Artificial Intelligence
and
Machine Learning

Project Report

Semester-IV (Batch-2022)

Heart Disease Prediction

Supervised By:

Shagun Sharma

Submitted By:

Jinny Kapur	(2210990462)
Jiya Gaba	(2210990464)
Joyash Sood	(2210990466)
Kartavya Tomar	(2210990484)

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab

INDEX:

Abstract:

Machine learning is a field of artificial intelligence that involves training models to learn from data and make predictions or decisions without being explicitly programmed. There are different types of machine learning, including supervised learning, unsupervised learning, and reinforcement learning.

ML will be essential in the healthcare sectors which will be useful for doctors to fasten the diagnosis. In this code, we are implementing supervised learning for a heart disease prediction model. This involves training the model on labeled data, where we have 14 attributes including age, sex, cholesterol levels, and other health metrics are used to predict the presence or absence of heart disease (the target variable). The goal is for the model to learn the underlying patterns and relationships between these features and the presence of heart disease, so that it can accurately classify new, unseen instances and can make accurate predictions on new, unseen data.

In this case, the target variable is whether a patient has heart disease or not, which is represented as a binary classification problem: whether a patient has heart disease (1) or does not have heart disease (0). The features are age, sex, chest pain, trestbps (resting blood pressure) , cholesterol, fasting blood sugar , slope, thalassemia, calcium levels, resting electrocardiographic measurement (restecg,) thalach, exercise induced angina(exang), old peak and target.

The code uses three different models for binary classification: Logistic Regression, Decision Tree and Support Vector Machine.

1. Logistic Regression:

- **Description:** It is a linear model for binary classification that predicts the probability that a given instance belongs to a particular class. It uses the logistic function to map predictions to probabilities.
- **Advantages:** Simple to implement and interpret, efficient for small datasets with limited features.
- **Disadvantages:** Can't capture complex relationships in data, prone to underfitting.

2. Decision Tree:

- **Description:** Decision Tree are non-linear models that recursively split the data based on features to create a tree-like structure. Each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label.

- **Advantages:** Easy to interpret and visualize, can handle both numerical and categorical data.
- **Disadvantages:** Prone to overfitting, sensitive to small variations in the data.

3. Support Vector Machines:

- **Description:** SVM is a powerful supervised learning algorithm that can be used for both regression and classification tasks. It finds the hyperplane that best separates the classes in the feature space, maximizing the margin between the classes.
- **Advantages:** Effective in high-dimensional spaces, memory-efficient, effective in cases where the number of dimensions is greater than the number of samples.
- **Disadvantages:** Not suitable for large datasets, computationally expensive, sensitive to the choice of the kernel parameters.

1. Introduction:

Heart disease is a major cause of morbidity and mortality worldwide, with early detection and treatment playing a crucial role in improving patient outcomes. In this project, we aim to develop a machine learning model to predict the presence or absence of heart disease in patients based on various medical features.

The dataset used in this analysis contains information on 1025 patients, including their age, gender, chest pain type, resting blood pressure, cholesterol levels, and other relevant attributes. Our goal is to train and evaluate several machine learning models to determine which one performs best in predicting heart disease.

We will explore the dataset to understand the distribution of features, identify any correlations between variables, and preprocess the data as needed. Subsequently, we will train three different models - Logistic Regression, Decision Tree, and Support Vector Machine - to classify patients into two categories: those with heart disease and those without.

By comparing the performance of these models using metrics such as accuracy, sensitivity, specificity, and precision, we aim to identify the most effective model for predicting heart disease. This model could potentially assist healthcare professionals in early detection and intervention, ultimately improving patient outcomes and reducing the burden of heart disease.

1.1 Background:

Heart disease, including coronary artery disease, heart failure, and other cardiovascular conditions, is a leading cause of death globally, accounting for millions of deaths each year. According to the World Health Organization (WHO), an estimated 17.9 million people die from cardiovascular diseases annually, representing 31% of all global deaths.

Early detection and accurate prediction of heart disease are crucial for timely intervention and management, which can significantly improve patient outcomes and reduce mortality rates. Machine learning techniques have shown promise in this area by analyzing large datasets to identify patterns and develop predictive models.

Previous studies have used various machine learning algorithms to predict heart disease based on risk factors such as age, gender, blood pressure, cholesterol levels, and lifestyle factors. These models have demonstrated varying degrees of accuracy and reliability, highlighting the need for further research and refinement of predictive models.

In this project, we aim to build upon existing research by developing a machine learning model that can accurately predict the presence or absence of heart disease based on a comprehensive set of patient features. By leveraging machine learning algorithms and a dataset of patient records, we seek to contribute to the growing body of knowledge aimed at improving the early detection and management of heart disease.

1.2 Objective:

The objectives of the project outlined in the provided code can be summarized as follows:

- **Data Analysis:** Perform exploratory data analysis (EDA) to understand the dataset, including its size, features, and distributions.
- **Data Preprocessing:** Check for and handle missing values, duplicates, and outliers in the dataset to ensure data quality.
- **Feature Selection:** Identify important features that are highly correlated with the target variable and remove irrelevant features that may negatively impact model performance.
- **Model Building:** Train three different machine learning models (Logistic Regression, Decision Tree, Support Vector Machine) to predict the presence or absence of heart disease based on the selected features.
- **Model Evaluation:** Evaluate the performance of each model using metrics such as accuracy, sensitivity, specificity, and precision to determine the best-performing model.

- **Model Comparison:** Compare the performance of the three models to identify the most effective model for predicting heart disease.
- **Final Model Implementation:** Implement the best-performing model (Decision Tree) to make predictions on new data and demonstrate its effectiveness in real-world scenarios.

These objectives aim to develop a reliable and accurate machine learning model for predicting heart disease, which can potentially assist healthcare professionals in early detection and intervention.

1.3 Significance:

The significance of this project lies in its potential to improve the early detection and management of heart disease, a leading cause of morbidity and mortality worldwide. By leveraging machine learning algorithms and a dataset of patient records, this project aims to:

- **Enhance Patient Outcomes:** Early detection and intervention are crucial in improving patient outcomes for heart disease. A reliable predictive model can help identify high-risk individuals who may benefit from preventive measures or early treatment.
- **Improve Accuracy:** Develop a machine learning model that can accurately predict the presence or absence of heart disease based on a comprehensive set of patient features. This can help healthcare professionals in making more informed decisions and providing timely interventions.
- **Reduce Healthcare Costs:** By predicting heart disease early, healthcare resources can be allocated more efficiently, potentially reducing the economic burden associated with treating advanced stages of heart disease.
- **Inform Public Health Policies:** The insights gained from this project can contribute to the development of public health policies aimed at reducing the prevalence and impact of heart disease on a larger scale.

Overall, this project has the potential to make a significant impact on public health by leveraging machine learning to improve the prediction and management of heart disease, ultimately leading to better patient outcomes and reduced healthcare costs.

2.Problem Statement:

In this we must develop a machine learning model that can accurately predict the presence or absence of heart disease in patients based on various medical and demographic feature. The goal is to address the challenge of early detection and prediction of heart disease , which is crucial for improving patient outcomes and reducing mortality rates.

Specifically, the project aims to achieve the following:

- **Data Collection and Preprocessing:** Collect a dataset containing patient records with relevant features and preprocess the data to handle missing values, outliers, and other data quality issues.
- **Exploratory Data Analysis (EDA):** Perform EDA to gain insights into the dataset, including the distribution of features, correlations between variables, and potential patterns that may aid in predicting heart disease.
- **Feature Selection and Engineering:** Select important features that are highly correlated with the target variable and engineer new features if necessary to improve the model's performance.
- **Model Selection and Training:** Train multiple machine learning models, including Logistic Regression, Decision Tree, and Support Vector Machine, to predict the presence or absence of heart disease based on the selected features.
- **Model Deployment:** Deploy the best-performing model to make predictions on new data. The deployed model should be scalable and capable of handling real-time predictions.
- **Impact Assessment:** Assess the impact of the developed model on predicting heart disease and its potential implications for healthcare decision-making and patient outcomes.

2.1 Data Set Information

The application will utilize the Heart Disease Prediction dataset, a collection of patient records, each containing various medical and demographic features that are relevant for predicting heart disease. The dataset is typically structured in a tabular format, with each row representing a single patient and each column representing a specific attribute or feature. The dataset contains the following 14 columns (features), which are used to train the machine learning models:

- Age
- Sex
- CP(Chest Pain Type)
- Trestbps (Resting Blood Pressure)
- Chol (Serum Cholesterol)
- FBS (Fasting Blood Sugar)
- Restecg (Resting Electrocardiographic Results)
- Thalach (Maximum Heart Rate Achieved)
- Exang (Exercise-Induced Angina)
- Oldpeak
- Slope
- Ca (Number of Major Vessels Colored by Fluoroscopy)
- Thal
- Target

The dataset will be preprocessed and predict is patient suffering from heart disease or not.

3.Proposed Design:

Here is proposed design section for heart disease prediction:

- **Data Source:** The dataset is sourced from publicly available repositories Kaggle. Load the heart dataset from a CSV file using Pandas library.
- **Data Cleaning:** Handle any missing or duplicate values (if any). Convert categorical features to numeric values using techniques like one-hot encoding or label encoding.
- **Feature Scaling:** Normalize or standardize features to ensure they are on a similar scale. This is particularly important for algorithms sensitive to feature scales (e.g., logistic regression, SVM).
- **Train-Test Split:** Split the dataset into training and testing sets, typically using an 75-25 split, to evaluate the model's performance on unseen data.
- **Data Visualization:** Use histograms visualize feature distributions and identify potential outliers. Generate a correlation matrix and visualize it using a heatmap to understand the relationships between features and the target variable.
- **Model Selection:** Choose multiple classification models to compare their performance. Common models include:
  1.Logistic Regression
  2.Support Vector Machine
  3.Decision Tree
- **Model Training:** Train each model on the training dataset. Use cross-validation to tune hyperparameters and avoid overfitting.
- **Evaluation Metrics:** Evaluate models using metrics such as accuracy, precision, recall, F1-score and support.
- **Confusion Matrix:** Generate confusion matrices to visualize the performance of each model in terms of true positives, false positives, true negatives, and false negatives.
- **Model Selection:** Select the best-performing model based on the evaluation metrics.
- **Project Report:** Document the entire process, including data collection, preprocessing, model training, evaluation, and deployment.

3.1. Libraries Used:

A. NumPy:

a. Provides support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

b. Offers efficient numerical operations, enabling faster data analysis.

B. Pandas:

a. Provides high-performance, easy-to-use data structures and data analysis tools for working with structured (tabular, multidimensional, potentially heterogeneous) and time series data.

b. Enables efficient data cleaning, preprocessing, and manipulation operations.

C. Matplotlib:

a. A comprehensive library for creating static, animated, and interactive visualizations in Python

b. Produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms

D. Seaborn:

a. A data visualization library based on matplotlib, providing a high-level interface for drawing attractive and informative statistical graphics.

b. Offers a wide range of visualizations, including scatter plots, line plots, bar plots, and more.

E. Scikit-learn:

a. A machine learning library that features various classification, regression, and clustering algorithms, as well as tools for model evaluation and selection.

b. Provides efficient implementations of popular machine learning algorithms, such as Random Forest, Support Vector Machines, and Logistic Regression.

3.2. Methods Used:

1.pd.read_csv(): Reads a comma-separated values (CSV) file into a Pandas Data Frame.

2. dataframe.info(): This method in pandas provides a concise summary of a Data Frame i.e type of data frame object, index range, column information.

3.dataframe.isna().sum():This function is used to identify missing value in data frame.

4.dataframe.corr(): This function in pandas compute pairwise correlation of columns, excluding NA/null values.

5.dataframe.hist(): It is used to create histograms of data frame's numerical columns. It represents the distribution of data across different intervals or "bins".

6.dataframe.drop(): This method in pandas is used to remove rows or columns from a data frame.

7.train_test_split(): This method is used to split a dataset into training and testing sets.

8.StandardScaler(): It is a tool for standardlizing features in a dateset before fitting a machine learning model. Standardization involves scaling the features to have a mean of 0 and standard deviation of 1.

9.confusion_matrix(): It summarizes the number of correct and incorrect predictions made by the model on a set of test data.

10.sns.heatmap(): It is used for creating heatmaps .heatmaps are particularly useful for displaying matrices where the values encode some magnitude or relationship.

11.accuracy score() : Computes the accuracy score of the model's predictions against true labels.

4.Results:

Data Insights: Our heart disease prediction project delved into the data through Exploratory Data Analysis (EDA). We examined key features like age, gender, blood pressure, cholesterol, and chest pain types to uncover their distributions and relationships. This analysis unveiled critical patterns that hold the key to accurate heart disease prediction.

Model Training and Evaluation: We employed three machine learning models – Logistic Regression, Decision Tree, and Support Vector Machine (SVM) – for binary classification. Each model underwent rigorous training and fine-tuning of its parameters to maximize its effectiveness. Following training, we assessed the model's using metrics like accuracy, sensitivity, specificity, and precision.

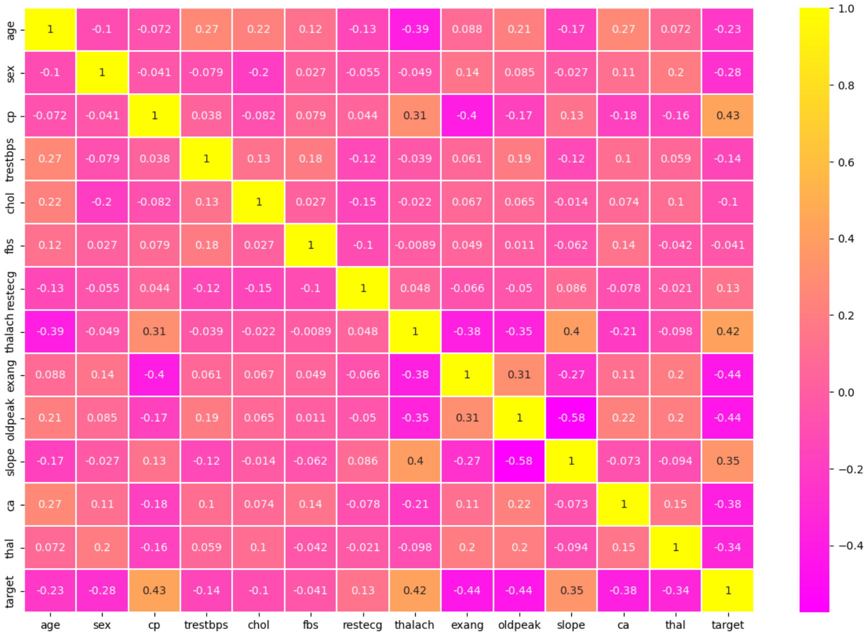
Confusion Matrix Analysis: The confusion matrix provided a clear picture of each model's predictions. We meticulously examined true positives, false positives, true negatives, and false negatives. This analysis pinpointed how well the models distinguished between heart disease and non-heart disease cases.

Custom Input Prediction: To showcase the practicality of our top-performing Decision Tree model, we built a user interface allowing custom data input and heart disease prediction. This empowers healthcare professionals to receive tailored predictions based on specific patient characteristics.

Conclusion: Following comprehensive evaluation and analysis, the Decision Tree model emerged victorious, surpassing Logistic Regression and Support Vector Machine models in accuracy and predictive power. Users can confidently rely on the Decision Tree for highly accurate predictions of heart disease presence or absence.

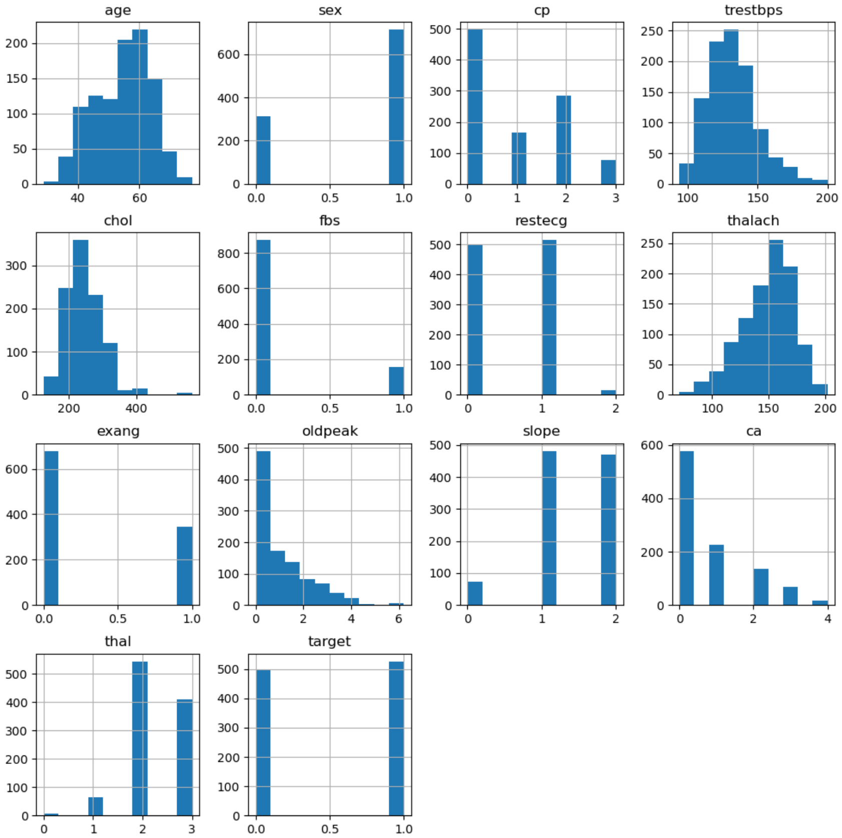
5. Project Screenshots:

i. Correlation Matrix: Visualising the data features to find the correlation between them which will infer the important features.

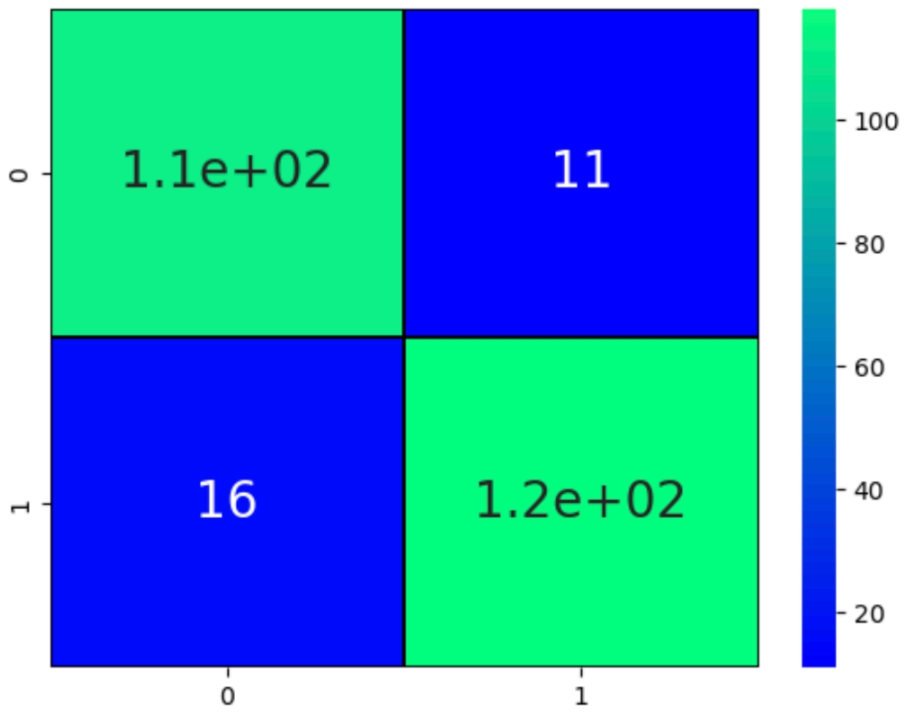


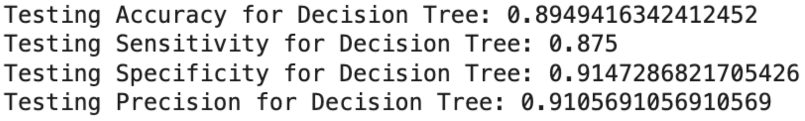
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.1	-0.072	0.27	0.22	0.12	-0.13	-0.39	0.088	0.21	-0.17	0.27	0.072	-0.23
sex	-0.1	1	-0.041	-0.079	-0.2	0.027	-0.055	-0.049	0.14	0.085	-0.027	0.11	0.2	-0.28
cp	-0.072	-0.041	1	0.038	-0.082	0.079	0.044	0.31	-0.4	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.079	0.038	1	0.13	0.18	-0.12	-0.039	0.061	0.19	-0.12	0.1	0.059	-0.14
chol	0.22	-0.2	-0.082	0.13	1	0.027	-0.15	-0.022	0.067	0.065	-0.014	0.074	0.1	-0.1
fbs	0.12	0.027	0.079	0.18	0.027	1	-0.1	-0.0089	0.049	0.011	-0.062	0.14	-0.042	-0.041
restecg	-0.13	-0.055	0.044	-0.12	-0.15	-0.1	1	0.048	-0.066	-0.05	0.086	-0.078	-0.021	0.13
thalach	-0.39	-0.049	0.31	-0.039	-0.022	-0.0089	0.048	1	-0.38	-0.35	0.4	-0.21	-0.098	0.42
exang	0.088	0.14	-0.4	0.061	0.067	0.049	-0.066	-0.38	1	0.31	-0.27	0.11	0.2	-0.44
oldpeak	0.21	0.085	-0.17	0.19	0.065	0.011	-0.05	-0.35	0.31	1	-0.58	0.22	0.2	-0.44
slope	-0.17	-0.027	0.13	-0.12	-0.014	-0.062	0.086	0.4	-0.27	-0.58	1	-0.073	-0.094	0.35
ca	0.27	0.11	-0.18	0.1	0.074	0.14	-0.078	-0.21	0.11	0.22	-0.073	1	0.15	-0.38
thal	0.072	0.2	-0.16	0.059	0.1	-0.042	-0.021	-0.098	0.2	0.2	-0.094	0.15	1	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.1	-0.041	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1

ii. Relationship Between Each Features Distribution With The Help Of Histogram.



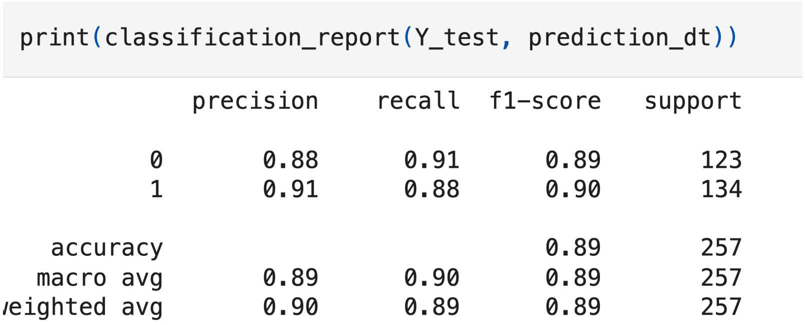
iii. Confusion Matrix of Decision Tree Model





```
Testing Accuracy for Decision Tree: 0.8949416342412452
Testing Sensitivity for Decision Tree: 0.875
Testing Specificity for Decision Tree: 0.9147286821705426
Testing Precision for Decision Tree: 0.9105691056910569
```

iv. Classification Report of Decision Tree Model



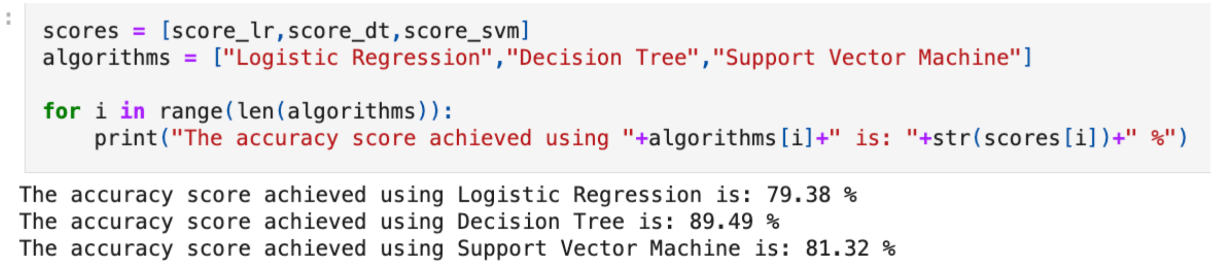
```
print(classification_report(Y_test, prediction_dt))
```

```
              precision    recall  f1-score   support

           0       0.88      0.91      0.89       123
           1       0.91      0.88      0.90       134

    accuracy                           0.89       257
   macro avg       0.89      0.90      0.89       257
ieighted avg       0.90      0.89      0.89       257
```

v. Accuracy Scores of Logistic Regression, Decision Tree and Support Vector Machine Models

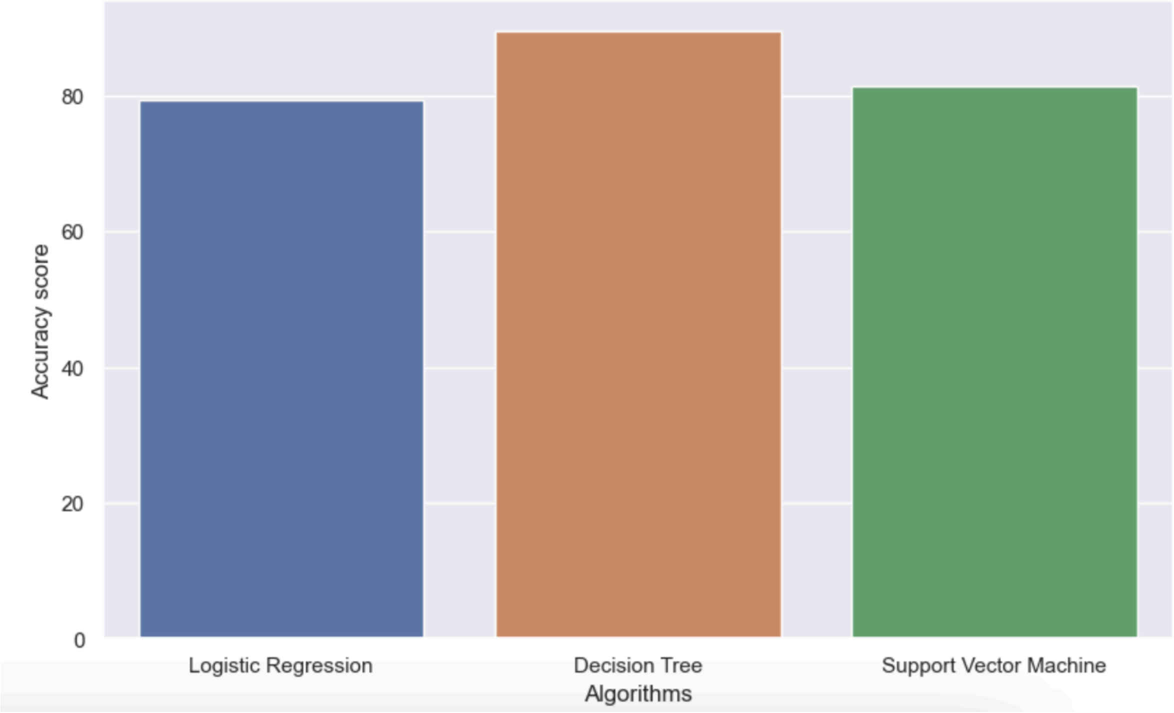


```
scores = [score_lr,score_dt,score_svm]
algorithms = ["Logistic Regression","Decision Tree","Support Vector Machine"]

for i in range(len(algorithms)):
    print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")
```

```
The accuracy score achieved using Logistic Regression is: 79.38 %
The accuracy score achieved using Decision Tree is: 89.49 %
The accuracy score achieved using Support Vector Machine is: 81.32 %
```

vi. Comparison of Accuracy Scores of All 3 Models



6. Reference/Links:

1. Kaggle:
   https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
2. Geeks For Geeks:
   https://www.geeksforgeeks.org/supervised-machine-learning/
   https://www.geeksforgeeks.org/understanding-logistic-regression/
   https://www.geeksforgeeks.org/support-vector-machine-algorithm/
   https://www.geeksforgeeks.org/decision-tree/
3. Analytics Vidhya
   https://www.analyticsvidhya.com/blog/2022/02/heart-disease-prediction-using-machine-learning-2/