Predicting Heart Disease Using Machine Learning

Adham mohamed Ibrahim

Faculty of computer science and information systems

October 6 university

Abstract: Heart disease is inflammation or damage to the heart and blood vessels over time. the disease can affect anyone of any age, gender, or social status. After many studies trying to overcome and learn about heart disease, in the end, this disease can be detected using machine learning systems. It predicts the likelihood of developing heart disease. The results of this system give the probability of heart disease as a percentage. Data collection using secret data mining. The data assets handled in python programming use two main algorithms for machine learning, Random forest and SVM algorithm which shows the best of both for heart disease accuracy. The results we get from this study show that the SVM algorithm is the algorithm with the most excellent precision. and the highest accuracy with a score of 95% in predicting heart disease using machine learning algorithms

Keywords: Heart disease, Machine learning, Support vector machine, Logistic Regression, Decision Tree Classifier, Random Forest Classifier,, Exploratory Data Analysis (EDA),Data Preprocessing, Evaluation metrics

INTRODUCTION:

Cardiovascular diseases (CVDs), or cardiovascular disease, are the world's leading cause of death, and are estimated to be responsible for 17.9 million deaths annually, according to the World Health Organization. CVDs comprise a broad group of disorders, including coronary artery disease, myocardial infarction, heart failure, and arrhythmias. Despite enhanced medical science, early detection and effective treatment of heart disease are still a daunting task since the role of the intricate interaction of genes, environment, and lifestyle makes the factors involved intricate.

Progress in data-driven technology has offered recent shining fields to improve the result of healthcare. Among them, machine learning (ML), a field of continued artificial intelligence, has shown unprecedented potential to reshape the future of cardiovascular medicine. With vast amounts of heterogeneous medical data, ML algorithms can find hidden patterns, identify risk factors, and support diagnostic and prognostic decision-making with unmatched accuracy.

In heart disease, machine learning techniques are being used more and more to predict the development of disease, risk-stratify patients, and assist clinical decision-making. These models utilize data such as electronic health records (EHRs), electrocardiograms (ECGs), imaging data, and laboratory results to identify subtle signs of cardiovascular dysfunction that cannot be detected using routine diagnostic tests. Machine learning in cardiology thus holds vast vows to enhance early diagnosis, enable tailored therapy, and therefore reduce the global burden of cardiovascular disease.

Related work:

In the realm of heart disease prediction research, the significance of the heart as a vital human organ cannot be overstated. Its role in blood circulation, akin to the importance of oxygen for human survival, underscores the need for its protection. Researchers across various domains, such as artificial intelligence, machine learning, and data mining, have devoted their efforts to this critical area. The field of heart disease prediction is not unexplored; a lot of people investigated the field and come up with satisfactory results. We will mention some of the papers we read and analyzed to assist us in our research, and all the papers mentioned will be referenced in the references section.

COMPARISON OF VARIOUS EXISTING ML METHODS:

Papers/year	Methods/Classifiers	datasets	Highest Accuracy	Best model	Technique
2025 [3]	KNN/Decision Tree/Random Forest	UCI Dataset	0.81	KNN	classification
2022 [1]	SVM/KNN/Decision Tree/Logistic regression/Naïve Bayes/neural networks	UCI Dataset	0.85	SVM	classification
2023 [6]	Gradient Boosting/ KNN/Decision Tree/Logistic regression/Naïve Bayes/neural networks	UCI Dataset	0.908	Gradient Boosting/ Logistic regression	classification
2023 [7]	SVM/KNN/Decision Tree/Logistic regression/Random Forest/XG boost	Cleveland	87.91%	SVM	classification
2024 [8]	SVM, Logistic Regression, Random Forest	UCI Kaggle	0.72	SVM	classification
2024 [9]	Random Forest	Framingham	86.52	Random Forest	classification
2023 [10]	SVM/KNN/Decision Tree/Logistic regression/Naïve Bayes/Random Forest	UCI dataset	94.79	Random Forest	classification

Literature Review:

Heart disease:

Heart disease is a term used to describe a variety of conditions that can affect your heart. When people think about heart disease, they may be thinking of the most common type — coronary artery disease (CAD) and

the heart attacks it can cause. But you can also have issues with other parts of your heart, like your heart muscle, valves or electrical system.

When your heart isn't functioning properly, it struggles to pump enough blood, oxygen and nutrients to your body. In a sense, your heart is supplying the fuel that keeps your body's systems going. If something goes wrong with the delivery of that fuel, it impacts everything your body's systems do.

-Heart disease symptoms:

Heart disease may present itself in a variety of ways, and symptoms do differ from one individual to another. Symptoms that many individuals have in common are: Chest pressure or pain (angina), Shortness of breath, Pain or discomfort in the neck, jaw, throat, upper abdomen, or back, Fatigue, Irregular heartbeat, Swelling in the legs, ankles, or feet (edema), Dizziness or lightheadedness.

Prompt identification of these symptoms is crucial as early identification and treatment can significantly improve the patient outcome. Machine learning algorithms trained to work with indicators of patient health attempt to recognize individuals with patterns in them suggestive of these symptoms.

Causes of Heart Disease:

Medical and lifestyle factors cause heart disease. Some of the most significant causes are:

Atherosclerosis: Fat plaque deposition in arteries

High blood pressure (hypertension): Places additional burden on the heart that gradually damages it

High cholesterol: Can cause plaque deposits in arteries

Diabetes: Typically accompanies an increased risk of cardiovascular disease

Smoking: Harms the inner wall of the arteries and leads to plaque deposits

Obesity: Associated with high blood pressure, cholesterol, and diabetes risk

Physical inactivity: Lowers cardiovascular fitness and leads to obesity

Poor diet: Diets high in cholesterol, saturated fat, and trans fat increase the risk of heart disease

Heavy drinking: Can lead to high blood pressure and damage to the heart muscle

-Machine learning:

Machine learning (ML) is a branch of artificial intelligence (AI) focused on enabling computers and machines to imitate the way that humans learn, to perform tasks autonomously, and to improve their performance and accuracy through experience and exposure to more data.

Machine learning is a branch of AI focused on building computer systems that learn from data. The breadth of ML techniques enables software applications to improve their performance over time.

Machine learning techniques classified into three techniques supervised, semi-supervised and unsupervised learning techniques

1-Supervised learning:

Supervised learning, also known as supervised machine learning, is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weight until it has been fitted appropriately. This occurs as part of the cross-validation process to ensure that the model avoids overfitting or underfitting. Supervised learning helps organizations solve a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, Naïve Bayes, linear regression, logistic regression, random forest, and support vector machine (SVM).

Classification:

Classification models are the second type of Supervised Learning techniques, which are used to generate conclusions from observed values in the categorical form. For example, the classification model can identify if the email is spam or not; a buyer will purchase the product or not, etc. Classification algorithms are used to predict two classes and categorize the output into different groups. In classification, a classifier model is designed that classifies the dataset into different categories, and each category is assigned a label like logistic regression.

Regression:

Regression in machine learning refers to a technique where the goal is to predict a continuous numerical value based on one or more independent features. It finds relationships between variables so that predictions can be made. We have two types of variables present in regression:

Dependent Variable (Target): The variable we are trying to predict e.g house price.

Independent Variables (Features): The input variables that influence the prediction e.g locality, number of rooms.

Regression analysis problem works with if output variable is a real or continuous value such as “salary” or “weight”. Many different regression models can be used but the simplest model in them is linear regression.

2-unsupervised learning:

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets (subsets called clusters). These algorithms discover hidden patterns or data groupings without the need for human intervention.

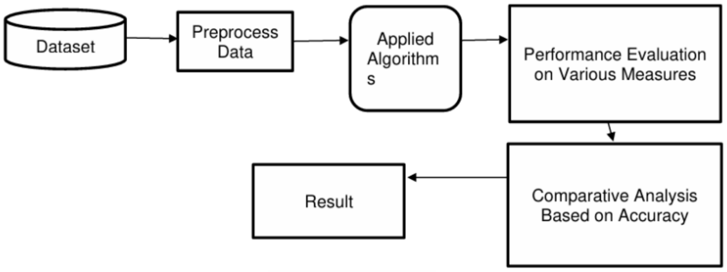
Unsupervised learning ability to discover similarities and differences in information make it ideal for exploratory data analysis, cross-selling strategies, customer segmentation, and image and pattern recognition. It’s also used to reduce the number of features in a model through the process of dimensionality reduction. Principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, and probabilistic clustering methods.

3-semi-supervised learning:

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of not having enough labeled data for a supervised learning algorithm. It also helps if it’s too costly to label enough data.

| METHODOLOGY

Here in this section, we are going to learn about the various classifiers used in machine learning for the prediction of heart disease. We are also going to present our proposed methodology for improving accuracy. Four different approaches have been used in this paper. The different approaches used are mentioned below. The output is the measure of accuracy of the machine learning models. Then, the model can be used in prediction.


Dataset
Preprocess Data
Applied Algorithm s
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Result

1.Dataset:

The Cleveland heart disease dataset is commonly used for heart disease prediction with supervised Machine Learning. The Cleveland dataset is obtained from the Kaggle Machine Learning repository. The Cleveland dataset was collected for use in a study in the field of health research by the Cleveland Clinic Foundation in 1988. In the original of this dataset, 76 different features of 303 subjects were recorded. However, it is known that most researchers use only 14 of these features, including the target class feature. These features include age, gender, blood pressure, cholesterol, blood sugar, and many more health metrics. The original Cleveland dataset has five class labels. It has integer values ranging from zero (no presence) to four. The Cleveland dataset experiments focused on just trying to discriminate between presence (Values 1, 2, 3, 4) and absence (Value 0). However, the number of samples for each class is not homogeneous (Values 0, 1, 2, 3 , 4—samples 164, 55, 36, 35, 13). Researchers suggest that the five class features of this data set be reduced to Two classes: 0 = no disease and 1 = disease. The target feature refers to the presence of heart disease in the subject. Table 1 shows the features included in the Cleveland heart disease dataset.

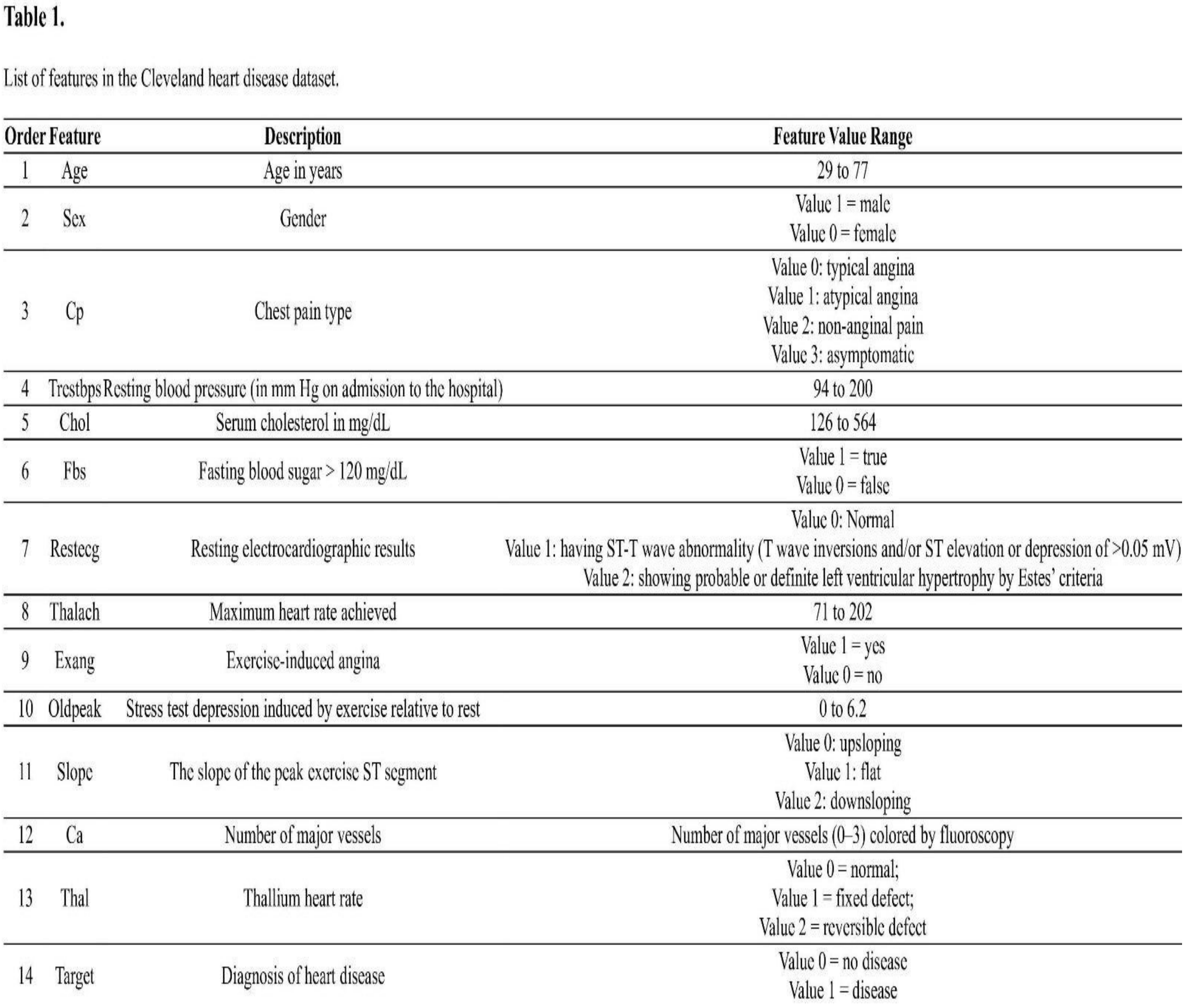


Table 1.

List of features in the Cleveland heart disease dataset.

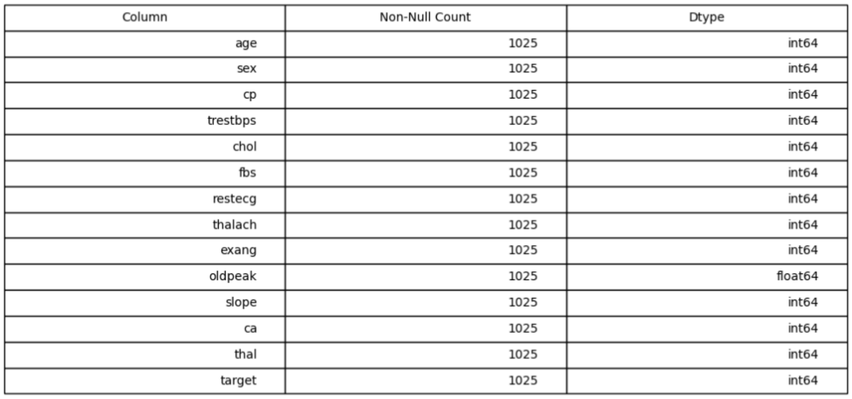
Order	Feature	Description	Feature Value Range
1	Age	Age in years	29 to 77
2	Sex	Gender	Value 1 = male Value 0 = female
3	Cp	Chest pain type	Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	94 to 200
5	Chol	Serum cholesterol in mg/dL	126 to 564
6	Fbs	Fasting blood sugar > 120 mg/dL	Value 1 = true Value 0 = false
7	Restecg	Resting electrocardiographic results	Value 0: Normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	Thalach	Maximum heart rate achieved	71 to 202
9	Exang	Exercise-induced angina	Value 1 = yes Value 0 = no
10	Oldpeak	Stress test depression induced by exercise relative to rest	0 to 6.2
11	Slope	The slope of the peak exercise ST segment	Value 0: upsloping Value 1: flat Value 2: downsloping
12	Ca	Number of major vessels	Number of major vessels (0–3) colored by fluoroscopy

13	Thal	Thallium heart rate	Value 0 = normal;
			Value 1 = fixed defect;
			Value 2 = reversible defect

14	Target	Diagnosis of heart disease	Value 0 = no disease Value 1 = disease

Dataset overview:

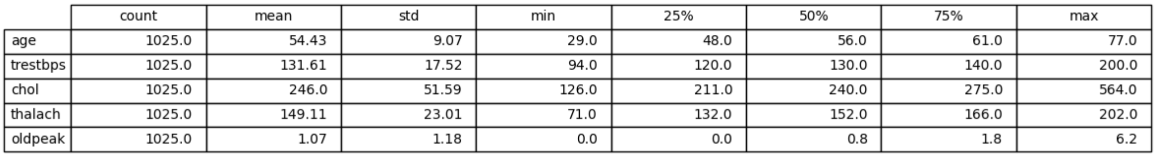
DataFrame Info Table



Column	Non-Null Count	Dtype
age	1025	int64
sex	1025	int64
cp	1025	int64
trestbps	1025	int64
chol	1025	int64
fbs	1025	int64
restecg	1025	int64
thalach	1025	int64
exang	1025	int64
oldpeak	1025	float64
slope	1025	int64
ca	1025	int64
thal	1025	int64
target	1025	int64

Statistics for Numerical Variables:

Descriptive Statistics Table

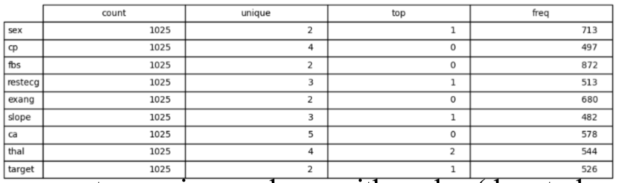


	count	mean	std	min	25%	50%	75%	max
age	1025.0	54.43	9.07	29.0	48.0	56.0	61.0	77.0
trestbps	1025.0	131.61	17.52	94.0	120.0	130.0	140.0	200.0
chol	1025.0	246.0	51.59	126.0	211.0	240.0	275.0	564.0
thalach	1025.0	149.11	23.01	71.0	132.0	152.0	166.0	202.0
oldpeak	1025.0	1.07	1.18	0.0	0.0	0.8	1.8	6.2

age: The average age of the patients is approximately 54.4 years, with the youngest being 29 and the oldest 77 years.
trestbps: The average resting blood pressure is about 131.62 mm Hg, ranging from 94 to 200 mm Hg.
Chol: The average cholesterol level is approximately 246.26 mg/dl, with a minimum of 126 and a maximum of 564 mg/dl.
thalach: The average maximum heart rate achieved is around 149.65, with a range from 71 to 202.
old peak: The average ST depression induced by exercise relative to rest is about 1.04, with values ranging from 0 to

Statistics for Categorical Variables:

Descriptive Stats for Object Columns



	count	unique	top	freq
sex	1025	2	1	713
cp	1025	4	0	497
fbs	1025	2	0	872
restecg	1025	3	1	513
exang	1025	2	0	680
slope	1025	3	1	482
ca	1025	5	0	578
thal	1025	4	2	544
target	1025	2	1	526

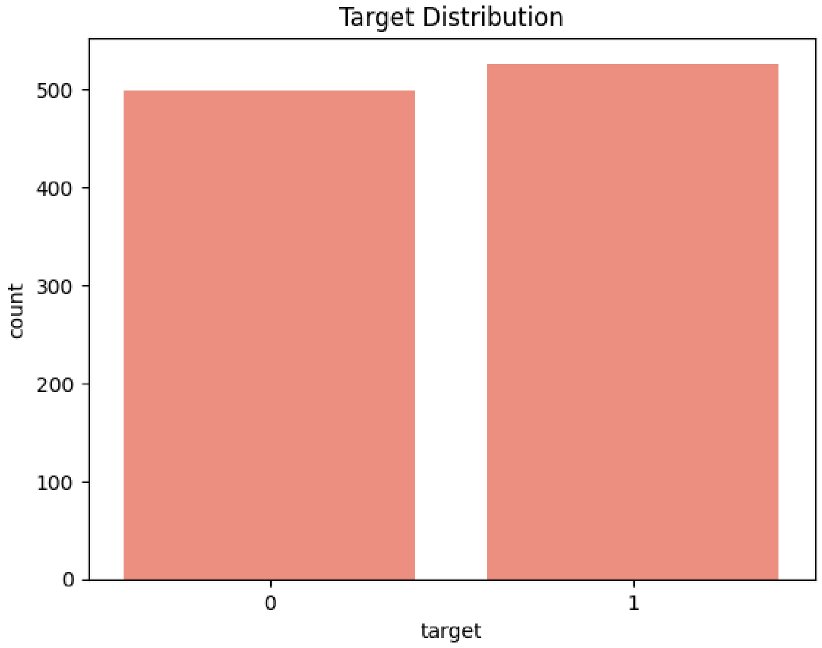
sex: There are two unique values, with males (denoted as 0) being the most frequent category, occurring 207 times out of 303 entries.
cp: Four unique types of chest pain are present. The most common type is "0", which occurs 143 times.
fbs: There are two categories, and the most frequent one is "0" (indicating fasting blood sugar less than 120 mg/dl), which appears 258 times.
restecg: Three unique results are present. The most common result is "1", appearing 152 times.
exang: There are two unique values. The most frequent value is "0" (indicating no exercise-induced angina), which is observed 204 times.
slope: Three unique slopes are present. The most frequent slope type is "2", which occurs 142 times.
ca: There are five unique values for the number of major vessels colored by fluoroscopy, with "0" being the most frequent, occurring 175 times.
thal: Four unique results are available. The most common type is "2" (indicating a reversible defect), observed 166 times.
target: Two unique values indicate the presence or absence of heart disease. The value "1" (indicating the presence of heart disease) is the most frequent, observed in 165 entries.

Exploratory Data Analysis (EDA):

The EDA process involved the following steps to get accustomed to the dataset and identify patterns that would be helpful for heart disease prediction:

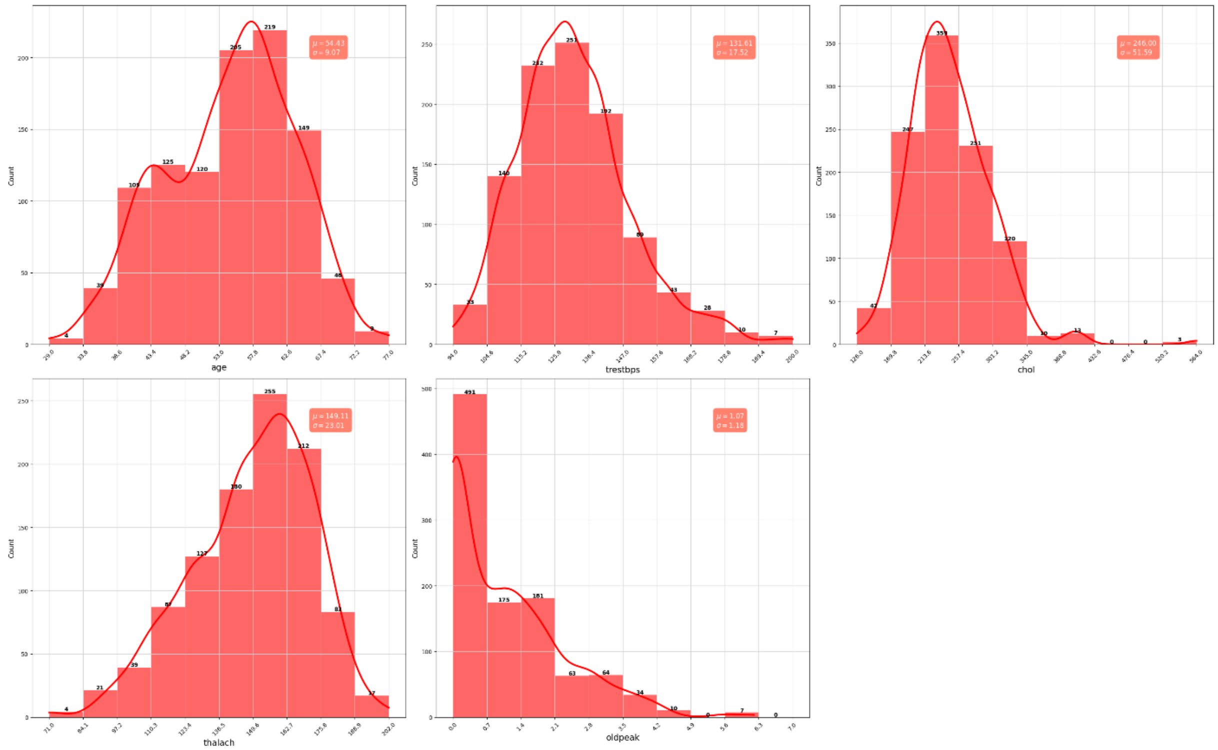
Target Variable Distribution:

A count plot was used to represent the target variable distribution with a predominantly balanced dataset between being influenced by heart disease (1) and not being influenced (0).



Univariate Analysis:

1-Numerical Features: Histograms and boxplots were used to track feature distributions like age, cholesterol, resting blood pressure, and maximal heart rate.



Age (age): The distribution is somewhat uniform, but there's a peak around the late 50s. The mean age is approximately 54.43 years with a standard deviation of 9.07 years.

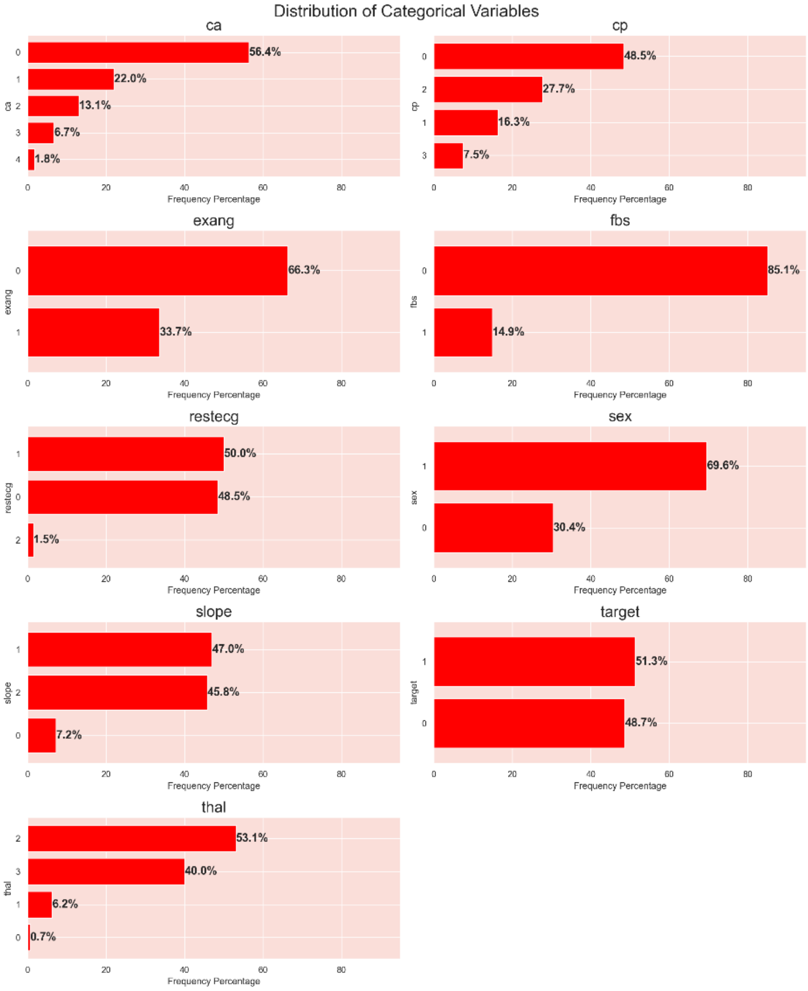
Resting Blood Pressure (trestbps): The resting blood pressure for most individuals is concentrated around 120-140 mm Hg, with a mean of approximately 131.61 mm Hg and a standard deviation of 17.52 mm Hg.

Serum Cholesterol (chol): Most individuals have cholesterol levels between 200 and 300 mg/dl. The mean cholesterol level is around 246 mg/dl with a standard deviation of 51.59 mg/dl.

Maximum Heart Rate Achieved (thalach): The majority of the individuals achieve a heart rate between 140 and 170 bpm during a stress test. The mean heart rate achieved is approximately 149.11 bpm with a standard deviation of 23.01bpm.

ST Depression Induced by Exercise (oldpeak): Most of the values are concentrated towards 0, indicating that many individuals did not experience significant ST depression during exercise. The mean ST depression value is 1.07 with a standard deviation of 1.18

2-Categorical Features: Count plots were used for features like chest pain type, fasting blood sugar, and electrocardiographic findings to determine class frequencies.



Gender (sex): The dataset is predominantly female, constituting a significant majority.

Type of Chest Pain (cp): The dataset shows varied chest pain types among patients. Type 0 (Typical angina) seems to be the most prevalent, but an exact distribution among the types can be inferred from the bar plots.

Fasting Blood Sugar (fbs): A significant majority of the patients have their fasting blood sugar level below 120 mg/dl, indicating that high blood sugar is not a common condition in this dataset.

Resting Electrocardiographic Results (restecg): The results show varied resting electrocardiographic outcomes, with certain types being more common than others. The exact distribution can be gauged from the plots.

Exercise-Induced Angina (exang): A majority of the patients do not experience exercise-induced angina, suggesting that it might not be a common symptom among the patients in this dataset.

Slope of the Peak Exercise ST Segment (slope): The dataset shows different slopes of the peak exercise ST segment. A specific type might be more common, and its distribution can be inferred from the bar plots.

Number of Major Vessels Colored by Fluoroscopy (ca): Most patients have fewer major vessels colored by fluoroscopy, with '0' being the most frequent.

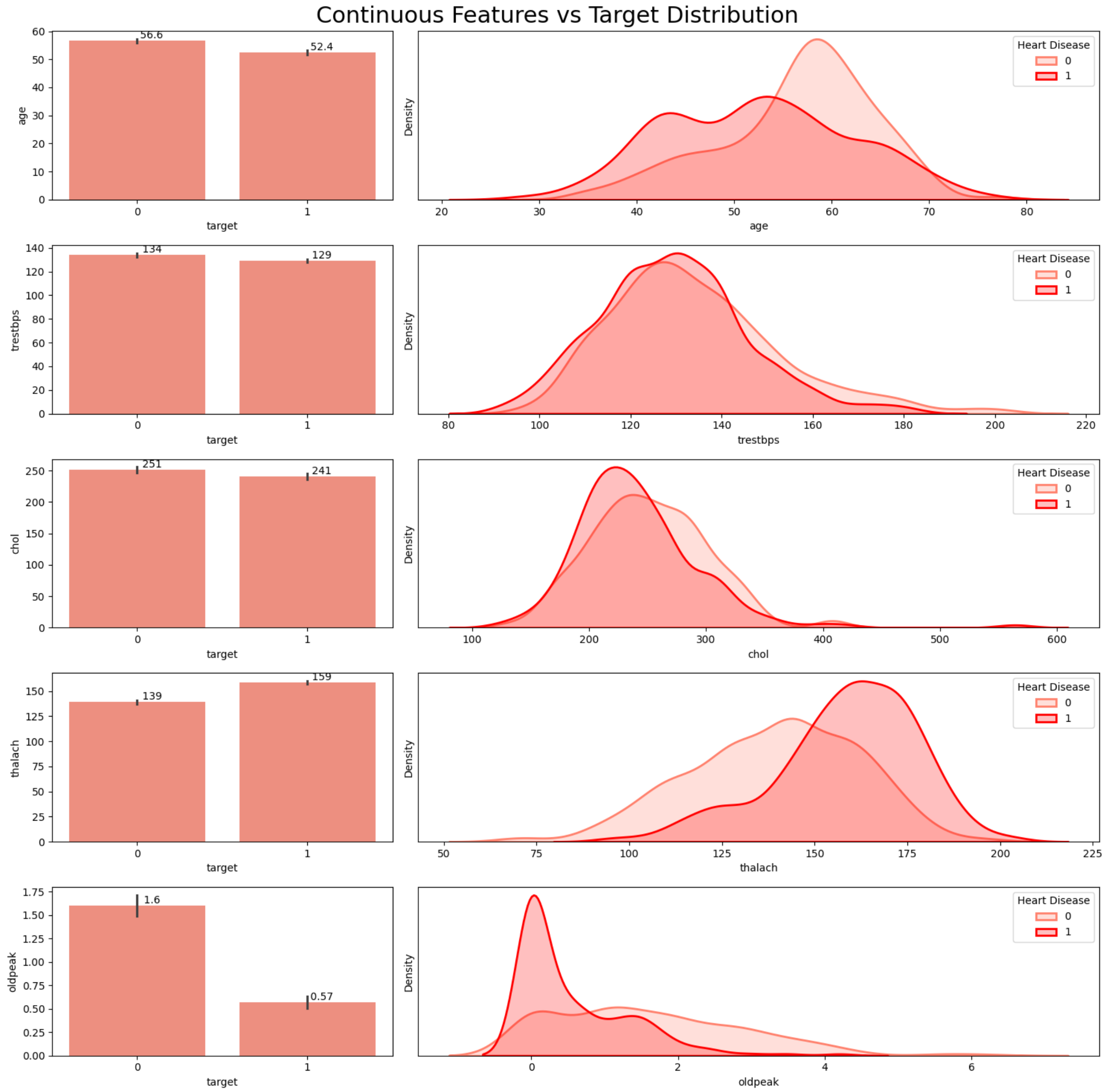
Thallium Stress Test Result (thal): The dataset displays a variety of thallium stress test results. One particular type seems to be more prevalent, but the exact distribution can be seen in the plots.

Presence of Heart Disease (target): The dataset is nearly balanced in terms of heart disease presence, with about 54.5% having it and 45.5% not having it.

Bivariate Analysis:

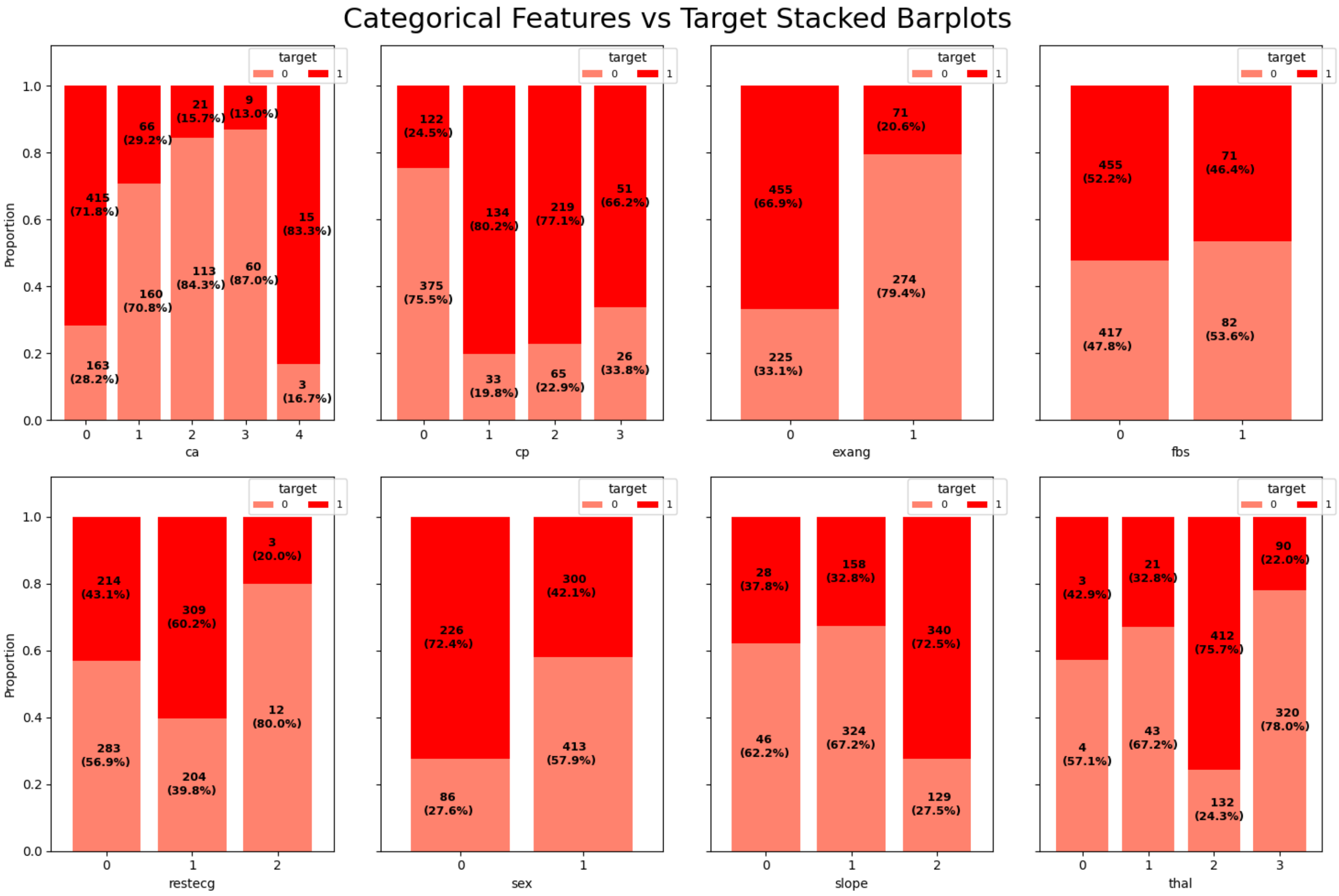
Association of each feature with the target variable was illustrated using boxplots and violin plots, i.e., for:

Continuous features vs. heart disease



Age (age): The distributions show a slight shift with patients having heart disease being a bit younger on average than those without. The mean age for patients without heart disease is higher.
Resting Blood Pressure (trestbps): Both categories display overlapping distributions in the KDE plot, with nearly identical mean values, indicating limited differentiating power for this feature.
Serum Cholesterol (chol): The distributions of cholesterol levels for both categories are quite close, but the mean cholesterol level for patients with heart disease is slightly lower.
Maximum Heart Rate Achieved (thalach): There's a noticeable difference in distributions. Patients with heart disease tend to achieve a higher maximum heart rate during stress tests compared to those without.
ST Depression (oldpeak): The ST depression induced by exercise relative to rest is notably lower for patients with heart disease. Their distribution peaks near zero, whereas the non-disease category has a wider spread.

Categorical features vs heart disease:



Number of Major Vessels (ca): The majority of patients with heart disease have fewer major vessels colored by fluoroscopy. As the number of colored vessels increases, the proportion of patients with heart disease tends to decrease. Especially, patients with 0 vessels colored have a higher proportion of heart disease presence.

Chest Pain Type (cp): Different types of chest pain present varied proportions of heart disease. Notably, types 1, 2, and 3 have a higher proportion of heart disease presence compared to type 0. This suggests the type of chest pain can be influential in predicting the disease.

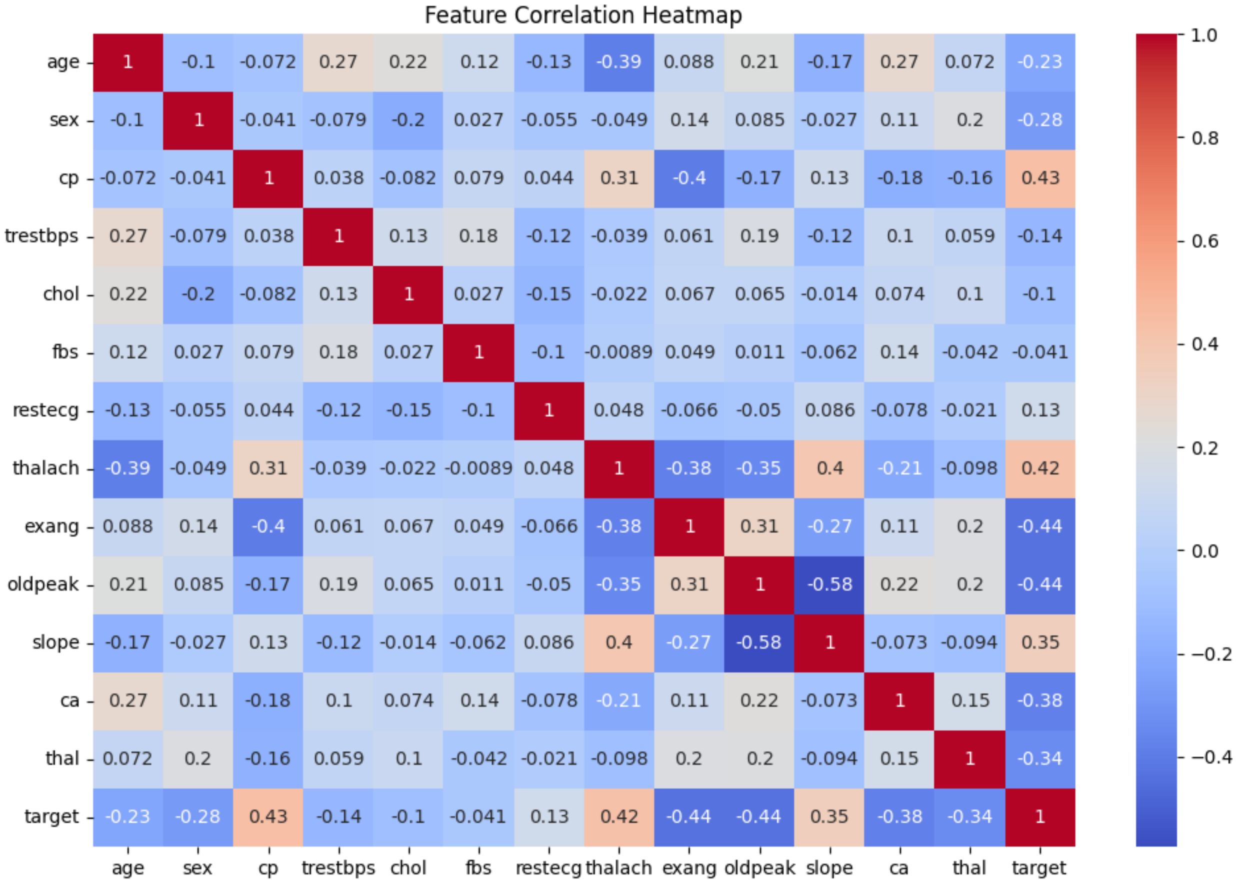
Exercise Induced Angina (exang): Patients who did not experience exercise-induced angina (0) show a higher proportion of heart disease presence compared to those who did (1). This feature seems to have a significant impact on the target.

Fasting Blood Sugar (fbs): The distribution between those with fasting blood sugar > 120 mg/dl (1) and those without (0) is relatively similar, suggesting fbs might have limited impact on heart disease prediction.

Resting Electrocardiographic Results (restecg): Type 1 displays a higher proportion of heart disease presence, indicating that this feature might have some influence on the outcome.
Sex (sex): Females (1) exhibit a lower proportion of heart disease presence compared to males (0). This indicates gender as an influential factor in predicting heart disease.

Slope of the Peak Exercise ST Segment (slope): The slope type 2 has a notably higher proportion of heart disease presence, indicating its potential as a significant predictor.

Thallium Stress Test Result (thal): The reversible defect category (2) has a higher proportion of heart disease presence compared to the other categories, emphasizing its importance in prediction.
Correlation Map:



cp (Chest Pain Type): +0.43 → Positive. Higher chest pain score is linked with higher probability of heart disease.

thalach (Max Heart Rate): +0.42 → Increased. Increased heart rate is linked with occurrence of heart disease.

exang (Exercise-induced angina): −0.44 → Negative. Exercise-induced angina is likely to indicate lower chance of having a healthy heart.

oldpeak (ST depression): −0.45 → Negative correlation. ST depression at exercise more → likely to have heart issues.

ca (Number of large vessels filled in by fluoroscopy): −0.38 → More vessels uncovered = poorer condition = less likely to have healthy heart.

thal: −0.34 → Certain thalassemia levels more connected with disease.

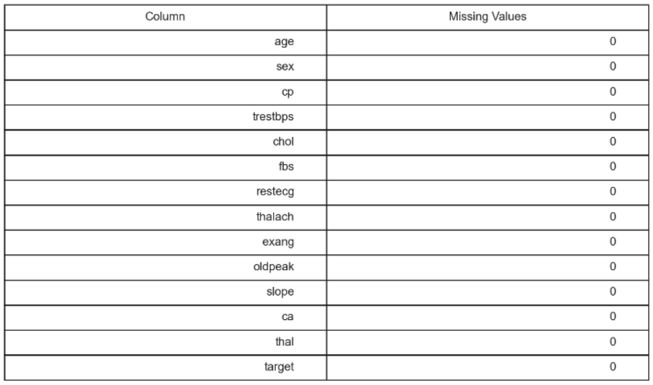
2- Data Preprocessing:

Below are the steps used for preprocessing to clean data for modeling:

Handling Missing Values:

Missing values in the data, if present, were identified and replaced or dropped to make the data complete. Although this data was relatively clean, some preprocessing was done to keep data consistent everywhere.

Missing Values Table

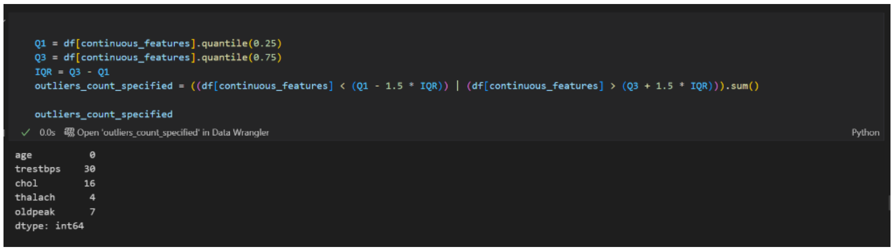


Column	Missing Values
age	0
sex	0
cp	0
trestbps	0
chol	0
fbs	0
restecg	0
thalach	0
exang	0
oldpeak	0
slope	0
ca	0
thal	0
target	0

Treatment of Outliers:

Outliers identified during EDA (like extremely high cholesterol levels) were capped or dropped to prevent the learning process of the model.

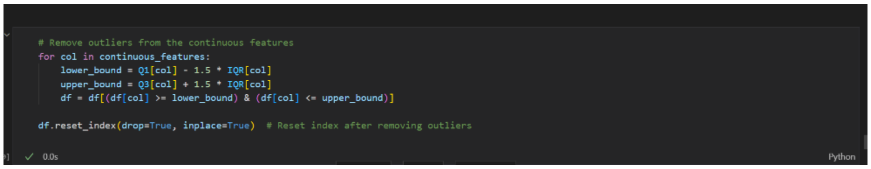
Outliers check:



```python
Q1 = df[continuous_features].quantile(0.25)
Q3 = df[continuous_features].quantile(0.75)
IQR = Q3 - Q1
outliers_count_specified = ((df[continuous_features] < (Q1 - 1.5 * IQR)) | (df[continuous_features] > (Q3 + 1.5 * IQR))).sum()

outliers_count_specified
```

```
age          0
trestbps    30
chol        16
thalach      4
oldpeak      7
dtype: int64
```

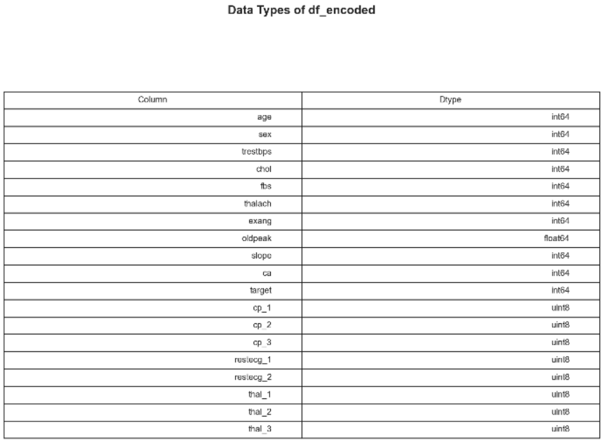
Treating outliers:



```python
# Remove outliers from the continuous features
for col in continuous_features:
    lower_bound = Q1[col] - 1.5 * IQR[col]
    upper_bound = Q3[col] + 1.5 * IQR[col]
    df = df[(df[col] >= lower_bound) & (df[col] <= upper_bound)]

df.reset_index(drop=True, inplace=True)  # Reset index after removing outliers
```

Encoding Categorical Variables:

Categorical attributes such as type of chest pain, thalassemia, and rest ECG result were converted into numerical form using one-hot encoding to be prepared for use by machine learning algorithms.



Data Types of df_encoded

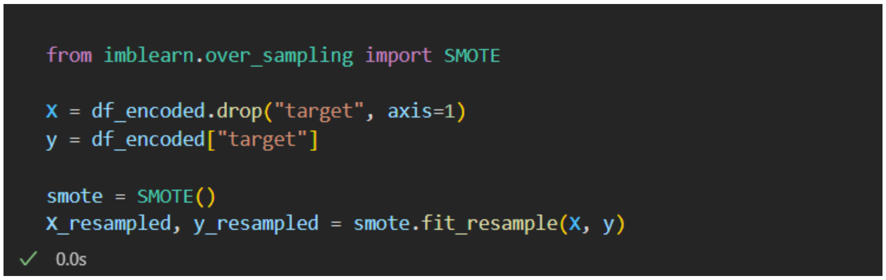
Column	Dtype
age	int64
sex	int64
trestbps	int64
chol	int64
fbs	int64
thalach	int64
exang	int64
oldpeak	float64
slope	int64
ca	int64
target	int64
cp_1	uint8
cp_2	uint8
cp_3	uint8
restecg_1	uint8
restecg_2	uint8
thal_1	uint8
thal_2	uint8
thal_3	uint8

Feature Scaling:

Numerical attributes were normalized using techniques like Min-Max normalization or Standardization (z-score scaling) to get features scalable. It is particularly beneficial for distance-based models like KNN and SVM.

SMOTE (Synthetic Minority Oversampling Technique):

SMOTE is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesizes new minority instances between existing minority instances. It generates the virtual training records by linear interpolation

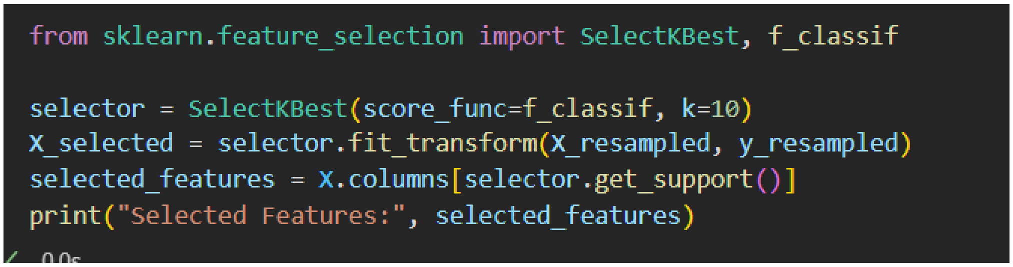


```
from imblearn.over_sampling import SMOTE

X = df_encoded.drop("target", axis=1)
y = df_encoded["target"]

smote = SMOTE()
X_resampled, y_resampled = smote.fit_resample(X, y)
```

✓ 0.0s

Feature Selection (Optional):

Features of low correlation or redundant features were monitored and perhaps removed to increase the efficiency of models and generalization.



```
from sklearn.feature_selection import SelectKBest, f_classif

selector = SelectKBest(score_func=f_classif, k=10)
X_selected = selector.fit_transform(X_resampled, y_resampled)
selected_features = X.columns[selector.get_support()]
print("Selected Features:", selected_features)
```

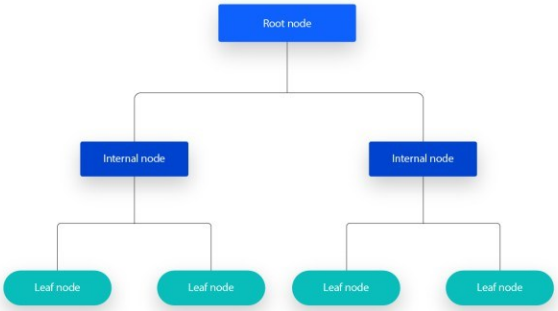
✓ 0.0s

modeling techniques:

. decision tree:

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root

Node , branches, internal nodes and leaf nodes.



Gini index:

The Gini Index or Impurity measures the probability for a random instance being misclassified when chosen randomly. The lower the Gini Index, the better the lower the likelihood of misclassification.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

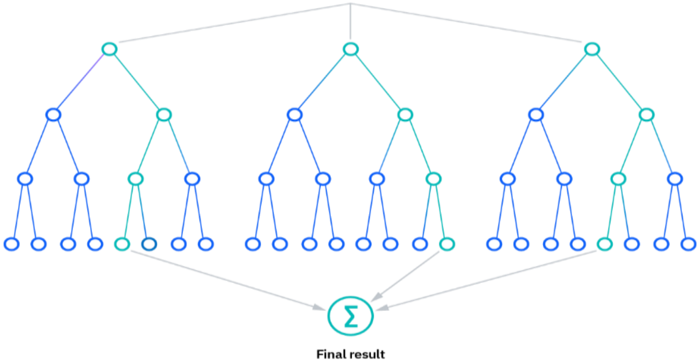
Entropy:

Entropy quantifies the randomness in the dataset. It's measured in bits and ranges from 0 to 1.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

. Random forest classifier:

Random forest is a commonly used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

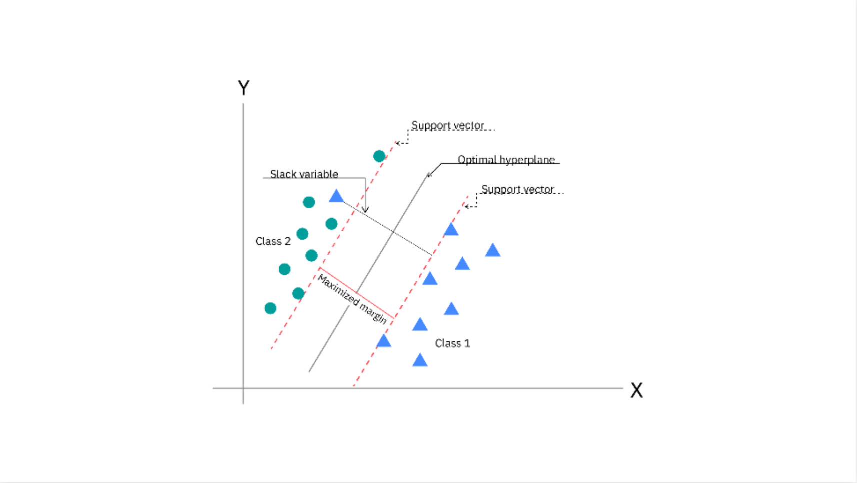


. Support vector machine (SVM):

A support vector machine (SVM) is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space.

SVMs were developed in the 1990s by Vladimir N. Vapnik and his colleagues, and they published this work in a paper titled "Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing"[1] in 1995.

SVMs are commonly used within classification problems. They distinguish between two classes by finding the optimal hyperplane that maximizes the margin between the closest data points of opposite classes. The number of features in the input data determine if the hyperplane is a line in a 2-D space or a plane in a n-dimensional space. Since multiple hyperplanes can be found to differentiate classes, maximizing the margin between points enables the algorithm to find the best decision boundary between classes. This, in turn, enables it to generalize well to new data and make accurate predictions. The lines that are adjacent to the optimal hyperplane are known as support vectors as these vectors run through the data points that determine the maximal margin.



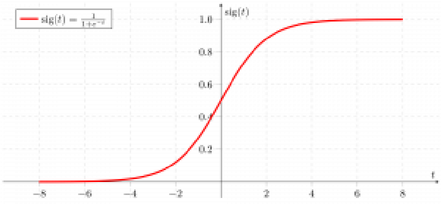
Logistic regression:

Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyzes the relationship between two data factors. The article explores the fundamentals of logistic regression, its types and implementations.

Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.

Sigmoid Function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



RESULTS AND DISCUSSION:

The results from the implementation of the machine learning classifier algorithm, Support Vector Machine (SVM),Logistic Regression Algorithm, Decision Tree Algorithm (DC Algorithm) and Random Forest are demonstrated in this section. The metrics Accuracy score, Precision (P), Recall/Sensitivity (R), and F1 Score are being used to analyze the algorithm's performance. The precision parameter is the ratio of the positive correct predictions (TP) to the overall positive results which the model has predicted . Recall is calculated as the proportion of the correct positive prediction (TP) to the total positive data. The weighted comparison of the average precision and recall is defined by the F1 Score . Accuracy The correlation between the prediction value and the total quantity of data

Model evaluation:

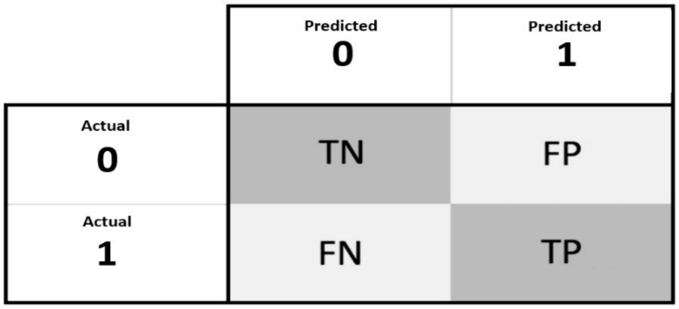
. Evaluation Metrics:

Evaluation metrics are quantitative measures used to assess the performance and effectiveness of a statistical or machine learning model. These metrics provide insights into how well the model is performing and help in comparing different models or algorithms.

A confusion/classification matrix:

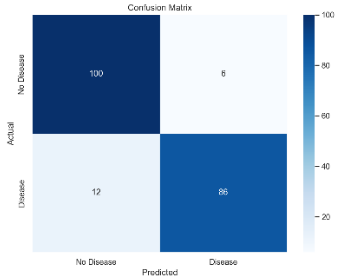
Popular tool for evaluating the classification model performance by comparing the predicted labels against the true labels.

Used to calculate several performance metrics: precision, recall, accuracy, F1-scorePrecision & recall are most frequently used to assess a binary model.

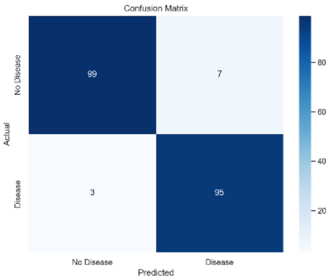


	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

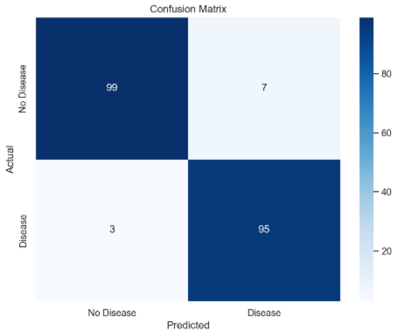
Decision tree confusion matrix:



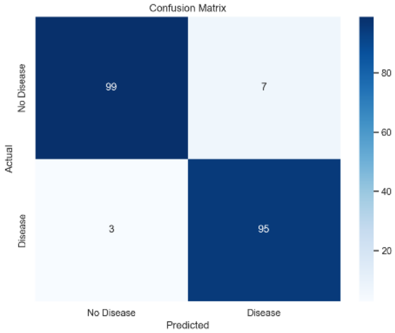
SVM confusion matrix:



Random forest confusion matrix:



Logistic regression confusion matrix



Accuracy:

Accuracy is the most common metric to be used in everyday talk. Accuracy answers the question **"Out of all the predictions we made, how many were true?"**

As we will see later, accuracy is a blunt measure and can sometimes be misleading.

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ negatives + false\ positives}$$

Precision

Precision is a metric that gives you the proportion of true positives to the amount of total positives that the model predicts. It answers the question **"Out of all the positive predictions we made, how many were true?"**

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Recall

Recall focuses on how good the model is at finding all the positives. Recall is also called true positive rate and answers the question "**Out of all the data points that should be predicted as true, how many did we correctly predict as true?**"

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

As you can see from the definitions of precision and recall they are tightly connected.

F1 Score

F1 Score is a measure that combines recall and precision. As we have seen there is a trade-off between precision and recall, F1 can therefore be used to measure how effectively our models make that trade-off.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Comparison between models' accuracy:

Decision Tree	91.1%
SVM	95.09%
Random forest	94.82%
Logistic regression	84.9%

Comparison between models' evaluation metrics for class 0:

model	Precision	Recall	F1-score
Decision tree	0.89	0.94	0.92
Random forest	0.96	0.94	0.95
Logistic Regression	0.87	0.82	0.84
SVM	0.97	0.93	0.95

Comparison between models' evaluation metrics for class 1:

model	Precision	Recall	F1-score
Decision tree	0.93	0.88	0.91
Random forest	0.94	0.96	0.95
Logistic Regression	0.83	0.88	0.86
SVM	0.93	0.97	0.95

Possible optimizations:

-Hyperparameter tuning:

When you're training machine learning models, each dataset and model needs a different set of hyperparameters, which are a kind of variable. The only way to determine these is through multiple experiments, where you pick a set of hyperparameters and run them through your model. This is called *hyperparameter tuning*. In essence, you're training your model sequentially with different sets of hyperparameters. This process can be manual, or you can pick one of several automated hyperparameter tuning methods.

Hyperparameters directly control model structure, function, and performance. Hyperparameter tuning allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success.

ACKNOWLEDGMENT:

We extend our heartfelt gratitude to all those who contributed to the successful completion of this research endeavor. First and foremost, we express our appreciation to the entire research community for their continuous efforts in advancing the field of heart disease prediction through machine learning. Special thanks are due to our academic mentors and advisors for their invaluable guidance, insights, and unwavering support throughout the research process. We also acknowledge the creators and maintainers of the datasets and tools that facilitated our analysis. Our thanks are also extended to the participants of this study for their cooperation and the individuals who inspired us with their commitment to improving healthcare. Finally, we express our profound appreciation for the encouragement and support from our families and loved ones

-Conclusion:

We determine the use of what machine learning algorithms will be effective to predict heart disease from appropriately formulated health information. With an ample range of supervised machine learning techniques applied on Cleveland Heart Disease, we checked on the predictability of the model through the usage of Decision Tree, Random Forest, Logistic Regression and Support Vector Machine (SVM). Among all of them, Random Forest classifier worked best on all the parameters of measurement precision, recall, and F1-score—and thus is the most reliable model to identify persons at risk.

Application of exploratory data analysis and adequate preprocessing of data, i.e., elimination of outliers, one-hot encoding of categorical attributes, and normalization of attributes, was crucial in improving model accuracy and generalizability. Reduction of recall also made the models extremely efficient at detecting true positive instances, a requirement which is extremely critical in medical diagnosis as not detecting a positive instance would be disastrous.

In summary, this research demonstrates the potential of machine learning to become a useful asset in early heart disease diagnosis. With continued evolution, clinically relevant application, and increasing access to large and heterogeneous data, the

models will facilitate clinical specialists to make improved and earlier-informed decisions, thus resulting in enhanced patient outcomes and reduced cardiovascular mortality.

References:

1. Luísa Soares, Tatiana Leal, Ana Lúcia Faria, Ana Aguiar and Cátia Carvalho. Cardiovascular Disease: A Review. Biomed J Sci & Tech Res 51(3)- 2023. BJSTR. MS.ID.008101.
2. https://my.clevelandclinic.org/health/diseases/16898-coronary-artery-disease
3. Muhammad, Bakhtawar & Umar, Hooria & Fatima Yousaf, Hoor & Nasir, Usama & Hussain, Muhammad Zunnurain & Hasan, Muhammad Zulkifl & Mustafa, Muzzamil & Yaqub, Muhammad. (2025). Heart Disease Prediction Using Machine Learning. 10.1109/IDICAIEI61867.2024.10842908.
4. https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/
5. Anderies, Anderies & Tchin, Jalaludin & Putro, Prambudi & Darmawan, Yudha & Gunawan, Alexander. (2022). Prediction of Heart Disease UCI Dataset Using Machine Learning Algorithms. Engineering, MAthematics and Computer Science (EMACS) Journal. 4. 87-93. 10.21512/emacsjournal.v4i3.8683.
6. AbdElminaam, D. S., Mohamed, N., Wael, H., Khaled, A., & Moataz, A. (2023). *MLHeartDisPrediction: Heart Disease Prediction using Machine Learning*. *Journal of Computing and Communication*, 2(1), 50–65.
7. Ahamad, G.N.; Shafiullah; Fatima, H.; Imdadullah.; Zakariya, S.; Abbas, M.; Alqahtani, M.S.; Usman, M. Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease. *Processes* **2023**, *11*, 734.
8. Bhowmik, P. K., Miah, M. N. I., Uddin, M. K., Sizan, M. M. H., Pant, L. P., Islam, R., & Gurung, N. (2024). Advancing Heart Disease Prediction through Machine Learning: Techniques and Insights for Improved Cardiovascular Health. *British Journal of Nursing Studies*, *4*(2), 35–50. https://doi.org/10.32996/bjns.2024.4.2.5

9. Mir, A., Ur Rehman, A., Ali, T. M., Javaid, S., Almufareh, M. F., Humayun, M., & Shaheen, M. (2024). A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques. *ESC Heart Failure*. https://doi.org/10.1002/ehf2.14942

10. Torthi, R., Marapatla, A. D. K., Mande, S., Gadiraju, H. K. V., & Kanumuri, C. (2023). *Heart disease prediction using random forest based hybrid optimization algorithms*

11. https://www.ibm.com/think/topics/random-forest

12. https://www.ibm.com/think/topics/support-vector-machine

13. https://www.ibm.com/think/topics/decision-trees

14. https://www.geeksforgeeks.org/understanding-logistic-regression/