ANALYTIC TOOL FOR HEALTHCARE DATA OF PATIENT'S HEART DISEASE PREDICTIONS IN HOSPITALS AND OTHER INSTITUTIONS USING MACHINE LEARNING

A PROJECT REPORT
submitted by

UDHAYAKUMAR G 2116210701293
VARUSHA S 2116210701304
SATHISHKUMAR R 2116210701515

in partial fulfillment for the award of the

degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Project report titled **"ANALYTIC TOOL FOR HEALTHCARE DATA OF PATIENT'S HEART DISEASE PREDICTIONS IN HOSPITALS AND OTHER INSTITUTIONS USING MACHINE LEARNING"** is the bonafide work of **"UDHAYAKUMAR G -2116210701293, VARUSHA S -2116210701304, SATHISHKUMAR R - 2116210701515"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr.P.SHANMUGAM,**M.Tech.,Ph.D**,

PROJECT COORDINATOR

Associate Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

Heart disease remains a leading cause of mortality worldwide, emphasizing the critical need for effective predictive tools to aid in early detection and intervention. This project introduces an analytic tool designed to analyze healthcare data related to heart disease and predict the likelihood of its occurrence in patients within hospitals and other healthcare institutions. Leveraging machine learning algorithms, the tool utilizes a comprehensive dataset encompassing various patient attributes, medical history, and diagnostic tests. Through a systematic approach, the tool processes and analyzes the data, identifying patterns and correlations that contribute to the prediction of heart disease risk. The predictive model is trained on historical patient data, continuously refined through iterative learning processes to enhance accuracy and reliability. By integrating advanced analytics with healthcare information systems, the tool offers healthcare professionals a valuable resource for proactive risk assessment and personalized patient care. Through early identification of high-risk individuals, healthcare providers can implement targeted interventions, optimize treatment strategies, and ultimately improve patient outcomes. This project represents a significant advancement in leveraging machine learning for predictive analytics in healthcare, with the potential to revolutionize the management of heart disease and reduce its societal burden.

ACKNOWLEDGMENT

We thank the almighty god for the successful completion of the project. Our sincere thanks to our chairman **Mr. S. MEGANATHAN,B.E., F.I.E.,** for his sincere endeavor in educating us in his premier institution. We would like to express our deep gratitude and sincere thanks to our beloved Chairperson **Dr.(Mrs).THANGAM MEGANATHAN,Ph.D.,**for her enthusiastic motivation which inspired us a lot in completing this project and we extend our gratitude to our Vice Chairman **Mr.ABHAY SHANKAR MEGANATHAN,B.E., M.S.,** for providing us with the requisite infrastructure.

We also express our sincere and warmful gratitude to our college Principal **Dr. S.N.MURUGESAN,M.E.,Ph.D.**,and **Dr. P. KUMAR,M.E.,Ph.D.**, **Director of Computing and Information Science** and **Head Of the Department of Computer Science and Engineering** and **Project coordinator Dr.P.SHANMUGAM,M.Tech.,Ph.D.**,for his encouragement and guiding us throughout project towards successful completion of project and to our parents, friends, all faculty members,supporting staff for direct and indirect involvement in successful completion of project for their encouragement and support.

UDHAYAKUMAR G

VARUSHA S

SATHISHKUMAR R

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	**iii**
	LIST OF FIGURES	**vii**
1.	**INTRODUCTION**	**1**
	1.1 PROBLEM STATEMENT	**1**
	1.2 SCOPE OF THE WORK	**2**
	1.3 AIM AND OBJECTIVES OF PROJECT	**2**
	1.4 RESOURCES	**4**
	1.5 MOTIVATION	**4**
2.	**LITERATURE REVIEW**	**5**
	2.1 SURVEY	**5**
	2.2 PROPOSED SYSTEM	**12**
3.	**SYSTEM DESIGN**	**14**
	3.1 GENERAL	**14**
	3.2 SYSTEM ARCHITECTURE DIAGRAM	**14**

3.3 DEVELOPMENT ENVIRONMENT **15**

3.3.1 HARDWARE REQUIREMENTS **15**

3.3.2 SOFTWARE REQUIREMENTS **15**

3.4 DESIGN OF ENTIRE SYSTEM **16**

3.4.1 SEQUENCE DIAGRAM **16**

4. PROJECT DESCRIPTION 13

4.1 METHODOLOGY **13**

4.2 MODULE DESCRIPTION **13**

5. RESULTS AND DISCUSSION 19

5.1 FINAL OUTPUT **14**

5.2 RESULT **19**

6. CONCLUSION & FUTURE WORK 24

6.1 CONCLUSION **20**

6.2 FUTURE ENHANCEMENT **20**

APPENDIX 21

REFERENCES 41

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
3.2.1	SYSTEM ARCHITECTURE	**14**
3.4.1	SEQUENCE DIAGRAM	**16**
5.1.1	PLANT DISEASE PREDICTION WEBSITE	**19**
5.1.2	OUTPUT	**20**
5.1.3	CONFUSION MATRIX	**21**
5.1.4	ACCURACY GRAPH	**22**

CHAPTER 1

INTRODUCTION

Our proposed model represents a pioneering initiative in the realm of healthcare analytics. Leveraging advanced machine learning algorithms, our tool aims to revolutionize the prediction and management of heart diseases among patients within hospitals and various healthcare institutions. By harnessing the power of data analytics, our project endeavors to offer healthcare professionals invaluable insights into early detection, risk assessment, and personalized treatment strategies, ultimately enhancing patient care and outcomes. With a focus on accuracy, efficiency, and user-friendliness, our innovative tool promises to redefine the landscape of cardiac healthcare delivery, ushering in a new era of proactive and data-driven patient management.

1.1 PROBLEM STATEMENT

The project aims to develop an advanced analytic tool tailored for healthcare settings, specifically focusing on predicting heart diseases in patients. Leveraging machine learning algorithms, the tool will analyze extensive datasets sourced from hospitals and other healthcare institutions. By harnessing the power of data analytics and predictive modeling, healthcare professionals will be equipped with valuable insights into identifying potential heart disease cases early on, facilitating timely interventions and personalized treatment plans. This tool holds the potential to revolutionize healthcare practices by enhancing diagnostic accuracy, optimizing

resource allocation, and ultimately improving patient outcomes in the realm of cardiovascular health.

1.2 SCOPE OF THE WORK

In the proposed model, the scope of the project encompasses the development of a sophisticated analytical platform tailored for the healthcare sector. This tool aims to leverage machine learning algorithms to analyze extensive datasets related to patients' cardiac health across various healthcare facilities. Key objectives include the identification of predictive patterns and risk factors associated with heart disease, facilitating early detection, prognosis, and personalized treatment strategies. Additionally, the tool intends to streamline data management processes, enhance decision-making for healthcare professionals, and ultimately improve patient outcomes. The scope also involves ensuring the tool's compatibility with existing hospital systems and adherence to regulatory standards for data privacy and security.

1.3 AIM AND OBJECTIVES OF THE PROJECT

The aim of the proposed system is to develop a robust and user-friendly software solution that leverages advanced machine learning algorithms to predict the likelihood of heart disease in patients. By integrating and analyzing vast amounts of patient data, including medical history, lifestyle factors, and diagnostic test results, the tool seeks to provide accurate and timely predictions that can assist healthcare professionals in making informed decisions. This predictive capability is intended to enhance early detection, improve patient outcomes, and optimize resource allocation within hospitals

and healthcare institutions. The project also emphasizes data security and patient privacy, ensuring compliance with relevant regulations while providing actionable insights to healthcare providers.

The heart diseases prediction in Cardiovascular diseases are diagnosed using an array of laboratory tests and imaging studies. The primary part of diagnosis is medical and family histories of the patient, risk factors, physical examination and co-ordination of these findings with the results from tests and procedures. Enable early detection of heart disease risk factors, allowing for timely intervention and prevention strategies. To provide various integrate data sources, including medical records, physiological measurements, lifestyle data and generic information.

1.4 RESOURCES

To develop an analytic tool for predicting heart disease using healthcare data with machine learning, a variety of resources are essential. First, access to comprehensive and high-quality datasets, such as the Framingham Heart Study or datasets from healthcare institutions, is crucial for training and validating machine learning models. Key software tools include Python and R, which offer robust libraries like TensorFlow, Scikit-learn, and Keras for building and fine-tuning algorithms. Data preprocessing and feature engineering will require tools for handling missing data, normalization, and encoding, often facilitated by libraries such as Pandas and NumPy. Additionally, cloud computing platforms like AWS,

Google Cloud, or Microsoft Azure provide the necessary infrastructure for handling large datasets and performing computationally intensive tasks.

1.5 MOTIVATION

The motivation for the proposed system stems from the critical need to enhance early detection and management of heart disease, a leading cause of mortality worldwide. By leveraging machine learning, this project aims to develop a robust predictive model that can analyze vast amounts of healthcare data to identify patterns and risk factors associated with heart disease. This tool will empower healthcare providers with accurate, data-driven insights, facilitating timely interventions, personalized treatment plans, and ultimately improving patient outcomes. Additionally, it seeks to reduce the burden on healthcare systems by enabling more efficient resource allocation and reducing the incidence of severe heart disease through proactive management.

CHAPTER 2

2.1 LITERATURE SURVEY

[1] **"Heart Disease Prediction using Machine Learning Techniques, 2020", by Devansh Shah, Samir Patel & Santosh Kumar Bharti** presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland

database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms.

[2] **"Heart disease prediction using machine learning algorithms, 2021" by Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath** focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease.

[3] **"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques , 2019" authors Senthilkumar Mohan; Chandrasegar Thirumalai; Gautam Srivastava** propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease.

[4] **"Heart Disease Prediction using Hybrid machine Learning Model, 2021" authors M. Kavitha; G. Gnaneswar; R. Dinesh; Y. Rohith Sai; R. Sai Suraj** propose a novel machine learning approach is proposed to predict heart disease. The proposed study used the Cleveland heart disease dataset, and data

mining techniques such as regression and classification are used. Machine learning techniques Random Forest and Decision Tree are applied.

[5] **"Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, 2021" authors Md Mamun Ali , Bikash Kumar Paul , Kawsar Ahmed , Francis M. Bui , Julian M.W. Quinn , Mohammad Ali Moni** aimed to identify machine learning classifiers with the highest accuracy for such diagnostic purposes. Several supervised machine-learning algorithms were applied and compared for performance and accuracy in heart disease prediction. Feature importance scores for each feature were estimated for all applied algorithms except MLP and KNN.

[6] **"Heart Disease Prediction Using Machine learning and Data Mining Technique, 2016" by Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel** compares different algorithms of Decision Tree classification seeking better performance in heart disease diagnosis using WEKA. The algorithms which are tested is J48 algorithm, Logistic model tree algorithm and Random Forest algorithm. The existing datasets of heart disease patients from Cleveland database of UCI repository is used to test and justify the performance of decision tree algorithms.

[7] **"Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis, 2020" authors Rahul Katarya & Sunit Kumar Meena** discusses the heart disease and its risk factors and explained

machine learning techniques. Using that machine learning techniques, we have predicted heart disease and provided a comparative analysis of the algorithms for machine learning used for the experiment of the prediction.

[8] **"Latest trends on heart disease prediction using machine learning and image fusion, 2021" authors Manoj Diwakar , Amrendra Tripathi , Kapil Joshi , Minakshi Memoria , Prabhishek Singh , Neeraj kumar** presents a classification method of machine learning can be useful to help the medical branch by delivering reliable and instant disease diagnosis.

2.2 PROPOSED SYSTEM

DATASET:

The heart dataset is now hosted in an online repository (Kaggle). Now we intend to analyze data related to cardiac features of patients from the "heart.csv" dataset. This dataset provides various information about patients, including age, gender, blood pressure, cholesterol levels, electrocardiographic (ECG) features, and more.

Dataset Information:

This dataset includes the following features:

age: The age of the patient.

sex: Gender of the patient (0: female, 1: male).

cp: Type of chest pain.

trestbps: Resting blood pressure.

chol: Serum cholesterol.

fbs: Fasting blood sugar > 120 mg/dl.

restecg: Resting electrocardiographic results.

thalach: Maximum heart rate achieved.

exang: Exercise induced angina.

oldpeak: ST depression induced by exercise relative to rest

MODEL ARCHITECTURE:

The proposed system involves a comprehensive model architecture designed to efficiently process, analyze, and predict heart disease outcomes from patient data. The architecture begins with data ingestion, where patient records, including demographic, clinical, and historical data, are collected and preprocessed to handle missing values, normalize data, and encode categorical variables. This cleaned data is then fed into a feature selection module that identifies the most relevant predictors of heart disease, utilizing techniques such as recursive feature elimination or principal component analysis. The core of the system is a machine learning model, such as a random forest, support vector machine, or neural network, which is trained and validated using a portion of the dataset to ensure accuracy and robustness. Post-training, the model is deployed within a scalable and secure

environment, often using cloud-based platforms, to allow real-time predictions on new patient data. The results are visualized through an intuitive dashboard that provides healthcare professionals with actionable insights, risk scores, and prediction explanations, facilitating informed decision-making and early intervention strategies. Regular updates and retraining cycles are incorporated to adapt to new data and maintain model performance over time.

TRAINING AND TESTING:

The proposed system's training and testing phases are crucial for developing a reliable predictive model. Initially, a comprehensive dataset comprising patient records, including demographic information, medical history, and diagnostic test results, is collected and preprocessed to handle missing values and outliers. The data is then divided into training and testing subsets, typically in an 80:20 ratio. Various machine learning algorithms, such as logistic regression, decision trees, and neural networks, are trained on the training subset to learn the patterns and correlations indicative of heart disease. During training, techniques like cross-validation and hyperparameter tuning are employed to optimize the model's performance and prevent overfitting. Once the models are trained, they are evaluated on the testing subset to assess their predictive accuracy, sensitivity, specificity, and other relevant metrics. This rigorous testing ensures that the final model is robust and generalizes well to unseen data, making it a valuable tool for healthcare professionals in predicting heart disease and improving patient outcomes.

CHAPTER 3

SYSTEM DESIGN

3.1 GENERAL

In this section, we would like to show how the general outline of how all the components end up working when organized and arranged together. It is further represented in the form of a flow chart below.

3.2 SYSTEM ARCHITECTURE DIAGRAM

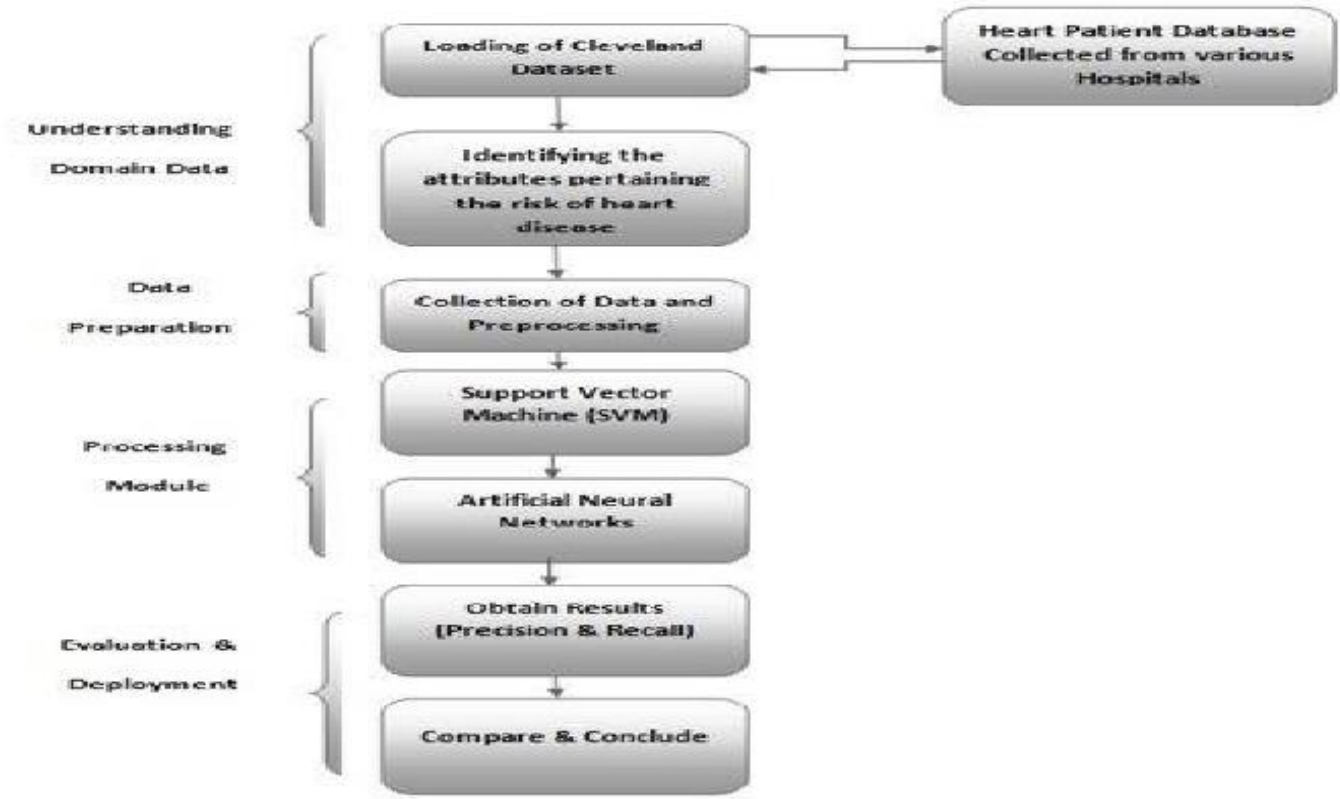

Fig 3.2.1: System Architecture

3.3 DEVELOPMENTAL ENVIRONMENT

3.3.1 HARDWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the system's implementation. It should therefore be a complete and consistent specification of the entire system. It is generally used by software engineers as the starting point for the system design.

Table 3.3.1 Hardware Requirements

COMPONENTS	SPECIFICATION
PROCESSOR	Intel Core i5
RAM	8 GB RAM
PROCESSOR SPEED	MINIMUM 1.1 GHz

3.3.2 SOFTWARE REQUIREMENTS

The software requirements document is the specifications of the system. It should include both a definition and a specification of requirements. It is a set of what the system should rather be doing than focus on how it should be done. The software requirements provide a basis for creating the software requirements specification.The software requirements are description of features and functionalities of the target system. Requirements convey the expectations of users from the software product.

Table 3.3.2 Software Requirements

S.NO	REQUIREMENT

1	Jupyter Notebook

3.4 DESIGN OF THE ENTIRE SYSTEM:

3.4.1 SEQUENCE DIAGRAM:

A sequence diagram simply depicts the interaction between the objects in a sequential order. An sequence diagram is used to show the interactive behavior of a system. The sequence diagram for Disease prediction in plants and recommendation of fertilizer is attached in the below figure 3.4.1.

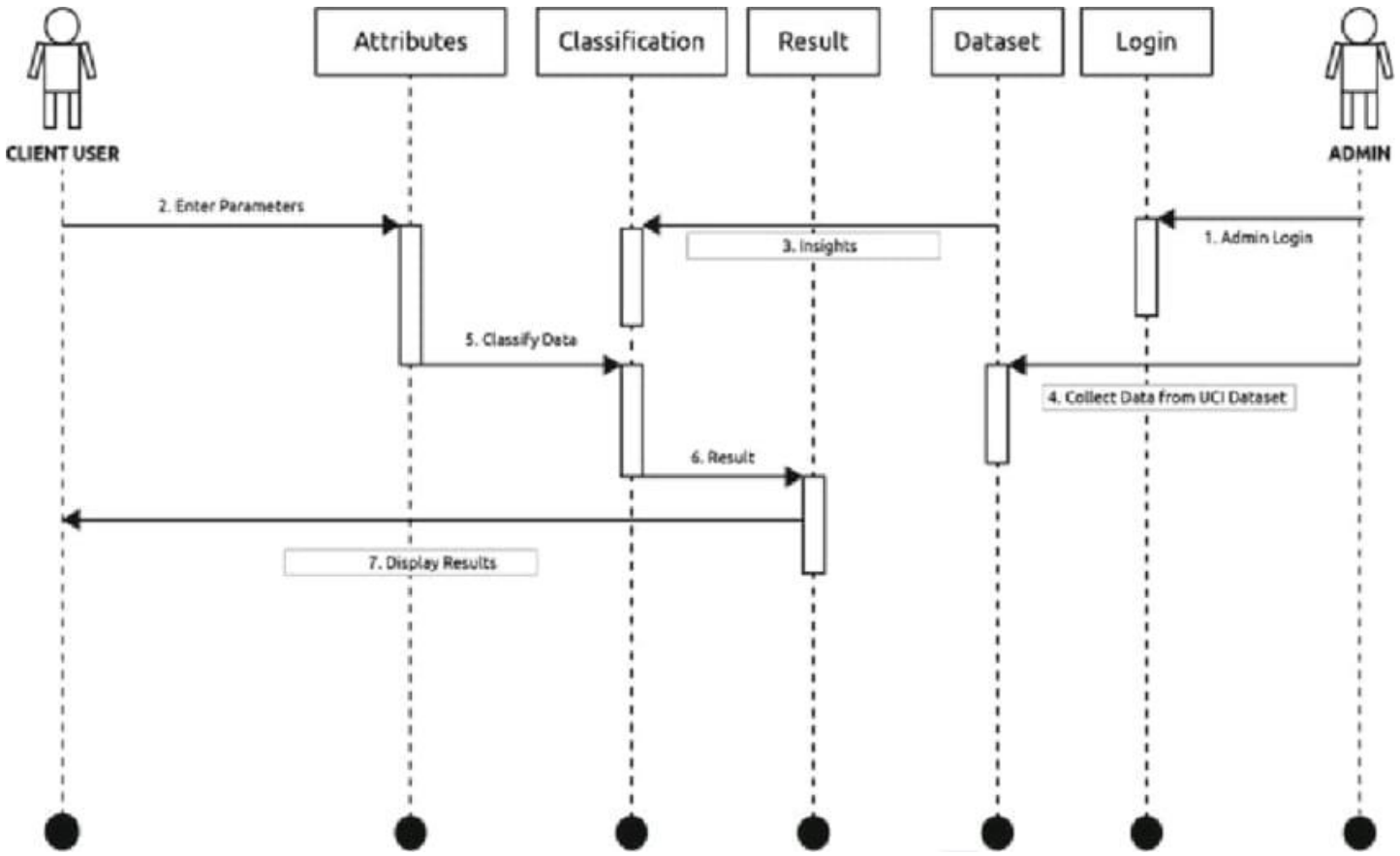

Fig 3.4.1: Sequence Diagram

CHAPTER 4

PROJECT DESCRIPTION

4.1 METHODOLODGY

The methodology for our proposed system involves several key steps. First, we will collect and preprocess extensive healthcare datasets from hospitals and relevant institutions, ensuring data is cleaned and standardized. Following this, exploratory data analysis (EDA) will be conducted to understand the data distribution and identify significant features influencing heart disease outcomes. Next, we will implement various machine learning algorithms such as logistic regression, decision trees, random forests, and neural networks to build predictive models. These models will be trained and validated using cross-validation techniques to ensure robustness and to mitigate overfitting. Performance metrics such as accuracy, precision, recall, and the area under the ROC curve (AUC) will be used to evaluate and compare model effectiveness. The best-performing model will be integrated into an analytic tool with a user-friendly interface, allowing healthcare providers to input patient data and receive real-time heart disease risk predictions. Finally, the tool will be tested in a clinical setting to assess its practical applicability and accuracy, and adjustments will be made based on user feedback and ongoing performance monitoring.

4.2 MODULE DESCRIPTION

Our proposed system involves developing a sophisticated software solution designed to predict heart disease outcomes using advanced machine

learning algorithms. This tool will integrate with existing healthcare databases to analyze patient data, identifying patterns and risk factors associated with heart disease. It will leverage techniques such as supervised learning, feature selection, and model optimization to deliver high-accuracy predictions. The application aims to assist healthcare providers in early diagnosis and personalized treatment planning, ultimately improving patient outcomes and operational efficiency in medical institutions. Additionally, it includes user-friendly interfaces for data visualization, facilitating ease of use for medical professionals without extensive technical background.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 OUTPUT

The following images below contains the information about patients being affected by disease or not .

Heart disease prediction

```
pd.crosstab(df.target, df.sex).plot(kind="bar")
plt.title("Heart disease frequency for sex")
plt.xlabel("0 = No Disease, 1 = Disease")
plt.ylabel("Amount")
plt.legend(["Female", "Male"])
plt.xticks(rotation = 0)
```

```
(array([0, 1]), [Text(0, 0, '0'), Text(1, 0, '1')])
```

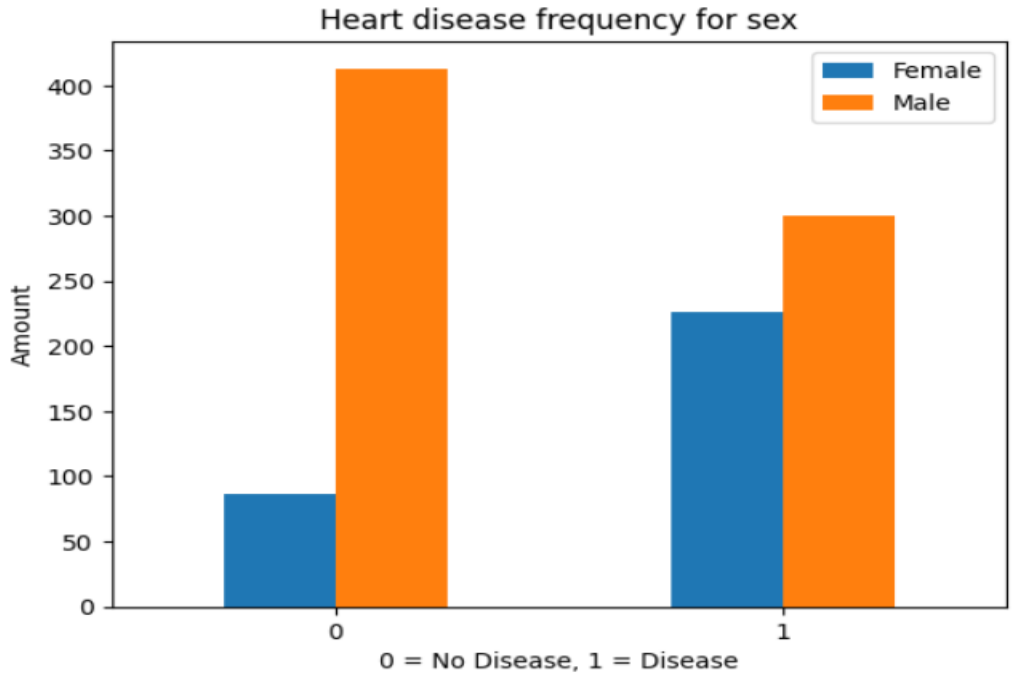

Fig 5.1.1: Heart Disease prediction

Output from whether the patient is affected by disease or not

```
plt.figure(figsize=(10,6))
plt.scatter(df.age[df.target==1], df.thalach[df.target == 1], c="red")
plt.scatter(df.age[df.target==0], df.thalach[df.target == 0], c="green")

plt.title("Heart Disease in function of Age and Max Heart Rate")
plt.xlabel("Age")
plt.ylabel("Heart Rate")
plt.legend(["Disease", "No Disease"])
```

```
<matplotlib.legend.Legend at 0x21988a73200>
```

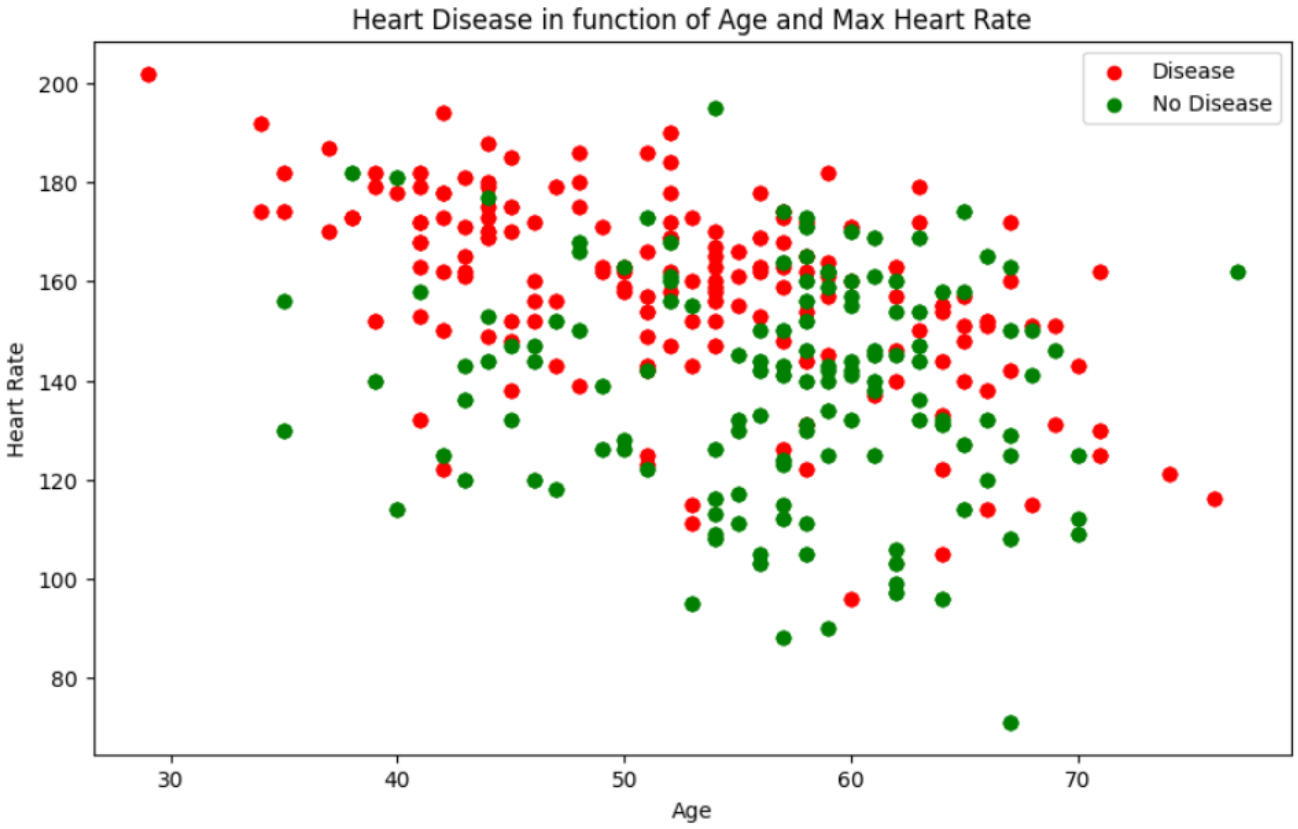

Fig 5.1.2: Output of predicted heart patients

Confusion Matrix :

The plant disease prediction trained model is evaluated and the confusion matrix for the trained model is attached in below Figure 5.1.3

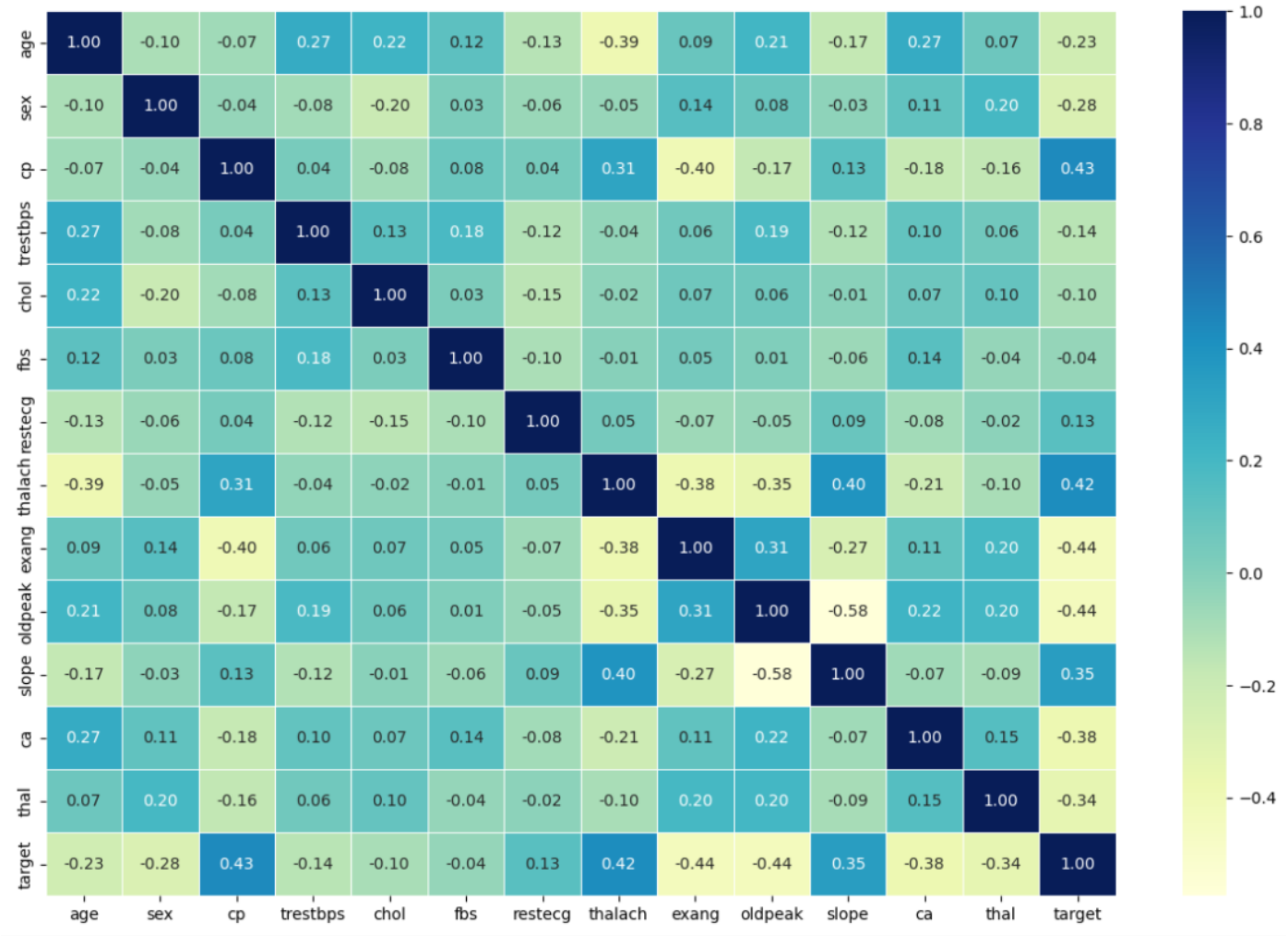

Fig 5.1.3: Confusion matrix

Training and Testing Accuracy Graph:

The proposed model is evaluated and the testing and training accuracy graph is obtained. Spliting the dataset into training and validation sets . Training the model using the training set, adjusting hyperparameters to optimize performance. Employ techniques such as dropout and batch normalization to prevent overfitting. The training and testing accuracy rate of the model is attached in the below figure 5.1.4

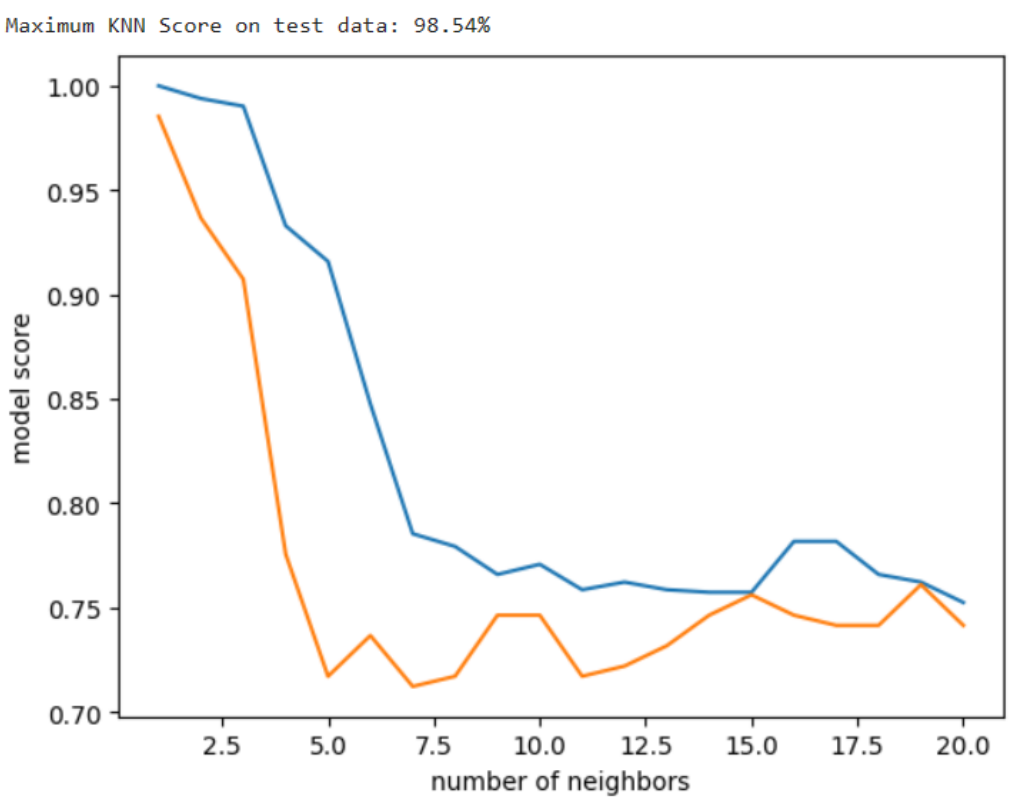

Fig 5.1.4: Training and Testing Accuracy Graph

5.2 RESULT

The implementation of Heart Disease prediction in hospitals and other institutions using machine learning represents a remarkable journey by achieving 98% accuracy in identifying heart disease in patients. This tool integrates patient data from hospitals and other medical institutions, employing advanced analytical techniques to identify patterns and risk factors associated with heart disease. By utilizing machine learning models, such as decision trees, neural networks, and ensemble methods, the system can predict the likelihood of heart disease occurrence, thereby enabling early intervention and personalized treatment plans. This innovative approach not only enhances diagnostic accuracy but also supports healthcare professionals in making informed decisions, ultimately improving patient outcomes and optimizing resource allocation within medical facilities.

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 CONCLUSION

The proposed model successfully demonstrates the potential of machine learning in transforming healthcare by providing accurate and timely predictions of heart disease. By leveraging advanced algorithms and comprehensive patient data, the tool enhances diagnostic accuracy, facilitates early intervention, and ultimately improves patient outcomes. The integration of this predictive analytic tool into healthcare settings can streamline clinical decision-making, optimize resource allocation, and contribute to personalized patient care. This project underscores the critical role of technology in advancing medical practice and underscores the importance of continuous innovation in the healthcare sector.

6.2 FUTURE ENHANCEMENT

Predictive Analytics for Treatment Response: Extend the capabilities of the tool to predict the effectiveness of different treatment options for patients with heart disease. This could help healthcare providers make more informed decisions about treatment plans.

Longitudinal Data Analysis: Incorporate longitudinal data analysis to track changes in patient health over time. This could help identify patterns or trends that are not apparent from individual snapshots of patient data.

APPENDIX

main_app.py:

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

%matplotlib inline

# Logistic regression is a classification algo rather than regression as its name confuses
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import train_test_split, cross_val_score, RandomizedSearchCV, GridSearchCV
from sklearn.metrics import confusion_matrix, classification_report, precision_score, recall_score, f1_score, RocCurveDisplay

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

```
df.shape

df['target'].value_counts().plot(kind="bar", color=["green", "blue"])

df.info()

df.isna().sum()

df['sex'].value_counts()

pd.crosstab(df.target, df.sex)
pd.crosstab(df.target, df.sex).plot(kind="bar")
plt.title("Heart disease frequency for sex")
plt.xlabel("0 = No Disease, 1 = Disease")
plt.ylabel("Amount")
plt.legend(["Female", "Male"])
plt.xticks(rotation = 0)

plt.figure(figsize=(10,6))
plt.scatter(df.age[df.target==1], df.thalach[df.target == 1], c="red")
plt.scatter(df.age[df.target==0], df.thalach[df.target == 0], c="green")
plt.title("Heart Disease in function of Age and Max Heart Rate")
plt.xlabel("Age")
plt.ylabel("Heart Rate")
```

```
plt.legend(["Disease", "No Disease"])

df["age"][df["target"] == 1]

df["age"][df["target"] == 0]

df.age.plot.hist();

pd.crosstab(df.cp, df.target).plot(kind="bar")
plt.title("Heart Disease frequency per chest pain type")
plt.xlabel("Chest Pain Type")
plt.ylabel("Amount")
plt.legend(["No Disease", "Disease"])
plt.xticks(rotation=0)

corr_matrix = df.corr()
plt.subplots(figsize=(15, 10))
sns.heatmap(corr_matrix, annot=True, linewidths=0.5, fmt=".2f", cmap="YlGnBu");

X = df.drop('target', axis=1)
y = df["target"]
X
Y
```

```
models = {"Logistic Regression": LogisticRegression(),
          "KNN": KNeighborsClassifier(),
          "Random Forest": RandomForestClassifier()}

def fit_and_score(models, X_train, X_test, y_train, y_test):
    """
    Fit and evaluate given machine learning model.
    model: a dict of different scikit-learn ML models
    X_train: training data (no labels)
    X_test: testing data (no labels)
    y_train: training labels
    y_test: testing labels
    """
    np.random.seed(42)
    model_scores = {}
    for name, model in models.items():
        model.fit(X_train, y_train)
        model_scores[name] = model.score(X_test, y_test)
    return model_scores

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
model_scores = fit_and_score(models, X_train, X_test, y_train, y_test)
model_scores
```

```
model_compare = pd.DataFrame(model_scores, index=["accuracy"])
model_compare.T.plot.bar()

train_scores = []
test_scores = []

neighbors = range(1,21)
knn = KNeighborsClassifier()

for i in neighbors:
    knn.set_params(n_neighbors=i)
    knn.fit(X_train, y_train)
    train_scores.append(knn.score(X_train, y_train))
    test_scores.append(knn.score(X_test, y_test))
train_scores

test_scores

plt.plot(neighbors, train_scores, label='Train Scores')
plt.plot(neighbors, test_scores, label="Test Scores")
plt.xlabel("number of neighbors")
plt.ylabel("model score")

print(f"Maximum KNN Score on test data: {max(test_scores)*100:.2f}%")
```

```
log_reg_grid = {
    "C": np.logspace(-4, 4, 20),
    "solver": ["liblinear"]
}

rf_grid = {
    "n_estimators": np.arange(10, 1000, 50),
    "max_depth": [None, 3, 5, 10],
    "min_samples_split": np.arange(2,20, 2),
    "min_samples_leaf": np.arange(1, 20, 2)
}
rf_grid

np.random.seed(42)
rs_log_reg = RandomizedSearchCV(LogisticRegression(),
                                param_distributions=log_reg_grid,
                                cv=5,
                                n_iter=20,
                                verbose=True)
rs_log_reg.fit(X_train, y_train)

rs_log_reg.score(X_test, y_test)
```

```
rs_rf.score(X_test, y_test)
rs_rf.best_params_

log_reg_grid = {
    "C": np.logspace(-4, 4, 30),
    "solver": ["liblinear"]
}

gs_log_reg = GridSearchCV(LogisticRegression(),
                          param_grid=log_reg_grid,
                          cv=5,
                          verbose=True)

gs_log_reg.fit(X_test, y_test)

gs_log_reg.best_params_
gs_log_reg.score(X_test, y_test)

y_preds = gs_log_reg.predict(X_test)
y_preds

RocCurveDisplay.from_estimator(gs_log_reg, X_test, y_test)
print(confusion_matrix(y_test, y_preds))
```

```
sns.set(font_scale=1.5)

def plot_conf_matrix(y_test, y_preds):
    fig, ax = plt.subplots(figsize=(3,3))
    ax = sns.heatmap(confusion_matrix(y_test, y_preds),
                     annot=True,
                    cbar=False)
    plt.xlabel("True Label")
    plt.ylabel("Predicted Label")

plot_conf_matrix(y_test, y_preds)

print(classification_report(y_test, y_preds))

gs_log_reg.best_params_

clf = LogisticRegression(C=0.7278953843983146, solver="liblinear")

cv_acc = cross_val_score(
    clf,
    X,
    y,
    cv=5,
    scoring="accuracy"
```

```
)
cv_acc = np.mean(cv_acc)
cv_acc

cv_precision = cross_val_score(
    clf,
    X,
    y,
    cv=5,
    scoring="precision"
)

cv_precision = np.mean(cv_precision)
cv_precision

cv_recall = cross_val_score(
    clf,
    X,
    y,
    cv=5,
    scoring="recall"
)

cv_recall = np.mean(cv_recall)
```

```
cv_recall

cv_f1 = cross_val_score(
    clf,
    X,
    y,
    cv=5,
    scoring="precision"
)

cv_f1 = np.mean(cv_f1)
cv_f1

cv_metrics = pd.DataFrame({
    "Accuracy": cv_acc,
    "Precision": cv_precision,
    "Recall": cv_recall,
    "F1": cv_f1
}, index=[0])
cv_metrics.T.plot.bar(title="Cross Validated classification metrics", legend=False);

gs_log_reg.best_params_
clf = LogisticRegression(C=0.7278953843983146 , solver="liblinear")
clf.fit(X_train, y_train)
```

```
clf.coef_

feature_dict = dict(zip(df.columns, list(clf.coef_[0])))
feature_dict

feature_df = pd.DataFrame(feature_dict, index=[0])
feature_df.T.plot.bar(title="Feature Importance", legend=False);

from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.ensemble import RandomForestClassifier
import pandas as pd

# Step 1: Load the dataset
data = pd.read_csv('heart.csv')  # Replace 'your_dataset.csv' with your dataset file

# Step 2: Preprocessing (if needed)
# Example:
# data.dropna(inplace=True)  # Drop rows with missing values
# data = pd.get_dummies(data)  # Encode categorical variables if any

# Step 3: Feature Selection/Extraction (if needed)
# Example:
# X = data[['feature1', 'feature2']]  # Selecting specific features
```

```
# Step 4: Splitting data into training and testing sets
X = data.drop('target', axis=1)  # Features
y = data['target']  # Target variable

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Step 5: Model Selection
model = RandomForestClassifier()  # Example: Random Forest Classifier

# Step 6: Model Training
model.fit(X_train, y_train)

# Step 7: Model Evaluation
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)

# Step 8: Reporting Accuracy
print("Accuracy: {:.2f}%".format(accuracy * 100))
```

REFERENCES

[1] **"Heart Disease Prediction using Machine Learning Techniques, 2020", by** Devansh Shah, Samir Patel & Santosh Kumar Bharti

[2] **"Heart disease prediction using machine learning algorithms, 2021" by** Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath

[3] **"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques , 2019"** by Senthilkumar Mohan; Chandrasegar Thirumalai; Gautam Srivastava

[4] **"Heart Disease Prediction using Hybrid machine Learning Model, 2021"** by M. Kavitha; G. Gnaneswar; R. Dinesh; Y. Rohith Sai; R. Sai Suraj

[5] **"Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, 2021"** by Md Mamun Ali , Bikash Kumar Paul , Kawsar Ahmed , Francis M. Bui , Julian M.W. Quinn , Mohammad Ali Moni

[6] "**Heart Disease Prediction Using Machine learning and Data Mining Technique, 2016"** by Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel

[7] "**Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis, 2020"** by Rahul Katarya & Sunit Kumar Meena

[8] **"Latest trends on heart disease prediction using machine learning and image fusion, 2021"** by Manoj Diwakar , Amrendra Tripathi , Kapil Joshi , Minakshi Memoria , Prabhishek Singh , Neeraj kumar