, 1 means possibility to have heart disease.

For Details: preliminary_EDA_Model_Notebook.ipynb

Methodology: Exploratory Data Analysis (Correlation Matrix) (11/11)



Correlation Matrix

	age	resting_BP	cholesterol	max_Heart_Rate	st_Depression	number_of_Vessels	heart_Disease_Severity
age	1.00	0.28	0.21	-0.39	0.20	0.36	0.22
resting_BP	0.28	1.00	0.13	-0.05	0.19	0.10	0.15
cholesterol	0.21	0.13	1.00	-0.00	0.05	0.12	0.09
max_Heart_Rate	-0.39	-0.05	-0.00	1.00	-0.34	-0.26	-0.42
st_Depression	0.20	0.19	0.05	-0.34	1.00	0.30	0.42
number_of_Vessels	0.36	0.10	0.12	-0.26	0.30	1.00	0.46
heart_Disease_Severity	0.22	0.15	0.09	-0.42	0.42	0.46	1.00

- Age has a positive correlation of 0.223 with heart disease severity, indicating a mild positive relationship.
- Resting BP (Blood Pressure) has a positive correlation of 0.158 with heart disease severity.
- Cholesterol has a positive correlation of 0.071 with heart disease severity.
- Max Heart Rate has a negative correlation of -0.415 with heart disease severity, indicating a moderate negative relationship.
- ST Depression has a positive correlation of 0.504 with heart disease severity, indicating a strong positive relationship.
- Number of Vessels has a positive correlation of 0.519 with heart disease severity, indicating a strong positive relationship.

For Details: preliminary_EDA_Model_Notebook.ipynb

Feature Selection(1/1)

Depending on EDA and Feature Importance, Selected Features are-

1. **'max_Heart_Rate',**
2. **'st_Depression',**
3. **'age',**
4. **'cholesterol',**
5. **'thalassemia_Type_normal',**
6. **'resting_BP'**
7. **chest_Pain**

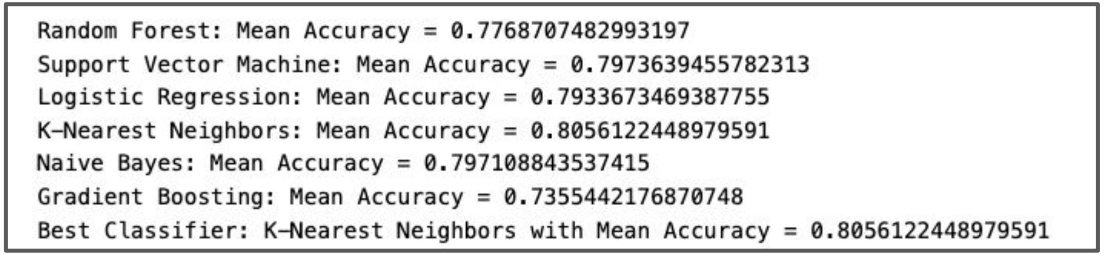
For Details: preliminary_EDA_Model_Notebook.ipynb

Preprocessing(1/1)

1. Split Dataset into Train - Test 80:20
2. Check Null Values and imputing the feature with values
   a. **'chest_Pain_Type': 'mode',**
   b. **'slope_of_ST': 'mode',**
   c. **'thalassemia_Type': 'mode'**
3. Categorical Feature Encoding
   a. One-hot Encoding.
4. Scaling (min-max; standardize)

For Details: preliminary_EDA_Model_Notebook.ipynb

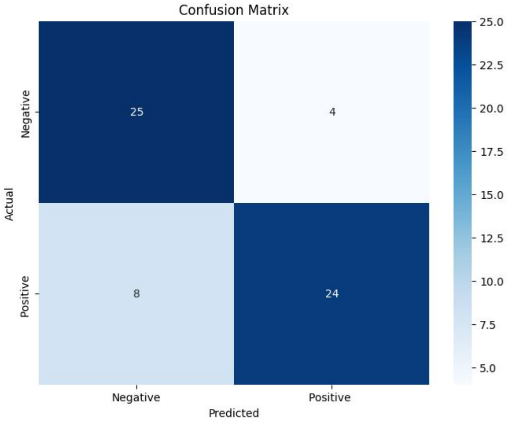
Results(1/1)

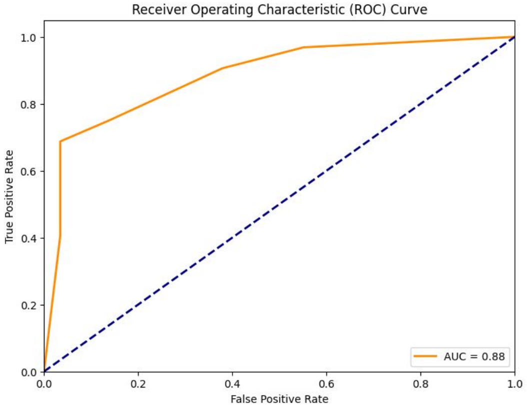


```
Random Forest: Mean Accuracy = 0.7768707482993197
Support Vector Machine: Mean Accuracy = 0.7973639455782313
Logistic Regression: Mean Accuracy = 0.7933673469387755
K-Nearest Neighbors: Mean Accuracy = 0.8056122448979591
Naive Bayes: Mean Accuracy = 0.797108843537415
Gradient Boosting: Mean Accuracy = 0.7355442176870748
Best Classifier: K-Nearest Neighbors with Mean Accuracy = 0.8056122448979591
```

- **Optimal Classifier:** The K-Nearest Neighbors (KNN) model emerged as the best-performing classifier.
- **Cross-Validation Performance:** The KNN model achieved a mean accuracy of 80.56% during cross-validation, demonstrating its robustness.
- **Performance Metrics:** Precision, recall, and F1 score were used to evaluate the model's balanced performance.

  - **Precision: 0.81**
  - **Recall: 0.80**
  - **F1 Score: 0.80**





For Details: preliminary_EDA_Model_Notebook.ipynb

Deployment(1/4)

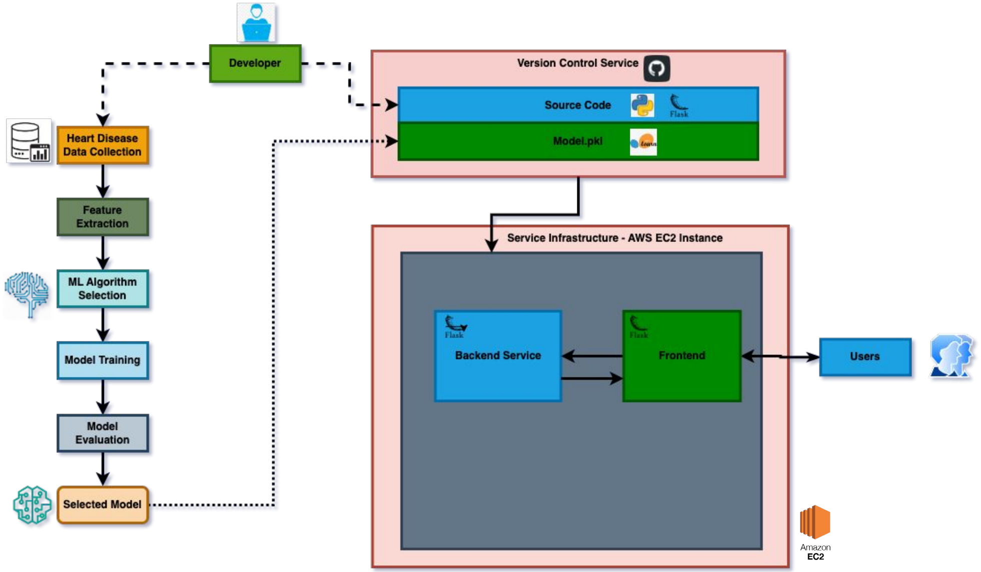
Flask

Flask is an open-source framework and deploying with Flask .To deploy using Flask, it's essential to ensure the repository follows this specific file structure:

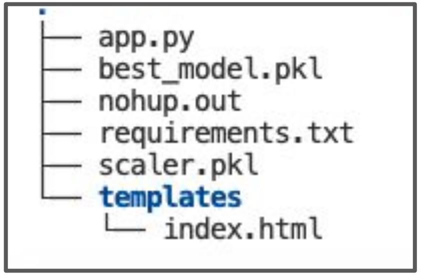
- **app.py:** This Python script serves as the core of our application and comprises the following functions-
    - Loads the scikit-learn model.
    - Defines the inference function.
    - Specifies input and output validation rules.
    - Handles preprocessing and postprocessing of input and output.
- **requirements.txt:** This text file includes a list of libraries required to run the application, such as scikit-learn, numpy, and others.
- **templates/index.html:** In Flask, the UI components are defined in an HTML file using HTML, CSS, and JavaScript.
- **best_model.pkl :** This is the scikit-learn model that has been trained in the previous steps and is prepared for deployment after rigorous testing.
- **scaler.pkl:** This pickle file contains the scaling ratio that we used to process numerical features during the training time.

Amazon EC2

- The application has been deployed in Amazon EC2 Instance.



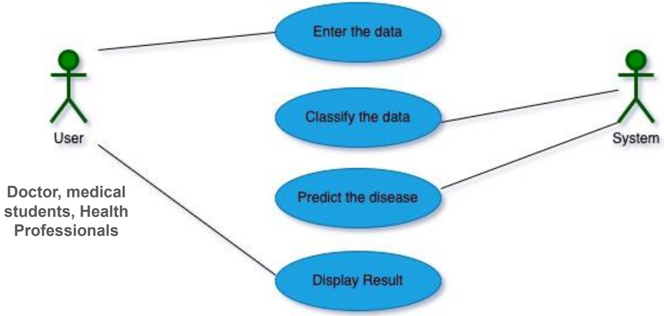
System Architecture



```
├── app.py
├── best_model.pkl
├── nohup.out
├── requirements.txt
├── scaler.pkl
└── templates
    └── index.html
```

Deployment Codebase File Structure

Deployment(2/4)

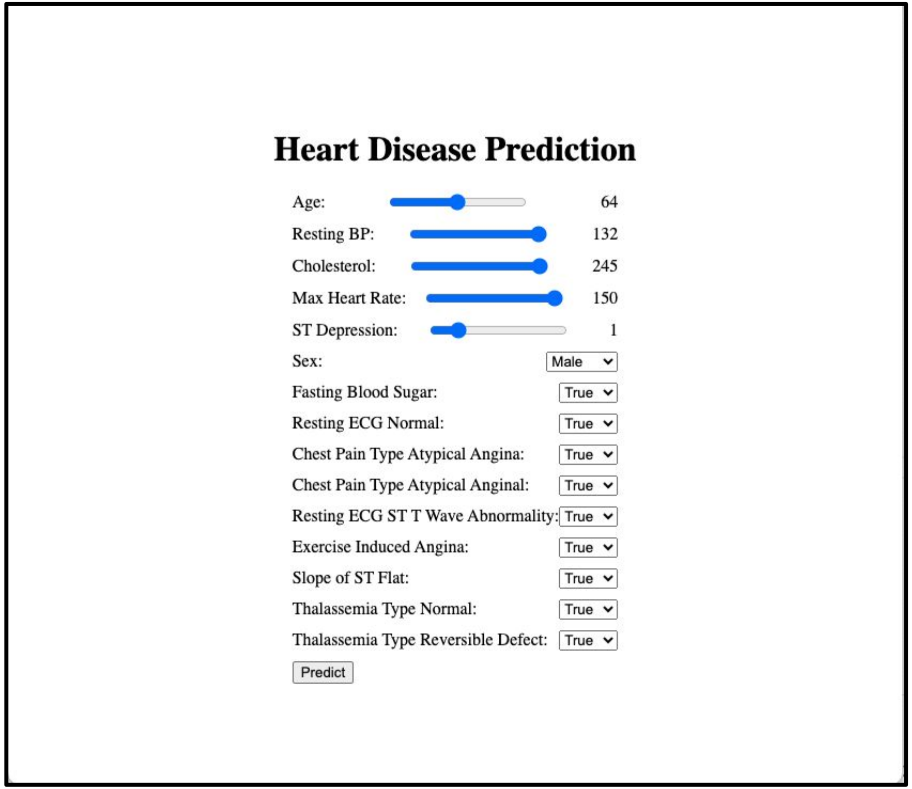


Use Case

- **Application URL:** http://ec2-44-216-245-219.compute-1.amazonaws.com:8080/

Default Value Setting Criterion of the Application

- **Numerical Features:**
  - Initialized using mean values for balance.
- **Data Range:**
  - Adjusted based on our dataset for realism.
- **Categorical Features:**
  - Set to maximum category for each feature.



Heart Disease Prediction

Age: 64

Resting BP: 132

Cholesterol: 245

Max Heart Rate: 150

ST Depression: 1

Sex: Male

Fasting Blood Sugar: True

Resting ECG Normal: True

Chest Pain Type Atypical Angina: True

Chest Pain Type Atypical Anginal: True

Resting ECG ST T Wave Abnormality: True

Exercise Induced Angina: True

Slope of ST Flat: True

Thalassemia Type Normal: True

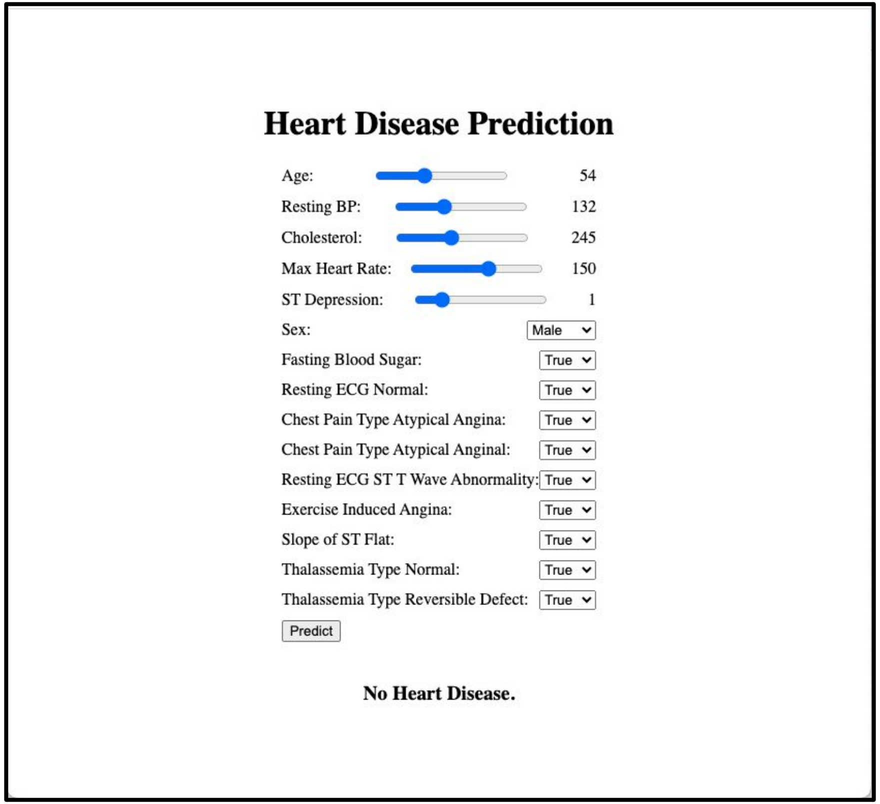
Thalassemia Type Reversible Defect: True

Application UI

Deployment(3/4)

- **Application URL:** **http://ec2-44-216-245-219.compute-1.amazonaws.com:8080/**

Sample Prediction for No Heart Disease →



Heart Disease Prediction

Age: 54

Resting BP: 132

Cholesterol: 245

Max Heart Rate: 150

ST Depression: 1

Sex: Male

Fasting Blood Sugar: True

Resting ECG Normal: True

Chest Pain Type Atypical Angina: True

Chest Pain Type Atypical Anginal: True

Resting ECG ST T Wave Abnormality: True

Exercise Induced Angina: True

Slope of ST Flat: True

Thalassemia Type Normal: True

Thalassemia Type Reversible Defect: True

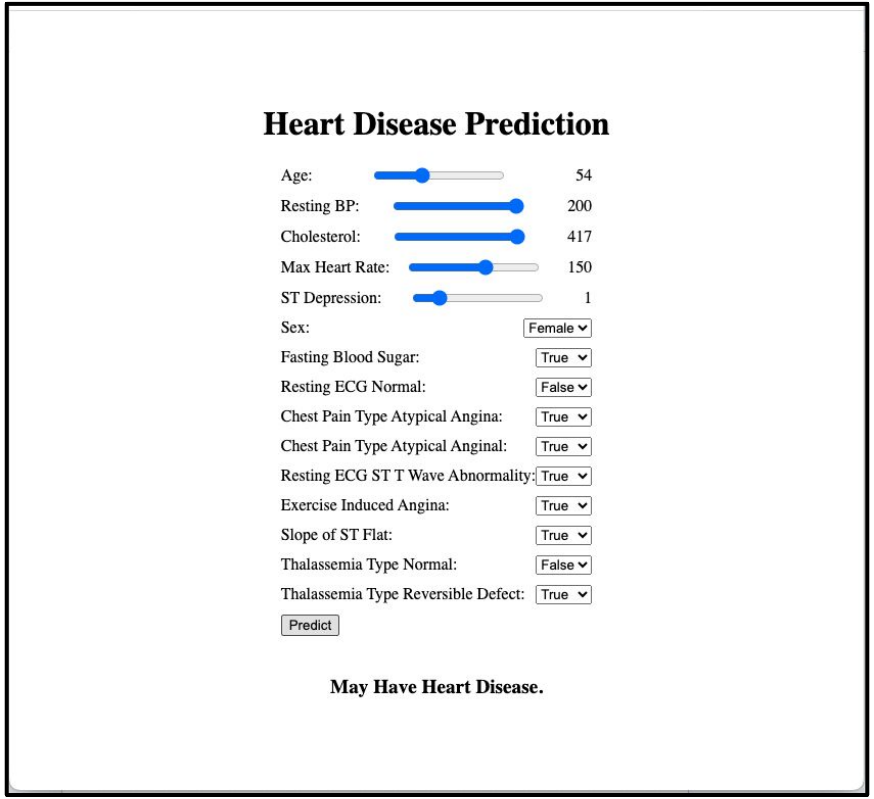
No Heart Disease.

Application UI

Deployment(4/4)

- **Application URL:** http://ec2-44-216-245-219.compute-1.amazonaws.com:8080/

Sample Prediction for May Have Heart Disease →



Heart Disease Prediction

Age:	54
Resting BP:	200
Cholesterol:	417
Max Heart Rate:	150
ST Depression:	1
Sex:	Female
Fasting Blood Sugar:	True
Resting ECG Normal:	False
Chest Pain Type Atypical Angina:	True
Chest Pain Type Atypical Anginal:	True
Resting ECG ST T Wave Abnormality:	True
Exercise Induced Angina:	True
Slope of ST Flat:	True
Thalassemia Type Normal:	False
Thalassemia Type Reversible Defect:	True

May Have Heart Disease.

Application UI

Conclusions:

This application serves as a tool for preliminary assessment, potentially aiding healthcare professionals in decision-making processes and risk stratification. The integration of the KNN model into a user-friendly app demonstrates our commitment to making health risk assessments more accessible and data-driven.

Future Works:

- Integrate additional predictive variables and health indicators to enhance the model's accuracy.
- Expand the dataset with more diverse populations to improve the model's generalizability.
- Improve the app's interface based on user feedback to ensure it is intuitive and meets the needs of its target audience.
- Consulting healthcare professionals with **domain knowledge** in heart disease for effective data analysis, clinical decision-making, and building up patient education.