A Project Report

on

Heart Disease Detection Using Machine Learning

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2022-23

in

Name of discipline

By

Dhirendra Yadav (1900290100057)

Devesh Chandra (1900290100056)

Utkarsh Singh (1900290100182)

Dharmendra Singh Yadav (1900290400054)

Under the supervision of

Asst. Prof. Pushpendra Kumar

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2023

(PCS23-15)

DECLARATION

We affirm that this submission is entirely our original creation and, as far as we know and believe, it does not include any content that has been previously published or authored by someone else. Furthermore, it does not incorporate material that has been substantially accepted for the attainment of any other academic degree or diploma from the university or any other institution of higher education, except where proper credit has been acknowledged within the text.

Signature:-
Name:- Dhirendra Yadav
Roll No.:- 1900290100057

Signature:-
Name:- Devesh Chandra
Roll No.:- 1900290100056

Signature:-
Name:- Utkarsh Singh
Roll No.:- 1900290100182

Signature:-
Name:- Dharmendra Singh Yadav
Roll No.:- 1900290400054

Place: KIET Groups Of Institution

Date:-

CERTIFICATE

I hereby confirm that the Project Report titled "Utilizing Machine Learning for Heart Disease Detection," authored by Dhirendra Yadav, Utkarsh Singh, Devesh Chandra and Dharmendra Singh Yadav, is a genuine representation of their independent work towards fulfilling the requirements for the B. Tech. degree in the Department of Computer Science & Engineering at Dr. A.P.J. Abdul Kalam Technical University, Lucknow. I have personally supervised their progress throughout this endeavor. The contents of this report are authentic and have not been previously presented for the attainment of any other academic qualification

Date:

Supervisor Name:

Asst. Prof. Pushpendra Kumar

(Assistant Prof.)

ACKNOWLEDGEMENT

We take immense joy in sharing the B. Tech Project report that was completed during our final year of the B. Tech program. We express our sincere appreciation to Prof. Pushpendra Kumar from the Department of Computer Science & Engineering at KIET, Ghaziabad, for his unwavering assistance and mentorship throughout our project. His dedication, meticulousness, and determination have continuously motivated us. It is solely due to his conscientious contributions that our efforts have come to fruition.

We would like to extend our sincere appreciation to Dr. Vineet Sharma, the esteemed Head of the Department of Computer Science & Engineering at KIET, Ghaziabad. His steadfast support and invaluable guidance played a crucial role in the development of our project. Furthermore, we would like to express our gratitude to the entire faculty of the department for their generous assistance and cooperative efforts throughout the project's implementation.

We also highly value recognizing the efforts made by the entire faculty team, including individuals from academia, industry, or any other relevant field, within our department. Their valuable assistance and cooperation were pivotal during the progression of our project. Finally, we extend our gratitude to our friends for their valuable contributions in successfully concluding the project.

Signature:-

Name:- Dhirendra Yadav

Roll No.:- 1900290100057

Signature:-

Name:-Devesh Chandra

Roll No.:- 1900290100056

Signature:-

Name:- Utkarsh Singh

Roll No.:- 1900290100056

Signature:-

Name:- Dharmendra Singh Yadav

Roll No.:- 1900290400054

ABSTRACT

Machine Learning has widespread applications worldwide, including in the healthcare sector. It can be instrumental in predicting various conditions such as locomotor disorders and heart diseases. These predictions, when made ahead of time, can offer valuable insights to doctors, enabling them to adjust their diagnoses and treatments on a per-patient basis. Our research focuses on utilizing Machine Learning algorithms to predict potential heart diseases in individuals. We conduct a comparative analysis of classifiers, including Decision trees, Logistic Regression, and K-Nearest Neighbor. Additionally, we propose an ensemble classifier that combines multiple approaches to achieve hybrid classification.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ML : Machine Learning

AI : Artificial Intelligence

NN : Neural Network

SVM : Support Vector Machine

XG : Extreme Gradient

EXANG : Exercise induced angina

CHAPTER 1

INTRODUCTION

1.1 Introduction

The World Health Organization reports that heart disease is responsible for 12 million deaths annually on a global scale. It is a significant contributor to both illness and mortality worldwide. Analyzing data to predict cardiovascular disease is considered a crucial area of study. Over the past few years, the burden of cardiovascular disease has been steadily increasing worldwide. Numerous research studies have been conducted to identify the most influential factors contributing to heart disease and to develop accurate risk prediction models. Heart disease is often referred to as a silent killer, as it can lead to a person's demise without obvious symptoms. Timely detection of heart disease plays a critical role in making informed decisions regarding lifestyle modifications for individuals at high risk, ultimately reducing complications.

Machine learning has demonstrated its effectiveness in aiding decision-making and forecasting within the healthcare sector by processing the substantial volume of data generated. The objective of this project is to employ machine learning algorithms to analyze patient data, thereby predicting future occurrences of Heart Disease. Utilizing machine learning techniques can prove advantageous in this context. While heart disease may manifest in various forms, there exists a common set of fundamental risk factors that determine an individual's susceptibility to the condition. By gathering data from diverse sources, organizing it into relevant categories, and subsequently conducting thorough analysis, it can be inferred that this approach is well-suited for heart disease prediction.

1.2 Motivation of Work

The primary objective of this research is to develop a heart disease prediction model that can accurately predict the occurrence of heart disease. Additionally, the study aims to determine the most effective classification algorithm for identifying the likelihood of heart disease in patients. To justify this research, a comparative analysis was conducted using three commonly employed classification algorithms: Naïve Bayes, Decision Tree, and Random Forest. Given the critical nature of heart disease prediction and the need for utmost accuracy, these algorithms were extensively evaluated using various levels and types of assessment strategies. The findings from this research will offer valuable insights to researchers and medical professionals, enabling them to establish improved approaches in this field.

1.3 Problem Statement

Detecting heart disease poses a significant challenge due to various factors. Existing tools capable of predicting heart disease are either prohibitively expensive or lack the necessary efficiency to accurately assess the risk in individuals. Timely identification of cardiac ailments plays a crucial role in reducing mortality rates and minimizing complications. However, constant monitoring of patients is often unfeasible, and round-the-clock doctor consultation is limited by the need for extensive knowledge, time, and expertise. Fortunately, with the abundance of data available today, we can leverage diverse machine learning algorithms to scrutinize this data for concealed trends. These hidden patterns can be harnessed to facilitate health diagnosis in the field of medicine.

CHAPTER 2

LITERATURE REVIEW

With growing development in the field of medical science alongside machine learning various experiments and research has been carried out in these recent years releasing the relevant significant papers.

[1] A study conducted by Purushottam et al. introduced a research paper titled "Efficient Heart Disease Prediction System." The authors employed hill climbing and decision tree algorithms in their approach. The Cleveland dataset was utilized, and data preprocessing was performed before applying the classification algorithms. To handle missing values in the dataset, the authors employed Knowledge Extraction based on Evolutionary Learning (KEEL), an open-source data mining tool. The decision tree was constructed in a top-down manner, with the hill-climbing algorithm selecting nodes at each level based on a test. The parameters used in the study included confidence, with a minimum value of 0.25. The system achieved an accuracy rate of approximately 86.7%.

[2] In their study titled "Heart Disease Prediction Using Machine Learning Algorithms," Santhana Krishnan. J, et al. introduced a research paper that employed decision tree and Naive Bayes algorithms to predict heart disease. The decision tree algorithm constructs a tree structure based on specific conditions, enabling True or False decisions. Other algorithms such as SVM and KNN utilize vertical or horizontal split conditions depending on the dependent variables. The decision tree algorithm, characterized by its tree-like structure with root nodes, leaves, and branches, also aids in comprehending the significance of attributes within the dataset. The researchers utilized the Cleveland dataset, dividing it into a 70% training set and a 30% testing set using certain methods. This algorithm achieved a 91% accuracy rate. The second algorithm employed was Naive Bayes, which is well-suited for classification tasks and capable of handling complex, nonlinear, and dependent data. Considering the intricate, dependent, and nonlinear nature of the heart disease dataset, Naive Bayes exhibited an accuracy of 87%.

[3] Sonam Nikhar and colleagues presented a research paper titled "Utilizing Machine Learning Algorithms for Heart Disease Prediction." In their study, they provide a comprehensive explanation of the Naïve Bayes and Decision Tree classifiers, which are commonly employed in heart disease prediction. The

researchers conducted an analysis to compare the performance of these predictive data mining techniques using the same dataset. The findings revealed that the Decision Tree classifier exhibited superior accuracy compared to the Bayesian classifier.

[4] In their publication titled "Forecasting Heart Disease through Machine Learning," Aditi Gavhane and colleagues introduced a novel approach to predicting heart disease. The study utilized the multi-layer perceptron neural network algorithm for training and testing the dataset. The algorithm incorporates an input layer, an output layer, and one or more hidden layers between them. Each input node is connected to the output layer through these hidden layers, with random weights assigned to the connections. Additionally, a bias input is included, and its weight can be adjusted according to the specific needs of the connection, allowing for both feedforward and feedback connections between the nodes.

[5] In their paper titled "Heart Disease Prediction Using Effective Machine Learning Techniques," Avinash Golande and colleagues introduced a novel approach to assist doctors in distinguishing heart diseases. The study employs various data mining techniques, including k-nearest neighbor, Decision tree, and Naïve Bayes algorithms. Additionally, distinctive characterization-based methods such as packing algorithm, Part thickness, consecutive minimal optimization, neural networks, linear Kernel self-organizing map, and Support Vector Machines (SVM) are utilized.

[6] In their paper titled "Heart Disease Prediction using Machine Learning Techniques," Lakshmana Rao and colleagues addressed the challenge of identifying heart disease due to the numerous contributing factors involved. To assess the severity of heart disease in individuals, various neural networks and data mining methods were employed.

[7] In their work titled "Heart Attack Prediction Utilizing Deep Learning," Abhay Kishore and colleagues introduced a system that employs deep learning techniques to predict heart attacks and identify potential factors associated with heart-related infections in patients. The model utilizes Recurrent Neural Systems to achieve accurate results and minimize errors. This study serves as a valuable point of reference for future developments in heart attack prediction models.

[8] In their study titled "Enhancing Accuracy in Cardiovascular Disease Prediction through Hybrid Machine Learning Approaches," Senthil Kumar Mohan and colleagues aimed to improve the precision of cardiovascular issue diagnosis. The

authors employed a combination of KNN, LR, SVM, and NN algorithms, resulting in an enhanced performance level. Their prediction model for heart disease, known as hybrid random forest with linear model (HRFLM), achieved a precision level of 88.7%.

[9] In their study, Anjan N. Repaka and colleagues presented a model that evaluates the predictive performance of two classification models. This model was analyzed and compared to prior research. The experimental findings demonstrate that our proposed method outperforms other models in accurately determining the percentage of risk prediction.

[10] In their study titled "Heart Disease Prediction through Evolutionary Rule Learning," Aakash Chauhan et al. introduced a method that utilizes electronic records to streamline data retrieval and reduce manual efforts. By employing frequent pattern growth association mining on patients' datasets, the researchers generated robust associations that significantly contributed to accurate heart disease prediction. This approach effectively decreased the number of services required for prediction, enhancing the overall efficiency of the process.

CHAPTER 3
METHODOLOGY

3.1 Existing System

Heart disease is increasingly recognized as a silent killer, causing death without obvious symptoms. This characteristic of the disease has led to growing concerns and anxiety about its impact. Therefore, ongoing efforts are being made to forecast the likelihood of this fatal illness in advance. Various tools and techniques are continuously being explored to meet present-day healthcare needs, and machine learning methods hold great potential in this regard. Although heart disease can manifest in different forms, there exists a common set of core risk factors that determine an individual's susceptibility. By gathering data from diverse sources, organizing it into appropriate categories, and conducting thorough analysis, we can draw conclusions. This approach can effectively be employed for predicting heart disease. Following the well-known adage, "Prevention is better than cure," early prediction and control measures can significantly aid in reducing heart disease-related mortality rates.

3.2 Proposed System

The system begins by gathering data and identifying significant attributes. Subsequently, the necessary data undergoes preprocessing to achieve the desired format. The data is then split into two sets: training and testing data. Algorithms are employed to train the model using the training data. To evaluate the system's accuracy, it is tested with the testing data. This implementation incorporates the following modules.

3.2.1 Collection of Dataset

In the beginning, we gather a dataset specifically for our heart disease prediction system. Following the dataset collection phase, we divide it into two subsets: training data and testing data. The training dataset is employed to train the prediction model, while the testing data is utilized to assess the performance of the model. In this project, we allocate 70% of the data for training and 30% for testing. We utilize the Heart Disease UCI dataset, which encompasses a total of 76 attributes. However, for our system, we only utilize 14 of these attributes.

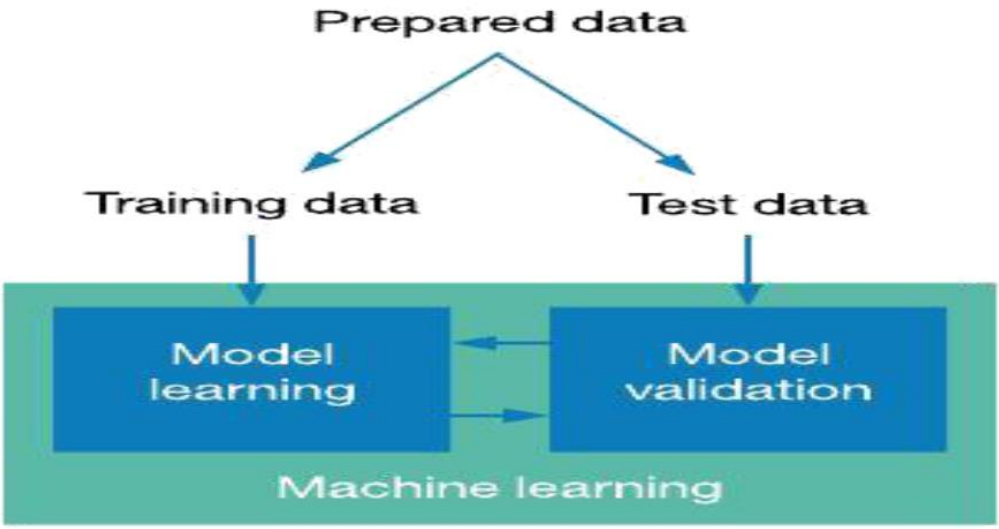


Figure: Collection of Data

3.2.2 Selection of Attributes

The process of attribute or feature selection involves choosing suitable attributes for the prediction system in order to enhance its efficiency. This selection includes various patient attributes such as gender, chest pain type, fasting blood pressure, serum cholesterol, exang, and more, which are utilized for prediction purposes. The model employs a correlation matrix to aid in the attribute selection process.

Figure: Correlation matrix

3.2.3 Pre-processing Data

Data preprocessing plays a crucial role in developing a machine learning model. At the outset, the data may be unrefined or not in the desired format, leading to potentially misleading results. To address this, data preprocessing involves converting the data into the necessary format. Its purpose is to handle issues such as noise, duplicates, and missing values within the dataset. Activities involved in data preprocessing include importing datasets, splitting them, and scaling attributes, among others. Ultimately, data preprocessing is essential for enhancing the model's accuracy.

Figure: Data Pre-processing

3.2.4 Balancing of Data

Two methods can be employed to address imbalanced datasets, namely Under Sampling and Over Sampling.

(a) Under Sampling involves reducing the size of the majority class in order to balance the dataset. This approach is used when there is a sufficient amount of data available.

(b) Over Sampling involves increasing the size of the minority class in order to achieve dataset balance. This approach is used when there is an insufficient amount of data available.

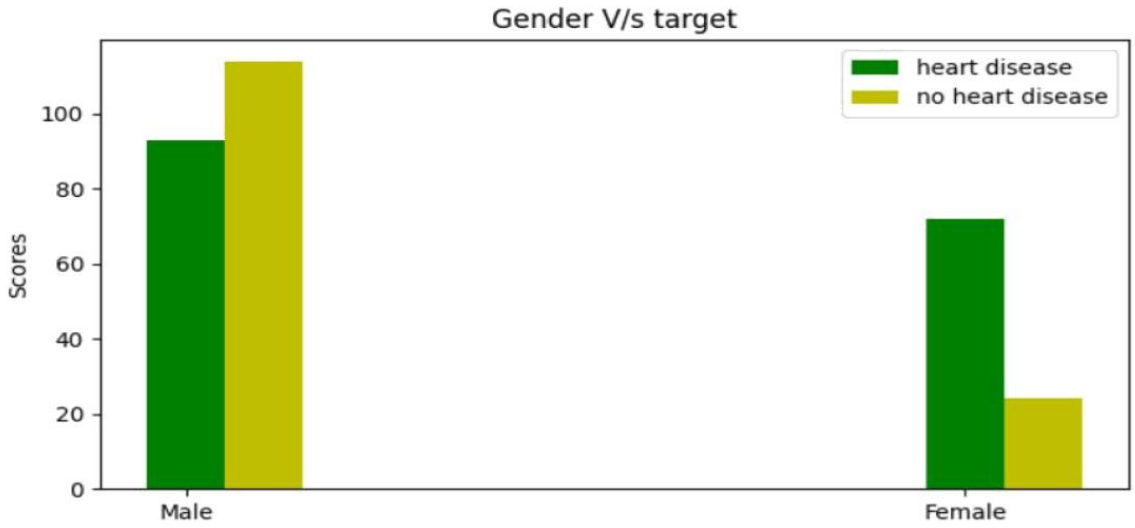


Figure: Data Balancing

3.2.5 Prediction of Disease

Different classification algorithms, including SVM, Naive Bayes, Decision Tree, Random Tree, and Logistic Regression, are utilized in machine learning for predicting heart disease. A comparative evaluation is conducted among these algorithms to determine the one that yields the highest accuracy in heart disease prediction.

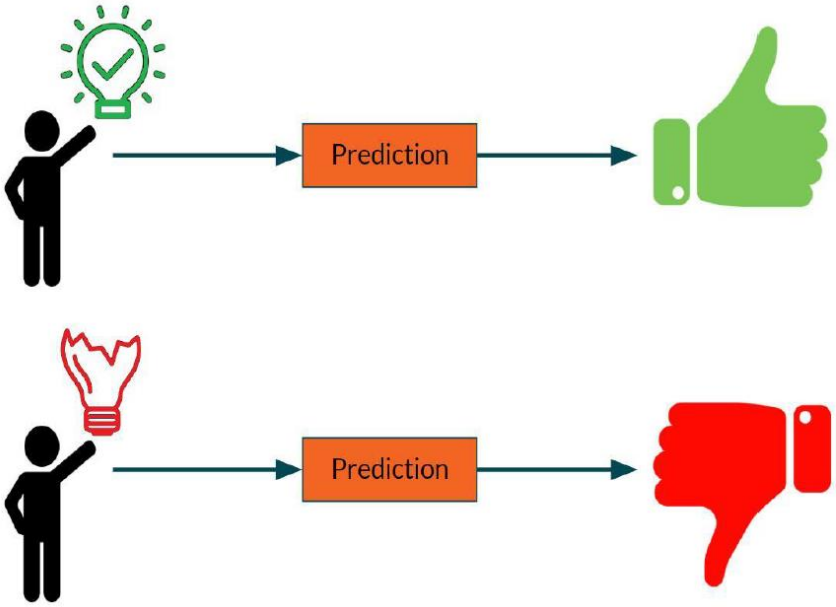


Figure: Prediction of Disease

CHAPTER 4
WORKING OF SYSTEM

4.1 System Architecture

The system architecture gives an overview of the working of the system.

The operational procedure of this system can be outlined as follows:
The initial step involves gathering a dataset comprising patient information. Next, a process is carried out to select the relevant attributes that are useful for predicting heart disease. Once the available data resources are identified, they undergo a selection and cleaning process to be transformed into the required format. Subsequently, various classification techniques mentioned earlier are applied to the preprocessed data in order to determine the accuracy of heart disease prediction. Finally, an accuracy measure is utilized to compare the performance of different classifiers.

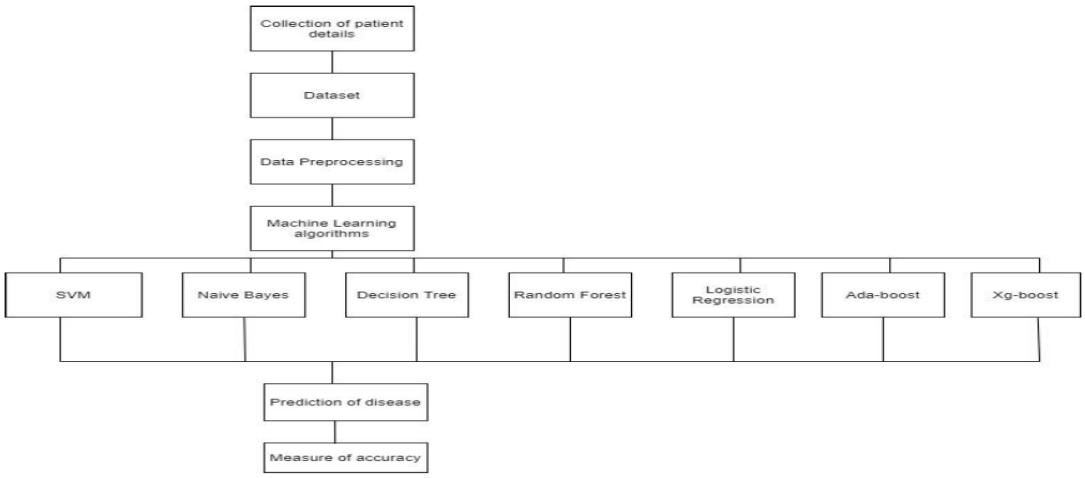


Figure:. SYSTEM ARCHITECTURE

4.2 Machine Learning

In the field of machine learning, the task of classification involves making predictions about the class label associated with a given input data instance.

- **Supervised Learning:**

Supervised learning refers to a type of machine learning where machines are trained using labeled training data, enabling them to predict outputs based on this information. Labeled data signifies input data that has been associated with the

correct output. In supervised learning, the provided training data acts as a supervisor, instructing the machines on how to accurately predict the output. This process parallels the way a student learns under the guidance of a teacher. The objective of a supervised learning algorithm is to discover a mapping function that establishes a relationship between the input variable (x) and the output variable (y), utilizing both input and correct output data provided to the machine learning model.

- **Unsupervised learning:**
  Unsupervised learning cannot be directly applied to regression or classification tasks because it lacks corresponding output data, unlike supervised learning. Instead, its objective is to uncover the inherent structure of a dataset, group similar data points, and represent the dataset in a condensed form.

  i) Unsupervised learning is valuable for extracting meaningful insights from data.
  ii) Unsupervised learning mimics the way humans learn through personal experiences, making it closer to true AI.
  iii) Unsupervised learning operates on unlabeled and uncategorized data, emphasizing its significance.
  iv) Unsupervised learning is necessary for situations where input data and corresponding outputs are unavailable in real-world scenarios.

- **Reinforcement Learning:**
  Reinforcement learning, a branch of Machine Learning, focuses on selecting optimal actions to maximize rewards within a given context. This approach is utilized by different software and machines to determine the most favorable behavior or path in specific situations. Unlike supervised learning, where the training data provides explicit answers, reinforcement learning relies on the reinforcement agent's decision-making to accomplish a given task without predetermined solutions. As a result, in the absence of training datasets, the agent learns through its own experiences.

4.3 Algorithms

An AI system relies on a machine learning algorithm to perform its task, typically making predictions based on input data. Classification and regression are the primary techniques employed by machine learning algorithms.

- **Decision Tree Algorithm:**
  The technique utilized by an AI system to perform its task is known as a machine learning algorithm. Typically, these algorithms predict output values based on provided input data. Machine learning algorithms primarily encompass two fundamental procedures: classification and regression.

  Decision nodes are utilized to make determinations and exhibit numerous branches, while Leaf nodes are the resulting outcomes of those determinations and do not possess additional branches. The determinations or tests are executed depending on characteristics found in the provided dataset. This form of graphical representation offers a comprehensive array of potential solutions for a problem or decision based on given conditions. It acquires the name "Decision Tree" due to its resemblance to a tree structure, commencing with a root node that extends into further branches. The construction of the tree involves the implementation of the CART algorithm, which stands for Classification and Regression Tree algorithm. Essentially, a Decision Tree poses a question and proceeds to split the tree into subtrees based on the response (Yes/No).

  The Decision Tree Algorithm is a type of supervised machine learning algorithm that can be utilized for both classification and regression tasks.

  The objective of this algorithm is to develop a predictive model for determining the value of a target variable. To solve the problem, a decision tree is employed, utilizing a tree structure where class labels are assigned to leaf nodes, and attributes are represented within the internal nodes of the tree.

  The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

- Decision Trees usually mimic human thinking ability while deciding, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

The primary difficulty in Decision Tree lies in determining the attribute for the root node at each level, which is referred to as attribute selection. There are two well-known measures for selecting attributes.

- **Information Gain:**
  The utilization of a node in a Decision Tree for dividing the training instances into smaller subsets results in a modification of entropy. Information gain quantifies the extent of this entropy alteration. Entropy represents the level of uncertainty associated with a random variable and defines the impurity of a given set of examples. A greater entropy value indicates a higher amount of information contained within.

- **Gini Index:**
  The Gini Index is a measurement used to assess the likelihood of misidentifying a randomly selected element. It indicates that attributes with lower Gini Index values are more desirable. Sklearn includes support for the "Gini" criteria in calculating the Gini Index, with the default value being set to "Gini".

The most notable types of Decision Tree algorithms are: -

- **IDichotomiser 3 (ID3):**
  This algorithm employs the concept of Information Gain to determine the most appropriate attribute for classifying the present subset of data. At each level of the tree, the information gain is computed iteratively for the remaining data.

- **C4.5:**
  The algorithm in question is a follow-up to the ID3 algorithm, and it employs either Information gain or Gain ratio to determine the attribute for classification. It represents a significant advancement over the ID3 algorithm, as it has the capability to manage continuous and missing attribute values.

- **Classification and Regression Tree (CART):**
  It is a dynamic learning algorithm which can produce a regression tree as well as a classification tree depending upon the dependent variable.

- **Working:**

When utilizing a Decision Tree to predict the class of a given dataset, the process begins at the root node of the tree. This involves comparing the attribute values of the root with the corresponding attributes of the dataset, and subsequently proceeding to the appropriate branch and advancing to the next node based on the outcome of the comparison.

This procedure is then repeated for each subsequent node, with the algorithm comparing attribute values to determine the path to follow. This iterative process continues until the algorithm reaches a leaf node in the tree. To gain a clearer understanding of the entire process, refer to the algorithm provided below.

Step 1: Start constructing the tree by establishing a root node called S, which encompasses the entire dataset.
Step 2: Determine the most suitable attribute from the dataset using an Attribute Selection Measure (ASM).
Step 3: Partition S into subsets based on the potential values associated with the selected attribute.
Step 4: Formulate a Decision Tree node that incorporates the chosen attribute.
Step 5: Repeat the procedure recursively for each subset generated in step 3, generating new decision trees. Continue this iterative process until reaching a stage where further classification of the nodes is not possible, designating the final node as a leaf node.

- **Logistic Regression Algorithm:**
  Logistic regression is a widely employed algorithm in Machine Learning that falls under the Supervised Learning category. Its purpose is to forecast the outcome of a categorical dependent variable by utilizing a specified group of independent variables. By employing logistic regression, the algorithm produces probabilistic values ranging between 0 and 1 instead of precise 0s and 1s, allowing for predictions of categorical outcomes such as Yes or No, 0 or 1, true or false, and similar.

  Logistic regression is similar to linear regression, but their applications differ. While linear regression is employed to address regression problems, logistic regression is utilized for classification problems. In logistic regression, we don't fit a regression line; instead, we fit a sigmoid or "S" shaped logistic function that predicts either 0 or 1 as the maximum values.

The logistic function's curve portrays the probability of various outcomes, such as determining if cells are cancerous or not, if a mouse is obese or not based on its weight, and so on. Logistic Regression holds immense importance in the field of machine learning as it enables the calculation of probabilities and facilitates the classification of new data using both continuous and discrete datasets.

- **Advantages:**
  Logistic Regression stands out as a straightforward machine learning technique that is both simple to implement and capable of delivering efficient training in certain scenarios. Moreover, the algorithm's effectiveness eliminates the need for substantial computational resources when training a model.

  The inferred parameters (trained weights) provide insight into the significance of each feature and also indicate the direction of the association, whether positive or negative. Therefore, Logistic Regression can be employed to determine the connection between the features.

  The ability of this algorithm to incorporate new data and update models sets it apart from Decision Tree or Support Vector Machine, with stochastic gradient descent being employed for the updating process.

  Logistic Regression offers the benefit of providing both well-calibrated probabilities and classification outcomes. This distinguishes it from models that solely produce final classifications. By considering the probabilities assigned to each class, we can determine the relative accuracy of training examples for the specific problem at hand. For instance, if one training example has a 95% probability for a certain class, while another has a 55% probability for the same class, we can make an inference about their respective accuracy levels.

- **Disadvantages:**
  Logistic Regression is a statistical model used for analyzing data and making predictions based on independent variables. However, when dealing with datasets that have a high number of dimensions, there is a risk of overfitting the model to the training data. Overfitting occurs when the model becomes too specialized to the training set, resulting in inflated accuracy during

prediction. Consequently, the model may struggle to provide accurate results when applied to the test set. This issue commonly arises when training data is limited, but there are numerous features. To address overfitting in high-dimensional datasets, it is advisable to employ regularization techniques. However, it is worth noting that regularization can increase the complexity of the model. Excessive regularization factors may even lead to the model underfitting the training data.

Logistic regression is limited in its ability to solve non-linear problems due to its reliance on a linear decision surface. In practical situations, it is uncommon to encounter data that is easily separable in a linear manner. To address this, transforming non-linear features becomes necessary. One approach is to increase the number of features, enabling the data to become linearly separable in higher dimensions.

Non-Linearly Separable Data: Logistic regression struggles to capture intricate relationships, making it challenging. However, more advanced and intricate algorithms like Neural Networks can surpass the performance of this method effortlessly.

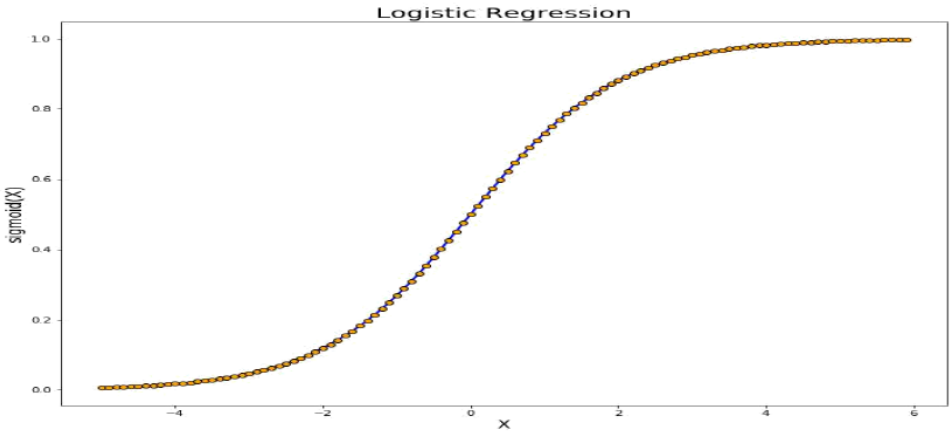


Figure: Logistic Regression

- **KNN (K-Nearest Neighbor):**

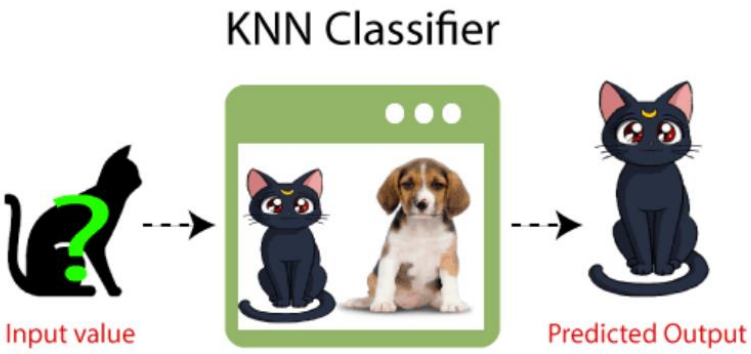
  The K-Nearest Neighbor (K-NN) algorithm is a straightforward Supervised Learning technique widely used in Machine Learning. It operates under the assumption of similarity between new and existing data points, assigning the new case to the category that closely matches the available categories. By

retaining all available data, the K-NN algorithm can effectively classify new data points into appropriate categories.

K-NN is a versatile algorithm applicable to both Regression and Classification tasks, although it is primarily employed for the latter. Being a non-parametric algorithm, K-NN does not rely on assumptions about the underlying data distribution. This characteristic has earned it the nickname of a "lazy learner" since it postpones learning from the training set. Instead, the algorithm stores the dataset and takes action when classifying new data points.

During the training phase, the K-NN algorithm simply stores the dataset. When encountering new data, it classifies it into the category that closely resembles the new data, based on similarity.

Example: Consider the following scenario: We possess an image of a creature bearing resemblance to both a cat and a dog, and our objective is to determine whether it belongs to the cat or dog category. In this situation, the KNN algorithm can be employed for identification purposes, as it operates on a measure of similarity. By utilizing the KNN model, we can discover the shared characteristics between the newly acquired dataset and the images of cats and dogs. Subsequently, based on the most analogous features, the algorithm will assign the creature to either the cat or dog classification.



Working:

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors.

Step-2: Calculate the Euclidean distance of **K number of neighbors**

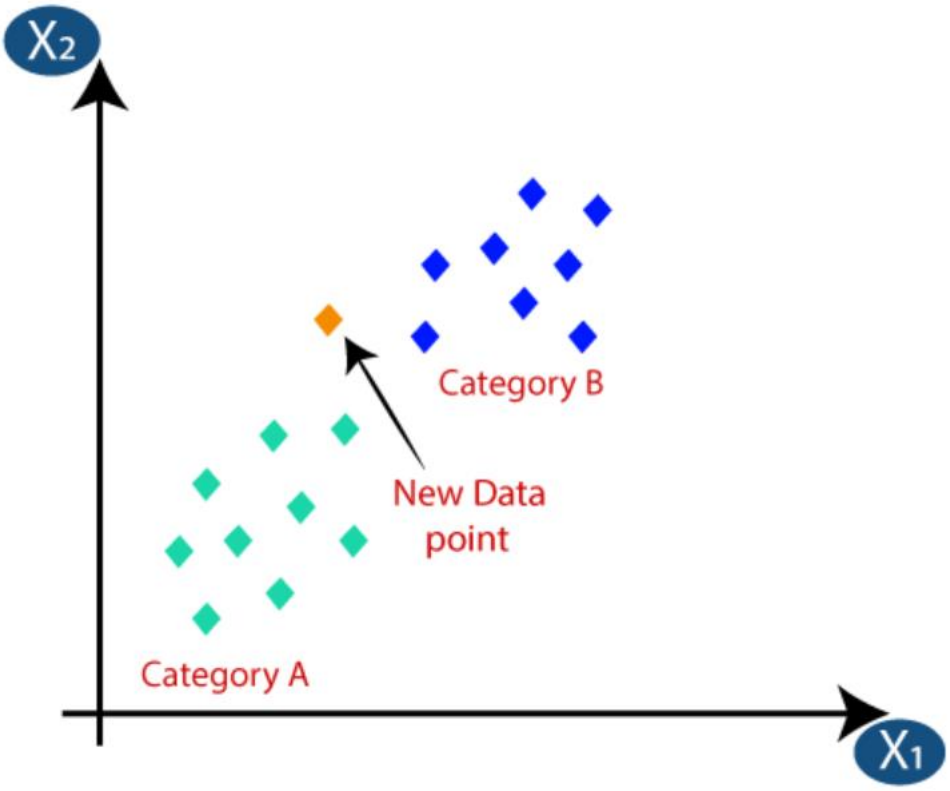
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

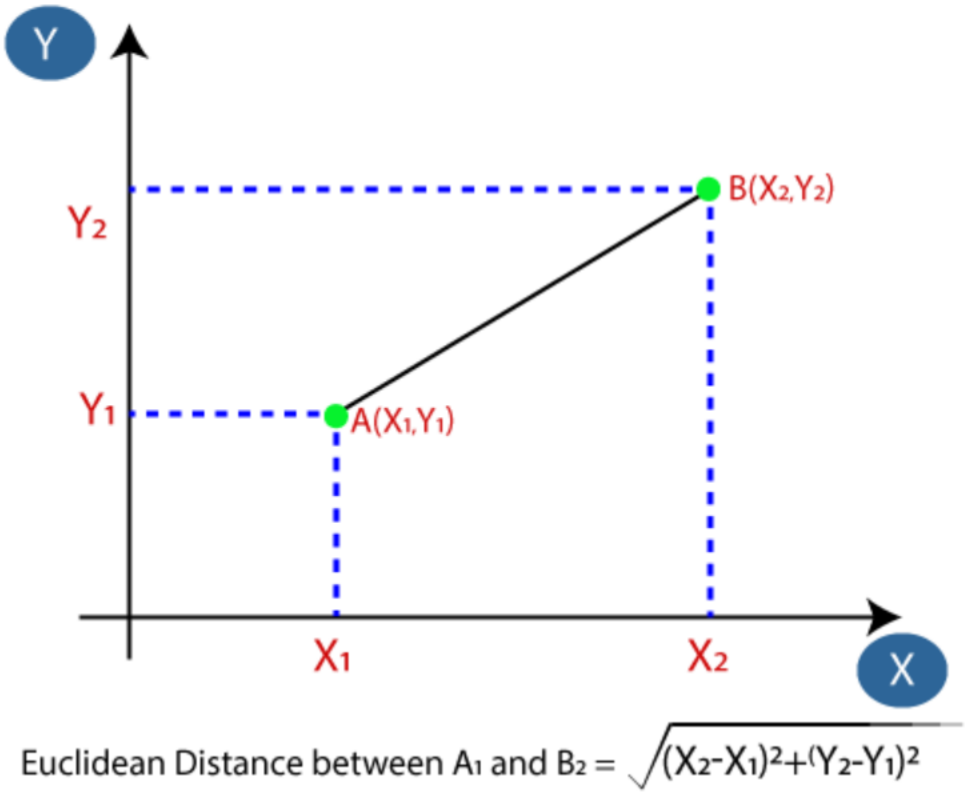
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

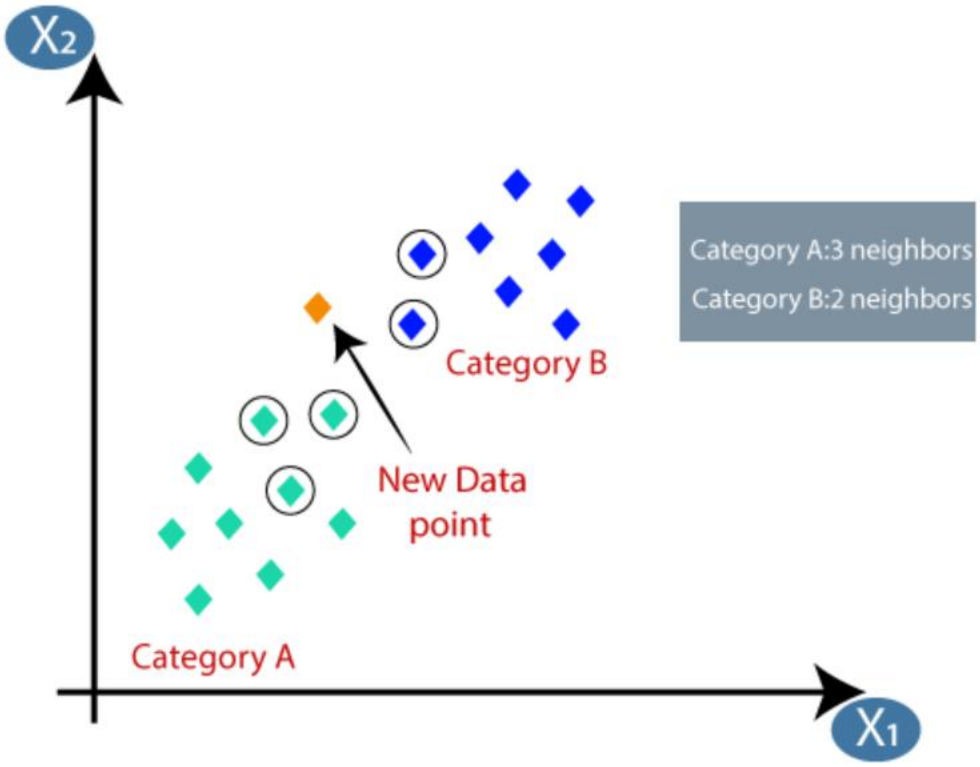
Suppose we have a new data point, and we need to put it in the required category. Consider the below image:



- Firstly, we will choose the number of neighbors, so we will choose the k=5.
- Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



- By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



- As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

ADVANTAGES OF KNN ALGORITHM:

- It is simple to implement.
- It is robust to the noisy training data.
- It can be more effective if the training data is large.

DISADVANTAGES OF KNN ALGORITHM:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

4.4 Code Implementation:

- **Import Libraries**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msn
```

- **Read CSV file**

```
df=pd.read_csv("heart.csv");
df.head()
df.tail()
df.shape

print('Sex', df['Sex'].unique())
print('ChestPainType', df['ChestPainType'].unique())
print('RestingECG', df['RestingECG'].unique())
print('ExerciseAngina', df['ExerciseAngina'].unique())
print('ST_Slope', df['ST_Slope'].unique())

sex_map = {'M':1, 'F':2}
chest_pain_type_map = {'ATA':1, 'NAP':2, 'ASY':3, 'TA':4 }
resting_ecg_map = {'Normal':1, 'ST':2, 'LVH':3}
excercise_angina_map = {'N':1, 'Y':2}
st_slop_map = {'Up':1, 'Flat':2, 'Down':3 }

df['Sex']=df['Sex'].map(sex_map)
df['ChestPainType']=df['ChestPainType'].map(chest_pain_type_map)
df['RestingECG']=df['RestingECG'].map(resting_ecg_map)
df['ExerciseAngina']=df['ExerciseAngina'].map(excercise_angina_map)
df['ST_Slope']=df['ST_Slope'].map(st_slop_map)

df.head()
```

```
sns.pairplot(df);

fig,ax=plt.subplots(figsize=(8,4));
sns.lineplot(data=df['Cholesterol']);

fig,ax=plt.subplots(figsize=(8,4));
sns.lineplot(data=df['Age']);

df.columns

X=df[['Age', 'Sex', 'ChestPainType', 'RestingBP', 'Cholesterol', 'FastingBS',
       'RestingECG', 'MaxHR', 'ExerciseAngina', 'Oldpeak', 'ST_Slope']]
X.head()
X = X.values

y=df['HeartDisease']
y.head()
y = y.values
```

- **Splitting Dataset into training and testing by train_test_split**

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_stat
e=22)
```

1. **Logistic Regression Model**

```
from sklearn.linear_model import LogisticRegression
model_lgr=LogisticRegression()
model_lgr.fit(X_train,y_train)

y_predict_lgr=model_lgr.predict(X_test)
```

- **Model Evaluation**

```
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
print("Accuracy of the model by Logistic Regression is :",accuracy_score(y_test,y_predict_lgr)*100)
print(confusion_matrix(y_test,y_predict_lgr))

print(classification_report(y_test,y_predict_lgr))
```

2. **KNN**

```
from sklearn.neighbors import KNeighborsClassifier
model_knn=KNeighborsClassifier()

model_knn.fit(X_train,y_train)

y_predict_knn=model_knn.predict(X_test)
```

- **Model Evaluation**

```
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
print("Accuracy of the model by KNN is :",accuracy_score(y_test,y_predict_knn)*100)

print(confusion_matrix(y_test,y_predict_knn))

print(classification_report(y_test,y_predict_knn))
```

3. **Decision Tree**

```
from sklearn import tree

model_tree = tree.DecisionTreeClassifier()  # empty model of the decision tree
```

```
model_tree.fit(X_train,y_train)

y_predict_tree=model_tree.predict(X_test)
```

- **Model Evaluation**

```
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
print("Accuracy of the model by Decision Tree is : ",accuracy_score(y_test,y_predict_tree)*100)

print(confusion_matrix(y_test,y_predict_tree))

print(classification_report(y_test,y_predict_tree))
```

- **Implementing Decision Tree, KNN & Logistic Regression**

```
import ipywidgets as widgets
import ipywidgets
from ipywidgets import interact

selected_algo = None
def onSelectAlgoritham(x):
    global selected_algo
    selected_algo = x

selected_sex = None
def onSelectSex(x):
    global selected_sex
    selected_sex = x

selected_chest_pain_type = None
def onSelectChestPainType(x):
    global selected_chest_pain_type
    selected_chest_pain_type = x
```

```
selected_resting_ecg = None
def onSelectRestingECG(x):
    global selected_resting_ecg
    selected_resting_ecg = x

selected_exercise_angina = None
def onSelectExerciseAngina(x):
    global selected_exercise_angina
    selected_exercise_angina = x


selected_st_slope = None
def onSelectST_Slope(x):
    global selected_st_slope
    selected_st_slope = x

interact(onSelectAlgoritham,
x=ipywidgets.Combobox(options=["LogisticRegression",
"KNeighborsClassifier", "DecisionTreeClassifier"]));

input_age=float(input("Enter age :-"))
interact(onSelectSex, x=ipywidgets.Combobox(options=['Male', 'Female']));
interact(onSelectChestPainType,   x=ipywidgets.Combobox(options=['ATA',
'NAP', 'ASY', 'TA']));
input_resting_bp=float(input("Enter RestingBP : "))
input_cholesterol=float(input("Enter Cholesterol : "))
input_fasting_bs=float(input("Enter FastingBS : "))
interact(onSelectRestingECG,   x=ipywidgets.Combobox(options=['Normal',
'ST', 'LVH']));
input_max_hr=float(input("Enter Max HR : "))
interact(onSelectExerciseAngina,    x=ipywidgets.Combobox(options=['Yes',
'No']));
input_oldpeak = float(input("Enter Oldpeak : "))
```

```
interact(onSelectST_Slope, x=ipywidgets.Combobox(options=['Up', 'Flat', 'Down']));

print(selected_algo)
print(input_age)
print(selected_sex)
print(selected_chest_pain_type)
print(input_resting_bp)
print(input_cholesterol)
print(input_fasting_bs)
print(selected_resting_ecg)
print(input_max_hr)
print(selected_exercise_angina)
print(input_oldpeak)
print(selected_st_slope)

sex_int = sex_map[selected_sex[0]]
sex_int = float(sex_int)

chest_pain_type_int = chest_pain_type_map[selected_chest_pain_type]
chest_pain_type_int = float(chest_pain_type_int)

resting_ecg_int = resting_ecg_map[selected_resting_ecg]
resting_ecg_int = float(resting_ecg_int)

selected_exercise_int = excercise_angina_map[selected_exercise_angina[0]]
selected_exercise_int = float(selected_exercise_int)

selected_st_slope_int = st_slop_map[selected_st_slope]
selected_st_slope_int = float(selected_st_slope_int)

if selected_algo == "LogisticRegression":
   result = model_lgr.predict([[input_age, sex_int, chest_pain_type_int, input_resting_bp
```

```
                        , input_cholesterol, input_fasting_bs, resting_ecg_int,
input_max_hr,
                            selected_exercise_int,          input_oldpeak,
selected_st_slope_int]])

    print("LogisticRegression:")
    print(result)
    if result[0]==0:
        print("Heart disease not deteacted!")
    elif result[0]==1:
        print("Heart disease deteacted!")

elif selected_algo == "KNeighborsClassifier":
    result = model_knn.predict([[input_age, sex_int, chest_pain_type_int,
input_resting_bp
                        , input_cholesterol, input_fasting_bs, resting_ecg_int,
input_max_hr,
                            selected_exercise_int,          input_oldpeak,
selected_st_slope_int]])
    print("KNeighborsClassifier:")
    print(result)
    if result[0]==0:
        print("Heart disease not deteacted!")
    elif result[0]==1:
        print("Heart disease deteacted!")

elif selected_algo == "DecisionTreeClassifier":
    result = model_tree.predict([[input_age, sex_int, chest_pain_type_int,
input_resting_bp
                        , input_cholesterol, input_fasting_bs, resting_ecg_int,
input_max_hr,
                            selected_exercise_int,          input_oldpeak,
selected_st_slope_int]])
    print("DecisionTreeClassifier:")
```

```
print(result)
if result[0]==0:
    print("Heart disease not deteacted!")
elif result[0]==1:
    print("Heart disease deteacted!")
```

CHAPTER 5

EXPERIMENTAL ANALYSIS

5.1 SYSTEM CONFIGURATION

5.1.1 Hardware requirements:

Processer	:	Any Update Processer
Ram	:	Min 4GB
Hard Disk	:	Min 100GB

5.1.2 Software requirements:

Operating System	:	Windows family
Technology	:	Python3.7
IDE	:	Jupiter notebook

5.2 Dataset Details:

```
Age,Sex,ChestPainType,RestingBP,Cholesterol,FastingBS,RestingECG,MaxHR,ExerciseAngina,Oldpeak,ST_Slope,HeartDisease
40,M,ATA,140,289,0,Normal,172,N,0,Up,0
49,F,NAP,160,180,0,Normal,156,N,1,Flat,1
37,M,ATA,130,283,0,ST,98,N,0,Up,0
48,F,ASY,138,214,0,Normal,108,Y,1.5,Flat,1
54,M,NAP,150,195,0,Normal,122,N,0,Up,0
39,M,NAP,120,339,0,Normal,170,N,0,Up,0
45,F,ATA,130,237,0,Normal,170,N,0,Up,0
54,M,ATA,110,208,0,Normal,142,N,0,Up,0
37,M,ASY,140,207,0,Normal,130,Y,1.5,Flat,1
48,F,ATA,120,284,0,Normal,120,N,0,Up,0
37,F,NAP,130,211,0,Normal,142,N,0,Up,0
58,M,ATA,136,164,0,ST,99,Y,2,Flat,1
39,M,ATA,120,204,0,Normal,145,N,0,Up,0
49,M,ASY,140,234,0,Normal,140,Y,1,Flat,1
42,F,NAP,115,211,0,ST,137,N,0,Up,0
54,F,ATA,120,273,0,Normal,150,N,1.5,Flat,0
38,M,ASY,110,196,0,Normal,166,N,0,Flat,1
43,F,ATA,120,201,0,Normal,165,N,0,Up,0
60,M,ASY,100,248,0,Normal,125,N,1,Flat,1
36,M,ATA,120,267,0,Normal,160,N,3,Flat,1
43,F,TA,100,223,0,Normal,142,N,0,Up,0
44,M,ATA,120,184,0,Normal,142,N,1,Flat,0
49,F,ATA,124,201,0,Normal,164,N,0,Up,0
44,M,ATA,150,288,0,Normal,150,Y,3,Flat,1
40,M,NAP,130,215,0,Normal,138,N,0,Up,0

36,M,NAP,130,209,0,Normal,178,N,0,Up,0
53,M,ASY,124,260,0,ST,112,Y,3,Flat,0
52,M,ATA,120,284,0,Normal,118,N,0,Up,0
53,F,ATA,113,468,0,Normal,127,N,0,Up,0
51,M,ATA,125,188,0,Normal,145,N,0,Up,0
53,M,NAP,145,518,0,Normal,130,N,0,Flat,1
56,M,NAP,130,167,0,Normal,114,N,0,Up,0
54,M,ASY,125,224,0,Normal,122,N,2,Flat,1
41,M,ASY,130,172,0,ST,130,N,2,Flat,1
43,F,ATA,150,186,0,Normal,154,N,0,Up,0
32,M,ATA,125,254,0,Normal,155,N,0,Up,0
65,M,ASY,140,306,1,Normal,87,Y,1.5,Flat,1
41,F,ATA,110,250,0,ST,142,N,0,Up,0
48,F,ATA,120,177,1,ST,148,N,0,Up,0
48,F,ASY,150,227,0,Normal,130,Y,1,Flat,0
54,F,ATA,150,230,0,Normal,130,N,0,Up,0
54,F,NAP,130,294,0,ST,100,Y,0,Flat,1
35,M,ATA,150,264,0,Normal,168,N,0,Up,0
52,M,NAP,140,259,0,ST,170,N,0,Up,0
43,M,ASY,120,175,0,Normal,120,Y,1,Flat,1
59,M,NAP,130,318,0,Normal,120,Y,1,Flat,0
37,M,ASY,120,223,0,Normal,168,N,0,Up,0
50,M,ATA,140,216,0,Normal,170,N,0,Up,0
36,M,NAP,112,340,0,Normal,184,N,1,Flat,0
41,M,ASY,110,289,0,Normal,170,N,0,Flat,1
50,M,ASY,130,233,0,Normal,121,Y,2,Flat,1
```

Input Dataset Attributes:

- Age in year
- Sex: value (M: Male, F: Female)

- Chest Pain Type
- Resting BP: Resting Blood Pressure
- Cholesterol
- FastingBS: Fasting Blood Sugar value (1: >120md/dl, 0: <120mg/dl)
- RestingECG: Resting electrocardiograph can detect coronary heart diseases.
- MaxHR: Maximum Heart Rate
- Exercise Angina
- OldPeak
- ST_Slope: The ST segment shift relative to exercise-induced increments in heart rate, the ST/heart rate slope, has been proposed as a more accurate ECG criterion for diagnosing significant coronary artery disease (CAD).

S. No.	Attributes	Description	Type
1.	Age	Patient's age (29 to 77)	Numerical
2.	Sex	Gender of patient	Nominal
3.	Chest Pain Type	Chest pain type	Nominal
4.	RestingBP	Resting blood pressure	Numerical
5.	Cholesterol	Serum cholesterol in mg/dl	Numerical
6.	FastingBS	Fasting blood sugar	Numerical
7.	RestingECG	Resting electrocardiograph	Nominal
8.	MaxHR	Maximum Heart Rate	Numerical
9.	Exercise Angina		Nominal
10.	OldPeak		Numerical
11.	ST_Slope	ST segment shift	Nominal

Attributes of the Dataset

5.3 Performance Analysis

In this study, heart disease prediction is conducted using different machine learning algorithms including Decision Tree, Logistic Regression, and KNN. The Heart Disease UCI dataset contains a total of 76 attributes, but only 11 attributes are taken into account for predicting heart disease. Patient attributes such as gender, chest pain type, fasting blood pressure, serum cholesterol, and exang are among the factors considered. Each algorithm's accuracy is calculated, and the algorithm with the highest accuracy is chosen for heart disease prediction. To evaluate the experiment, several metrics including accuracy, confusion matrix, precision, recall, and f1-score are utilized.

Accuracy is defined as the proportion of correct predictions to the total number of inputs in the dataset, and it can be calculated as follows:

Accuracy = (TP + TN) / (TP + FP + FN + TN)

Confusion Matrix: - It gives us a matrix as output and gives the total performance of the system.

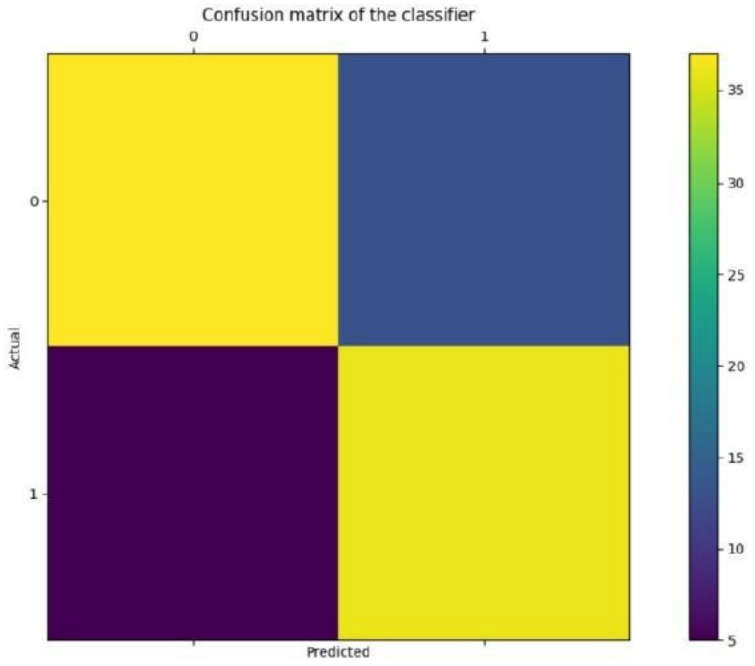


Figure: Confusion Matrix

where
TP: True positive
FP: False Positive
FN: False Negative
TN: True Negative

Correlation Matrix: - The correlation matrix in machine learning is used for feature selection. It represents dependency between various attributes.

Precision: - It is the ratio of correct positive results to the total number of positive results predicted by the system.
It is expressed as:

- Recall-It is the ratio of correct positive results to the total number of positive results.
- predicted by the system.
- It is expressed as:
- F1 Score-It is the harmonic mean of Precision and Recall. It measures the test accuracy. The range of this metric is 0 to 1.

5.4 Performance Measures

	Precision	**recall**	**F1-Score**	**Support**
0	0.80	0.84	0.82	122
1	0.87	0.84	0.85	154
accuracy			0.84	276
Micro avg.	0.83	0.84	0.84	276
Weighted avg	0.84	0.84	0.84	276

Logistic Regression Model Evaluation

	Precision	recall	F1-Score	Support
0	0.65	0.57	0.61	122
1	0.69	0.76	0.72	154
accuracy			0.68	276
Micro avg.	0.67	0.67	0.67	276
Weighted avg	0.68	0.68	0.67	276

KNN Model Evaluation

97	25
34	120

Decision Tree Confusion Matrix

	Precision	recall	F1-Score	Support
0	0.74	0.80	0.77	122
1	0.83	0.78	0.80	154
accuracy			0.79	276
Micro avg.	0.78	0.79	0.78	276
Weighted avg	0.79	0.79	0.79	276

Decision Tree Model Evaluation

- **The highest accuracy is given by Logistic Regression Model.**

5.5 Result

In comparison to other algorithms, we discover that the accuracy of the Logistic Regression is superior after applying the machine learning approach for training and testing. With the help of the confusion matrices for each algorithm, accuracy is calculated. Here, the number of TP, TN, FP, and FN is given. Using the equation for accuracy, value has been calculated. It is concluded that extreme gradient boosting is best with 84% accuracy.

TABLE: Accuracy comparison of algorithms Algorithm Accuracy

Algorithm	**Accuracy**
Logistic Regression	83.5%
KNN	67.5%
Decision Tree	78.5%

Accuracy Table

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

The utilization of promising technology, such as machine learning, in the initial prediction of heart problems could have a noteworthy societal impact, considering that heart diseases are a prominent cause of death both in India and worldwide. The early identification of cardiac ailments has the potential to assist high-risk individuals in making informed choices regarding lifestyle adjustments, representing a significant advancement in the field of medicine. Each year, an increasing number of individuals receive a diagnosis of cardiac illnesses, necessitating the need for early detection and appropriate intervention. The application of suitable technological support in this domain has the potential to greatly benefit both the medical community and patients. To assess performance, this study employed a range of machine learning algorithms, including SVM, Decision Tree, Random Forest, Naive Bayes, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting.

The dataset contains 76 features that pertain to various characteristics associated with heart disease in individuals. To evaluate the system, the author selects 14 significant features from this dataset. However, the author notices a decrease in system efficiency when considering all the features. In order to enhance efficiency, attribute selection is implemented. This involves choosing a specific number of characteristics (referred to as 'n') that yield higher accuracy for evaluating the model. Some features in the dataset exhibit nearly identical correlations, leading to their elimination. It is observed that system efficiency declines significantly when all attributes in the dataset are considered.

The accuracies of seven different machine learning methods were compared to determine the best prediction model. The objective was to employ various evaluation metrics, such as the confusion matrix, accuracy, precision, recall, and f1-score, to efficiently predict the disease. Among the seven methods, the extreme gradient boosting classifier yielded the highest accuracy of 84%.

APPENDIX

Python

Python, which was developed by Guido Van Rossum and made available in 1991, is a general-purpose programming language known for its interpretive nature and high-level capabilities. The language places a strong emphasis on code readability by utilizing significant whitespace. With its object-oriented methodology and various language constructs, Python aids programmers in producing coherent and well-structured code for projects of all sizes. Notably, Python is dynamically typed and features automatic garbage collection. It further facilitates diverse programming paradigms such as procedural, object-oriented, and functional programming.

Sklearn

Sklearn, also known as Scikit-learn, is an exceptionally valuable and reliable Python library extensively employed for machine learning purposes. It offers a comprehensive range of efficient resources for tasks such as classification, regression, clustering, and dimensionality reduction, all accessible through a unified Python interface. Primarily coded in Python, this library relies on the foundational frameworks of NumPy, SciPy, and Matplotlib.

Numpy

NumPy is a Python library that enhances the capabilities of the Python programming language by introducing extensive support for large, multi-dimensional arrays and matrices. It also provides a vast collection of high-level mathematical functions that enable efficient operations on these arrays. The precursor to NumPy, known as Numeric, was initially developed by Jim in collaboration with multiple other programmers. In 2005, Travis took the initiative to merge the features of the rival Num array with Numeric, resulting in the creation of NumPy with significant enhancements. Being open-source software, NumPy benefits from numerous contributors who actively contribute to its development and improvement.

Librosa

Librosa serves as a Python toolkit designed for music and audio analysis purposes. It proves particularly useful in various applications involving audio data, such as music generation with LSTMs and Automatic Speech Recognition. This package offers essential components for constructing music information retrieval systems. By employing diverse signal processing techniques, Librosa facilitates audio signal visualization and enables feature extraction.

Matplotlib

Matplotlib serves as a plotting library designed for Python and its numerical mathematics extension, NumPy. It offers an object-oriented interface that allows developers to incorporate plots into applications using various GUI toolkits such as Tkinter, wxPython, Qt, or GTK. Additionally, there exists a procedural "pylab" interface, which operates on a state machine similar to OpenGL, aiming to mimic MATLAB's interface, although its usage is not recommended.

Seaborn

Seaborn, a Python library for data visualization, is built upon matplotlib and closely integrated with pandas' data structures. It offers a user-friendly interface for creating visually appealing and informative statistical graphics. Seaborn primarily focuses on visualization, facilitating the exploration and comprehension of data.

SciPy

SciPy encompasses a range of modules catering to various scientific and engineering tasks, such as optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers, and more. In addition, there are dedicated conferences organized for both users and developers of these tools: SciPy (held in the United States), Euro SciPy (hosted in Europe), and SciPy.in (conducted in India). The original SciPy conference was initiated by Enthought in the United States and remains involved in sponsoring international conferences while also serving as the host for the SciPy website. Functioning as a scientific computation library, SciPy is built upon the foundation of NumPy and offers an expanded set of utility functions encompassing optimization, statistics, and signal processing.

REFERENCES

[1] Soni J, Ansari U, Sharma D, and Soni S (2011) provide an overview of heart disease prediction in their paper titled "Predictive data mining for medical diagnosis: an overview of heart disease prediction" published in the International Journal of Computer Applications.

[2] Dangare C S and Apte S S (2012) present an improved study on heart disease prediction system using data mining classification techniques in their article titled "Improved study of heart disease prediction system using data mining classification techniques" published in the International Journal of Computer Applications.

[3] Ordonez C (2006) discusses association rule discovery for heart disease prediction using the train and test approach in the IEEE Transactions on Information Technology in Biomedicine article titled "Association rule discovery with the train and test approach for heart disease prediction".

[4] Shinde R, Arjun S, Patil P, and Waghmare J (2015) propose an intelligent heart disease prediction system utilizing k-means clustering and Naïve Bayes algorithm in their paper titled "An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm" published in the International Journal of Computer Science and Information Technologies.

[5] Bashir S, Qamar U, and Javed M Y (2014, November) present an ensemble-based decision support framework for intelligent heart disease diagnosis in their paper titled "An ensemble-based decision support framework for intelligent heart disease diagnosis" published in the International Conference on Information Society (i-Society 2014) proceedings.

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W, and Yun Y D (2014) propose a coronary heart disease prediction model based on the Korean Heart Study in their article titled "A coronary heart disease prediction model: the Korean Heart Study" published in BMJ open.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J, and Ingelsson E (2013) discuss multilogues genetic risk scores for coronary heart disease prediction in the Arteriosclerosis, thrombosis, and vascular biology journal article titled "Multilogues genetic risk scores for coronary heart disease prediction".

[8] Jabbar M A, Deekshatulu B L, and Chandra P (2013, March) propose heart disease prediction using lazy associative classification in their paper titled "Heart disease prediction using lazy associative classification" published in the 2013 International Mutli Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) proceedings.

[9] Brown N, Young T, Gray D, Skene A M, and Hampton J R (1997) analyze inpatient deaths from acute myocardial infarction in the Nottingham heart attack register in their article titled "Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register" published in BMJ.

[10] Folsom A R, Prineas R J, Kaye S A, and Soler J T (1989) investigate the association between body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women in their article titled "Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women" published in the International Journal of Epidemiology.

[11] Chen et al. (2011) developed a heart disease prediction system called HDPS, as presented in the 2011 Computing in Cardiology conference proceedings by the IEEE, with the reference citation details provided.

[12] Parthiban, Latha, and R Subramanian (2008) proposed an intelligent heart disease prediction system that employs CANFIS and genetic algorithm techniques, as published in the International Journal of Biological, Biomedical and Medical Sciences.

[13] Wolgast et al. (2016) introduced a wireless body area network designed for heart attack detection, as described in the IEEE antennas and propagation magazine's Education Corner section.

[14] Patel and Chauhan (2014) presented a heart attack detection and medical attention system that utilizes a motion sensing device called Kinect, as documented in the International Journal of Scientific and Research Publications.

[15] Piller et al. (2002) conducted a validation study on heart failure events in participants assigned to doxazosin and chlorthalidone within the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT), as reported in the Current Controlled Trials in Cardiovascular Medicine.

[16] Raihan et al. (2016) proposed a prototype design for a smartphone-based ischemic heart disease risk prediction system, utilizing clinical data and data mining approaches. Their work was presented in the 2016 19th International Conference on Computer and Information Technology (ICCIT) by the IEEE.

[17] Aldallal and Al-Moosa (2018) employed data mining techniques to predict diabetes and heart diseases, as discussed in the 2018 4th International Conference on Frontiers of Signal Processing (ICFSP).

[18] Takci (2018) investigated the enhancement of heart attack prediction through feature selection methods, as published in the Turkish Journal of Electrical Engineering & Computer Sciences.

[19] Dewan and Sharma (2015) proposed a hybrid technique in data mining classification for heart disease prediction, as presented in the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).

[20] Methaila et al. (2014) conducted a study on early heart disease prediction using data mining techniques, as published in the Computer Science & Information Technology Journal.

Research Paper

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, March 2023
Impact Factor: 7.301

Heart Disease Detection using Machine Learning

Pushpendra Tyagi[1], Devesh Chandra[2], Dhirendra Yadav[3], Dharmendra Singh Yadav[4], Utkarsh Singh[5]
Assistance Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3,5]
Student, Department of Mechanical Engineering[4]
KIET Groups of Institution, Ghaziabad, UP, India

Abstract: *Heart disease is the major reason of death in the entire world. Early detection and diagnosis of heart disease can benefits patient results and reduce mortality rates. ML is a powerful device for analyzing medical data and detecting patterns that may indicate heart disease. This paper presents a rundown of the use of ml algorithms for heart disease detection. The methods discussed include decision trees, SVM, and ANN. The performance of these algorithms was evaluated using various parameters, such as accuracy, sensitivity, and specificity. The results indicate that machine learning algorithms can provide accurate and efficient detection of heart disease. However, further research is needed to optimize these algorithms and address issues related to data quality and interpretability. Despite these challenges, machine learning has the potential to revolutionize the way heart disease is diagnosed and treated, improving patient outcomes and reducing mortality rates.*

Keywords: Neural Network, Machine Learning, Supervised learning, Support vector machine, Random Forest

I. INTRODUCTION

Heart disease is a prime global health concern and the major cause of death worldwide. Early detection and prompt intervention are critical for improving patient outcomes and reducing the burden of disease. From past few years, ml algorithms have been broadly used in medical diagnosis, including heart disease diagnosis. The objective of this study is to analyze and evaluate the current state of the art in the use of ml algorithms for heart disease detection.

The study focuses on the various approaches, datasets, and performance metrics used in the literature. Results indicate that machine learning algorithms can accurately diagnose heart disease with high sensitivity and specificity and can be a valuable tool for healthcare providers in the early detection of the condition. The most used algorithms in heart disease diagnosis include LR, decision trees, random forests, and SVM. These algorithms have applied to a variety of datasets, including demographic data, electrocardiogram signals, and imaging data.

One of the key challenges in heart disease detection using ml algorithms is the restricted presence of good quality and informative datasets. To address this challenge, some studies have proposed methods for synthesizing synthetic datasets or augmenting existing datasets to increase the size and diversity of the training data. Additionally, feature selection and feature engineering techniques have been used to identify the most relevant and informative features for heart disease diagnosis.

The performance of ml algorithms for heart disease diagnosis is typically evaluated using a range of parameters, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve. Results indicate that machine learning algorithms can achieve high levels of accuracy and AUC, with some studies reporting performance comparable to that of human experts.

The study concludes with a discussion of the restriction on current approaches and directions for future research. One of the important limitations is the limited generalizability of machine learning algorithms, as they are typically trained and evaluated on specific datasets and may not perform well on other datasets. Additionally, there is a need for further research to analyze the outcome of ml algorithms in real-world settings, with real patients and real data.

This study suggest that ml algorithms have the possibility to play an important role in the early detection of heart disease. With continued improvement in ml and the increasing availability of patient data, it is likely that these algorithms will continue to improve and become increasingly integrated into clinical practice. However, it is important to strictly consider the restriction of these algorithms and make sure that they are used in a fair and ethical manner.

DOI: 10.48175/568

Research Paper Certificates


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER
ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Dhirendra Yadav
KIET Groups of Institution, Ghaziabad, UP, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Heart Disease Detection using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 1, MARCH 2023
Certificate No: 032023-A151
www.ijarsct.co.in
Crossref
DOI: 10.48175/568
www.doi.org
www.crossref.org
7.301
Scientific Journal Impact Factor
Journal Impact Factor
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER
ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Dharmendra Singh Yadav
KIET Groups of Institution, Ghaziabad, UP, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Heart Disease Detection using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 1, MARCH 2023
Certificate No: 032023-A152
www.ijarsct.co.in
Crossref
DOI: 10.48175/568
www.doi.org
www.crossref.org
7.301
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER
ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Utkarsh Singh
KIET Groups of Institution, Ghaziabad, UP, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Heart Disease Detection using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 1, MARCH 2023
Certificate No: 032023-A153
www.ijarsct.co.in
Crossref
DOI: 10.48175/568
www.doi.org
www.crossref.org
7.301
www.sjifactor.com
Editor-in-Chief

Plagiarism Report Of Report Paper

turnitin

Similarity Report ID: oid:21301:36295087

● 18% Overall Similarity

Top sources found in the following databases:

- 12% Internet database
- Crossref database
- 17% Submitted Works database
- 5% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	cse.anits.edu.in Internet	7%
2	University of North Texas on 2022-10-08 Submitted works	2%
3	University of Essex on 2022-07-25 Submitted works	1%
4	irjmets.com Internet	1%
5	Meerut Institute of Engineering & Technology on 2023-05-14 Submitted works	<1%
6	ijraset.com Internet	<1%
7	eprajournals.com Internet	<1%
8	Queen Mary and Westfield College on 2023-03-14 Submitted works	<1%

Sources overview