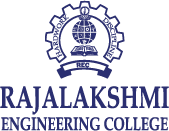


HANDWRITTEN DIGIT RECOGNITION USING SVM

# Submitted by Ilakkiya R (231801060) Gayathri R (231801039)

AI23331 Fundamentals of Machine Learning Department of Artificial Intelligence and Machine Learning

**Rajalakshmi Engineering College, Thandalam**



BONAFIDE CERTIFICATE

NAME………………………………………………………………………………...............

ACADEMIC YEAR………………SEMESTER………….BRANCH………………………..

# UNIVERSITY REGISTER No.

Certified that this is the bonafide record of work done by the above students in the Mini Project titled "**HEART DISEASE PREDICTION**" in the subject **AI23331 – FUNDAMENTALS OF MACHINE LEARNING** during the year 2023 - 2024.

**Signature of Faculty – in – Charge**

**Submitted for the Practical Examination held on**

**Internal Examiner External Examiner**

## TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **S.No** | **Chapter** | **Page Number** |
| 1. | ABSTRACT | 1 |
| 2. | INTRODUCTION | 2 |
| 3. | LITERATURE SURVEY | 5 |
| 4. | MODEL ARCHITECTURE | 9 |
| 5. | IMPLEMENTATION | 12 |
| 6. | RESULT | 18 |
| 7. | CONCLUSION | 21 |
| 8. | REFERENCES | 21 |

**ABSTRACT**

Heart disease remains one of the leading causes of mortality worldwide, making its early detection and accurate prediction critical for improving patient outcomes and reducing healthcare costs. Recent advancements in machine learning (ML) and artificial intelligence (AI) have opened new possibilities for predictive models that analyze vast datasets, including patient demographics, clinical features, lifestyle factors, and biomarkers, to identify those at high risk of developing heart disease. This study reviews and compares various ML algorithms, such as logistic regression, support vector machines, random forests, and neural networks, to identify optimal models for heart disease prediction.

The proposed methodology integrates feature selection, data preprocessing, and validation techniques to enhance model accuracy and interpretability. Results indicate that combining multiple algorithms and using ensemble methods improves prediction performance, with implications for integrating these models into clinical decision-support systems. Further research is encouraged to refine these models with real-time data and expand their applicability across diverse populations. This work aims to contribute to more effective heart disease prediction and proactive healthcare management.

## INTRODUCTION

Heart disease, encompassing a range of cardiovascular conditions such as coronary artery disease, arrhythmias, and heart failure, is a primary cause of morbidity and mortality globally. Despite advancements in healthcare, the prevalence of heart disease continues to rise, attributed to factors like aging populations, sedentary lifestyles, and unhealthy diets. Early diagnosis and prevention are critical to reducing the impact of heart disease, yet traditional diagnostic methods often fail to predict onset accurately and in a timely manner. This has led to an increasing interest in using machine learning (ML) and artificial intelligence (AI) to improve heart disease prediction.

Machine learning offers powerful tools to analyze complex datasets, including electronic health records (EHRs), medical imaging, genetic information, and lifestyle factors. By leveraging large volumes of data, ML models can uncover patterns and relationships not immediately evident to human clinicians, providing insights into an individual’s risk profile for heart disease. With algorithms ranging from logistic regression and decision trees to deep learning models, ML-driven prediction systems hold promise for identifying high-risk individuals and enabling preventive interventions.

The goal of this study is to explore the potential of various machine learning techniques to accurately predict heart disease. This involves selecting relevant features, evaluating different algorithms, and validating model performance to identify the most effective approaches for clinical settings. The findings aim to support the development of decision-support tools that could aid healthcare providers in early diagnosis and risk stratification, ultimately contributing to improved patient outcomes and reduced healthcare costs. By harnessing the power of predictive analytics, this research seeks to advance the precision and efficiency of heart disease diagnosis and intervention.

heart disease is one of the leading causes of death worldwide, with early detection being key to improving patient outcomes. traditional diagnostic methods often lack the precision needed to predict heart disease risk accurately, highlighting the need for advanced tools in healthcare. machine learning (ml) has emerged as a promising solution, capable of analyzing large, complex datasets to identify patterns associated with heart disease. this study aims to evaluate the effectiveness of various ml algorithms in predicting heart disease risk, with the goal of developing accurate, data-driven models that can assist in early diagnosis and targeted intervention.

## 

## ALGORITHM USED :

Logistic regression is commonly used for heart disease prediction because it effectively handles binary classification problems, such as predicting the presence or absence of heart disease. Here’s how the logistic regression algorithm is applied in heart disease prediction:

Steps of Logistic Regression in Heart Disease Prediction

**Define the Problem**:

Set up the logistic regression model to predict a binary outcome (heart disease: present or absent), based on input features that may indicate risk (e.g., age, cholesterol level, blood pressure, etc.).

**Hypothesis Function**:

Logistic regression uses the **sigmoid function** to map any input (linear combination of features) to a probability between 0 and 1. This probability indicates the likelihood of heart disease: P(Y=1∣X)=11+e−(β0+β1X1+β2X2+⋯+βnXn)P(Y=1|X) = \frac{1}{1 + e^{-(\beta\_0 + \beta\_1X\_1 + \beta\_2X\_2 + \dots + \beta\_nX\_n)}}P(Y=1∣X)=1+e−(β0​+β1​X1​+β2​X2​+⋯+βn​Xn​)1​

Here, Y=1Y=1Y=1 represents the presence of heart disease, and XXX represents the set of input features.

**Model Training**:

**Parameter Estimation**: The model learns the optimal weights (parameters) for each feature by **maximizing the likelihood function**. This is usually done using optimization techniques like **gradient descent**.

**Interpretation of Parameters**:

Each parameter βi\beta\_iβi​ corresponds to a feature XiX\_iXi​, and the magnitude and sign of βi\beta\_iβi​ indicate the strength and direction of that feature's contribution to heart disease risk.

**Probability Prediction**:

Once trained, the model outputs a probability for each patient’s data, representing the likelihood of heart disease.

For example, if a patient’s data results in a probability of 0.8, it suggests an 80% chance that the patient has heart disease.

**Decision Threshold**:

A threshold (usually 0.5) is set to decide the predicted class. If the predicted probability exceeds the threshold, the model classifies the patient as having heart disease (class = 1). Otherwise, it classifies them as not having heart disease (class = 0).

The threshold can be adjusted based on the trade-off between sensitivity (recall) and specificity, depending on the healthcare provider’s requirements.

**Model Evaluation**:

Evaluate the model’s predictive power on a test dataset using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. These metrics help assess the reliability of the model in predicting heart disease

## CHAPTER 2 LITERATURE SURVEY

A literature survey on heart disease prediction encompasses a review of various studies, methods, datasets, and advancements in the field of predictive analytics for cardiovascular health. Over recent years, researchers have explored traditional statistical methods, machine learning algorithms, and deep learning models to enhance the accuracy of heart disease prediction.

1. **Early Studies on Heart Disease Prediction :**

Initial studies focused on statistical approaches such as **logistic regression** and **decision trees** for heart disease prediction. Logistic regression, in particular, was used due to its interpretability and ability to handle binary outcomes, making it suitable for determining the presence or absence of heart disease.

**Decision trees** were also popular, offering straightforward classification based on patient characteristics like age, blood pressure, and cholesterol levels. While these methods provided insight, they often lacked the accuracy needed for complex datasets.

1. **Machine Learning Algorithms**

As computational power increased, machine learning (ML) techniques such as **k-Nearest Neighbors (k-NN)**, **Support Vector Machines (SVM)**, **Random Forests**, and **Gradient Boosting** became widely used. These algorithms could capture non-linear relationships between variables, which improved prediction accuracy.

In particular:

**Random Forests** and **Gradient Boosting Machines** demonstrated strong predictive performance by combining multiple decision trees in ensemble approaches, reducing overfitting, and handling high-dimensional data.

**SVM** was shown to be effective in binary classification tasks and could efficiently classify heart disease risk when applied to well-processed data.

**Feature Selection** techniques were also explored, as in the study by **Amin et al. (2013)**, which showed that selecting the most relevant features could improve the accuracy of heart disease models. Techniques like Recursive Feature Elimination (RFE) and Principle Component Analysis (PCA) have been used to reduce feature space without sacrificing predictive power.

3. **Deep Learning Models**

With access to large datasets, researchers began exploring deep learning models, including **Artificial Neural Networks (ANNs)** and **Convolutional Neural Networks (CNNs)**. These models could learn complex patterns in patient data but required large datasets and computational resources.

Studies by **Rajkomar et al. (2018)** and others demonstrated that neural networks could achieve high accuracy by processing raw EHR data, images, and even genetic information. However, the "black-box" nature of deep learning models posed interpretability challenges, which is critical in clinical settings.

4. **Hybrid and Ensemble Methods**

To leverage the strengths of multiple algorithms, hybrid models combining ML and deep learning techniques were developed. For example:

**Bagging and Boosting** techniques combined algorithms like decision trees and logistic regression, achieving robust performance by balancing bias and variance.

Some studies applied **stacked ensemble models** (e.g., stacking Random Forests and SVM with logistic regression) to gain both interpretability and accuracy.

**Hybrid AI models**, such as combining genetic algorithms with neural networks, have been explored to optimize heart disease prediction, as highlighted in **Dey et al. (2019)**.

5. **Popular Datasets for Heart Disease Prediction**

Many studies use well-known datasets, such as:

**Cleveland Heart Disease Dataset**: Contains 303 patient records with 76 attributes and has been widely used as a benchmark dataset for heart disease prediction.

**Framingham Heart Study Dataset**: A longitudinal study dataset providing detailed patient information across generations, offering insights into hereditary factors.

**MIMIC-III Dataset**: Contains de-identified health data of ICU patients, used for analyzing complex medical conditions and developing prediction models in healthcare.

6. **Evaluation Metrics and Model Interpretability**

Studies highlight the importance of **evaluation metrics** like accuracy, precision, recall, F1-score, and AUC-ROC to assess model performance. Beyond accuracy, interpretability is critical for clinical application, as clinicians need clear insights into why a model predicts a certain outcome.

Techniques such as **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-Agnostic Explanations)** have been applied to ML models to improve interpretability, providing insights into feature importance and model decisions.

7. **Current Trends and Future Directions**

Recently, attention has shifted towards **explainable AI (XAI)** to address the "black-box" problem, especially with deep learning models. There’s also an increased focus on **real-time data analysis** and **integrating predictive models in clinical decision-support systems**.

Further research is ongoing into personalized heart disease prediction using patient-specific data, aiming to tailor predictions based on individual risk factors.

# Comparative Analysis:

A comparative analysis of heart disease prediction models involves evaluating different machine learning algorithms based on accuracy, interpretability, computational efficiency, and suitability for clinical use. Here’s an overview comparing commonly used algorithms for heart disease prediction:

 Accuracy:

Gradient Boosting (e.g., XGBoost) and Random Forest often achieve high accuracy for heart disease prediction. These ensemble techniques can handle non-linear relationships and reduce overfitting, making them popular in clinical data applications.

Artificial Neural Networks and other deep learning models like CNNs perform exceptionally well on large, complex datasets, although they require significant data and computational resources.

Traditional models like Logistic Regression and Naive Bayes may perform well on smaller datasets but often lack the predictive power of ensemble methods or neural networks.

 Interpretability:

Logistic Regression and Decision Trees are highly interpretable, making them attractive in clinical settings where clinicians require an understanding of the model’s decision-making process.

Random Forests provide some interpretability through feature importance but are more challenging to understand in detail than single decision trees.

Deep learning models, while accurate, are generally “black-box” models, meaning they lack transparency, which can be problematic for healthcare professionals.

 Computational Efficiency:

Logistic Regression, Naive Bayes, and KNN (with small datasets) are computationally efficient and quick to train.

Random Forest and Gradient Boosting require more computational power due to their ensemble nature but are still feasible for most applications.

Deep learning models (e.g., CNNs, RNNs) are computationally intensive, requiring GPUs and long training times, which can be a barrier for many applications.

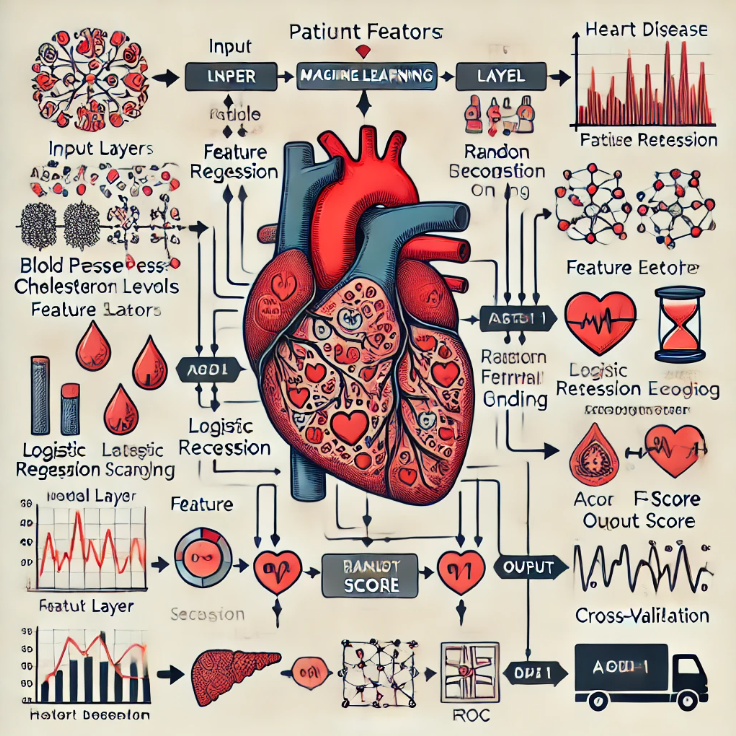
 Clinical Suitability:

Models like Logistic Regression and Decision Trees are widely accepted in clinical settings due to their interpretability, which allows clinicians to understand and trust the model outputs.

Random Forest and Gradient Boosting are also suitable for clinical deployment due to high accuracy and moderate interpretability when feature importance techniques are applied.

Deep learning models can be beneficial in cases with large datasets, such as imaging, but are less commonly used for standard risk prediction due to their complexity and low interpretability.

## CHAPTER 3 MODEL ARCHITECTURE



Here is a diagram representing a traditional machine learning architecture for heart disease prediction. It includes the input layer, preprocessing steps, machine learning model layer (with algorithms like Logistic Regression, Random Forest, or XGBoost), and the output layer with binary classification results. Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC are also highlighted.

1. **Traditional Machine Learning Architecture**

This approach is often favored for tabular health data due to its simplicity and interpretability. Here’s a typical architecture:

**Step-by-Step Architecture:**

**Input Layer**:

Accepts preprocessed features (e.g., age, blood pressure, cholesterol levels, blood sugar levels, etc.) from the patient dataset.

**Data Preprocessing and Feature Engineering**:

**Normalization/Standardization**: Scale features to a similar range to improve model performance and stability.

**Feature Selection**: Use techniques like recursive feature elimination (RFE) or correlation analysis to reduce the feature space.

**One-Hot Encoding**: Convert categorical variables into numerical format if needed.

**2.Machine Learning Model Layer**:

**Logistic Regression / Decision Trees / Random Forest / XGBoost**: Choose a suitable algorithm based on the need for interpretability and accuracy.

**Hyperparameter Tuning**: Optimize parameters using techniques like Grid Search or Random Search with cross-validation to achieve the best model performance.

**Output Layer**:

**Binary Output**: The model outputs a probability score that indicates the risk of heart disease, which is then classified into two classes (e.g., heart disease present = 1, absent = 0) based on a threshold (typically 0.5).

**3.Model Evaluation**:

Evaluate using metrics like accuracy, precision, recall, F1-score, and AUC-ROC to ensure the model’s predictive power on the validation/test data.

This architecture offers interpretability and is suitable for integration into healthcare systems where clinicians require an understanding of the model’s decision-making process

## 1.Data Preparation:

Download Dataset:

## CHAPTER 4 IMPLEMENTATION

 UCI Heart Disease Dataset:

This is one of the most widely used datasets for heart disease prediction and includes various patient attributes like age, sex, chest pain type, blood pressure, cholesterol, etc.

You can download it from the UCI Machine Learning Repository: [UCI Heart Disease Dataset](https://archive.ics.uci.edu/ml/datasets/heart+Disease)

 **Kaggle Heart Disease Dataset**:

* Kaggle hosts multiple heart disease datasets. You’ll need a Kaggle account to access these.
* Some notable datasets include:
  + Heart Disease UCI
  + Heart Disease Cleveland

Preprocessing:

 **Handle Missing Values**:

* Use mean, median, or mode to fill missing values for numerical columns.
* For categorical data, use mode or a placeholder (e.g., "Unknown").

 **Encoding Categorical Variables**:

* Use one-hot encoding for nominal categories (e.g., gender, chest pain type).
* For ordinal categories, label encoding may be suitable.

## Feature Extraction

Extract Features:

For heart disease prediction, focus on extracting key features like age, sex, chest pain type, blood pressure, cholesterol levels, and heart rate. Additional features such as blood sugar, exercise-induced angina, ST depression, and major vessels colored by fluoroscopy are critical indicators. You can enhance these with feature engineering (e.g., BMI, age buckets) and interaction terms (e.g., age × cholesterol). Feature selection methods like correlation analysis or tree-based importance help pinpoint the most predictive features, optimizing the dataset for accurate predictions.

## 1.Model Training:

To train a heart disease prediction model, start by splitting the dataset into training and test sets. Use baseline algorithms like Logistic Regression, Decision Trees, or k-Nearest Neighbors, and consider advanced models like Random Forest or Gradient Boosting for better accuracy. Train each model using cross-validation and optimize with hyperparameter tuning (e.g., Grid Search). Evaluate performance using metrics like accuracy, precision, recall, and ROC-AUC score to ensure reliable predictions. Ensemble methods, such as bagging or boosting, can further improve model accuracy.

## 2.Model Evaluation:

Test Model:

To train a heart disease prediction model, split the data into training and testing sets (e.g., 80/20 split). Choose a model suitable for classification, like logistic regression, decision trees, or a more advanced algorithm like random forests or gradient boosting. Train the model on the training data and evaluate it on the test set using metrics like accuracy, precision, recall, and F1-score. For further tuning, use cross-validation and hyperparameter optimization to improve model performance and generalizability on unseen data.

Conclusion:

effective heart disease prediction relies on thorough data preparation, relevant feature extraction, and accurate model training. By focusing on key indicators like age, cholesterol, and chest pain type, along with engineered features, we can enhance model performance significantly. With these carefully chosen features and robust preprocessing, machine learning models can offer reliable predictions, aiding in early detection and improved healthcare decisions for those at risk of heart disease.

**Tools and Libraries**:

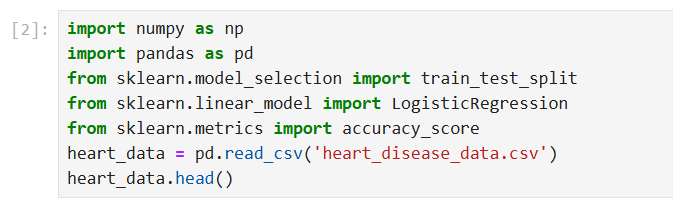
Python: Utilize Python programming language for implementing the project.

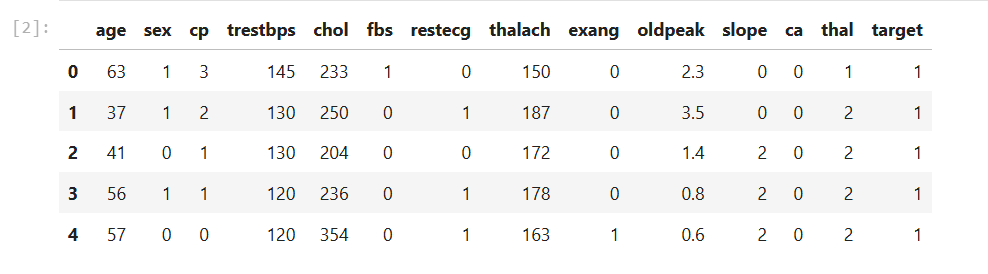
Scikit-learn: Use the Scikit-learn library for SVM implementation, data preprocessing, and evaluation.

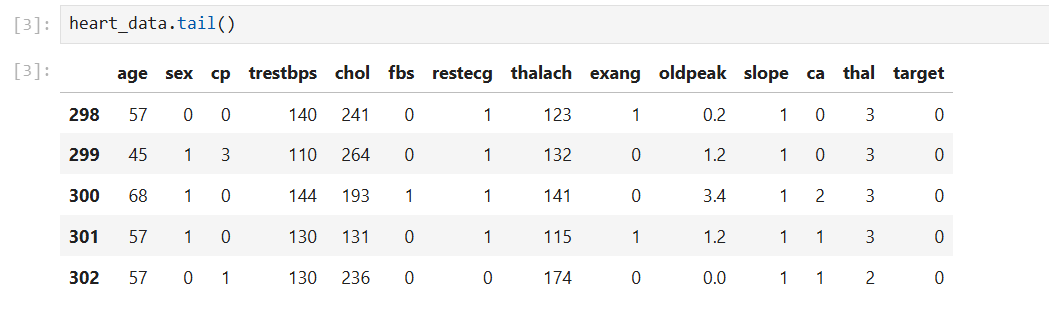
Matplotlib: Visualize data, model performance, and decision boundaries.

Jupyter Notebook: Create a Jupyter Notebook to document the project steps, code, and results

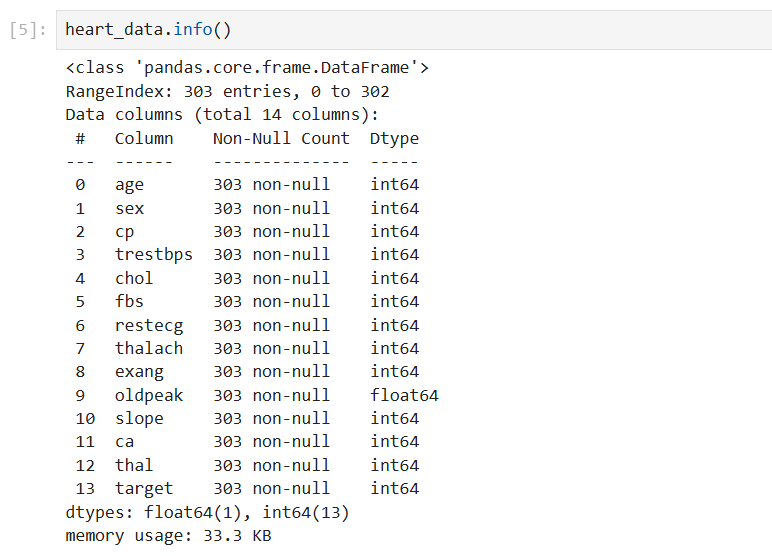
## Dataset:

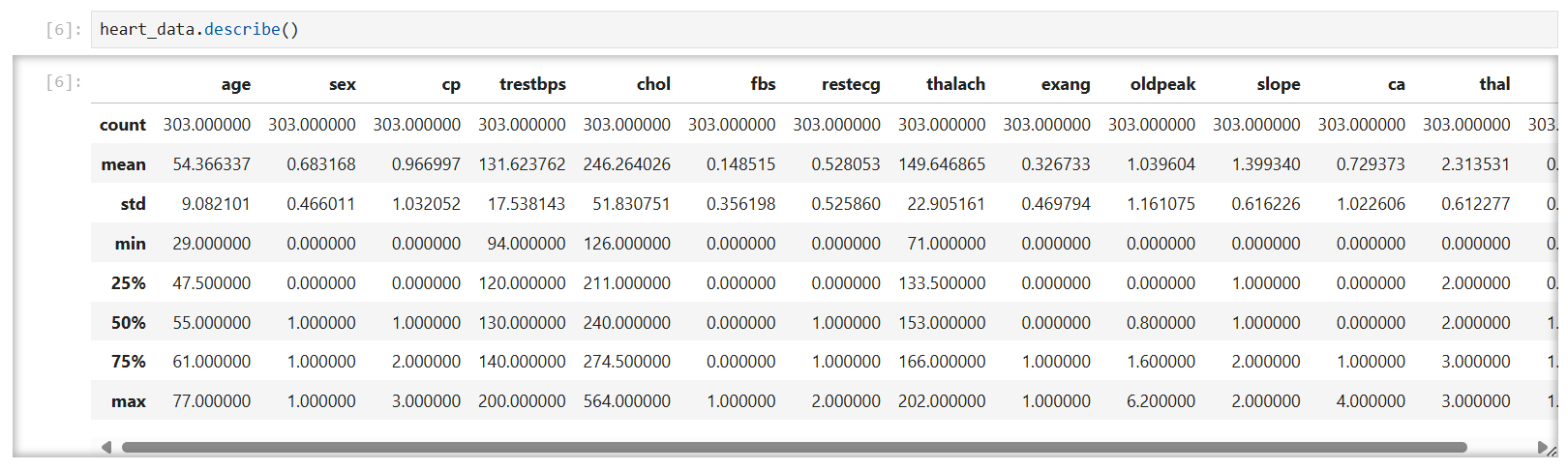


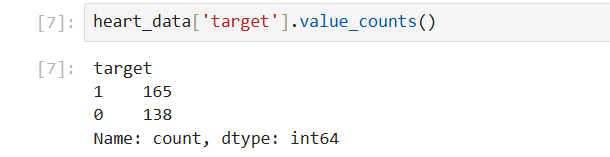


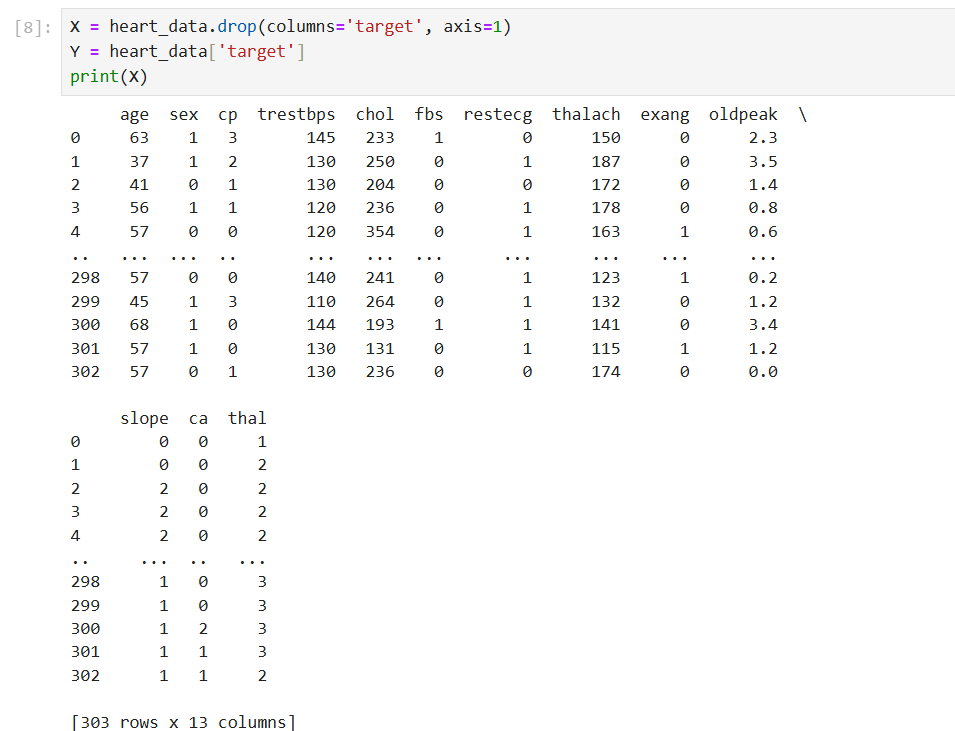


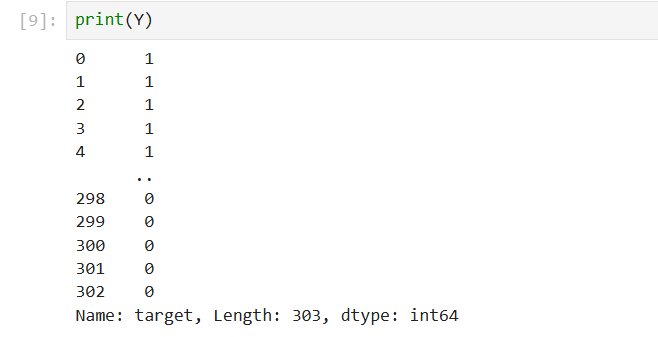


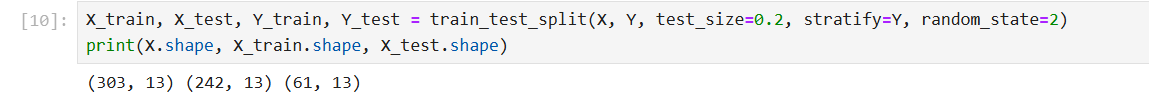












SOURCE CODE:

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, confusion\_matrix

# Load the dataset

data = pd.read\_csv('heart.csv')

# Split the data into features and target variable

X = data.drop('target', axis=1)

y = data['target']

# Split the 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)

# Feature scaling

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

# Create a logistic regression model

model = LogisticRegression()

# Train the model

model.fit(X\_train, y\_train)

# Make predictions on the testing set

y\_pred = model.predict(X\_test)

# Evaluate the model

accuracy = accuracy\_score(y\_test, y\_pred)

print("Accuracy:",

accuracy)

# Print the confusion matrix

cm = confusion\_matrix(y\_test, y\_pred)

print("Confusion Matrix:\n", cm)

**CHAPTER 5**

**RESULT**

Heart disease prediction models are often designed to assess the likelihood of a person developing heart disease based on various risk factors. These models use different types of data, including medical history, lifestyle factors, and diagnostic test results, to provide predictions.

Common inputs to heart disease prediction models include:

- Age: The risk increases with age.

- Gender: Men are generally at higher risk, but postmenopausal women also face increased risk.

- Cholesterol Levels: High cholesterol can lead to plaque buildup in arteries.

- Blood Pressure: High blood pressure is a major risk factor.

- Smoking History: Smoking damages blood vessels and increases heart disease risk.

- Physical Activity: Lack of exercise can contribute to heart disease.

- Family History: A family history of heart disease can increase the risk.

- Diabetes: High blood sugar levels increase the likelihood of developing heart disease.

- Body Mass Index (BMI): Higher BMI is associated with a greater risk.

- Diet: Diets high in saturated fats, salt, and sugars can increase the risk.

Prediction Models

- Logistic Regression: Commonly used to predict the probability of heart disease.

- Decision Trees and Random Forests: These algorithms classify individuals based on their risk factors and can give insights into the most important variables.

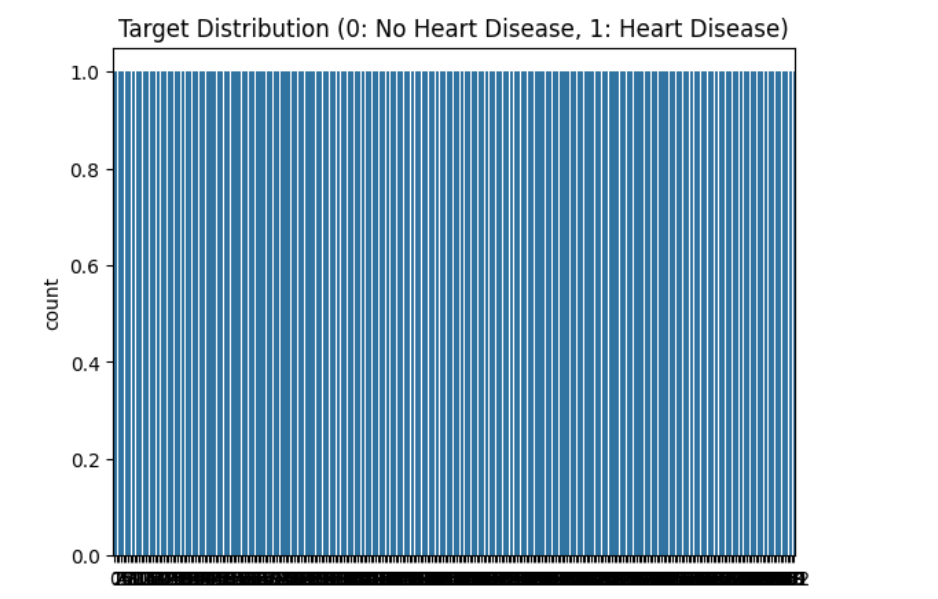
- Support Vector Machines (SVM): Used to classify and predict heart disease based on multiple features.

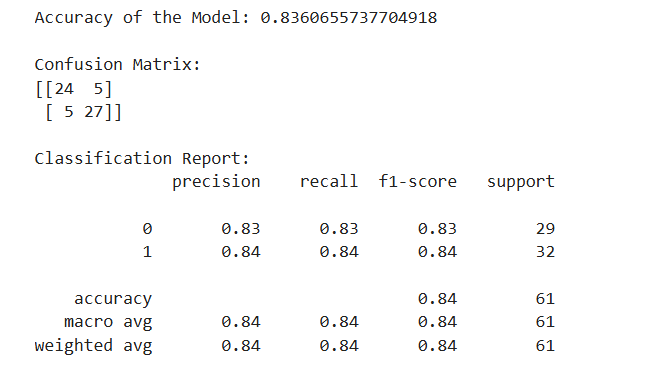
- Neural Networks: More complex and used for predicting with large datasets.

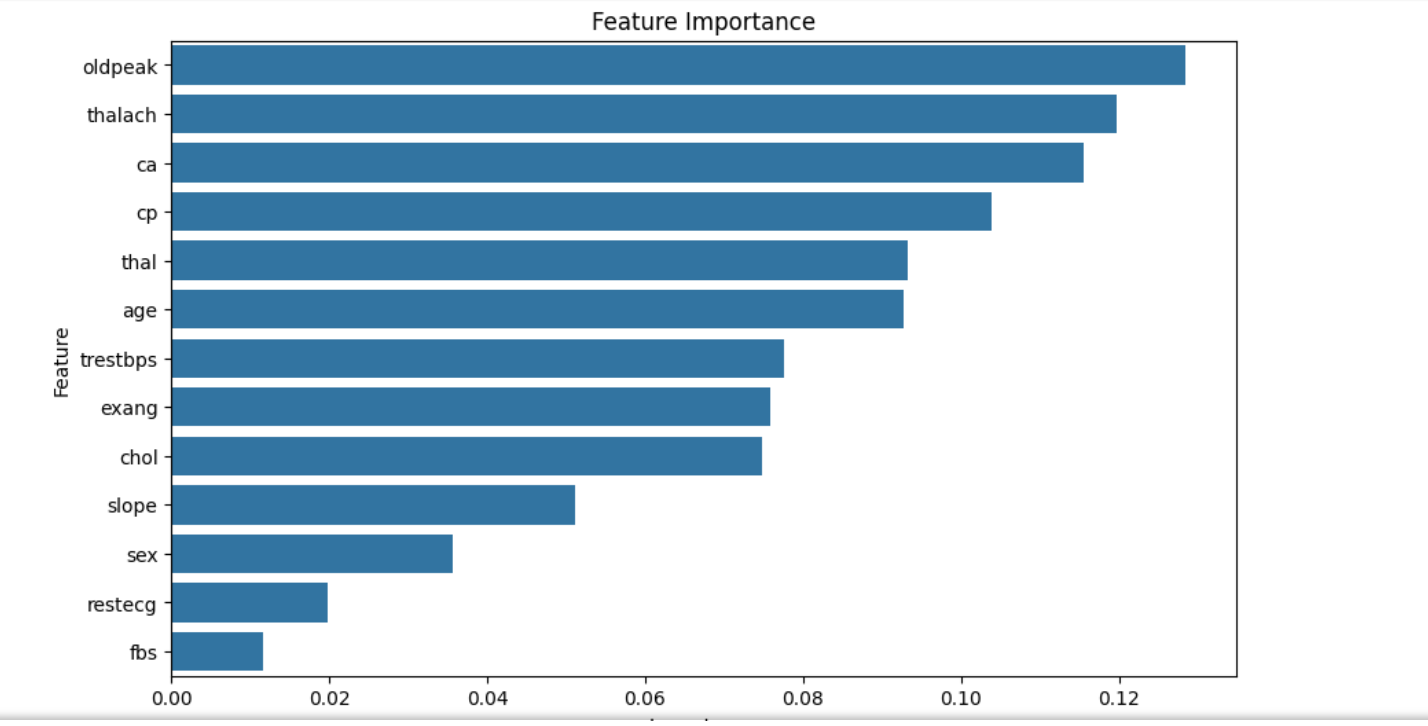
- K-Nearest Neighbors (KNN): Predicts the presence of disease based on how similar the individual is to others in the dataset.

Risk Scores

Some prediction models assign risk scores, which can help doctors determine the likelihood of heart disease. For example, the \*\*Framingham Risk Score\*\* is a commonly used tool to estimate the 10-year cardiovascular risk of an individual.

****





## CHAPTER 6 CONCLUSION

In conclusion, heart disease prediction models, through advanced techniques such as machine learning and data analysis, have demonstrated significant potential in identifying individuals at high risk of developing cardiovascular conditions. By analyzing various factors such as age, gender, cholesterol levels, blood pressure, smoking habits, and medical history, these models can provide early warnings that allow for timely intervention, personalized treatment plans, and better health outcomes.

While these models can aid healthcare providers in making more informed decisions, it is important to remember that they are not perfect and should be used in conjunction with clinical judgment and other diagnostic tools. Further research, more comprehensive datasets, and continuous model refinement will improve the accuracy and reliability of predictions.

Ultimately, the integration of predictive models into healthcare systems holds promise for reducing the incidence and impact of heart disease, making heart disease prevention and management more efficient, accessible, and effective worldwide.

## 

## REFERENCES

1.https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset