Heart Disease Diagnostic Analysis

Report By: Aditya Chikte

Introduction

Health data analysis is indispensable in modern healthcare, offering invaluable insights into disease trends, treatment efficacy, and population health. By deciphering complex datasets, healthcare professionals can make informed decisions, optimize resource allocation, and drive evidence-based practices.

From enhancing disease surveillance to guiding public health policies, health data analysis plays a pivotal role in advancing healthcare delivery and improving patient outcomes.

Objectives

- Conduct comprehensive exploratory data analysis to understand the distribution of heart disease rates, demographics, and other relevant factors.
- Utilize various visualization techniques to illustrate key insights and trends in heart disease diagnostic data.
- Train machine learning models to predict heart disease diagnosis, evaluate their performance, and select the best-performing model.
- Develop an interactive dashboard to visualize and communicate the analysis results effectively.
- Derive actionable insights from the analysis to inform healthcare decision-making and future preparedness strategies.

Project Overview

A. Dataset Detail:

- Source: The dataset has been provided by Unified Mentors and the file name was “Heart Disease Data.csv”.
- Format: CSV (Comma-Separated Values)
- Size: 1025

B. Domain: Health Care

C. Technology and tools used:

- Python: Used for data preprocessing, analysis, modeling and visualization.
- Google Colab: Utilized as the primary development environment for running Python code

Methodology and Process

- Data Extraction and Loading
- Data Cleaning and Preprocessing
- Exploratory Data Analysis (EDA)
- Model Training and Evaluation
- Dashboard Development
- Insights and Conclusion
- Summary

Data Extraction and Loading

- The dataset has been provided by Unified Mentors in a CSV format.
- By using the **gdown** library, we extracted this dataset file.
- Using the pandas library, the extracted file was loaded into the pandas **dataframe** for further processing.

Data Cleaning and Preprocessing

- The dataset was examined for any missing values, but no missing values were found.
- For numerical columns (‘**age**’, ‘**trestbps**’, ‘**chol**’, ‘**thalach**’, ‘**oldpeak**’), missing values were replaced with the median of each respective column to preserve central tendency and mitigate the impact of outliers.
- For categorical columns (‘**sex**’, ‘**cp**’, ‘**fbs**’, ‘**restecg**’, ‘**exang**’, ‘**slope**’, ‘ca’, ‘thal’), missing values were filled with the mode values to maintain the distribution of categorical variables.

Exploratory Data Analysis (EDA)

- Calculated mean, standard deviation, and quartiles for numerical features.

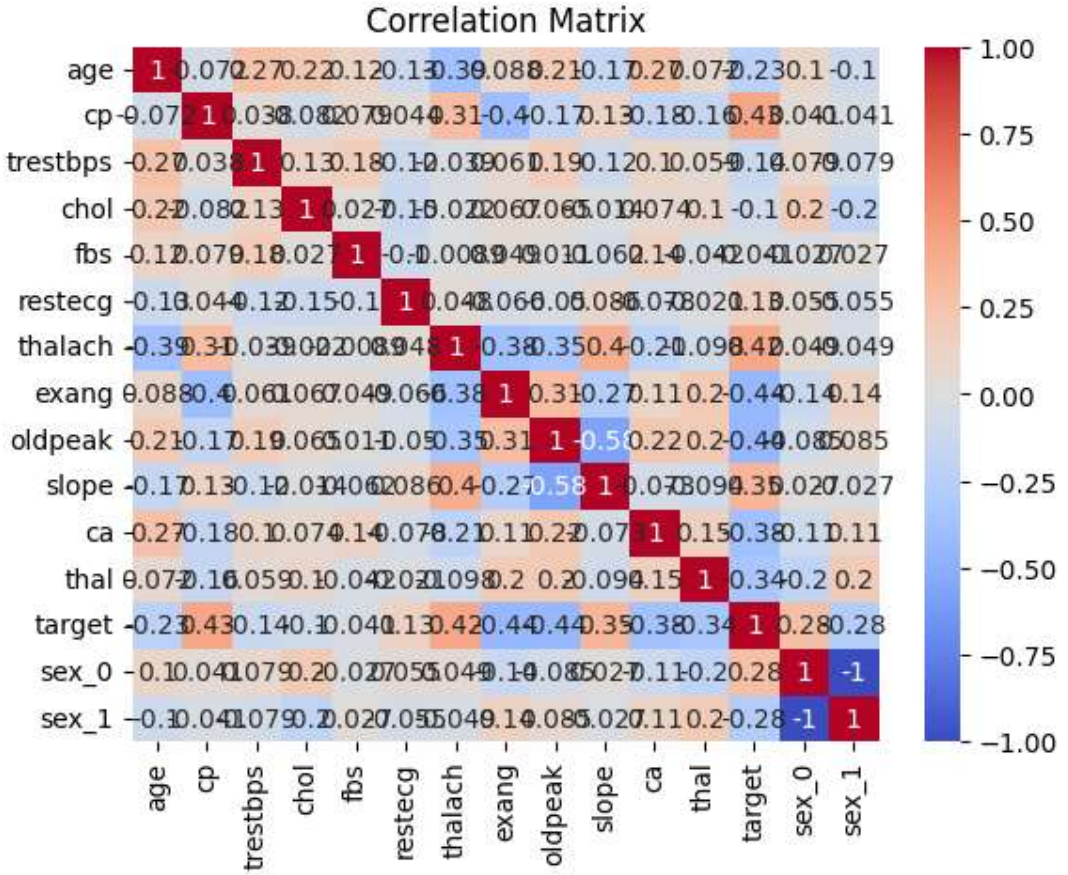
	age	cp	trestbps	chol	fbs
count	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000
mean	54.434146	0.942439	131.611707	246.00000	0.149268
std	9.072290	1.029641	17.516718	51.59251	0.356527
min	29.000000	0.000000	94.000000	126.00000	0.000000
25%	48.000000	0.000000	120.000000	211.00000	0.000000
50%	56.000000	1.000000	130.000000	240.00000	0.000000
75%	61.000000	2.000000	140.000000	275.00000	0.000000
max	77.000000	3.000000	200.000000	564.00000	1.000000

	restecg	thalach	exang	oldpeak	slope
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	0.529756	149.114146	0.336585	1.071512	1.385366
std	0.527878	23.005724	0.472772	1.175053	0.617755
min	0.000000	71.000000	0.000000	0.000000	0.000000
25%	0.000000	132.000000	0.000000	0.000000	1.000000
50%	1.000000	152.000000	0.000000	0.800000	1.000000
75%	1.000000	166.000000	1.000000	1.800000	2.000000
max	2.000000	202.000000	1.000000	6.200000	2.000000

	ca	thal	target
count	1025.000000	1025.000000	1025.000000
mean	0.754146	2.323902	0.513171
std	1.030798	0.620660	0.500070
min	0.000000	0.000000	0.000000
25%	0.000000	2.000000	0.000000
50%	0.000000	2.000000	1.000000
75%	1.000000	3.000000	1.000000
max	4.000000	3.000000	1.000000

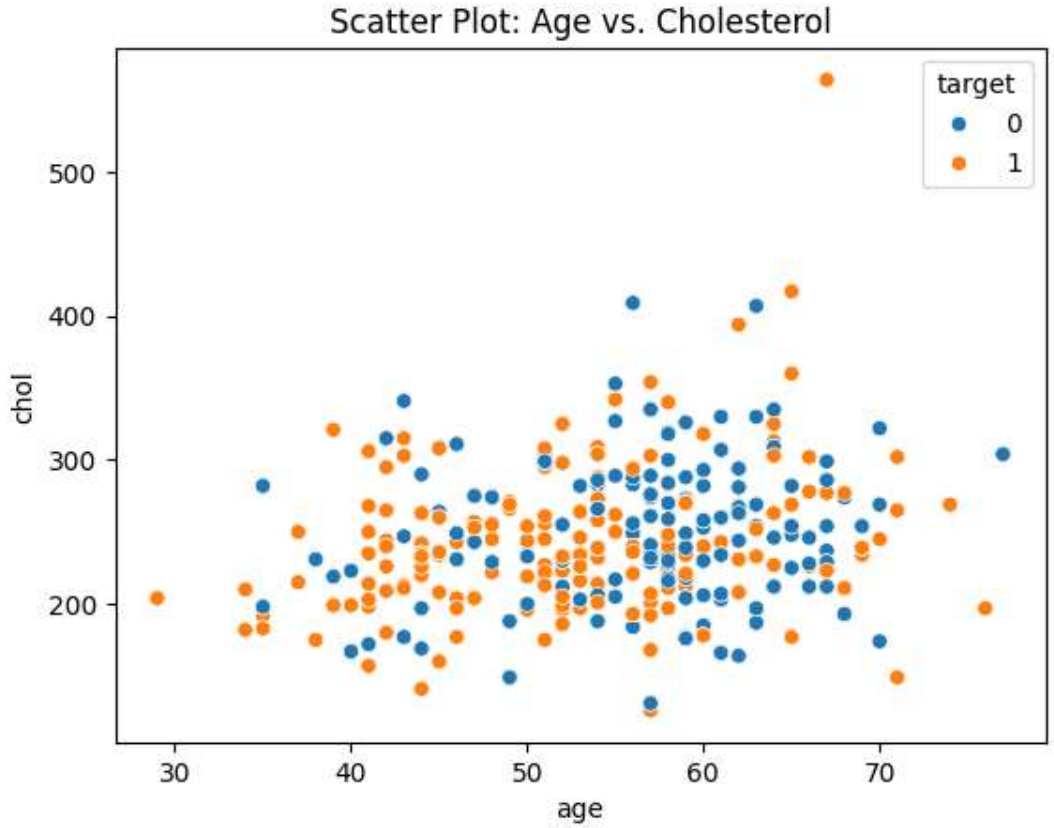
Exploratory Data Analysis (EDA)

- Examined linear relationships between features using correlation matrices and heatmaps.



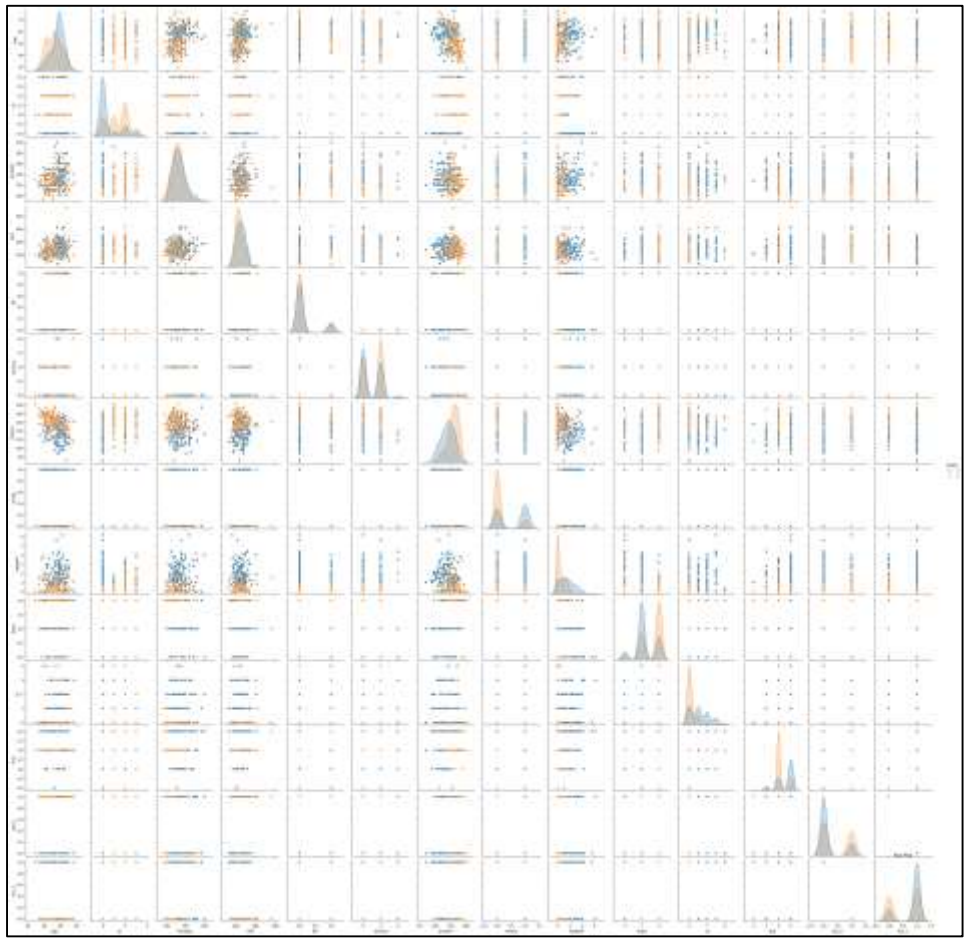
Exploratory Data Analysis (EDA)

- Employed histograms, scatter plots, pair plots, and box plots to identify patterns and distributions in the data.

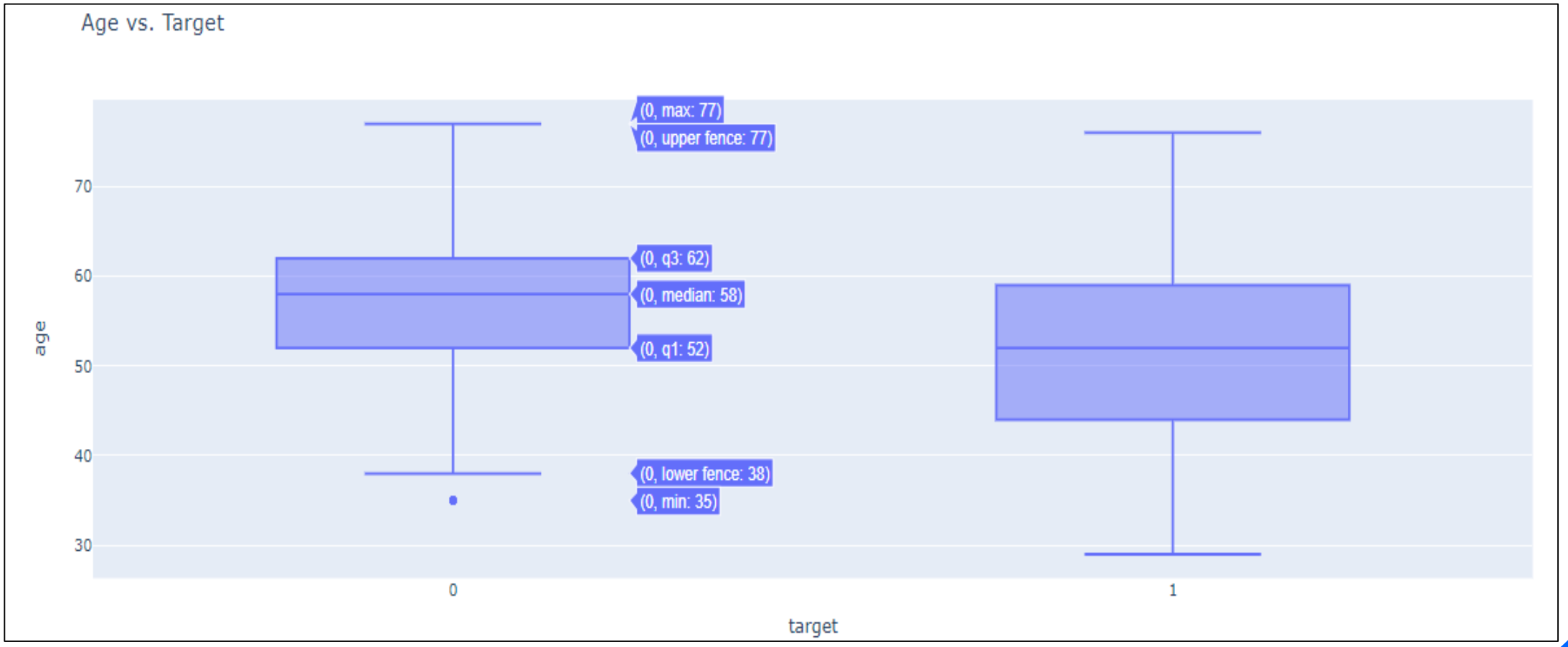


Exploratory Data Analysis (EDA)

- Utilized Matplotlib, seaborn, Plotly, and Plotly Express for visualization.

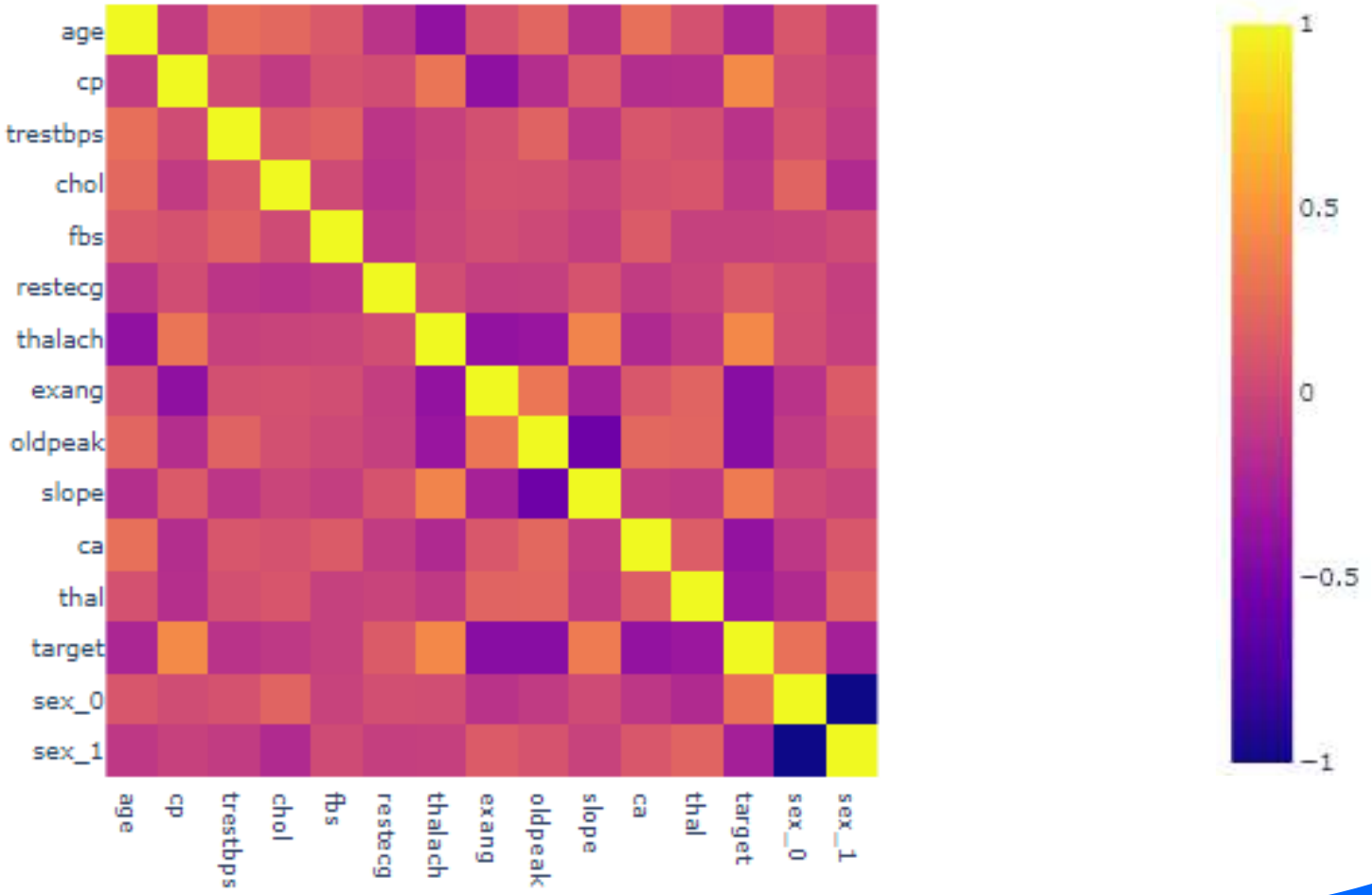


Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)

- Explored correlations between features using correlation matrices or heatmaps.



Model Training and Evaluation

- Partitioned the dataset into training and testing sets using a ratio of 80:20.
- Utilized a Random Forest Classifier to train the model for heart disease prediction.
- Implemented Logistic Regression as an alternative model for heart disease diagnosis.
- Assessed model performance using accuracy metrics, including precision, recall, and F1-score.

```
Accuracy: 0.9853658536585366
```

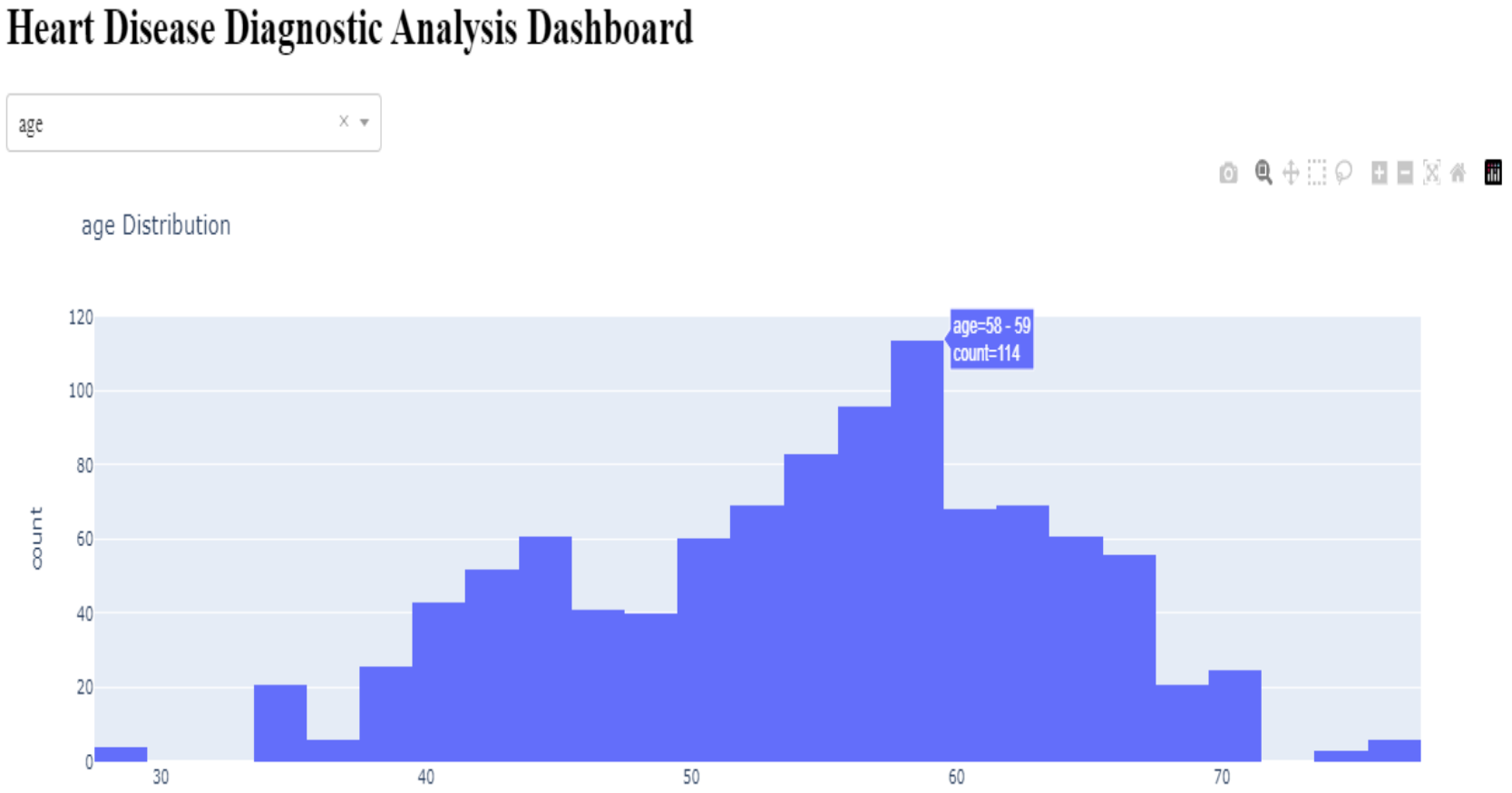
```
Precision: 0.7563025210084033
Recall: 0.8737864077669902
F1-score: 0.8108108108108107
```

Dashboard Development

- Developed a **Dash application** to create an interactive dashboard using the plotly dash library.
- Designed the dashboard layout with HTML components and Dash core components.
- Integrated a dropdown menu to select features for visualization.
- Incorporated two graph components to display histograms and box plots dynamically based on user-selected features.
- Added a **range slider** for age selection to filter data for visualization.

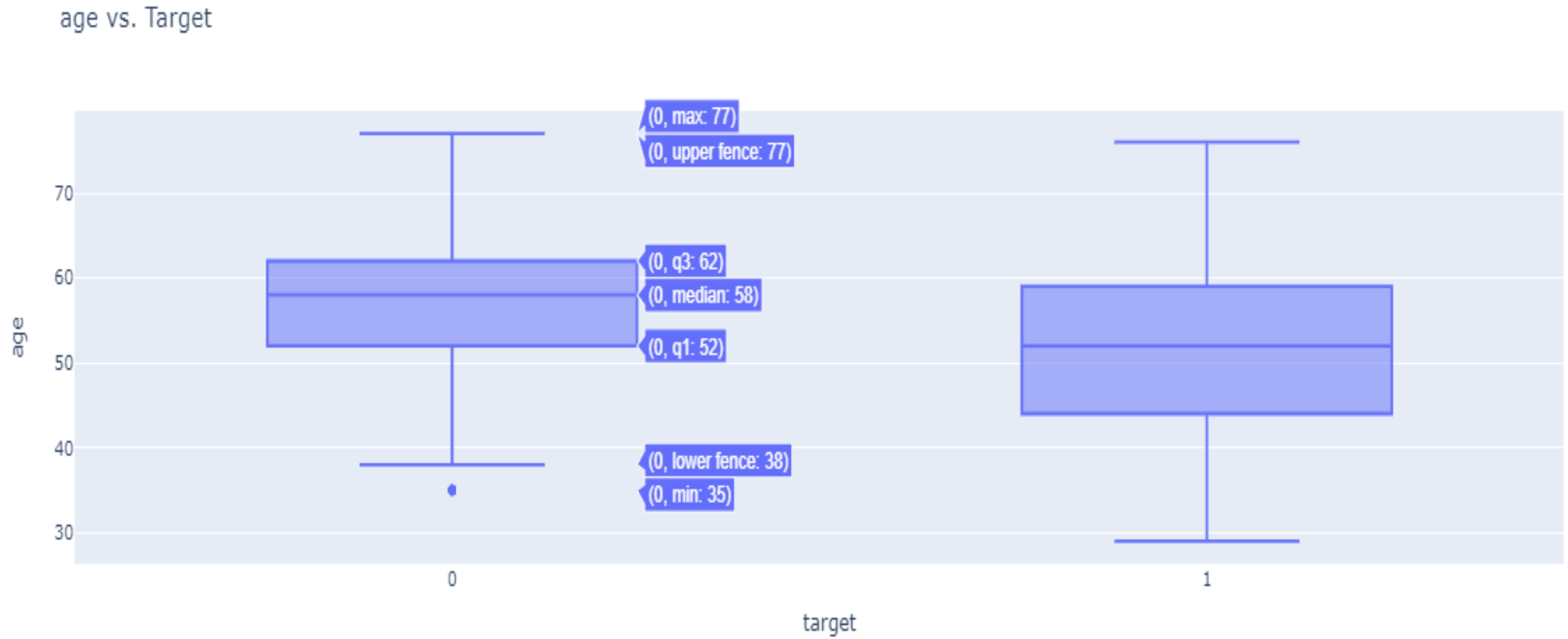
Dashboard Development

- Dashboard Interface:



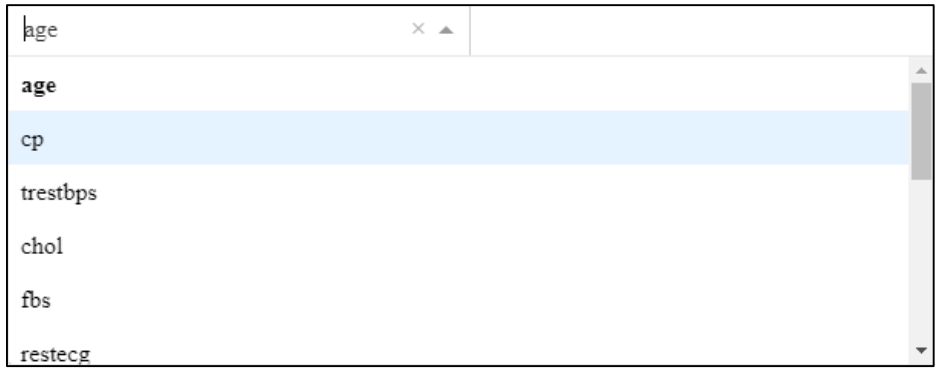
Dashboard Development

- Dashboard Interface:



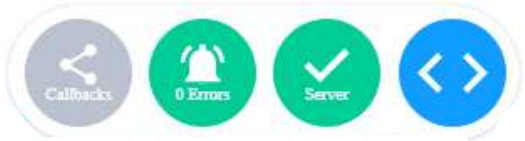
Dashboard Development

- Feature Selection:



- Range Selection Bar:

- Toggles & buttons:



Insights

- The age of patients ranges from **29** to **77** years, with a mean age of approximately **54.43** years and a standard deviation of **9.07** years.
- The majority of patients experienced **chest pain (cp)**, with the distribution skewed towards lower values. The mean value is approximately **0.94**.
- The average resting blood pressure is around **131.61** mm Hg, with a standard deviation of **17.52** mm Hg. The minimum and maximum values are **94** mm Hg and **200** mm Hg, respectively.
- The mean cholesterol level is **246** mg/dl, with a standard deviation of **51.59** mg/dl. Cholesterol levels range from **126** mg/dl to **564** mg/dl.

Insights

- Approximately **14.93%** of patients had high fasting blood sugar (fbs), as indicated by a value of **1**.
- The distribution of resting electrocardiographic results is fairly evenly distributed, with a mean value of approximately **0.53**.
- The average maximum heart rate achieved is **149.11** beats per minute (bpm), with a standard deviation of **23.01** bpm.
- About **33.66%** of patients experienced exercise-induced angina, as indicated by a value of **1**.
- The mean value of ST depression induced by exercise relative to rest is approximately **1.07**, with a standard deviation of **1.18.**

Insights

- The majority of patients have a slope value of **1**, indicating down sloping, followed by a value of **2**, indicating flat.
- The mean number of major vessels is approximately **0.75**, with a standard deviation of **1.03**.
- The majority of patients have thalassemia type **2**, followed by type **3**.
- Approximately **51.32%** of patients were diagnosed with heart disease, indicating a balanced dataset with roughly equal instances of **positive** and **negative** cases.

Conclusion

- The project successfully conducted exploratory data analysis, data cleaning, and preprocessing of heart disease diagnostic data.
- Machine learning models were trained to predict heart disease diagnosis with considerable accuracy.
- An interactive dashboard was developed to visualize key insights and trends from the data.
- Through this project, valuable insights were derived regarding demographic patterns, clinical indicators, and predictive factors associated with heart disease.
- These findings provide actionable insights for healthcare professionals to enhance diagnostic processes and inform future healthcare strategies aimed at mitigating the impact of heart disease.

Summary

- The dataset comprises individuals across a wide age range, with an average age of approximately **54** years. This suggests that heart disease is not limited to any specific age group and affects individuals across various stages of life.
- Key clinical parameters such as **resting blood pressure**, **cholesterol levels**, and **maximum heart rate** vary considerably among patients. Understanding these variations is crucial for accurate diagnosis and treatment planning.
- Factors such as **high fasting blood sugar** and **exercise-induced angina** are prevalent among a subset of patients, indicating potential risk factors for heart disease development.

Summary

- Features like **ST depression** induced by exercise and the **number of major vessels** provide valuable diagnostic insights. Higher values of these features may indicate a greater likelihood of heart disease.
- The dataset contains a balanced distribution of **positive** and **negative** instances of heart disease, facilitating model training and evaluation without bias towards either class.
- Thalassemia **type 2** appears to be the most common type among patients with heart disease. Exploring the relationship between **thalassemia** and **heart disease** could provide further insights into disease mechanisms and potential treatment strategies.

Thank you