PROJECT REPORT

Heart Disease Prediction

Submitted towards the partial fulfillment of the criteria for award

Of

Post Graduation Analytics

Batch 32

AMRIT KUMAR

ABSTRACT

The application of Machine Learning is highly successful in fields like business, retail and marketing has led to a vast spread in its application in various other industries. Among these sectors is the healthcare industry. This industry is known to be information rich and has a great scope for effective decision making & discovering hidden patterns. Advanced machine learning techniques can help us with this situation. Talking about the healthcare industry, one of the most talked about is the heart diseases. Heart is a vital organ. It pumps blood and supplies it to all organs to function. Prediction of the occurrences of heart diseases is significant work in this field. In this work, eight machine leaning algorithms which include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and K-Nearest Neighbors. These algorithms are used to unfold a prediction system which will analyze and predict whether the particular patient is pertaining to any heart disease or not. The main objective is to identify the best algorithm suitable for this, and which provides maximum accuracy.

Keywords

1. **P**rediction
2. **V**ariable
3. **T**arget
4. **M**anagement
5. **A**nalyze
6. **C**lassification
7. **M**achine Learning

Acknowledgements

I am using this opportunity to express my gratitude to everyone who supported us throughout the course of this project. I thankful for their aspiring guidance,invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, I am fortunate to have Purushottom Sharma and Vikas Atray as our mentor. He has readily shared his immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills. I wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the PGA program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: August 20, 2023

Amrit Kumar

Place: New Delhi

Certificate of Completion

I hereby certify that the project titled “Heart Disease Prediction” was undertaken and completed under my supervision by Amrit Kumar from the Batch of PGA 32

Mentor: Purushottam Sharma

Date: August 20, 2023

Place – New Delhi

Dataset:

The data set contain the following fields.

- 1. Gender : Male and Female.
- 2. Age: Person's age in Years.
- 3. Education : Levels of Education.
- 4. Current Smoker : Currently a Smoker.
- 5. Cigs Per Day : Number of Cigarettes Smoked per Day.
- 6. BP Meds : Blood Pressure Medications.
- 7. Prevalent Stroke : Prevalent Stroke.
- 8. Prevalent Hyp : Prevalent Hypertension.
- 9. Diabetes : Diabetes.
- 10. Tot Chol : Total Cholesterol Level (units mg/dL).
- 11. Sys BP : Systolic Blood Pressure (units mm Hg).
- 12. Dia BP : Diastolic Blood Pressure (units mm Hg).
- 13. BMI : Body Mass Index.
- 15. Heartrate : Heart Rate (pulse), Beats per minute (BPM)
- 16. Glucose : Blood Glucose level (units mg/dL).
- 17. Heart_ stroke : Target Variable (Risk of a Stroke).

	Gender	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
0	Male	39	postgraduate	0	0.0	0.0	no	0	0	195.0	106.0	70.0	26.97	80.0	77.0
1	Female	46	primaryschool	0	0.0	0.0	no	0	0	250.0	121.0	81.0	28.73	95.0	76.0
2	Male	48	uneducated	1	20.0	0.0	no	0	0	245.0	127.5	80.0	25.34	75.0	70.0
3	Female	61	graduate	1	30.0	0.0	no	1	0	225.0	150.0	95.0	28.58	65.0	103.0
4	Female	46	graduate	1	23.0	0.0	no	0	0	285.0	130.0	84.0	23.10	85.0	85.0
5	Female	43	primaryschool	0	0.0	0.0	no	1	0	228.0	180.0	110.0	30.30	77.0	99.0
6	Female	63	uneducated	0	0.0	0.0	no	0	0	205.0	138.0	71.0	33.11	60.0	85.0
7	Female	45	primaryschool	1	20.0	0.0	no	0	0	313.0	100.0	71.0	21.68	79.0	78.0
8	Male	52	uneducated	0	0.0	0.0	no	1	0	260.0	141.5	89.0	26.36	76.0	79.0
9	Male	43	uneducated	1	30.0	0.0	no	1	0	225.0	162.0	107.0	23.61	93.0	88.0

Table of Contents

Abstract 2
Acknowledgements 3
Certificate of Completion 4
Dataset 5,6
CHAPTER 1: INTRODUCTION 9
1.1 Title & Objective of the study 9
1.2 Need of the Study 10
1.3 Data Sources 11
1.4 Tools & Techniques 11
CHAPTER 2: DATA PREPARATION AND UNDERSTANDING 12
2.1 Phase I – Data Extraction and Cleaning: 12
2.2 Phase II - Feature Engineering 12
2.3 Data Dictionary: 13
2.4 Exploratory Data Analysis: 14
CHAPTER 3: DEFINITIONS 15
3.1 Phase I – Supervised ML 15
3.1.1 Classification 15
3.1.2 Regression 15
3.2 Phase II – Unsupervised ML 16
3.2.1 Association 16
3.2.2 Clustering 16

CHAPTER 4: Model Definitions 17
4.1 Decision Tree Regression 17
4.2 Logistic Regression 17
4.3 Random Forest Regression 17
4.4 Support Vector Machine 18
4.5 K – Nearest Neighbors 18
CHAPTER 5: FITTING MODELS TO DATA..... 19
5.1 Decision Tree Classification MODEL..... 19
5.2 Logistic Regression MODEL..... 19
5.2.1 Area Under the ROC Curve & Correaltion Coefficient..... 20
5.3 Random Forest MODEL..... 21
5.4 Support Vector Machine MODEL..... 21
5.5 K – Nearest Neighbors MODEL..... 21
CHAPTER 6: KEY FINDINGS 22
6.1 Accuracy Table..... 22
6.2 Model Performance Accuracy(Bar Graph)..... 23
CHAPTER 7: Conclusion and Recommendations..... 24,25

CHAPTER 1 : INTRODUCTION

1.1 Title & Objective of the study

Title of the project is Heart Disease Prediction.

Objective: The objective of this study is to investigate the relationship between various lifestyle factors and the risk of heart disease.

The objective of this machine learning project is to build and deploy predictive models for risk of Heart Disease .

The specific aims of the study include:

1: Predictive Analytics - Discuss the implementation of predictive models for assessment, such as classification models for identifying and regression models for estimating risk of heart disease. Evaluate the performance of these models using appropriate metrics.

2: Prescriptive Analytics – Introduce the concept of prescriptive analytics and the various techniques used for optimization. Explain how prescriptive analytics can leverage the predictive model's outputs to make informed decisions.

1.2 Need of the Study:

1. **Data Quality:** Discuss the data quality aspects of the dataset, such as completeness, accuracy, and consistency. Address any challenges or limitations related to data quality that might impact the analysis or results.

2. **Data Preprocessing Considerations:** Discuss any preprocessing steps required to clean and prepare the data for analysis. This may include handling missing values, outlier detection, feature engineering, and data transformation.

3. **Comparative Analysis:** If there were other potential insurance datasets considered, briefly explain why the chosen dataset was selected over others and how it compares in terms of relevance, data quality, and alignment with research objectives.

4. **Potential Insights:** Highlight the potential insights that can be gained from the dataset, such as patterns in claim frequencies, factors affecting claim amounts, and customer behavior related to insurance coverage.

5. **Practical Significance:** Explain how the findings and analyses derived from the insurance dataset will contribute to the practical significance of the study, such as improving insurance assessment practices, enhancing underwriting decisions, or optimizing management.

1.3 Data Sources:

The success of our Heart Disease Risk prediction project relies on the availability of reliable and relevant data. In this section, we detail the sources from which we obtained the data for our analysis and model development.

1.4 Tools & Techniques

Tools : Jupyter Notebook, Tableau

Techniques : Python Libraries, Machine Learning(Correlation Analysis, Feature Engineering, Data Preprocessing, Model Evaluation, Data Visualization.

CHAPTER 2: DATA PREPARATION AND UNDERSTANDING

One of the first steps we engaged in was to outline the sequence of steps that we will be following for our project. Each of these steps are elaborated below.

2.1 Phase I – Data Extraction and Cleaning:

- Missing Value Analysis and Treatment
- Handling Outliers
- Feature Extraction

2.2 Phase II - Feature Engineering:

- Encoding (Lable Encoding)
- EDA (Exploratory Data Analysis)
- Feature Scaling
- Handling Imbalance Data (Dependent Variable)
- Feature Selection

2.3 Data Dictionary:

	Gender	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
0	Male	39	postgraduate	0	0.0	0.0	no	0	0	195.0	106.0	70.0	26.97	80.0	77.0
1	Female	46	primaryschool	0	0.0	0.0	no	0	0	250.0	121.0	81.0	28.73	95.0	76.0
2	Male	48	uneducated	1	20.0	0.0	no	0	0	245.0	127.5	80.0	25.34	75.0	70.0
3	Female	61	graduate	1	30.0	0.0	no	1	0	225.0	150.0	95.0	28.58	65.0	103.0
4	Female	46	graduate	1	23.0	0.0	no	0	0	285.0	130.0	84.0	23.10	85.0	85.0
5	Female	43	primaryschool	0	0.0	0.0	no	1	0	228.0	180.0	110.0	30.30	77.0	99.0
6	Female	63	uneducated	0	0.0	0.0	no	0	0	205.0	138.0	71.0	33.11	60.0	85.0
7	Female	45	primaryschool	1	20.0	0.0	no	0	0	313.0	100.0	71.0	21.68	79.0	78.0
8	Male	52	uneducated	0	0.0	0.0	no	1	0	260.0	141.5	89.0	26.36	76.0	79.0
9	Male	43	uneducated	1	30.0	0.0	no	1	0	225.0	162.0	107.0	23.61	93.0	88.0

0. Gender - Object
1. Age - Integer
2. Education - Object
3. CurrentSmoker - Integer
4. CigsPerDay – Float
5. BPMeds - Float
6. PrevalentStroke - Object
7. PrevalentHyp - Integer
8. Diabetes - Integer
9. TotChol - Float
10. SysBP – Float
11. DiaBP - Float
12. BMI – Float
13. HeartRate - Float
14. Glucose – Float
15. Heart_ stroke - Object

2.4 Exploratory Data Analysis:

Missing Data Analysis:

Identify and analyze any missing values in the dataset. Visualize the missing data patterns and consider strategies for handling missing values during preprocessing.

Outlier Detection:

Detect and examine outliers in numerical variables. Plot box plots or use statistical methods to identify potential outliers and understand their impact on the data.

Data Visualization:

Use various data visualization techniques to explore the dataset visually. Create visualizations like histograms to understand the distribution and relationships between variables.

Data Preprocessing:

Data preprocessing is an important step in the data mining process. It refers to the cleaning,transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

CHAPTER 3. Definitions

3. 1 Supervised Learning:

Supervised learning is more widely used among the mentioned types. It is called supervised learning because in this approach, the algorithm is trained on the input data to get the desired result. Another way to define it is an algorithm is trained under a supervision of training data. In this case, the data used to train the algorithm is a labeled data. When the target or the result is known, the data is called labeled data. In a typical supervised machine learning problem, the data has two parts. The first part consists of input variables which are called features. The second part is actual target variable or label. Features helps to find out the target.

3.1.1 Classification:

The target variable or the label in a classification problem holds categorical values. That means the target variable is discrete. The categorical values of this target variable are usually finite. The mapping function given in the previous section then would try to predict one of the categorical values, the target variable has with the help of input variables.

3.1.2 Regression:

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

3.2 Unsupervised Learning:

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings, without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis.

3.2.1 Clustering:

Clustering is a data mining technique which groups unlabeled data based on their similarities or differences. Clustering algorithms are used to process raw, unclassified data objects into groups represented by structures or patterns in the information.

3.2.2 Association:

An association rule is a rule-based method for finding relationships between variables in a given dataset. These methods are frequently used for market basket analysis, allowing companies to better understand relationships between different products.

CHAPTER 4 : Model Definition

4.1 Decision tree regression:

Regression or classification models in decision treeregression builds in the form of a tree structure. The dataset is divided or segmented into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision tree with decision nodes and leaf nodes is obtained as a final result.

4.2 Logistic Regression:

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not. It is a kind of statistical algorithm, which analyze the relationship between a set of independent variables and the dependent binary variables. It is a powerful tool for decision-making.

4.3 Random Forest:

Random Forest is a robust machine learning algorithm that can be used for a variety of tasks including regression and classification. It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction.

4.4 Support Vector Machine:

A support vector machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far a part as possible. SVMs are used in text categorization, image classification, handwriting recognition and in the sciences.

4.4 K-Nearest Neighbors:

K-nearest neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closet to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes of 'K' training data and class holds the highest probability will be selected. In the case of regression, the value is the mean of the 'K' selected training points.

5. FITTING MODELS TO DATA

Decision Tree
Logistic Regression
Random Forest
Support Vector Machine
K-Nearest Neighbors

5.1 Decision Tree Classification MODEL

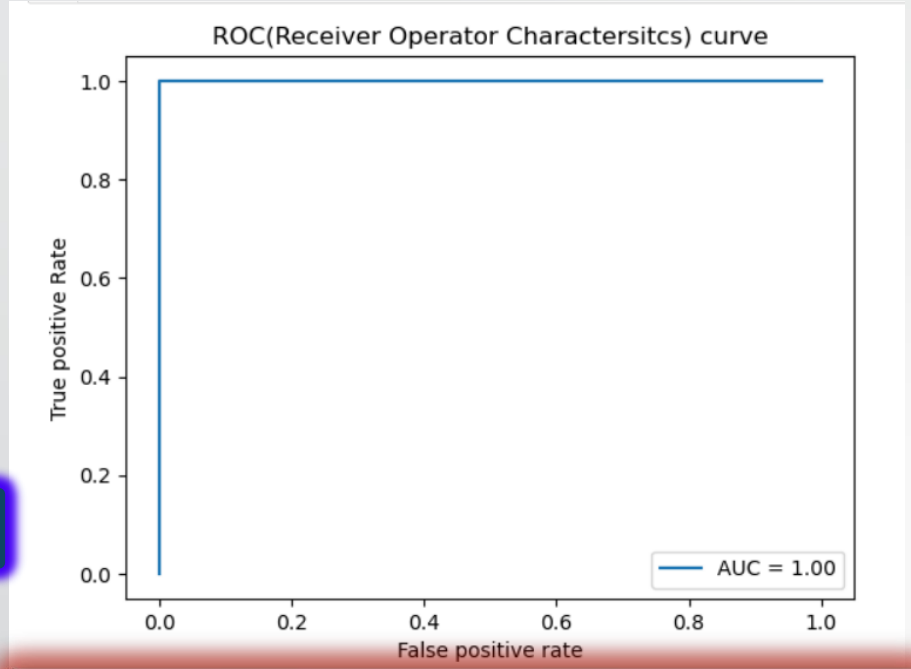
We applied Decision Tree on the Training data set. Below were the parameters which were applied for Decision Tree.

- Confusion Matrix
- Classification Report
- Accuracy Score

5.2 Logistic Regression MODEL

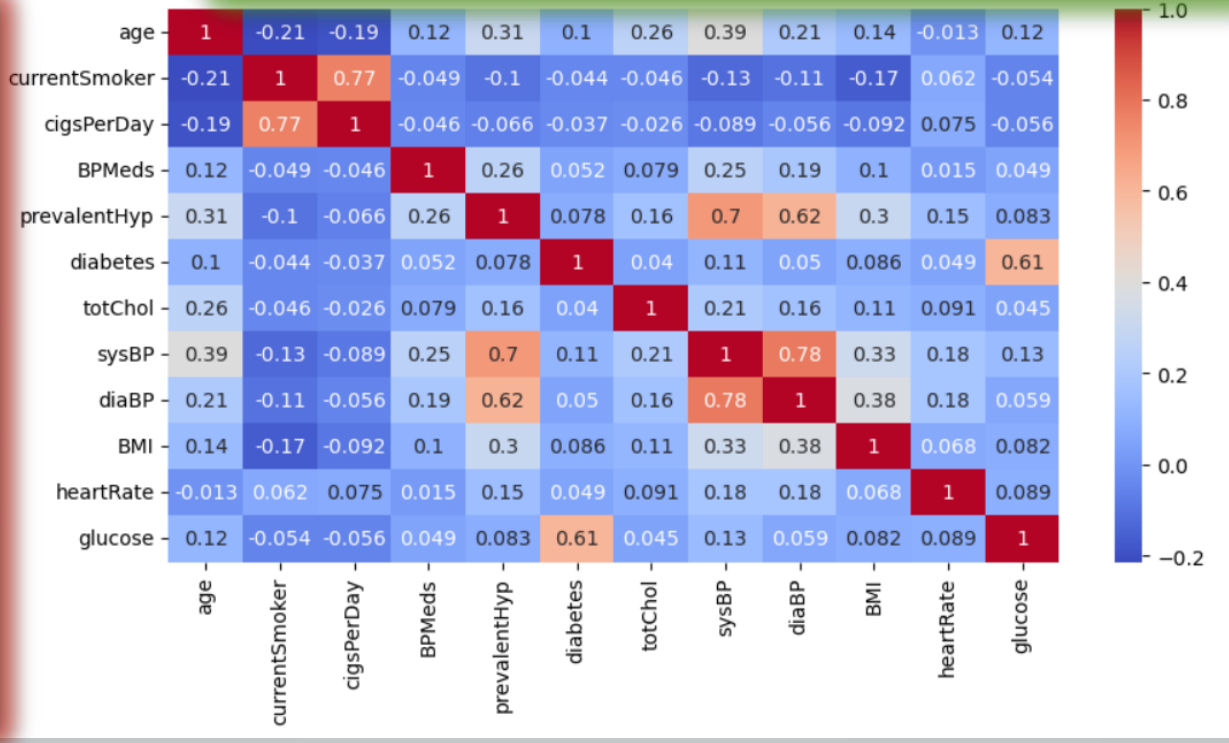
- Confusion Matrix,
- Classification Report,
- Accuracy Score
- AUC
- Cross Validation

5.2.1 Area Under the ROC Curve



- **AUC : Area Under the ROC Curve Accuracy score of 100%**
- **The model's predictions completely align with the true labels, and there are True Positive or True Negative.**
- **The model is very simple and does not overfit the training data.**
- **The model is very robust to noise.**

- **The Correlation between current Smoker and cigs Per Day is 77% Or sys BP and dia BP is 78%. Which is a Strong Positive Correlation. This means that there is a Positive relationship between the two variables.**



5.3 Random Forest MODEL

- Confusion Matrix
- Classification Report
- Accuracy Score

5.4 Support Vector Machine MODEL

- Confusion Matrix
- Classification Report
- Accuracy Score

5.5 K-Nearest Neighbors MODEL

- Confusion Matrix
- Classification Report
- Accuracy Score

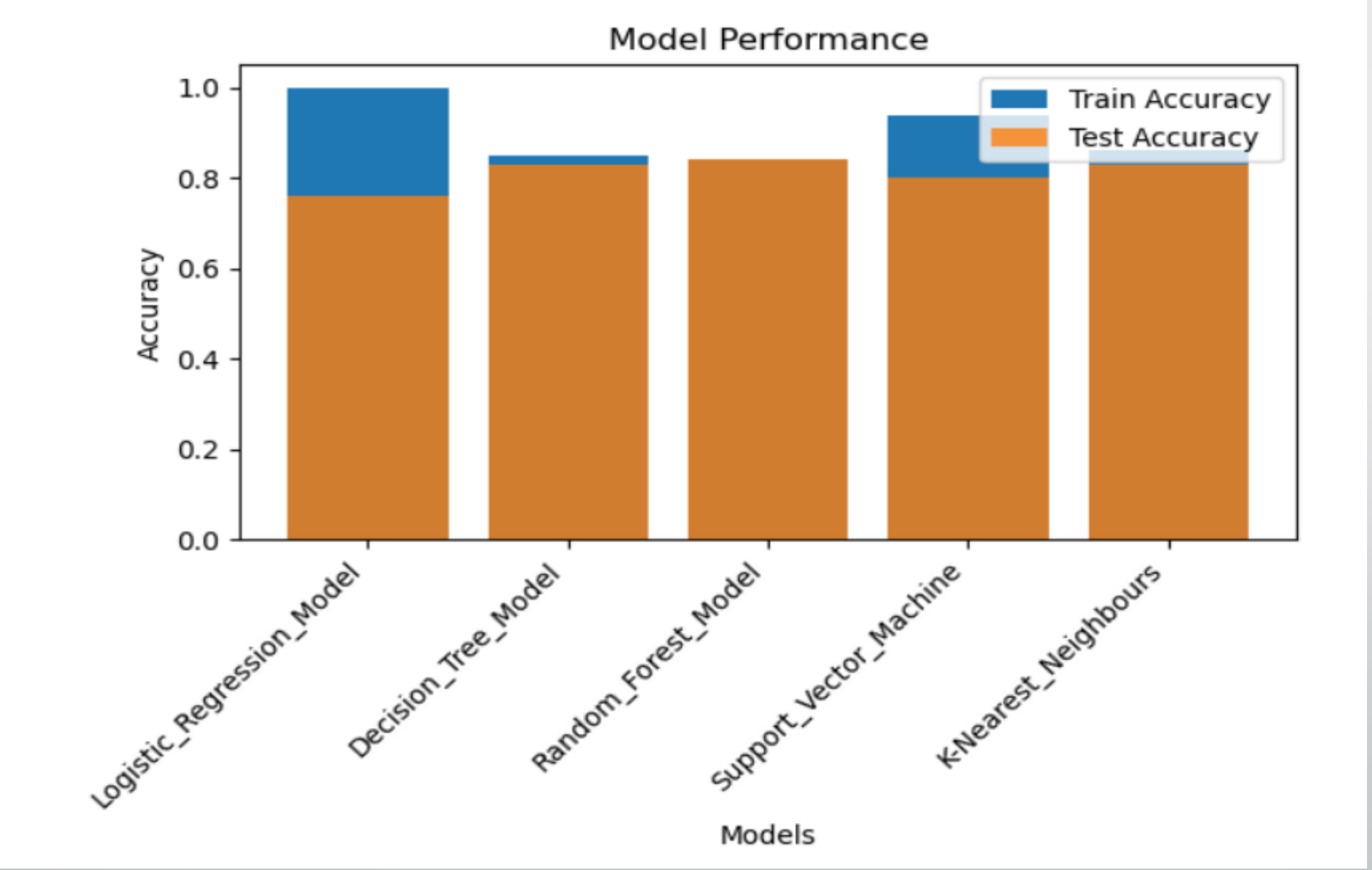
CHAPTER 6 : KEY FINDINGS

6.1 ACCURACY TABLE

	Models	Train	Test
0	Logistic_Regression_Model	1.00	0.76
1	Decision_Tree_Model	0.85	0.83
2	Random_Forest_Model	0.84	0.84
3	Support_Vector_Machine	0.94	0.80
4	K-Nearest_Neighbours	0.86	0.83

Sr No.	Model Name	Accuracy	OutPut
0	Logistic_Regrssion_Model	0.76	Fair Accuracy
1	Decision_Tree_Model	0.83	Good Accuracy
2	Random_Forest_Model	0.84	Good Accuracy
3	Support_Vector_Machine	0.80	Acceptable Accuracy
4	K-Nearest Neighbors	0.83	Good Accuracy

6.2 Model Performance Accuracy:



CHAPTER 7: Conclusion and Recommendations:

- The logistic regression model appears to overfit the training data, as evidenced by the perfect training accuracy but a lower test accuracy. It may benefit from regularization techniques to improve generalization.

- Decision tree, random forest, and KNN models demonstrate relatively good generalization performance, with test accuracies around 83-84%. These models could be suitable for heart disease risk prediction based on the given features.

- The SVM model shows signs of overfitting, as indicated by the gap between training and test accuracy. Fine-tuning hyperparameters or considering other kernel functions may help improve its generalization.

- Further investigation into feature engineering, including feature selection and engineering new features, could potentially enhance model performance.

- It's essential to consider additional evaluation metrics, such as precision, recall, and F1-score, especially if there is class imbalance or different costs associated with false positives and false negatives.

- To improve the robustness of the models, more extensive datasets and cross-validation techniques could be considered. In summary, the decision tree, random forest, and KNN models appear to perform reasonably well for heart disease prediction based on the provided dataset, while the logistic regression and SVM models may require further optimization to enhance their generalization capabilities.

Thank You !