**INTRODUCTION TO STATISTICAL LEARNING**

HEART DISEASE

DATA ANALYSIS

REPORT

VYTLA SAI KALYAN

16273044

[SV3T9@MAIL.UMKC.EDU](mailto:SV3T9@MAIL.UMKC.EDU)

**INTRODUCTION**

There are many life-threatening diseases in the world and among them heart disease has gained a great deal of importance in medical field. The diagnosis of heart disease and prediction of heart disease is a challenging task. The diagnosis of heart disease is based on some parameters. In this we take a dataset which contains all the set of parameters that are used in heart disease prediction.

**DATASET DISCRIPTION**

* Heart disease dataset contains 76 attributes (76 rows) and 14 fields (columns).
* Database used is a Cleveland database.
* The ‘Target’ field refer to predicted values of heart disease or not. If the value is 1 it is positive or if it is 0 then it is negative.

**DATASET CONTENTS**

* AGE – in years.​
* SEX – 0 or 1 ( 0 = female , 1 = male )​
* Cp -  chest pain type​
* Trestbps – blood pressure ( in mm hg )​
* Chol -  serum cholestoral in mg/dl​
* Fbs - (fasting blood sugar > 120 mg/dl) (0=false, 1=true)
* Restecg - resting electrocardiographic​
* thalach - maximum heart rate ​
* 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 ​
* ca - number of major vessels  colored by Flourosopy​
* thal - 3 = normal; 6 = fixed defect; 7 = reversable defect ​
* target - have disease or not (1=yes, 0=no)

APPROACH

Here we used three approaches for analyzing Logistic Regression , K-nearest neighbor,

Decision Tree.

**STEP-1:**

We read the data from dataset. Print the Target field which contains heart disease or not and represent that in a graph.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**STEP-2**

Then, we display the sex field and calculate the percentage of female, male patients and percentage of patients have and haven’t.

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**STEP-3**

We analysis the data present in the dataset and represent that by ages, sex. Here we use bar representation.

A screenshot of a social media post

Description automatically generated

A picture containing object

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

**STEP-4**

We also visualization by ST Segment and FBS.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

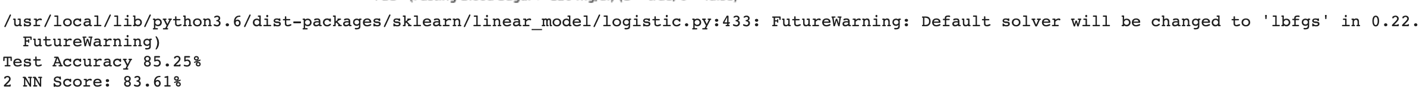
**STEP-5**

**2**

Later, Target field is splitted into training and testing sets. Here we take 20% data as test data and remaining 80% as Training data then initialize dimensions and add sigmoid function. Now using logistic Regression which is one of the statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome and here we predict the accuracy of the model.

A screenshot of a social media post

Description automatically generated



**STEP-6**

At last we use KNN which is a non-parametric method which uses feature similarity and also Decision Tree which is a Tree structure.

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**ACCURACY**

Logistic Regression - **85.25%**

K-nearest Neighbor - **90.12%**

Decision Tree - **73.77%**

From above accuracy we can say that KNN has highest accuracy. So, it is more preferable.

**SUMMARY**

* In this heart disease analysis we take a dataset from the Kaggle and apply some models like KNN, Logistic Regression and decision Tree algorithms and predict the accuracy.
* First, we load the dataset and visualize from the parameters present in the dataset and abstract different interesting patterns and facts.

* Then, we divide the dataset into training and testing and initialize dimensions and sigmoid function.
* At last we apply the logistic regression , KNN, Decision Tree and measure the accuracy.