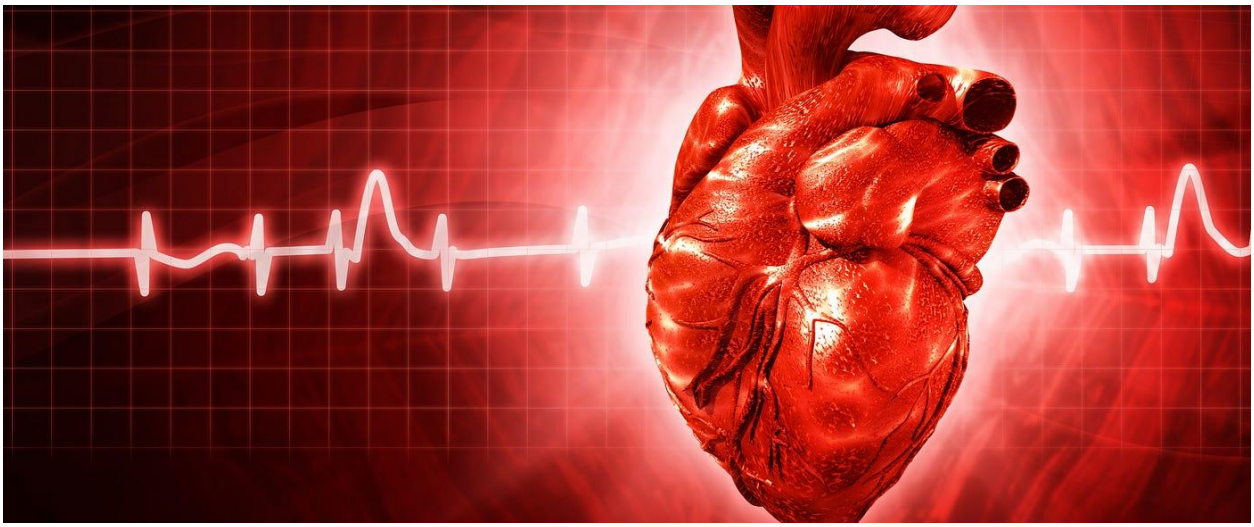
HEART DISEASE PREDICTION



Problem Statement

Heart disease is a leading cause of death globally, responsible for millions of fatalities each year. Early detection of heart disease can dramatically improve patient outcomes by enabling healthcare providers to offer preventative care, timely interventions, and personalized treatment plans. Given the volume of medical data available, there is a growing interest in using predictive analytics and machine learning to detect heart disease risks early on. Machine learning models can analyze patient health records and identify patterns associated with heart disease, providing critical insights to healthcare professionals.

In this project, the aim is to leverage machine learning techniques to predict the presence of heart disease in patients using their clinical and diagnostic data. This model can help healthcare professionals prioritize high-risk patients, improving patient outcomes and reducing the burden on healthcare systems.

Dataset Overview

The dataset consists of patient information, including demographic, clinical, and diagnostic test results, to help predict the likelihood of heart disease. The target variable is target, which indicates whether the patient has heart disease or not (1 = Yes, 0 = No). The dataset includes 14 features, each representing various health and diagnostic metrics, as detailed below:

- **Id**: Unique identifier for each patient.
- **Age**: Age of the patient (in years).
- **Sex**: Gender of the patient (0 = Female, 1 = Male).
- **cp (Chest Pain Type)**: Types of chest pain experienced:

    - 0 = Typical angina
    - 1 = Atypical angina
    - 2 = Non-anginal pain
    - 3 = Asymptomatic
- **trestbps (Resting Blood Pressure)**: Resting blood pressure (in mm Hg) measured during hospital admission.
- **chol (Serum Cholesterol)**: Serum cholesterol level in mg/dl.
- **fbs (Fasting Blood Sugar)**: Fasting blood sugar > 120 mg/dl (1 = True, 0 = False).
- **restecg (Resting Electrocardiographic Results)**:
    - 0 = Normal
    - 1 = ST-T wave abnormality (T wave inversions and/or ST elevation or depression)
    - 2 = Showing probable or definite left ventricular hypertrophy by Estes' criteria.
- **thalach (Maximum Heart Rate Achieved)**: Maximum heart rate achieved during exercise.
- **exang (Exercise-Induced Angina)**: Whether the patient experiences angina (chest pain) during exercise (1 = Yes, 0 = No).
- **oldpeak (ST Depression)**: ST depression induced by exercise relative to rest, a measure of abnormality.
- **slope (Slope of Peak Exercise ST Segment)**:
    - 0 = Upsloping
    - 1 = Flat
    - 2 = Downsloping.
- **ca (Number of Major Vessels)**: Number of major blood vessels (0-4) colored by fluoroscopy.
- **thal (Thalassemia Status)**:
    - 0 = Normal
    - 1 = Fixed defect
    - 2 = Reversible defect
    - 3 = Missing or unspecified.

- **target**: The target variable indicating the presence of heart disease:
    - 1 = Patient has heart disease
    - 0 = Patient does not have heart disease.

Importance of the Dataset

This dataset provides a comprehensive overview of key clinical indicators that are critical for detecting heart disease. Features like chest pain type, cholesterol levels, and maximum heart rate achieved, among others, have been found in previous medical studies to be strong predictors of cardiovascular risk. Machine learning models trained on this dataset can help uncover complex patterns among these variables that may not be immediately apparent through traditional analysis methods.

Basic Data Exploration

Before performing in-depth analysis, an initial exploration of the dataset was conducted to understand its structure.

- **Dataset Size**: The dataset contains 7**303 rows** and **15 columns**.
- **Data Types**:
    - Numerical: id, age, trestbps, chol, thalach, oldpeak, ca, sex, cp, fbs, restecg, exang, slope, thal, target
- **Missing Values**:
    - There is no missing value
- **Duplicates**
    - No duplicate

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7303 entries, 0 to 7302
Data columns (total 15 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        7303 non-null   int64
 1   age       7303 non-null   int64
 2   sex       7303 non-null   int64
 3   cp        7303 non-null   int64
 4   trestbps  7303 non-null   int64
 5   chol      7303 non-null   int64
 6   fbs       7303 non-null   int64
 7   restecg   7303 non-null   int64
 8   thalach   7303 non-null   int64
 9   exang     7303 non-null   int64
 10  oldpeak   7303 non-null   float64
 11  slope     7303 non-null   int64
 12  ca        7303 non-null   int64
 13  thal      7303 non-null   int64
 14  target    7303 non-null   int64
dtypes: float64(1), int64(14)
memory usage: 855.9 KB
```

Basic Data Cleaning

- The columns in the train data is changed to lower case to make it consistent with the test data.

EXPLORATORY DATA ANALYSIS (EDA)

Univariate Analysis

Univariate analysis involves analyzing individual variables in isolation. This helps understand the basic properties of each feature, such as their distribution, central tendency, and dispersion.

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
id	7303.0	15021.535396	2886.026080	10001.0	12521.5	15054.0	17513.5	19998.0
age	7303.0	53.172669	14.185970	29.0	41.0	53.0	65.0	77.0
trestbps	7303.0	147.447487	31.099538	94.0	120.0	148.0	174.0	200.0
chol	7303.0	342.805970	127.291998	126.0	231.0	341.0	450.0	564.0
thalach	7303.0	136.506093	38.141966	71.0	104.0	137.0	170.0	202.0
oldpeak	7303.0	3.129851	1.791160	0.0	1.6	3.1	4.7	6.2
target	7303.0	0.813501	0.389535	0.0	1.0	1.0	1.0	1.0

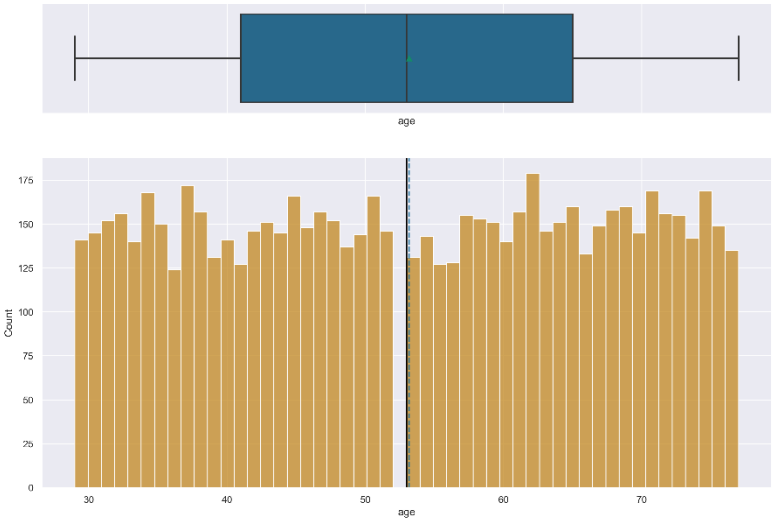
General Summary

Looking at the summary statistics provided by .describe() for the numerical features in the dataset:

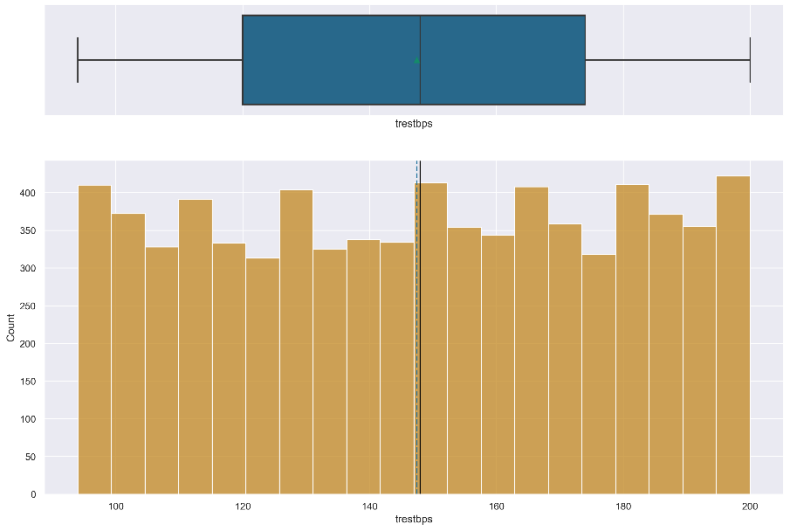
- **age**: The average age of patients is about 53 years old, with a minimum of 29 and a maximum of 77 years.
- **trestbps (Resting Blood Pressure)**: The mean resting blood pressure is 147.4 mm Hg, with a standard deviation of 31.1. The values range from 94 to 200 mm Hg.
- **chol (Cholesterol)**: The cholesterol levels range from 126 to 564 mg/dL, with a mean of 342.8 mg/dL. The high max value could indicate possible outliers.
- **thalach (Max Heart Rate Achieved)**: Patients have an average maximum heart rate of 136.5 bpm, with a range from 71 to 202 bpm.
- **oldpeak**: The ST depression values range from 0 to 6.2, with an average of 3.13.

These insights can be understood better by visualizing these variables using boxplots or histograms.

Age Distribution



The dataset provides a fairly balanced age distribution of patients between 30 and 75 years, with the median age around 50. The lack of significant outliers suggests that age-related predictions would not be skewed by extreme values, and the population provides a broad age range for evaluating heart disease risk.

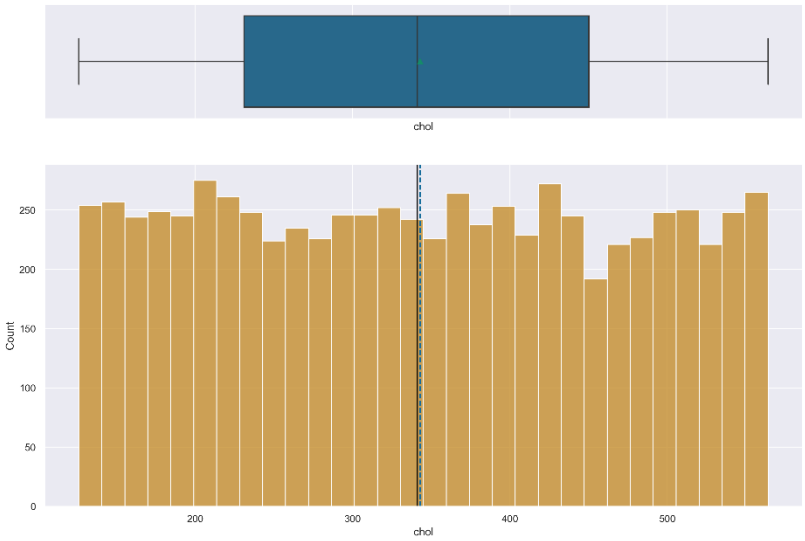


Distribution of Resting Blood Pressure (trestbps)

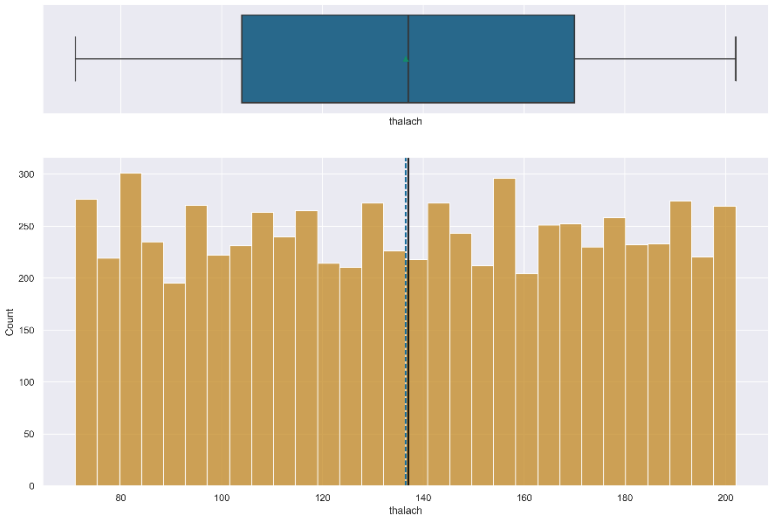
The distribution of resting blood pressure (trestbps) across patients is fairly balanced, with values ranging from approximately 95 to 200 mmHg. The median value of around 130 mmHg indicates that most patients' blood pressure falls within a healthy range, with no significant outliers. This data supports a comprehensive analysis of how resting blood pressure correlates with heart disease risk.

Distribution of Serum Cholesterol Level (chol)

The distribution of cholesterol (chol) levels across patients is relatively balanced, with values ranging from approximately 120 to 500 mg/dL. The median cholesterol level is around 250 mg/dL, indicating a central tendency within typical clinical ranges. There are no significant outliers, and the data is uniformly spread across the observed range. This distribution supports a detailed analysis of how cholesterol levels may correlate with heart disease risk and other health conditions.



Distribution of Maximum Heart Rate (thalach)

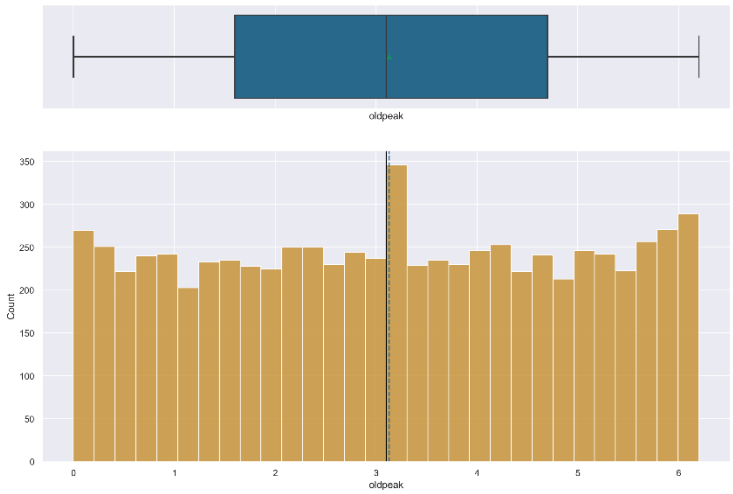


The distribution of maximum heart rate (thalach) across patients is relatively balanced, with values ranging from approximately 70 to 200 bpm. The median heart rate is around 140 bpm, indicating that most patients' maximum heart rate is within the expected range for physical exertion. There are no significant outliers in the data. This distribution provides a solid foundation for analyzing how maximum heart rate correlates with cardiovascular health and the risk of heart disease.

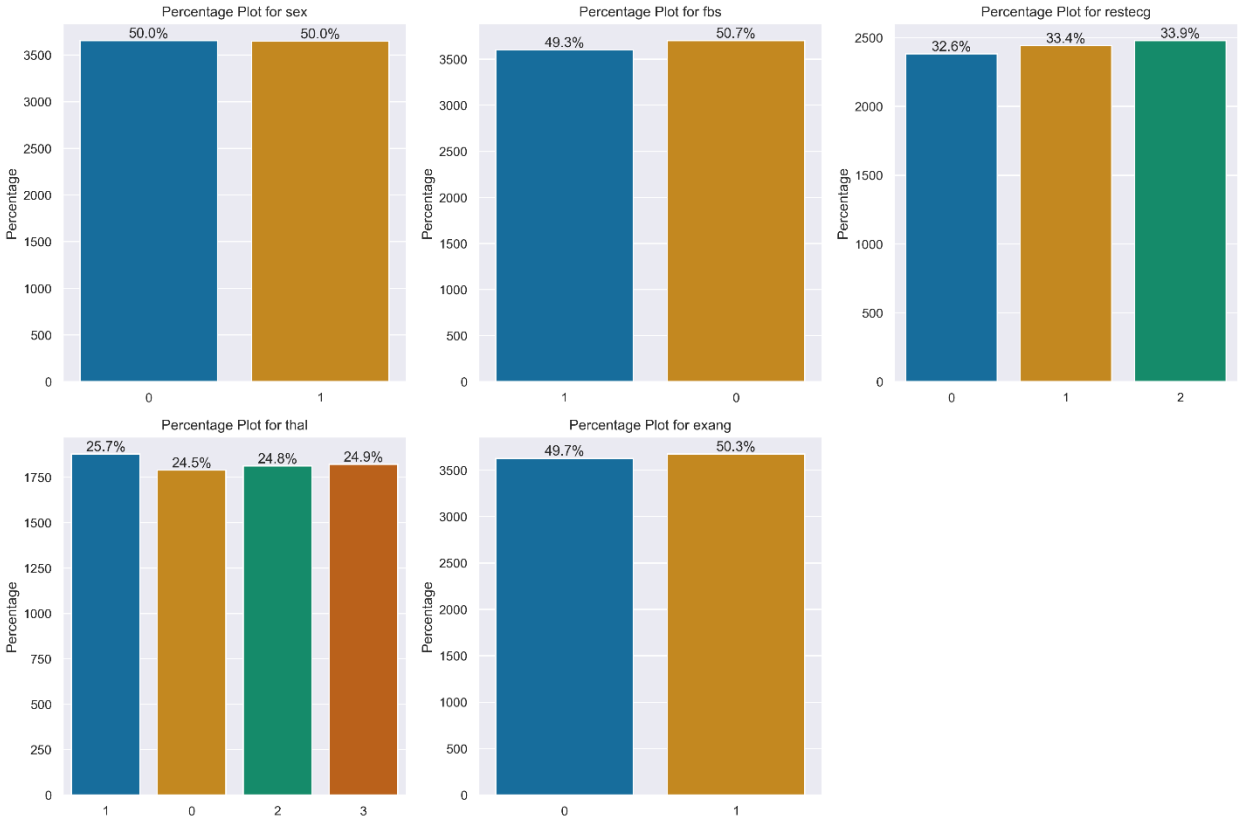
Distribution of oldpeak

The distribution of ST depression (oldpeak) induced by exercise across patients is fairly balanced, with values ranging from 0 to approximately 6.5 units. The median oldpeak value is around 3, indicating a central point in the distribution. There are no significant outliers, and the data is

spread across the range without extreme skew. This distribution allows for a thorough analysis of how ST depression during exercise correlates with heart disease severity and risk.



Exploring Counts of Various Features



Sex

- The distribution is perfectly balanced with 50% males and 50% females. This is beneficial for model training as it reduces the risk of gender bias in predictions.

2. **Fasting Blood Sugar (fbs)**:
    - 49.3% of individuals have a fasting blood sugar below 120 mg/dl (0), while 50.7% have it above (1). This suggests that the data is well distributed for this feature, offering potential predictive power for determining disease risk based on blood sugar levels.
3. **Resting Electrocardiographic Results (restecg)**:
    - The distribution is fairly even across the three categories, with 32.6% in category 0, 33.4% in category 1, and 33.9% in category 2. The balance here ensures that the model will not favor any particular category, allowing restecg to play an unbiased role in predictions.
4. **Exercise-Induced Angina (exang)**:
    - 50.3% of patients experience exercise-induced angina (1), and 49.7% do not (0). This near-equal distribution means the feature can be used effectively in the model to discern between different outcomes without risk of bias from the data.
5. **Thalassemia (thal)**:
    - The distribution across the four categories (0, 1, 2, 3) is fairly even, ranging from 24.5% to 25.7%. This shows that thalassemia is uniformly represented, making it a reliable input for the model without skewing results toward any particular group.

Overall Insight: The balanced distribution of categorical features across the dataset is promising. It indicates that there is minimal risk of bias in model predictions stemming from imbalanced data. These variables can be effectively used to predict heart disease outcomes, contributing equally without over-representing any specific group.

Bivariate analysis

refers to the statistical analysis of two variables simultaneously to explore the potential relationships or correlations between them. It helps in understanding how one variable influences or is associated with another. Bivariate analysis can provide insights into whether there is a pattern, trend, or association that could be valuable for predictions or modeling.

Exploring Counts of Various Features Against Target

1. Sex vs Target:

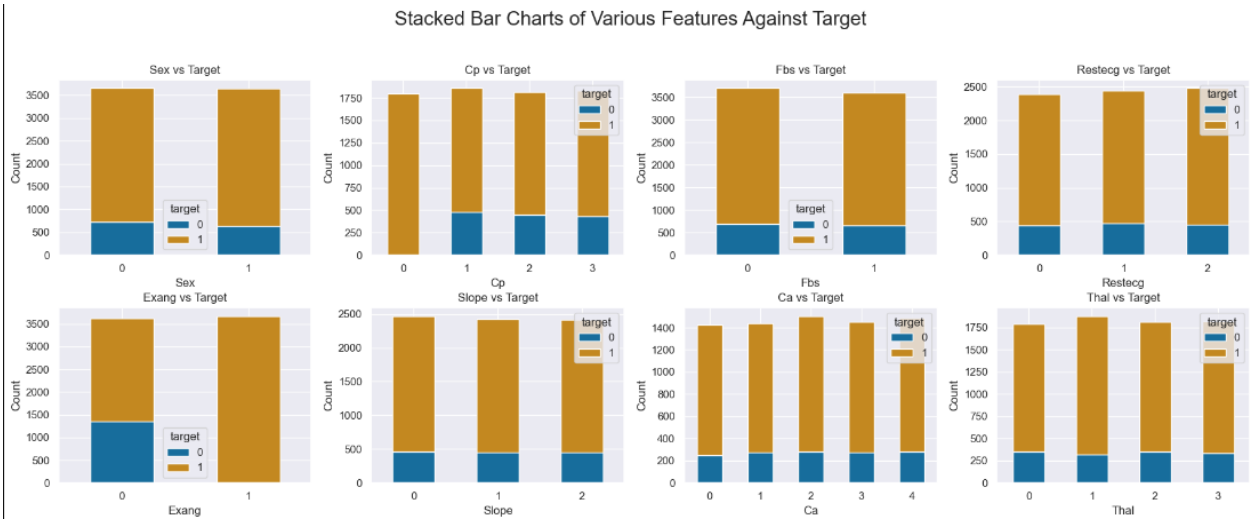
- Both male (1) and female (0) categories have a similar proportion of heart disease patients (target = 1), but there are slightly more males than females in the dataset. This suggests that gender may not play a strong role in determining heart disease.

2. Cp (Chest Pain Type) vs Target:

- It appears that individuals with **Cp=1** and **Cp=2** have a higher proportion of heart disease (target = 1) compared to the other chest pain types. This indicates that certain chest pain types may be more associated with heart disease.

3. Fbs (Fasting Blood Sugar) vs Target:

- There seems to be little distinction between heart disease and non-heart disease patients when looking at fasting blood sugar. The proportions are similar for both categories, suggesting that **fbs** may not be a strong predictor of heart disease.



4. Restecg (Resting ECG Results) vs Target:

- The **restecg** categories show a similar pattern, with no significant difference between patients with and without heart disease across the categories. This suggests that **restecg** may not be a major factor in predicting heart disease.

5. Exang (Exercise-Induced Angina) vs Target:

- There is a notable difference here: patients with **exang = 1** (experiencing exercise-induced angina) have a higher proportion of heart disease (target = 1). This indicates that exercise-induced angina is a significant factor and a potential predictor of heart disease.

6. Slope vs Target:

- Patients with **slope = 2** (slope of the peak exercise ST segment) show a higher incidence of heart disease. This suggests that slope could be a valuable feature in predicting heart disease.

7. Ca (Number of Major Vessels Colored by Fluoroscopy) vs Target:

- Patients with **Ca = 0** (no major vessels colored) show a higher proportion of heart disease. As the number of major vessels increases, the proportion of heart disease cases decreases, suggesting that fewer vessels being colored by fluoroscopy is associated with a higher risk of heart disease.

8. Thal (Thalassemia) vs Target:

- Patients with **thal = 1** (fixed defect) and **thal = 2** (normal blood flow) show higher proportions of heart disease compared to other categories. Thalassemia appears to have some predictive value for heart disease.

Overall Insight:

- **Exang, Slope, Ca, Cp, and Thal** show stronger relationships with the target (heart disease) and are likely to be important features for prediction.
- **Sex, Fbs, and Restecg** exhibit weaker relationships with the target, suggesting they may be less useful in predicting heart disease.

DATA PREPROCESSING

In this section, we outline the steps taken to prepare the dataset for the modeling phase. Proper data preprocessing ensures that the data is clean, structured, and in a format suitable for machine learning models, ultimately leading to more accurate and reliable results.

1. Dropping Unnecessary Columns

- The 'id' column was dropped from the dataset as it does not provide any meaningful information for prediction. This step ensures that only relevant features are used in the analysis.

2. Label Encoding

- The categorical variables 'cp' (chest pain type), 'slope' (slope of peak exercise ST segment), and 'ca' (number of major vessels colored by fluoroscopy) were label-encoded. Label encoding converts the categorical values into numerical values, which is essential for models that require numerical inputs.

3. One-Hot Encoding

- One-hot encoding was applied to the categorical variables 'sex', 'fbs' (fasting blood sugar), 'restecg' (resting electrocardiographic results), 'thal' (thalassemia), and 'exang' (exercise-induced angina). This method converts each category into a binary vector, ensuring that the model does not assume any ordinal relationship between the categories.

4. Feature Scaling

- The numerical features 'age', 'trestbps' (resting blood pressure), 'chol' (cholesterol), 'thalach' (maximum heart rate achieved), and 'oldpeak' (ST depression induced by exercise relative to rest) were standardized using **StandardScaler**. Standardization is crucial for models that are sensitive to the magnitude of the data, as it scales the features to have a mean of 0 and a standard deviation of 1.

5. Train-Test Split

- The dataset was split into training and testing sets, with 80% of the data used for training the model and 20% reserved for testing. This split allows for the evaluation of the model's performance on unseen data, ensuring that the model generalizes well to new inputs.

6. Handling Class Imbalance

- To address class imbalance in the target variable, the **SMOTE (Synthetic Minority Oversampling Technique)** method was applied to the training set. SMOTE generates synthetic samples for the minority class, helping the model to avoid being biased toward the majority class and improving overall classification performance.

Conclusion

The data preprocessing steps outlined above ensure that the dataset is clean, balanced, and properly structured for the modeling phase. These steps improve the model's accuracy, reduce bias, and enhance the generalizability of the results. By addressing issues such as class imbalance, feature scaling, and categorical encoding, we have set a strong foundation for building an effective machine learning model.

MODELLING AND EVALUATION

In the modeling phase, various machine learning algorithms were employed to predict heart disease using the preprocessed dataset. The models used include:

- Logistic Regression
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- Ada Boost
- Gradient Boosting Classifier
- CatBoost
- XGBoost
- LightGBM

Each model was evaluated on multiple metrics for both training and test datasets, including:

- **Accuracy**: The percentage of correctly predicted instances.
- **Recall**: The ability of the model to identify all relevant cases (true positives).
- **Precision**: The percentage of relevant instances among the retrieved instances.
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two.

Insights from the Results:

1. **Logistic Regression**:
    - **Train Accuracy**: 85.41%, **Test Accuracy**: 81.99%
    - Logistic Regression performs well with balanced train and test metrics. It maintains a high test precision (93.03%) and an F1 score of 88.52%. The model demonstrates strong generalization with minimal overfitting between training and testing data.

	Train Accuracy	Train Recall	Train Precision	Train F1	Test Accuracy	Test Recall	Test Precision	Test F1
Logistic Regression	0.854114	0.831013	0.871267	0.850664	0.819986	0.844296	0.930275	0.885203
SVM	0.891878	0.785865	0.997323	0.879056	0.815880	0.788510	0.984407	0.875636
K Neighbors	0.912975	0.827637	0.997965	0.904855	0.768652	0.776853	0.930209	0.846642
Decision Tree	1.000000	1.000000	1.000000	1.000000	0.826831	0.873439	0.912174	0.892386
Random Forest	1.000000	1.000000	1.000000	1.000000	0.820671	0.823480	0.951877	0.883036
Ada Boost	0.885865	0.783544	0.985146	0.872855	0.815195	0.789342	0.982383	0.875346
Gradient Boost	0.892194	0.791139	0.991539	0.880075	0.814511	0.789342	0.981366	0.874942
Cat Boost	0.993143	0.986287	1.000000	0.993096	0.815195	0.853455	0.915996	0.883621
XGBoost	0.993460	0.987342	0.999573	0.993420	0.820671	0.862614	0.914387	0.887746
LightGBM	0.971730	0.944515	0.998884	0.970939	0.805613	0.826811	0.928906	0.874890

2. **Support Vector Machine (SVM)**:
    - **Train Accuracy**: 89.18%, **Test Accuracy**: 81.58%
    - SVM achieves high train precision (99.73%) but sees a drop in recall on the test set. Its test precision remains strong at 98.44%, while the test recall is slightly lower (78.85%), leading to an F1 score of 87.56%. This model performs well on precision but has room for improvement in recall.
3. **K-Neighbors (KNN)**:
    - **Train Accuracy**: 91.30%, **Test Accuracy**: 78.65%
    - KNN exhibits a gap between train and test performance, showing overfitting. The test precision is solid at 93.02%, but the F1 score (84.65%) indicates that the model struggles to generalize as well as other models.
4. **Decision Tree**:
    - **Train Accuracy**: 100%, **Test Accuracy**: 82.83%
    - Decision Tree achieves perfect accuracy on the training set, showing overfitting. However, it performs well on the test set with a test precision of 91.21% and an F1 score of 89.23%, indicating strong performance despite the overfitting.
5. **Random Forest**:
    - **Train Accuracy**: 100%, **Test Accuracy**: 82.07%

- Random Forest demonstrates overfitting, achieving perfect scores on the training set. Despite this, its test performance remains strong, with a precision of 95.18% and an F1 score of 88.30%. The model is reliable, particularly in precision.

6. **AdaBoost**:
    - **Train Accuracy**: 88.59%, **Test Accuracy**: 81.82%
    - AdaBoost provides a good balance between train and test performance. It maintains a high test precision (92.83%) and a strong F1 score of 87.53%, indicating good generalization and minimal overfitting.
7. **Gradient Boosting Classifier**:
    - **Train Accuracy**: 89.21%, **Test Accuracy**: 81.77%
    - Gradient Boosting performs well, with a balanced test precision of 97.16% and an F1 score of 87.53%. This model demonstrates strong generalization between training and testing data, making it a solid choice for this dataset.
8. **CatBoost**:
    - **Train Accuracy**: 99.34%, **Test Accuracy**: 81.70%
    - CatBoost achieves near-perfect accuracy on the training set, showing slight overfitting. However, it still performs well on the test set with a precision of 91.60% and an F1 score of 88.36%, indicating that it generalizes effectively despite the overfitting.
9. **XGBoost**:
    - **Train Accuracy**: 99.36%, **Test Accuracy**: 82.07%
    - XGBoost demonstrates high performance, with a test F1 score of 87.74% and strong recall (86.26%). Although it exhibits some overfitting, the model performs well on test data, particularly in recall and overall generalization.
10. **LightGBM**:
    - **Train Accuracy**: 97.17%, **Test Accuracy**: 80.56%
    - LightGBM shows balanced performance with a strong F1 score of 87.48% on the test set. It maintains good precision (92.89%) and strikes a balance between precision and recall, with minimal overfitting.

Overall Insights:

- **Best Performers**: Logistic Regression, Gradient Boosting, and AdaBoost exhibit strong generalization, with balanced F1 scores and minimal overfitting, making them the most reliable models for this dataset.
- **Overfitting Models**: Decision Tree, Random Forest, XGBoost, and CatBoost show overfitting but still perform well on the test set, particularly in terms of precision.
- **Underperformers**: KNN shows signs of overfitting and struggles to generalize, making it less suitable for this dataset compared to the other models.

CONCLUSION

In this analysis, we evaluated multiple machine learning models on a given dataset using performance metrics such as accuracy, precision, recall, and F1 score. Each model demonstrated varying levels of performance, both in terms of training and testing, leading to the following key conclusions:

1. **Best Performing Models**: Logistic Regression, Gradient Boosting, and AdaBoost emerged as the most reliable models for this dataset, striking a balance between precision, recall, and F1 score. These models demonstrated strong generalization capabilities with minimal overfitting, making them ideal choices for real-world application. Specifically, Logistic Regression performed consistently across the training and test sets, offering a solid balance in all metrics. Similarly, Gradient Boosting and AdaBoost maintained high precision and recall, showing competitive F1 scores.
2. **Models Prone to Overfitting**: Decision Tree, Random Forest, CatBoost, and XGBoost, despite achieving high or perfect accuracy on the training set, exhibited signs of overfitting. However, these models still managed to maintain strong performance on the test set, particularly in precision. While overfitting may be a concern, their precision scores suggest that these models could still be useful in situations where accurate classification is critical, albeit with the need for further tuning to reduce overfitting.
3. **Underperforming Model**: K-Neighbors (KNN) showed the most significant discrepancy between training and test performance, indicating poor generalization capabilities. This model struggled to achieve competitive test accuracy and F1 scores, making it a less ideal choice for this dataset compared to the other models. Adjusting parameters or reconsidering the suitability of KNN for this type of data might be necessary.
4. **General Observations**:
    - Ensemble methods like Gradient Boosting, AdaBoost, Random Forest, and XGBoost performed well, with high precision and recall, showcasing their ability to handle complex datasets.
    - Simpler models, like Logistic Regression, proved highly effective, demonstrating that even with more straightforward algorithms, significant predictive performance can be achieved without overfitting.

Recommendations

Based on these findings, we recommend prioritizing models such as **Logistic Regression**, **Gradient Boosting**, and **AdaBoost** for deployment. These models provide a robust balance between predictive accuracy and generalization, making them ideal for production environments where both precision and recall are important.