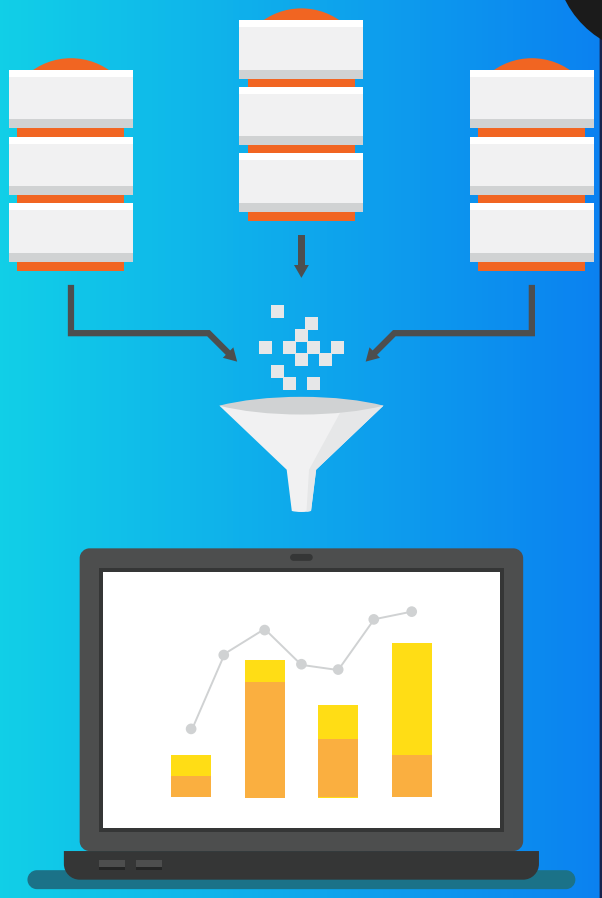
Data Mining

Assignment# 02



NAME: SHAMEER HASSAN
19011556-123
MUHAMMAD FASIAL
19011556-163
ZEESHAN AHMAD
19011556-061

SUBMITTED TO:
SIR USMAN ZIA

TOPIC:
MACHINE LEARNING

INTRODUCTION

Heart disease is a common and serious health issue that affects millions of people worldwide. It refers to a range of conditions that affect the heart, including coronary artery disease, heart attacks, and heart failure. Heart disease can lead to significant health problems, such as stroke, arrhythmias, and even death. It is a leading cause of death in many countries and has a significant impact on the quality of life of those who suffer from it. The Heart Disease Dataset is a valuable resource for researchers and healthcare professionals who are interested in understanding the risk factors associated with heart disease and developing effective diagnostic and treatment strategies.

Heart disease is a complex and multifactorial condition, with various risk factors contributing to its development, such as age, genetics, lifestyle, and underlying health conditions. Recent advances in machine learning and data analysis techniques have provided new opportunities for predicting and detecting heart disease. By analysing large datasets of patient attributes and health outcomes, researchers can develop predictive models that can identify patients at high risk of heart disease and improve treatment outcomes.

Machine learning is a subfield of artificial intelligence that focuses on the development of algorithms and models that can learn from data and make predictions or decisions. In the context of heart disease prediction, machine learning algorithms can analyse large datasets of patient attributes and health outcomes to identify patterns and risk factors associated with heart disease. By learning from these patterns, the algorithms can develop predictive models that can identify patients at high risk of developing heart disease and improve treatment outcomes.

Data analysis techniques are also used to predict heart disease by examining large datasets of patient information. These techniques include statistical methods and data mining, which can identify patterns and relationships in the data that are associated with heart disease. By using these techniques, researchers can identify new risk factors and develop predictive models that can be used in clinical practice.

BACKGROUND

In the case of heart disease, there are several important background factors that motivate the need for this type of research. First of all, the number one death in the world is heart disease. The World Health Organization estimates that 17.9 million deaths worldwide from cardiovascular disease occurred in 2019, accounting for 32% of all fatalities. The necessity for efficient prevention and treatment methods is highlighted by the high prevalence of heart disease.

Secondly, heart disease is a complex and multifactorial condition. While certain risk factors, such as smoking and high blood pressure, are well-established, the underlying mechanisms that contribute to heart disease are not fully understood. There is a need for further research to better understand the risk factors and underlying mechanisms of heart disease. Thirdly, early detection and management of heart disease are crucial for improving outcomes. Traditional methods of detecting heart disease, such as medical history, physical examination, and blood tests, have limitations in their ability to detect heart disease at an early stage or predict the risk of future cardiovascular events.

In the past, the diagnosis of heart disease was primarily based on medical history, physical examination, and electrocardiography. Physicians would ask patients about their symptoms, medical history, and family history of heart disease, and perform a physical examination to assess their heart function. Electrocardiography, or ECG, was developed in the early 20th century and became a standard tool for diagnosing heart disease. ECG measures the electrical activity of the heart and can detect abnormal heart rhythms, damage to the heart muscle, and other signs of heart disease.

There is a need for more accurate and reliable methods of detecting heart disease and predicting its risk. Machine learning and data analysis techniques have the potential to address these issues by analysing large datasets of patient attributes and health outcomes. The heart disease dataset provides a valuable resource for researchers to develop predictive models that can identify patients at high risk of developing heart disease and improve treatment outcomes.

In summary, heart disease is a significant health issue that requires effective prevention and treatment strategies. The heart disease dataset and machine learning techniques provide an opportunity for researchers

to gain insights into the risk factors and underlying mechanisms of heart disease and develop more accurate and reliable methods of detecting and predicting heart disease.

DESCRIPTION

Heart disease is a significant health problem that affects millions of people worldwide. The dataset includes information about various demographic and clinical variables that are associated with heart disease. For example, age is a known risk factor for heart disease, and the dataset includes age information for each patient. Similarly, sex is also a known risk factor, with men being at a higher risk than women. The dataset includes information about the sex of each patient, allowing researchers to explore the relationship between sex and heart disease.

Other variables included in the dataset are clinical measures that are commonly used to assess the risk of heart disease. These measures include blood pressure, cholesterol levels, and electrocardiographic results. The dataset also includes information about exercise-induced angina and the number of major vessels coloured by fluoroscopy. These variables can help researchers identify patients who are at a higher risk of developing heart disease and develop more effective screening strategies. The Heart Disease Dataset is a collection of data from patients who have undergone various tests and procedures to assess their risk of developing heart disease. The dataset includes a total of 303 instances or observations and 14 variables or features that describe the patients' demographic, clinical, and laboratory characteristics. The variables included in the dataset are as follows:

AGE: The age of the patient in years.

SEX: The sex of the patient (1 = male; 0 = female).

CP: Chest pain type, which can take four values: typical angina, atypical angina, non-anginal pain, or asymptomatic.

TRESTBPS: The resting blood pressure (in mm Hg) of the patient.

CHOL: The serum cholesterol (in mg/dl) of the patient.

FBS: Fasting blood sugar (in mg/dl) greater than 120 mg/dl or not (1 = true; 0 = false).

RESTECG: Resting electrocardiographic results, which can take three values: normal, having ST-T wave abnormality, or showing probable or definite left ventricular hypertrophy.

THALACH: Maximum heart rate achieved during exercise. exang: Exercise-induced angina (1 = yes; 0 = no).

OLDPEAK: ST depression induced by exercise relative to rest.

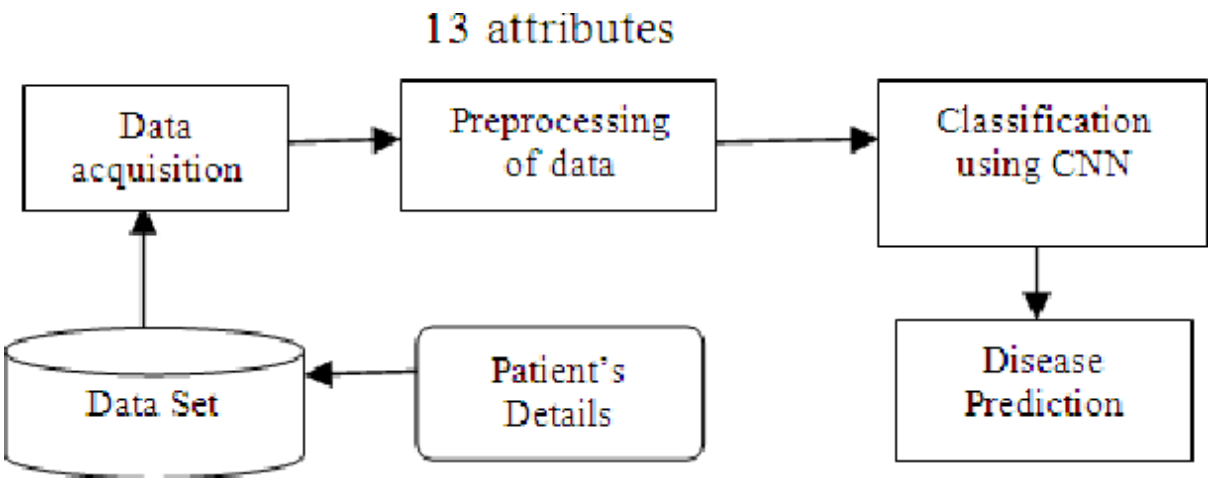
SLOPE: The slope of the peak exercise ST segment, which can take three values: upsloping, flat, or downsloping.

CA: The number of major vessels (0-3) colored by fluoroscopy.

THAL: A blood disorder called thalassemia, which can take three values: normal, fixed defect, or reversible defect.

TARGET: The presence of heart disease (1 = yes; 0 = no).

Attribute	Description	Data Type
age	Age of the patient in years	Numerical
sex	Sex of the patient (0 = female, 1 = male)	Categorical
cp	Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)	Categorical
trestbps	Resting blood pressure (mm Hg)	Numerical
chol	Serum cholesterol (mg/dl)	Numerical
fbs	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)	Categorical
restecg	Resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = probable or definite left ventricular hypertrophy)	Categorical
thalach	Maximum heart rate achieved during exercise	Numerical
exang	Exercise-induced angina (1 = yes, 0 = no)	Categorical
oldpeak	ST depression induced by exercise relative to rest	Numerical
slope	The slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping)	Categorical
ca	Number of major vessels colored by fluoroscopy (0-3)	Numerical
thal	A blood disorder called thalassemia (0 = normal, 1 = fixed defect, 2 = reversible defect)	Categorical
target	Presence of heart disease (0 = no, 1 = yes)	Categorical



RELATED ARTICLES

- "Predicting Cardiovascular Disease Using Machine Learning Techniques: A Systematic Review" by A. M. Alsharqi et al. (2021). This article provides a comprehensive review of the use of machine learning techniques for predicting cardiovascular disease, including heart disease.
- "Machine learning for predicting cardiovascular disease 10-year risk using routine clinical data from the Electronic Health Record: A systematic review and meta-analysis" by H. Abdi et al. (2021). This article reviews the use of machine learning for predicting cardiovascular disease risk using electronic health record data and provides insights into the performance and limitations of various algorithms.
- "A Deep Learning Approach for Heart Disease Diagnosis" by J. Abawajy et al. (2020). This article presents a deep learning approach for diagnosing heart disease using a combination of convolutional and recurrent neural networks.

LITRETURE REVIEW

Several studies have explored the use of machine learning techniques for predicting cardiovascular disease, including heart disease, using clinical data from patients. For example, Alsharqi et al. (2021) conducted a systematic review of the use of machine learning techniques for predicting cardiovascular disease risk. They found that machine learning

algorithms, such as logistic regression, decision trees, and artificial neural networks, can be used to develop predictive models for cardiovascular disease with high accuracy. Abdi et al. (2021) conducted a systematic review and meta-analysis of the use of machine learning for predicting cardiovascular disease risk using electronic health record data. They found that machine learning algorithms can provide accurate predictions of cardiovascular disease risk using routinely collected clinical data from electronic health records. Abawajy et al. (2020) developed a deep learning approach for diagnosing heart disease using a combination of convolutional and recurrent neural networks. They found that their approach achieved high accuracy in diagnosing heart disease using clinical data. The Heart Disease Dataset provides valuable information about various demographic and clinical variables that are associated with heart disease, including age, sex, blood pressure, cholesterol levels, electrocardiographic results, and exercise-induced angina. By exploring the relationships between these variables and the likelihood of developing heart disease, researchers can develop more effective prevention and treatment strategies for this condition. In summary, the Heart Disease Dataset is a valuable resource for researchers and healthcare professionals interested in studying the risk factors for heart disease and developing effective diagnostic and treatment strategies. Machine learning techniques, such as logistic regression, decision trees, and artificial neural networks, can be used to develop predictive models for heart disease with high accuracy using clinical data from patients. The development of more accurate and effective predictive models for heart disease will help to identify individuals at risk of developing this condition and develop more effective prevention and treatment strategies.

OBJECTIVES

- To provide a collection of data related to patients who have undergone cardiac diagnostic tests.
- To include a variety of clinical and demographic features of patients, such as age, sex, chest pain type, resting blood pressure, serum cholesterol levels, maximum heart rate achieved, and the presence or absence of coronary artery disease.
- To enable researchers and data analysts to use this information to develop and test predictive models to identify individuals at risk of heart disease.

- To discover patterns and relationships between various risk factors and the likelihood of developing heart disease.
- To aid in the development of effective prevention and treatment strategies for heart disease.
- To identify potential areas for further research, such as investigating the relationship between lifestyle factors and heart disease risk.

MY GOAL

My Goal is to train a model which predicts if a patient have heart disease or not.

STEPS TO ACHIEVE THIS GOAL

- Load the dataset into your preferred programming environment or machine learning platform.
- Explore and pre-process the data, including handling missing values, converting categorical features into numerical values, and scaling the data if necessary.
- Split the data into training and testing sets. Select an appropriate machine learning algorithm for binary classification, such as logistic regression, decision tree, or support vector machine (SVM).
- Train the model on the training data and evaluate its performance on the testing data using appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score.
- Once you are satisfied with the model's performance, use it to make predictions on new data to identify patients who may have heart disease.

HEART DIESES PREDICTOR PYTHON CODE

```
import numpy as np
import pandas as py
from sklearn.model_selection import train_test_split
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LinearRegression


heartdata=py.read_csv("heart.csv")

heartdata.head()
heartdata.tail()
# heartdata.shape



heartdata.info()
heartdata.describe()
targets=heartdata['target'].value_counts()

#all columns
X=heartdata.drop(columns='target',axis=1)
#target column
Y=heartdata['target']


X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_si
ze=0.2,stratify=Y,random_state=2)
# print(X.shape,X_train.shape,X_test.shape)



dt = DecisionTreeClassifier()
```

```
dt.fit(X_train, Y_train)
y_pred_dt = dt.predict(X_test)
acc_dt = accuracy_score(Y_test, y_pred_dt)
print("Decision Tree accuracy:", acc_dt)

# Random Forest model
rf = RandomForestClassifier()
rf.fit(X_train, Y_train)
y_pred_rf = rf.predict(X_test)
acc_rf = accuracy_score(Y_test, y_pred_rf)
print("Random Forest accuracy:", acc_rf)

# Linear Regression model
lr = LinearRegression()
lr.fit(X_train, Y_train)
y_pred_lr = lr.predict(X_test)
y_pred_lr[y_pred_lr < 0.5] = 0
y_pred_lr[y_pred_lr >= 0.5] = 1
acc_lr = accuracy_score(Y_test, y_pred_lr)
print("Linear Regression accuracy:", acc_lr)



# model = LogisticRegression()
# model.fit(X_train, Y_train)
# y_pred_lr = model.predict(X_test)
# acc_lr = accuracy_score(Y_test, y_pred_lr)
# print("Logistic Regression accuracy:", acc_lr)
```

```
model=LogisticRegression()
model.fit(X_train,Y_train )


X_train_prediction=model.predict(X_train)
trainigdataaccuracy=accuracy_score(X_train_prediction,Y_tr
ain)
# print( trainigdataaccuracy)




X_test_prediction=model.predict(X_test)
testdataaccuracy=accuracy_score(X_test_prediction,Y_test)
print( testdataaccuracy)



input_from_user=(71,0,0,112,149,0,1,125,0,1.6,1,0,2)
input_from_user_array=np.asarray(input_from_user)
input_from_user_reshaped=input_from_user_array.reshape(1,-
1)
prediction=model.predict(input_from_user_reshaped)


if prediction[0]==0:
    print("Patient Doesnot have  Any Heart Dieseas")
else:
    print("Patient Has hear dieseas he needs more
tests")
```

OUTPUT

```
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
Decision Tree accuracy: 1.0
Random Forest accuracy: 1.0
Linear Regression accuracy: 0.8048780487804879
C:\Users\Meer\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\linear_model\_logistic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
0.8048780487804879
C:\Users\Meer\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\base.py:409: UserWarning: X does not have valid feature names, but LogisticRegression was fitted
 with feature names
  warnings.warn(
Patient Has hear dieseas he needs more tests
PS M:\Programming\Python Projects 2023\Heart Dieseas Predictor>
```