FULL REPORT ON HEART DISEASE ML MODEL

Designed for the Health Care Industry

OVERVIEW

This project has been a remarkable journey, culminating in the creation of a powerful tool with the potential to save lives. Our machine learning model stands as a beacon of hope in the fight against heart disease, the leading cause of death globally.

Contributor(s)

Kehinde Adeniran

TABLE OF CONTENTS

Business Understanding

For many years, heart disease has been a prevalent cause of mortality among many, both the young and old. This project supports various intervention programs led by data scientists to tackle this ailment globally. One key factor, however, has remained constant: that is···"Early detection can save lives". This project hopes to create a machine-learning model that predicts a patient's risk of heart disease using historical medical data. By uncovering patterns in vital health indicators like age, chest pain type, cholesterol, maximum heart rate, exercise responses, and many others, I aim to empower healthcare providers to make timely, data-driven decisions, ultimately transforming patient care and enhancing lives. Let's delve in.

To help you quickly understand and maintain focus throughout your reading, I have broken down each discussion area into smaller sessions below.

Industry

This project focuses on the **healthcare industry**, specifically targeting the prevention and management of heart disease. It is a no-brainer that accurately predicting the risk of heart disease is essential for helping doctors make timely decisions, which in turn improves patient care and saves lives. This model promotes a more informed and effective approach to heart health in a world where proactive health solutions are crucial.

Problem Statement

Heart disease continues to be a significant global health challenge, often going undetected until it is too late. Our goal is to prevent and treat heart disease through early diagnosis by developing a machine-learning algorithm that can predict the risk of heart disease. By doing so, it will assist doctors, nurses, and all stakeholders involved in taking early actions to improve patient outcomes and save lives.

Goals and Objectives

Based on the backstory provided up to this point, I have outlined three main goals that summarize the vision for our ML model.

1. Build a precise model that empowers the identification of individuals at risk of heart disease.
2. Leverage essential health indicators to unveil patterns that inspire diagnosis and prevention.
3. Deliver a trustworthy, data-driven tool that fosters proactive and compassionate patient care.

Who are the Stakeholders?

The key stakeholders for this project include:

1. **Healthcare Providers**: Doctors and clinicians who will use the model to identify patients at risk of heart disease and provide timely care.
2. **Patients**: Individuals benefiting from early detection and improved treatment options.
3. **Healthcare Organizations**: Institutions aiming to enhance patient outcomes and streamline preventative care.

Why is it Worth Solving?

Early detection can save countless lives, highlighting its significance.

Real-Life Applications

Upon completing this model, we hope it can be applied in various real-life scenarios where awareness and timely action are critical.

a. **Preventive Healthcare**: Helping doctors identify patients at risk of heart disease for early intervention and personalized treatment plans.
b. **Community Health Programs**: Supporting public health initiatives by identifying high-risk populations and tailoring awareness campaigns.
c. **Hospital Resource Management**: Assisting hospitals and clinics in prioritizing care and efficiently allocating medical resources.
d. **Insurance Risk Assessment**: Helping insurance companies assess health risks and design fair, data-driven premium policies.
e. **Remote Monitoring Systems**: Integrating into wearable or telemedicine devices to provide real-time risk alerts for users.

By empowering these applications, the model improves patient care, advances health equity, and facilitates better decision-making in healthcare systems.

Data Collection

The data was collected from Kaggle, an open-source project datasets hub for data scientists. It is also worth mentioning that the dataset was originally sourced from the UCI Machine Learning Repository on the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/. It contains 12 columns and multiple rows of historical heart disease information.

Acknowledgments

Creators:

We would like to formally announce our recognition of the creators of this dataset, whose contributions have significantly impacted the machine learning field, particularly in research.

1. Hungarian Institute of Cardiology, Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor:

We would like to express our gratitude and appreciation to the donor, David W. Aha (aha '@' ics.uci.edu) (714) 856-8779.

Data Understanding

This section provides an overview of the data collected for this project. To fully understand the variety and key characteristics of the dataset used, it is common practice in data analysis and data science to examine the composition of each column, feature, or input of the data. This investigation helps provide a foundational analysis and gives a clearer picture of the overall dataset.

Before we explore the visualization analysis of our dataset, here is a comprehensive breakdown of the data collected for this project.

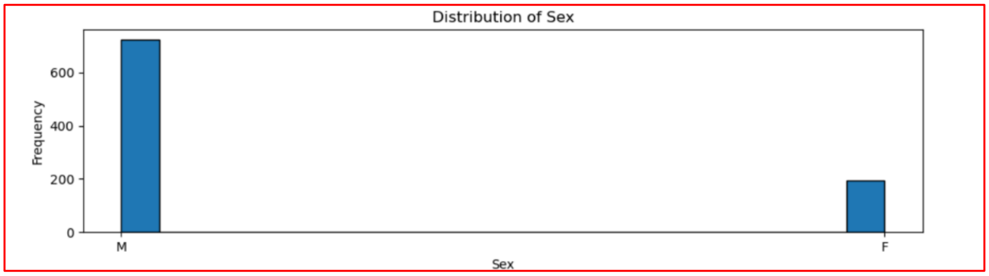
S/N	Columns or Features	Description	Data Types
1	Age	Age of the patient (years)	Numeric
2	Sex	Sex of the patient (M: Male, F: Female)	Categorical
3	ChestPainType	Chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)	Categorical
4	RestingBP	Resting blood pressure (mm Hg)	Numeric
5	Cholesterol	Serum cholesterol (mm/dl)	Numeric
6	FastingBS	Fasting blood sugar (1: if FastingBS > 120 mg/dl, 0: otherwise)	Numeric
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]	Categorical
8	MaxHR	Maximum heart rate achieved (Numeric value between 60 and 202)	Numeric
9	ExerciseAngina	Exercise-induced angina (Y: Yes, N: No)	Categorical

10	Oldpeak	ST (Numeric value measured in depression)	Numeric
11	ST_Slope	The slope of the peak exercise ST segment (Up: upsloping, Flat: flat, Down: downsloping)	Categorical
12	HeartDisease	Output or predicted class (1: heart disease, 0: Normal)	Numeric

Now that we understand the composition of our data, it's time to compare the columns using graphs. This will help us examine the distribution and breakdown of values within each column, providing a clearer understanding of the dataset. The images below offer a concise view of each column in our dataset.

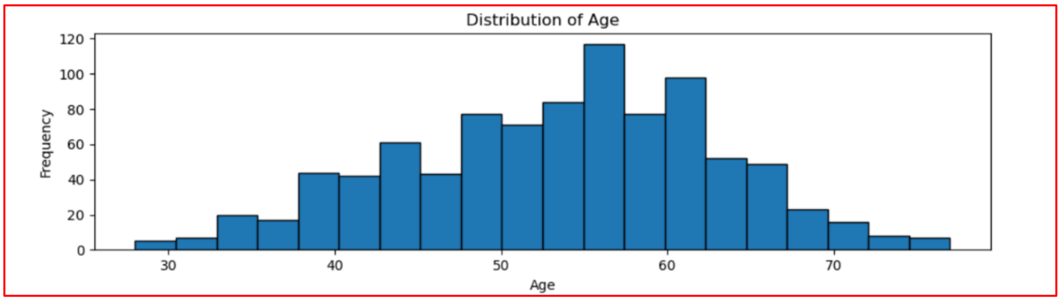
a) Data At Glance

Image 1 – Column 1 (Sex)



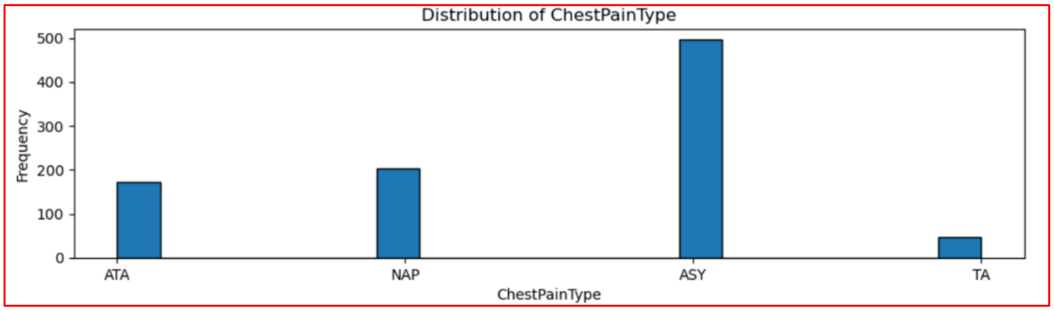
In the diagram above, the first bar graph represents the male gender (M), while the second represents the female gender (F). This shows the distribution of data points between both sex groups

Image 2 – Column 2 (Age)



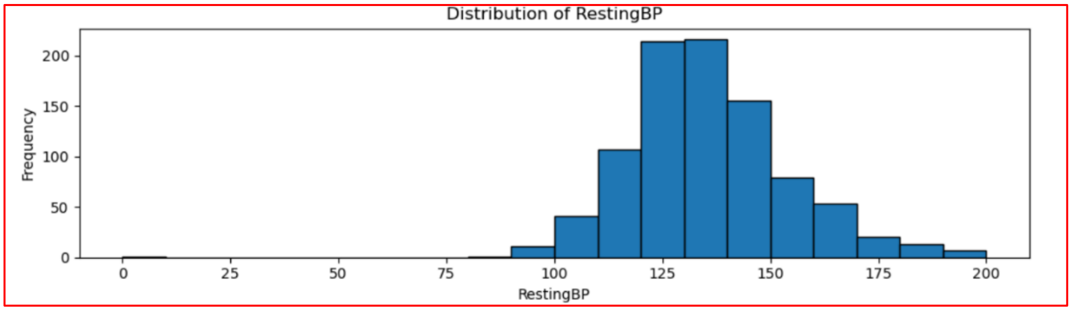
By examining the age column, we can confirm that the data points are close to a normal distribution, also known as a ***Gaussian distribution*** in data science.

Image 3 – Column 3 (Chest Pain Type)



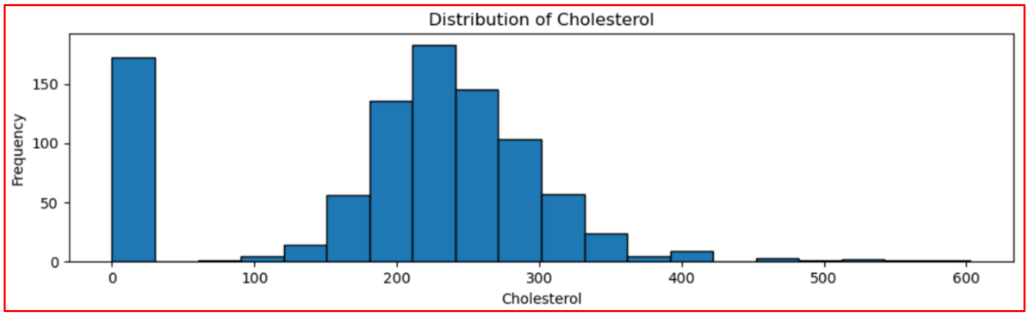
The chest pain type column contains four (4) categories namely TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic. The graph above shows the representation for each category in the column.

Image 4 – Column 4 (Resting Blood Pressure)



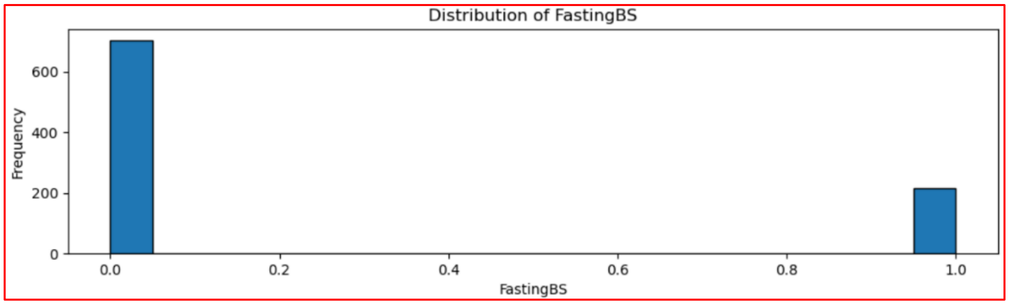
Notice there are some values recorded as zeros in the data points shown above. This is scientifically incorrect as the resting blood pressure of a human can never be zero. This may be null values replaced by zeros during the data entry process. Additionally, we can observe that the data in this column is right skewed as shown in the graph. This observation is significant to note.

Image 5 – Column 5 (Cholesterol)



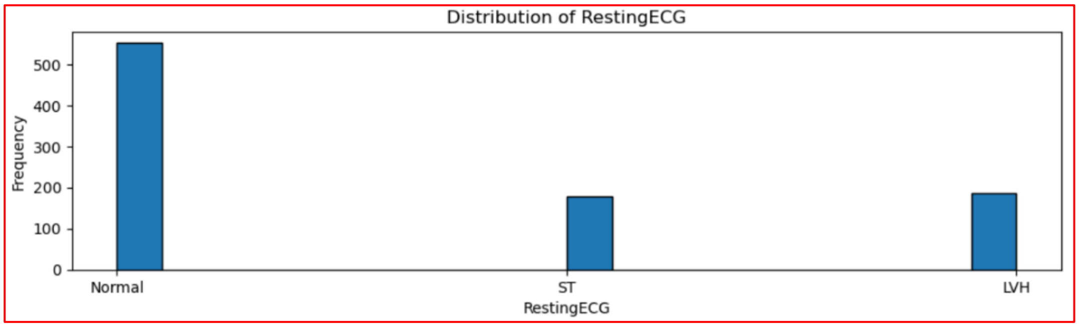
Notice there are values recorded as zeros in the data distribution as well. This is similar to the resting blood pressure and in many instances, these zeros may indicate null values that were replaced with zero in the dataset. Furthermore, the data points in the cholesterol column resemble a normal or Gaussian distribution, except for the previously mentioned null values.

Image 6 – Column 6 (Fasting Blood Sugar)



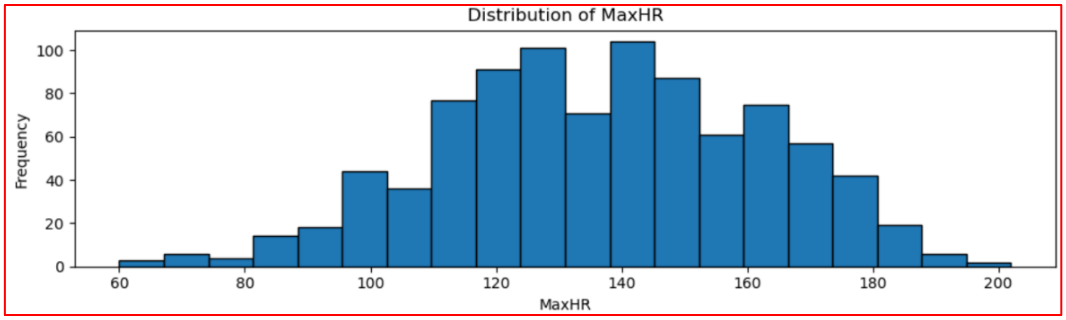
This column presents the fasting blood sugar levels of patients in this dataset. The first bar graph indicates "1" for Fasting BS > 120 mg/dl, while the second indicates "0" for levels otherwise.

Image 7 – Column 7 (Resting electrocardiogram results)



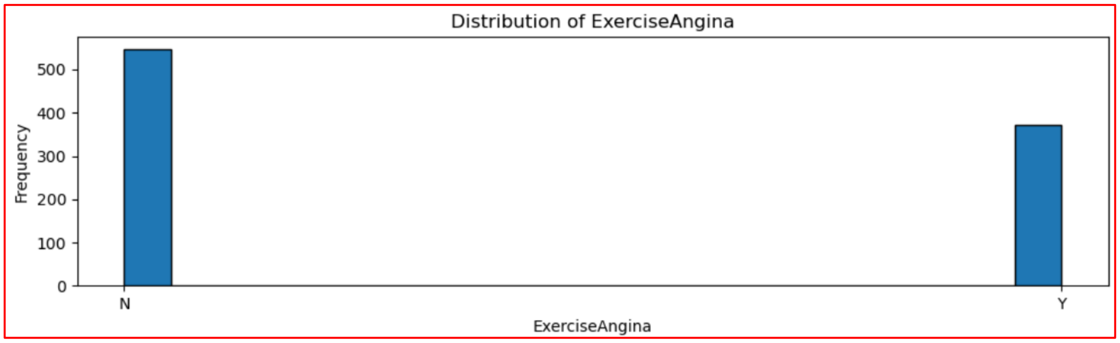
The graph above illustrates three categories of resting electrocardiogram (ECG) results obtained from the patients in the dataset, arranged in the correct order. "**Normal**" indicates a standard electrocardiogram result, "**ST**" refers to the presence of ST-T wave abnormalities (including T wave inversions and/or ST elevation or depression greater than 0.05 mV), and "**LVH**" denotes probable or definite left ventricular hypertrophy according to Estes' criteria.

Image 8 – Column 8 (Maximum heart rate)



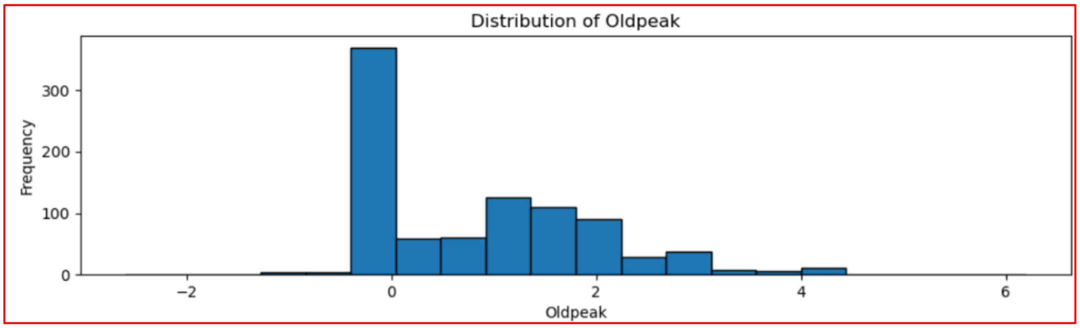
The data points in the maximum heart rate achieved column follow a normal distribution.

Image 9 – Column 9 (Exercise Angina)



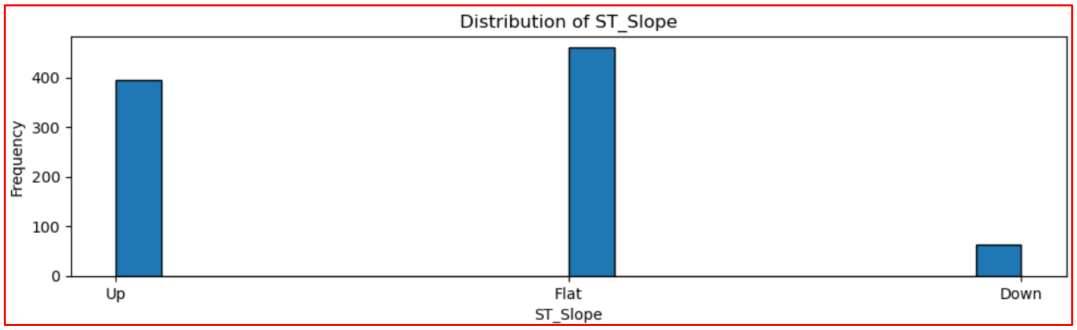
The graph above illustrates the two categories of exercise-induced angina: Yes and No.

Image 10 – Column 10 (Old Peak)



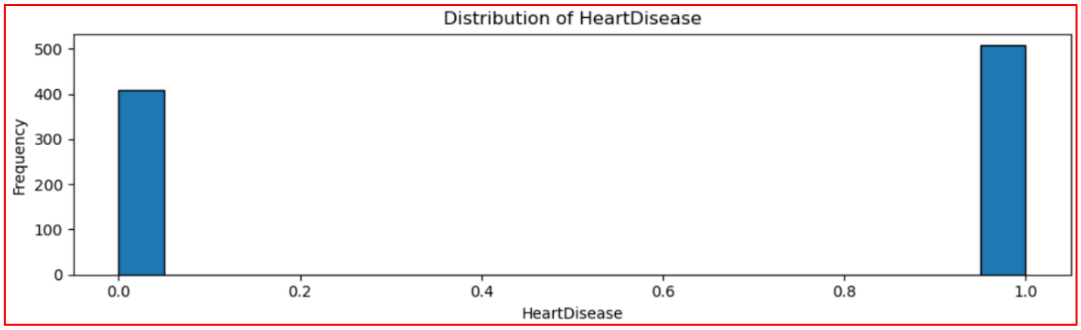
As illustrated in the graph above, the data points in this column exhibit a right skew.

Image 11 – Column 11 (ST Slope)



The graph above illustrates the slope of the peak exercise ST segment from the dataset, which is divided into three categories displayed in the correct order: first **Up** (upsloping), followed by **Flat** (flat), and lastly **Down** (downsloping).

Image 12 – Column 12 (heart disease – *Note this is the Output class*)



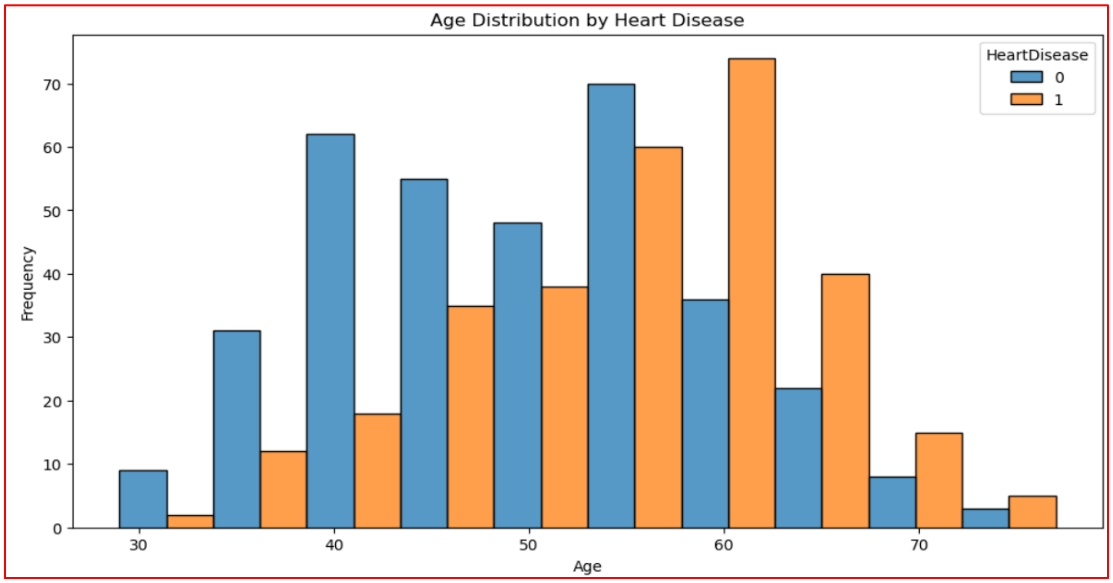
This is the current output class or results based on the existing data, which is divided into two categories. A value of 1 indicates that the patient has heart disease, as illustrated in the second bar graph, while a value of 0 signifies that there is no heart disease. This accurately represents our data.

b) Exploratory Data Analysis of the Dataset (EDA)

Exploratory Data Analysis (EDA) is like the first conversation we're having with the dataset introduced for this project. This is where we learn more about the data, its quirks, strengths, and secrets. It's the foundation for getting a deep insight into the data before it's then used in building a reliable model.

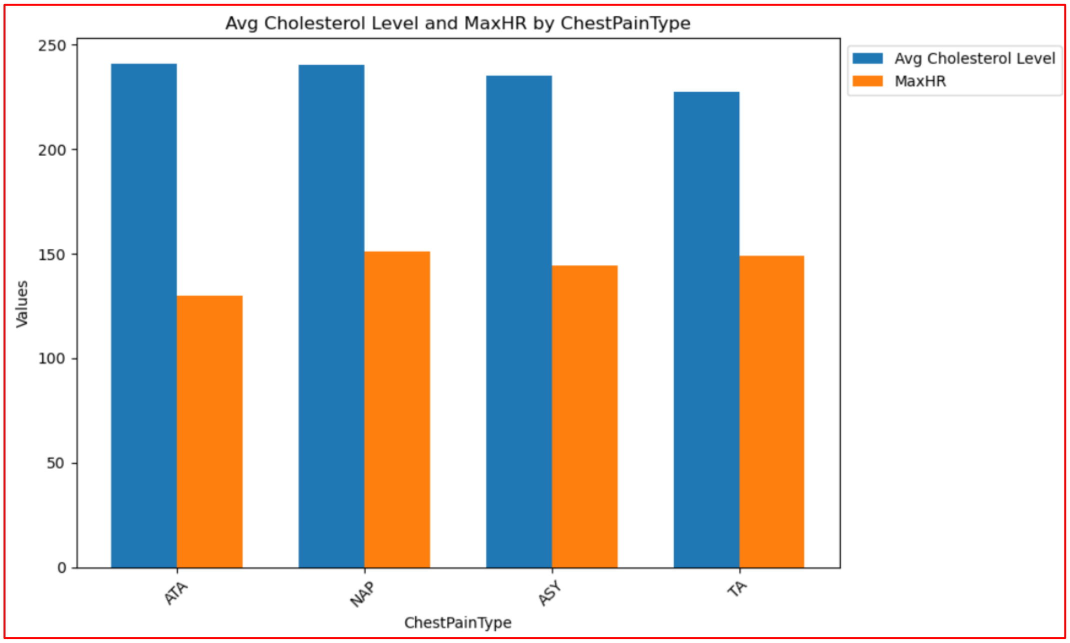
The images below help us uncover hidden patterns in the dataset while also giving us a solid understanding of the relationships between variables in the data. Especially how a feature correlates with another or with the actual outcome.

Image 1−Bivariate Analysis: Comparing ‘Age’ and ‘Heart Disease’ with each other.



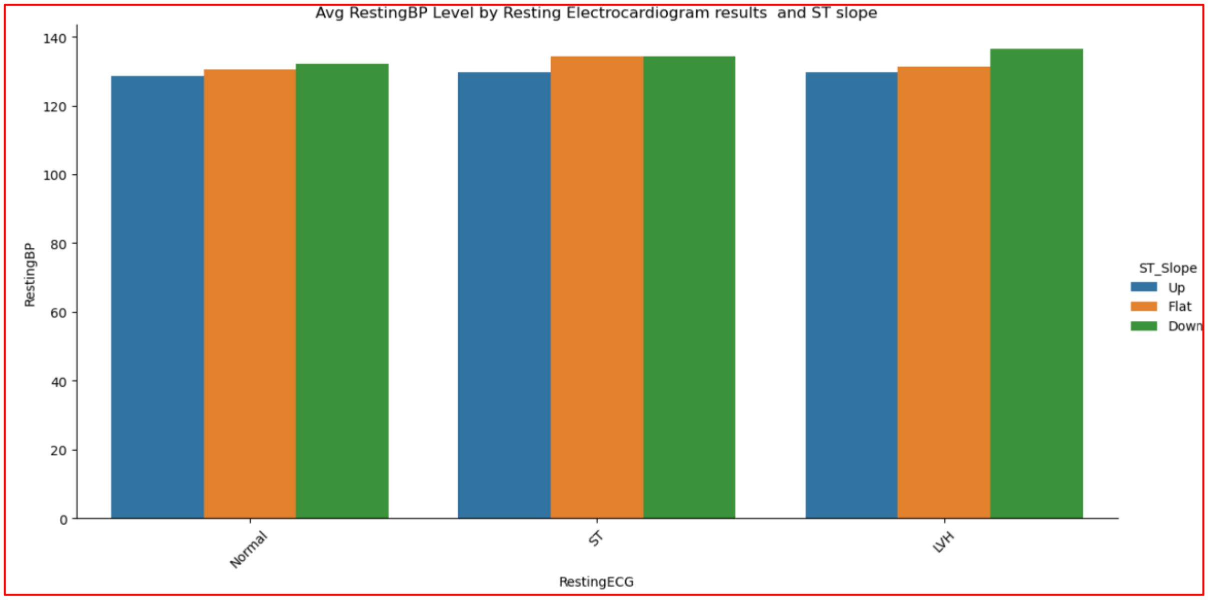
A quick overview shows that age is strongly correlated with heart disease. As individuals get older, their risk or likelihood of developing heart disease increases. The blue color represents individuals without heart disease, while the orange color indicates those with heart disease.

Image 2–Multivariate Analysis (i): Comparing ‘Cholesterol Level’ and ‘Maximum Heart Rate’ By ‘Chest Pain Type’



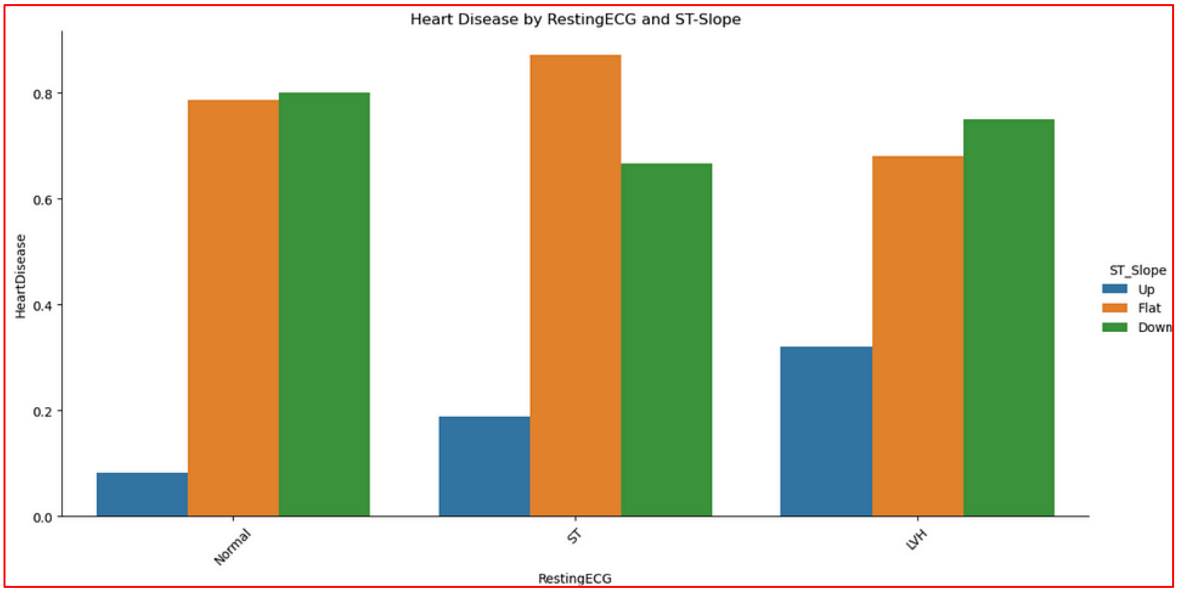
An analysis of the image above reveals a breakdown of the frequencies of chest pain types in relation to cholesterol levels and maximum heart rates in patients. Specifically, it shows that cholesterol levels have a stronger correlation with all four types of chest pain. This illustrates how exploratory data analysis (EDA) deepens our understanding of the patterns present in the data.

Image 3–Multivariate Analysis (ii): Comparing ‘Resting Electrocardiogram results’ and ‘ST Slope ’ By ‘Resting Blood Pressure



Based on the analysis above, it is evident that many patients in our dataset exhibited ST-T wave abnormality and LVH: showing probable or definite left ventricular hypertrophy. Additionally, flat and down-slopping rank as the two serious risk indicators that correlate with severe heart issues and require immediate attention.

Image 4−Multivariate Analysis (iii): Comparing ‘Heart Disease’ occurrences By ‘ST Slope’ and ‘ Resting Electrocardiogram results’



As shown in the image above, the ST Slope variables, flat and down-sloping, represent the two most common categories for many patients. Both categories are important predictors of heart disease, as they provide insight into how well the heart functions under stress and whether there may be blockages in the coronary arteries.

This concludes our overview and exploratory data analysis of the dataset.

Data Preparation

Data preparation involves all the necessary steps and procedures that data goes through before being used to build a machine-learning model. This series of processes ensures that the data is appropriately processed and ready for use.

In this section, we will discuss each step in detail.

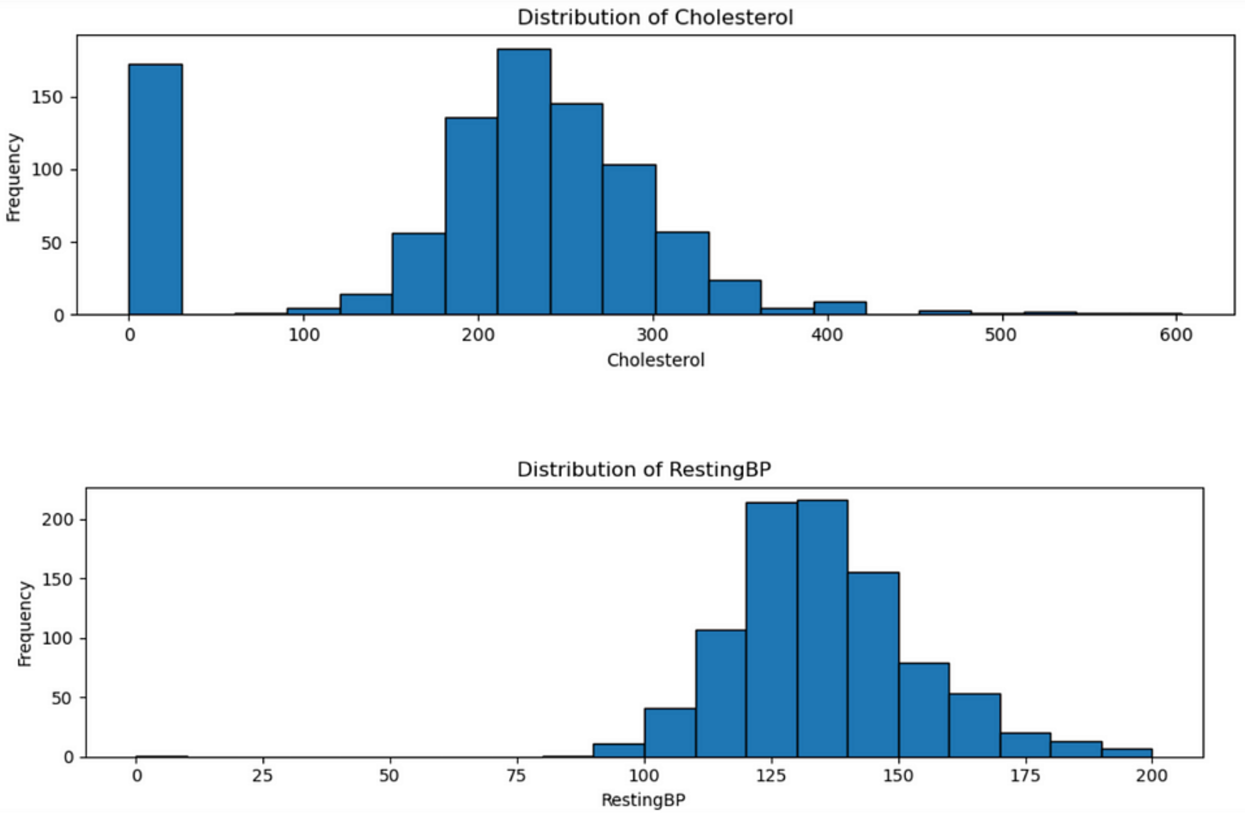
1. Data Wrangling
2. Define Dataset
3. Data Engineering
4. Data Splitting

Data Wrangling

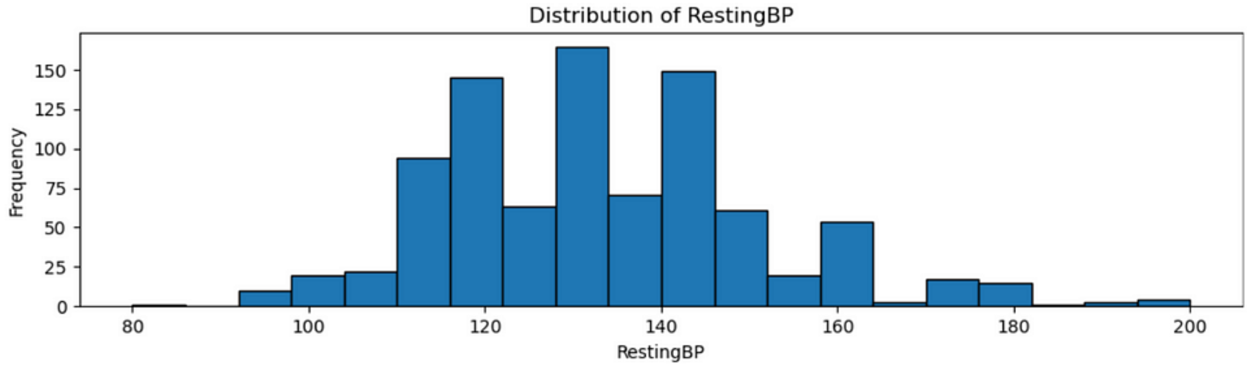
Wrangling the data connotes the necessary transformation or cleaning done on the data. The following were the data cleansing or transformation procedures used on our historical data.

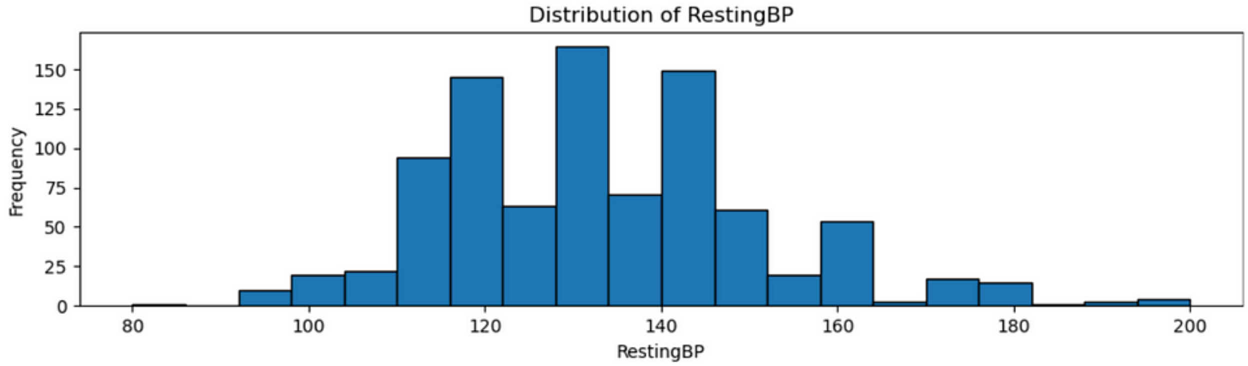
a) **Dealing with missing values or zeros in columns like Cholesterol and Resting Blood Pressure**: Earlier, I mentioned these columns had some rows filled with zeros. Scientifically a human’s resting blood pressure and cholesterol level can’t be at zero. We filled in missing values or zeros using the mean or average of these columns for better model performance.

Before: Cholesterol and Resting BP



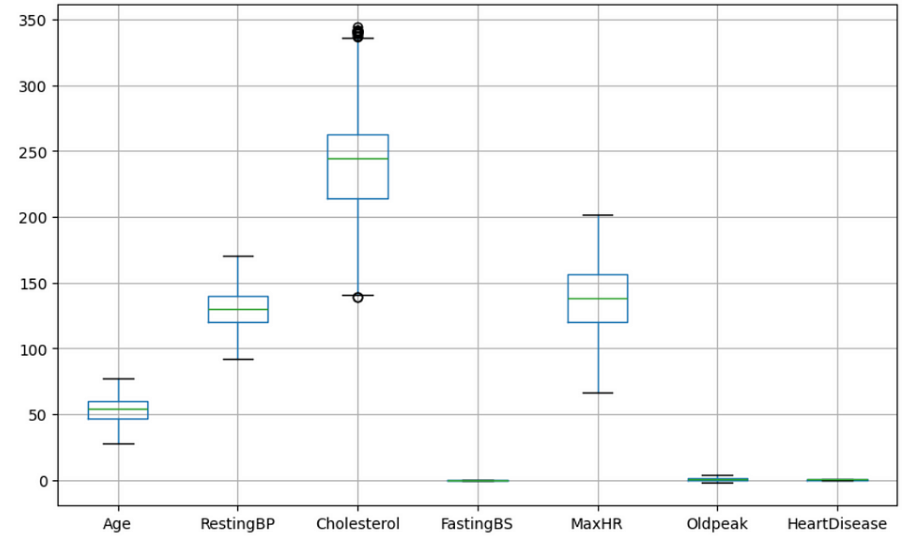
After: Cholesterol and Resting BP





b) Eliminating Outliers and Noise from the Numerical Columns in Our Data

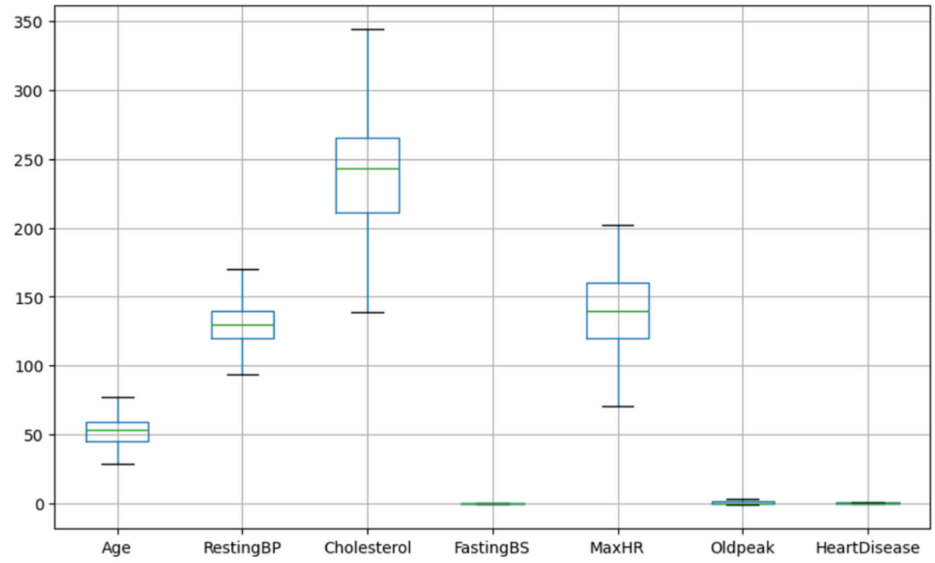
Before:



Outliers Identified in Cholesterol Levels

Outliers are like extreme values in a dataset that don't fit well with the rest. For example, in the chart above, you can see that some cholesterol levels are unusually high compared to most others. These outliers can skew our analysis and make it harder for the machine learning model to learn patterns correctly.

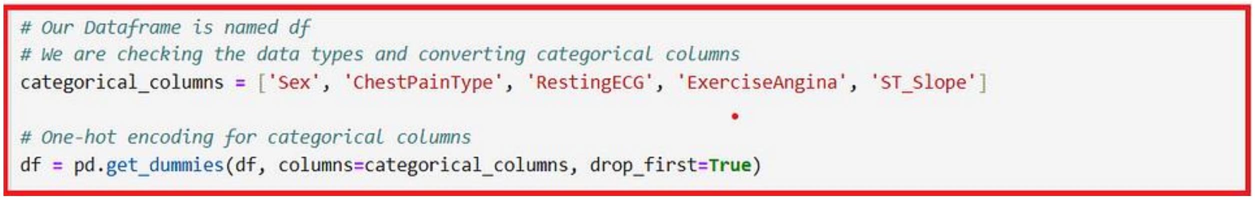
After:



Outliers Removed from Cholesterol Levels

To fix this, we removed those extreme values to make the data more balanced, as shown in the second chart. This step helps ensure our model gets a clearer picture of the true relationships in the data, leading to more accurate predictions. It' s like cleaning up noise to make a song sound better!

c) **Encoding Categorical Variables:** Our dataset contains several categorical columns, which need to be encoded numerically for a machine learning model to process them—examples: Sex, Chest Pain Type, Resting ECG, and many others.



```python
# Our Dataframe is named df
# We are checking the data types and converting categorical columns
categorical_columns = ['Sex', 'ChestPainType', 'RestingECG', 'ExerciseAngina', 'ST_Slope']

# One-hot encoding for categorical columns
df = pd.get_dummies(df, columns=categorical_columns, drop_first=True)
```

The code above is how we encoded categorical columns of our Data.

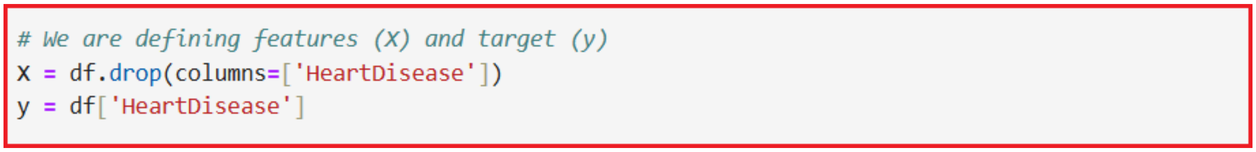
Define Dataset

We are organizing how the model will utilize the dataset. Which columns will serve as the independent variables for predicting the expected outcome? The dataset is divided into two categories:

1. Features, which also refer to columns or variables.
2. The target class (or output).

We have organized our dataset to include 11 variables that will assist us in predicting our only outcome which is: heart disease, as shown in the image below.

The image on Define Dataset

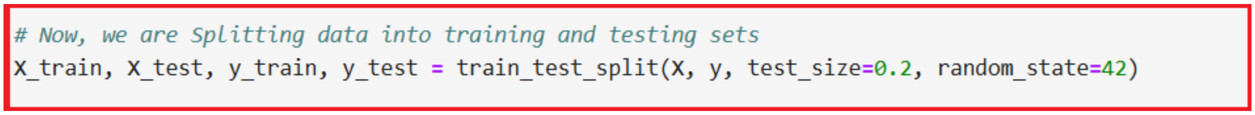


```python
# We are defining features (X) and target (y)
X = df.drop(columns=['HeartDisease'])
y = df['HeartDisease']
```

Data Splitting

This phase is the last preparatory stage any data will undergo in machine learning model building. We are dividing the data into two sets: 80% for training and 20% for testing. This step is like preparing for a big game where you train hard with most of your data (80%) to build a strong model, and then test it with the remaining 20% to see how well it performs in real-life scenarios. This balanced approach ensures our model is both well-trained and reliable.

The image on Data Splitting

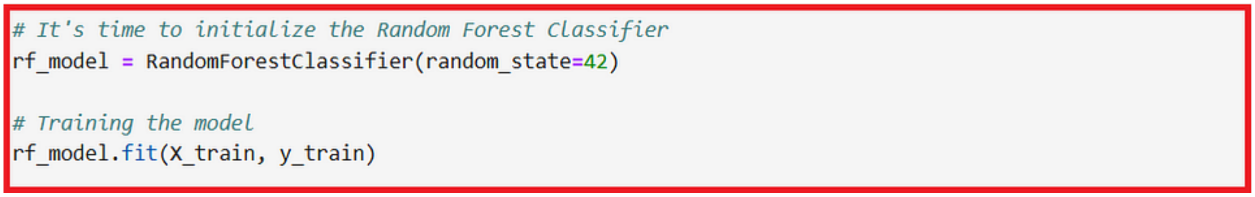


```
# Now, we are Splitting data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Model Building (Training dataset)

Remember that we previously divided our dataset into training sets (80%) and testing sets (20%). In this section, we will focus on building a model using only the training datasets. Additionally, we will review the methodology used for this process, as well as the algorithm implemented within that methodology.

For this project, I am implementing a supervised machine learning classification model. Specifically, we will build a ***random forest classification model*** to predict the risk of heart disease in future patients.

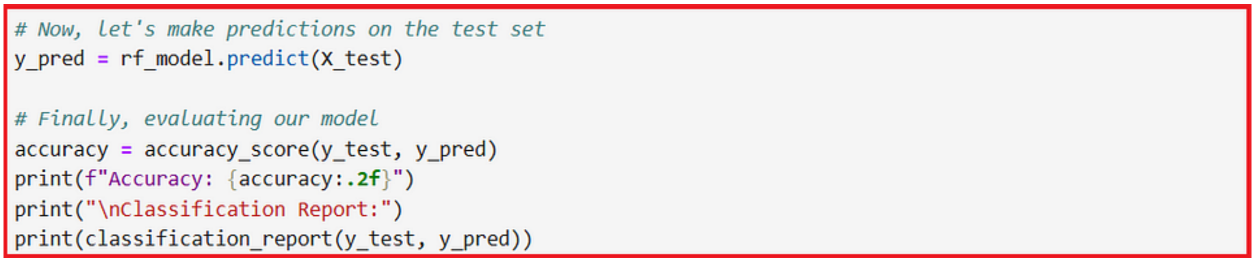


```
# It's time to initialize the Random Forest Classifier
rf_model = RandomForestClassifier(random_state=42)

# Training the model
rf_model.fit(X_train, y_train)
```

Model Evaluation (Test datasets)

As a follow-up to the successful development of our model in the model-building part, we need to evaluate its performance to ensure it accurately predicts our outcomes. This section will focus on testing the model using our testing set, which consists of 20% of the data.

Let's proceed with a comprehensive review of the model's performance based on the results below.



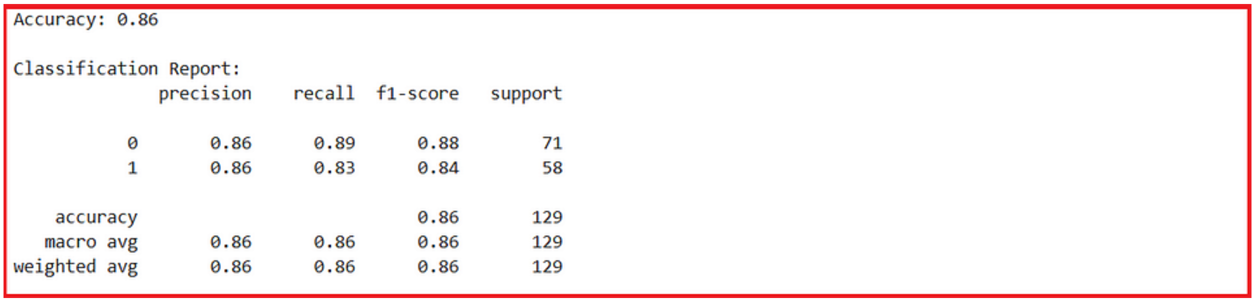
```
# Now, let's make predictions on the test set
y_pred = rf_model.predict(X_test)

# Finally, evaluating our model
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.2f}")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

The first part of the code evaluates our model by testing it on the new 20% of the dataset reserved for testing.

The final section of the code evaluates the model's overall performance, as shown in the image below.

Presently, the model boasts a strong accuracy rate of 86%. This has been accomplished through a series of iterations and adjustments to the parameters tuning pre-model building.

Image focusing on Accuracy, Precision and Recall



```
Accuracy: 0.86

Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.89      0.88        71
           1       0.86      0.83      0.84        58

    accuracy                           0.86       129
   macro avg       0.86      0.86      0.86       129
weighted avg       0.86      0.86      0.86       129
```

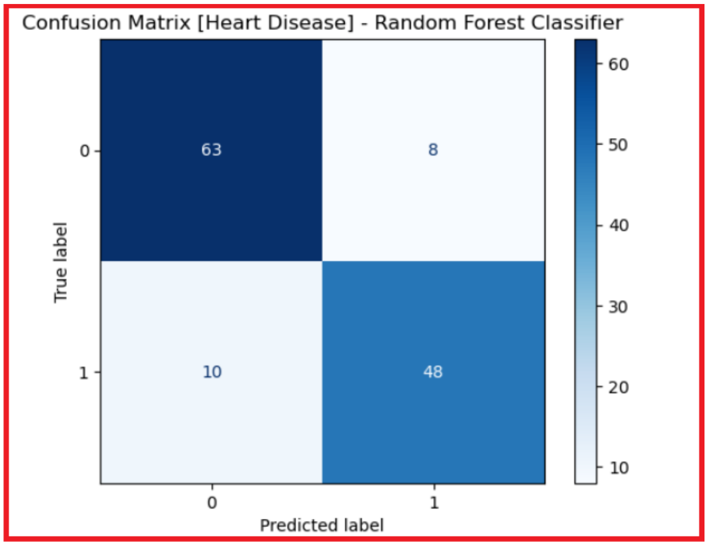
Let's explore additional metrics that highlight the importance of model evaluation in machine learning.

A. Target Heart Disease Precision Vs. Recall

Precision and recall are key metrics for evaluating our heart disease prediction model.

1. **Precision:** measures the accuracy of predicted heart disease cases, and our model currently has a very high precision score, indicating a good model performance.
2. **Recall:** measures how many actual cases of heart disease were correctly identified. Although a recall rate with 10 missed cases is acceptable, there is still room for

improvement if our goal is to minimize the risk of misdiagnosing patients with heart disease. Failing to identify these cases can pose significant risks.



The image on the Target Heart Disease Confusion Matrix

B. Target Heart Disease Confusion Matrix

The confusion matrix above shows how well our model predicted heart disease outcomes. It has two key sections:

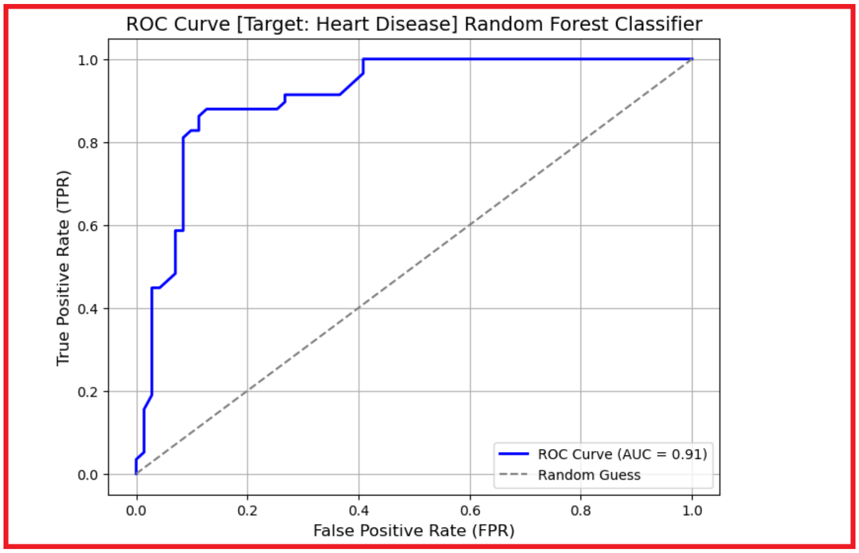
1) The top-left box shows 63 correct predictions for patients who did not have heart disease (True Negatives).
2) The bottom-left box shows 10 cases where the model missed predicting a heart disease (False Negatives).

This matrix demonstrates that the model is highly effective at predicting both cases of no heart disease and heart disease, with only a few errors that can be improved in the future.

C. Target Heart Disease ROC Curve

The ROC (Receiver Operating Characteristic) curve is a tool used to assess how effectively our model differentiates between cases of heart disease and those without. It illustrates the trade-off between the True Positive Rate (the proportion of actual heart disease cases correctly identified) and the False Positive Rate (the proportion of non-disease cases incorrectly identified as having heart disease). A curve that is closer to the top left corner indicates better performance, suggesting that our model performs well.

The image on the Heart Disease ROC Curve



This concludes the article. The next phase of this project is the ***deployment*** phase, which is beyond the scope of this discussion.

Conclusion

This project has yielded a powerful machine-learning model that can accurately predict the risk of heart disease. By leveraging key health indicators, this model empowers healthcare professionals to identify individuals at risk early on, enabling timely interventions and personalized treatment plans.

As we continue to refine and improve this model, we are hopeful that it will play a significant role in reducing the global burden of heart disease and ultimately saving lives.