HEART DISEASE CLASSIFICATION USING MACHINE LEARNING AND PYTHON

A Mini Project Report Submitted in Partial Fulfilment of the Requirement for the Award of the Degree of

BACHELOR OF TECHNOLOGY

in

ELECTRONICS AND COMMUNICATION ENGINEERING

by

SHAIK ABDUL KAREEM	**19PA1A04E7**
T.G.N. VAMSI KRISHNA	**19PA1A04F4**
UNAMATLA HARSHA KUMAR	**19PA1A04G4**
THOTE PRAKASH BABU	**19PA1A04D7**
Y. JASWANTH	**19PA1A04H9**

Under the Esteemed Guidance of

MS.V. PRAVEENA

Assistant Professor , Dept of ECE.

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

VISHNU INSTITUTE OF TECHNOLOGY (Autonomous)

(Accredited by NBA, NAAC, Approved by AICTE & Affiliated to JNTU Kakinada)

Vishnupur, Bhimavaram - 534202.

2021 - 2022

VISHNU INSTITUTE OF TECHNOLOGY (AUTONOMOUS)

(Accredited by NBA, NAAC, Approved by AICTE & Affiliated to JNTU Kakinada) Vishnupur, Bhimavaram-534202

DEPARTMENT OF ELECTRONICS AND COMMUNATION ENGINEERING

CERTIFICATE

This is to certify that the Mini Project entitled **"Heart disease classification using Machine Learning "** is being submitted by SHAIK ABDUL KAREEM , T.G.N. VAMSI KRISHNA, THOTE PRAKASH BABU,U. HARSHA KUMAR , Y.JASWANTH in partial fulfilment for the award of the degree of **Bachelor of Technology in Electronics and Communication Engineering** is a record of the bonafide work carried out by them under my guidance and supervision during academic year 2021-2022 and it has been found worthy of acceptance according to the requirements of the university.

MS. V. PRAVEENA
Assistant Professor,ECE Dept.

Dr. K. SRINIVAS
Professor & Head, ECE Dept.

External Examiner

ABSTRACT

The World Health Organization (WHO) lists cardiovascular diseases as the leading cause of death globally with 17.9 million people dying every year. The risk of heart disease increases due to harmful behavior that leads to overweight and obesity, hypertension, hyperglycemia, and high cholesterol. Prediction of Heart diseases helps practitioners make more accurate decisions regarding patient's health. useful parameters such as cp(chest pain type),12 chol(cholesterol level), blood pressure etc.. Machine learning models such as Support Vector Machines(SVM),k-Nearest Neighbor(KNN),Principal component analysis (PCA),etc.. can be used to build the model which can classify the disease and predict the type of disease. svm ,knn, etc are classifier models and we can predict the heart disease from the result and confusion matrix for accuracy calculation, Thus preventing Heart diseases has become more than necessary. Good data-driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that more people can live healthy lives. This is where Machine Learning comes into play. Machine Learning helps in predicting the Heart diseases, and the predictions made are quite accurate.

Keywords:
SVM,KNN,NumPy,Pandas,ScikitLearn.

CONTENTS

Topic		Page No
CONTENTS		
Abstract		**1**
1. INTRODUCTION		
1.1	Introduction.	2
1.2	Thought behind the work.	2
1.3	Problem Statement.	2
1.4	Literature Survey	3
2. METHODOLOGY		
2. 1	Data set collection and pre-processing	5
2.2	Model.	6
2.3	Required Libraries	7
2.4	Dataset and models	8-15
2.5	Evaluation metrics	16
3. RESULTS AND DISCUSSION		
3.1	Advantages and Applications	18
3.2	Results and discussion	20
4. CONCLUSION		**22**
REFERENCES		**23**
APPENDIX		**24**

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithms..

1.2 Thought behind the work:

Machine learning techniques have been around us and has been compared and used for analysis for many kinds of data science applications. The major motivation behind this research-based project was to explore the feature selection methods, data preparation and processing behind the training models in the machine learning. With first hand models and libraries, the challenge we face today is data where beside their abundance, and our cooked models, the accuracy we see during training, testing and actual validation has a higher variance. Hence this project is carried out with the motivation to explore behind the models, and further implement Logistic Regression model to train the obtained data..

1.3 Problem statement :

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data..

LITERATURE SURVEY:

There is number of works has been done related to disease prediction systems using different machine learning algorithms in medical Centres.

Senthil Kumar Mohan et al,[1] proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques in which strategy that objective is to finding critical includes by applying Machine Learning bringing about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. We produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) they likewise educated about Diverse data mining approaches and expectation techniques, Such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease.

Sonam Nikhar et al [2] has built up the paper titled as Prediction of Heart Disease Using Machine Learning Algorithms by This exploration plans to give a point by point portrayal of NaÃ¯ve Bayes and decision tree classifier that are applied in our examination especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the equivalent dataset, and the result uncovers that Decision Tree beats over Bayesian classification system.

Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD), [3] Prediction of Heart Disease Using Machine Learning, In this paper proposed system they used the neural network algorithm multi-layer

perceptron (MLP) to train and test the dataset. In this algorithm there will be multiple layers like one for input, second for output and one or more layers are hidden layers between these two input and output layers. Each node in input layer is connected to output nodes through these hidden layers. This connection is assigned with some weights. There is another identity input called bias which is with weight b, which added to node to balance the perceptron. The connection between the nodes can be feedforwarded or feedback based on the requirement.

CHAPTER 2

METHODOLOGY

2 METHODOLOGY :

.

2.1 DATASET COLLECTION AND PREPROCESSING :

The dataset is publicly available on the Kaggle Website at [4] which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 4000 records and 14 attributes. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2, where 0 is absence of heart disease. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
2	28	1	2	130	132	0	2	185	0	0	?	?	?	0
3	29	1	2	120	243	0	0	160	0	0	?	?	?	0
4	29	1	2	140	?	0	0	170	0	0	?	?	?	0
5	30	0	1	170	237	0	1	170	0	0	?	?	6	0
6	31	0	2	100	219	0	1	150	0	0	?	?	?	0
7	32	0	2	105	198	0	0	165	0	0	?	?	?	0
8	32	1	2	110	225	0	0	184	0	0	?	?	?	0
9	32	1	2	125	254	0	0	155	0	0	?	?	?	0
10	33	1	3	120	298	0	0	185	0	0	?	?	?	0
11	34	0	2	130	161	0	0	190	0	0	?	?	?	0
12	34	1	2	150	214	0	1	168	0	0	?	?	?	0
13	34	1	2	98	220	0	0	150	0	0	?	?	?	0
14	35	0	1	120	160	0	1	185	0	0	?	?	?	0
15	35	0	4	140	167	0	0	150	0	0	?	?	?	0
16	35	1	2	120	308	0	2	180	0	0	?	?	?	0
17	35	1	2	150	264	0	0	168	0	0	?	?	?	0
18	36	1	2	120	166	0	0	180	0	0	?	?	?	0
19	36	1	3	112	340	0	0	184	0	1	2	?	3	0
20	36	1	3	130	209	0	0	178	0	0	?	?	?	0
21	36	1	3	150	160	0	0	172	0	0	?	?	?	0

Fig: Snapshot of the data set.

2.2 MODEL :

Logistic Regression :

Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between the dependent variable (TenyearCHD) and the one or more independent variables (risk factors) by estimating probabilities using underlying logistic function (sigmoid function). Sigmoid function is used as a cost function to limit the hypothesis of logistic regression between 0 and 1 (squashing) i.e. $0 \leq h\theta(x) \leq 1$.

In logistic regression cost function is defined as:

$$Cost(h\theta(x), y) = \begin{cases} -\log(h\theta(x)) & if\ y = 1 \\ -\log(1 - h\theta(x)) & if\ y = 0 \end{cases}$$

Logistic Regression relies highly on the proper presentation of data. So, to make the model more powerful, important features from the available data set are selected using Backward elimination and recursive elimination techniques.

Random Forest Classifier:

The random forest algorithm provides flexibility and robustness for classification tasks using tabular data, which few other standard models can. Given its simplicity and versatility, the random forest classifier is widely used for fraud detection, loan risk prediction, and predicting heart diseases.

With the ensemble learning theorem, the random forest classifier combines results from several decision trees and optimizes training. It aims to utilize different subsets and find the best combinations to increase the dataset's predictive accuracy. The first step is building, optimizing, mixing, and matching several decision trees. Next, it uses these trees for prediction and ensembles their results to yield the final output prediction.

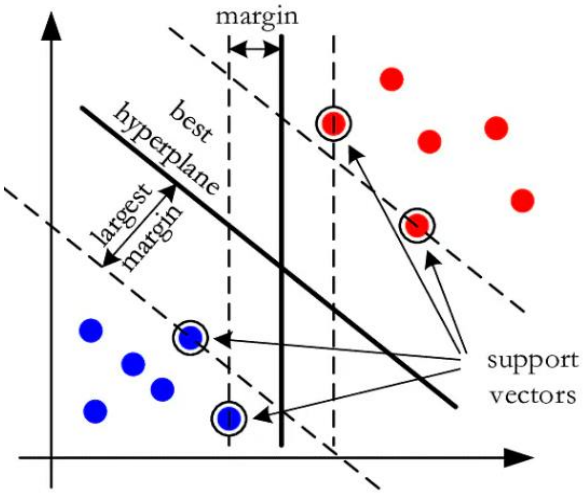
Decision Tree Classifier :

Decision Trees are the individual models that make a random forest after ensembling. Each decision tree classifier uses the dataset's attributes to create a tree. As shown in the image below, the branches end up in the leaves that are made up of target values. Using visual components and an information gain index, the tree identifies the leading features of the labels of each class. Thus, the branches are created that maximize the information gained in each split and lead up to the leaf node of that class. Decision trees are fast and robust for disease prediction if the dataset has powerful features for a simple use-case.

Support Vector Machine :

A Support Vector Machine (SVM) algorithm is a non-probabilistic classifier aiming to generate hyperplanes that divide the data points of two classes in the vector space. For N number of features and M targets, SVM creates M-1 N-dimensional hyperplanes that separate data points of different classes from each other. The image below shows how "support" vectors are calculated such that the margin (or distance) between the vectors of two classes is the most. SVM optimizes this margin metric to find the best hyperplane for all the categories.

Thus, SVMs are popular for disease prediction since they can effectively categorize tabular data into different categories



2.3 Required Libraries:

We used wide range of libraries for deep learning model

- For mathematical and scientific computing, we use NumPy.
- For High performance Analysis and Handling larger datasets we use Pandas
- For visualising the data, we use matplotlib.
- For importing classifier models we use sckitlearn.

IMPORT REQUIRED LIBRARIES:

Libraries

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

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC



from sklearn.metrics import mean_squared_error , accuracy_score,confusion_matrix
from sklearn.metrics import plot_confusion_matrix
```

READING DATASET:

Dataset

```
[ ] df = pd.read_csv("D:\intern classes\Project - 3\HeartAttack.csv",na_values = "?")
```

```
df.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	28	1	2	130.0	132.0	0.0	2.0	185.0	0.0	0.0	NaN	NaN	NaN	0
1	29	1	2	120.0	243.0	0.0	0.0	160.0	0.0	0.0	NaN	NaN	NaN	0
2	29	1	2	140.0	NaN	0.0	0.0	170.0	0.0	0.0	NaN	NaN	NaN	0
3	30	0	1	170.0	237.0	0.0	1.0	170.0	0.0	0.0	NaN	NaN	6.0	0
4	31	0	2	100.0	219.0	0.0	1.0	150.0	0.0	0.0	NaN	NaN	NaN	0

```
[ ] df.tail()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
289	52	1	4	160.0	331.0	0.0	0.0	94.0	1.0	2.5	NaN	NaN	NaN	1
290	54	0	3	130.0	294.0	0.0	1.0	100.0	1.0	0.0	2.0	NaN	NaN	1

2.4 OPERATION:

INFO OF THE DATA:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294 entries, 0 to 293
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   age         294 non-null    int64  
 1   sex         294 non-null    int64  
 2   cp          294 non-null    int64  
 3   trestbps    293 non-null    float64
 4   chol        271 non-null    float64
 5   fbs         286 non-null    float64
 6   restecg     293 non-null    float64
 7   thalach     293 non-null    float64
 8   exang       293 non-null    float64
 9   oldpeak     294 non-null    float64
 10  slope       104 non-null    float64
 11  ca          3 non-null      float64
 12  thal        28 non-null     float64
 13  num         294 non-null    int64  
dtypes: float64(10), int64(4)
memory usage: 32.3 KB
```

DESCRIBING MATHEMATICAL CALCULATION OF THE DATA:

```
df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
count	294.000000	294.000000	294.000000	293.000000	271.000000	286.000000	293.000000	293.000000	293.000000	294.000000	104.000000	3.0	28.000000	294.000000
mean	47.826531	0.724490	2.982993	132.583618	250.848708	0.069930	0.218430	139.129693	0.303754	0.586054	1.894231	0.0	5.642857	0.360544
std	7.811812	0.447533	0.965117	17.626568	67.657711	0.255476	0.460868	23.589749	0.460665	0.908648	0.338995	0.0	1.615074	0.480977
min	28.000000	0.000000	1.000000	92.000000	85.000000	0.000000	0.000000	82.000000	0.000000	0.000000	1.000000	0.0	3.000000	0.000000
25%	42.000000	0.000000	2.000000	120.000000	209.000000	0.000000	0.000000	122.000000	0.000000	0.000000	2.000000	0.0	5.250000	0.000000
50%	49.000000	1.000000	3.000000	130.000000	243.000000	0.000000	0.000000	140.000000	0.000000	0.000000	2.000000	0.0	6.000000	0.000000
75%	54.000000	1.000000	4.000000	140.000000	282.500000	0.000000	0.000000	155.000000	1.000000	1.000000	2.000000	0.0	7.000000	1.000000
max	66.000000	1.000000	4.000000	200.000000	603.000000	1.000000	2.000000	190.000000	1.000000	5.000000	3.000000	0.0	7.000000	1.000000

CREATING DUMMIES USING PANDAS:

Creating dummies using pandas

+ Code + Text

```
df = pd.get_dummies(df,columns = ["cp","restecg"])
```

```
df.head()
```

	age	sex	trestbps	chol	fbs	thalach	exang	oldpeak	num	cp_1	cp_2	cp_3	cp_4	restecg_0.0	restecg_1.0	restecg_2.0
0	28	1	130.0	132.0	0.0	185.0	0.0	0.0	0	0	1	0	0	0	0	1
1	29	1	120.0	243.0	0.0	160.0	0.0	0.0	0	0	1	0	0	1	0	0
3	30	0	170.0	237.0	0.0	170.0	0.0	0.0	0	1	0	0	0	0	1	0
4	31	0	100.0	219.0	0.0	150.0	0.0	0.0	0	0	1	0	0	0	1	0
5	32	0	105.0	198.0	0.0	165.0	0.0	0.0	0	0	1	0	0	1	0	0

RENAME FEATURES OF DATASET :

```
[ ]  df  = df.rename(columns = {"num       ":"target"})

[ ]  numerical_columns = ["age","trestbps","chol","thalach","oldpeak"]
     cat_cols = list(set(df.columns) - set(numerical_columns) - {"target"})

     cat_cols

     ['exang',
      'sex',
      'cp_1',
      'restecg_2.0',
      'fbs',
      'restecg_1.0',
      'cp_4',
      'restecg_0.0',
      'cp_2',
      'cp_3']
```

SPLITTING TRAIN AND TEST DATA:

```
splitting DATA

[ ]  df_train,df_test = train_test_split(df,test_size = 0.2,random_state = 42)

[ ]  len(df_train),  len(df_test)

     (208, 53)
```

CREATING FUNCTION TO GET FEATURES AND TARGET ARRAYS:

```
scaler = StandardScaler()

def get_features_and_target_arrays(df,numerical_columns,cat_cols,scaler):

    x_numerical_scaled = scaler.fit_transform(df[numerical_columns])
    x_categorical = df[cat_cols].to_numpy()
    x = np.hstack((x_categorical,x_numerical_scaled))


    y = df["target"]


    return x,y
```

LOGISTIC REGRESSION MODEL:

Logestic Regression

```
[ ]  #train the model


    clf = LogisticRegression()
    clf.fit(x_train,y_train)
```

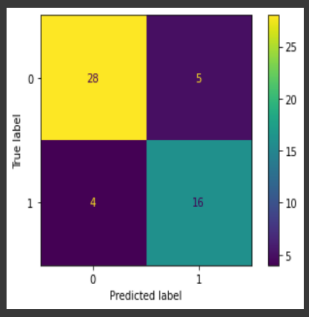
```
D:\ANACONDA_NAVIGATOR\lib\site-packages\sklearn\linear_model\_logistic.py:763: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

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(
LogisticRegression()
```

CORRESPONDING CALCUATIONS OF ACCURACY AND MEAN SQAURE ERROR:

```
[ ] x_test, y_test = get_features_and_target_arrays(df_test,cat_cols,numerical_columns,scaler)

[ ] test_pred = clf.predict(x_test)

[ ] mean_squared_error(y_test,test_pred)

    0.16981132075471697

[ ] accuracy_score(y_test,test_pred)

    0.8301886792452831

[ ] confusion_matrix(y_test,test_pred)

    array([[28,  5],
           [ 4, 16]], dtype=int64)
```

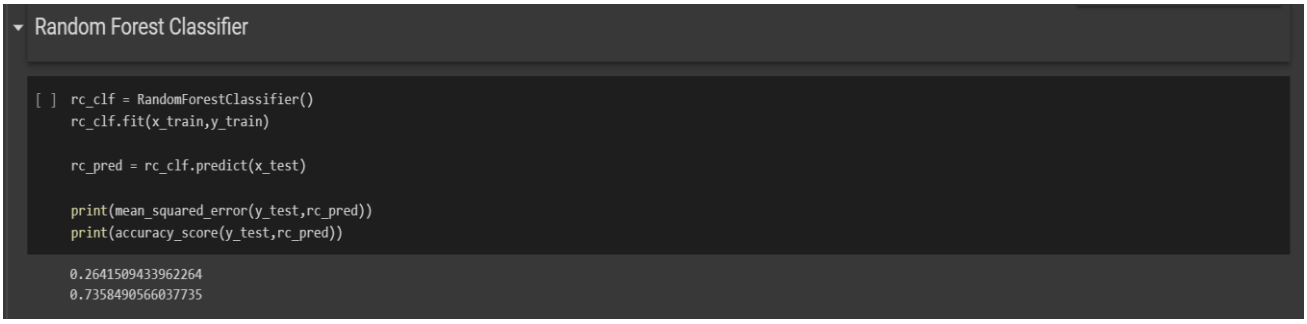
PLOTTING CONFUSION MATRIX:

```
[ ] plot_confusion_matrix(clf, x_test, y_test)
    plt.show()
```

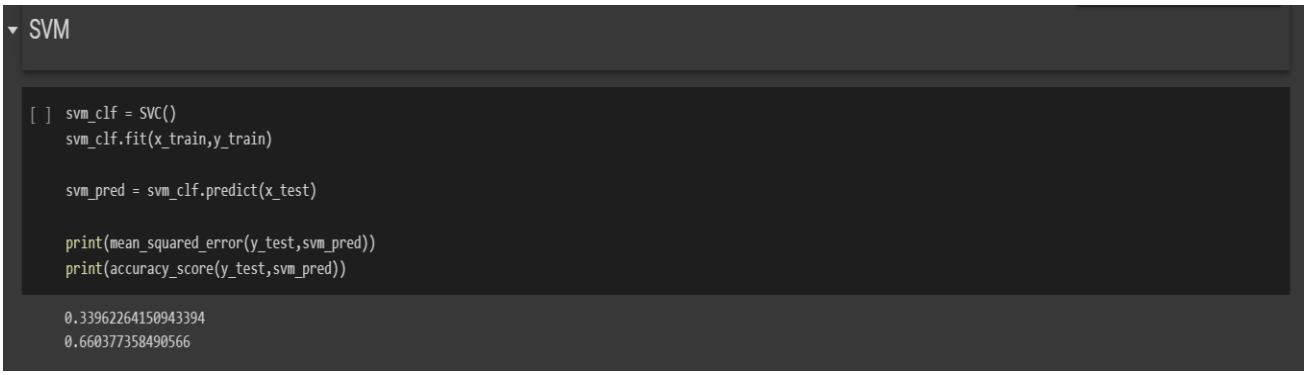


DECISION TREE CLASSIFIER MODEL FUNCTION:

RANDOM FOREST CLASSIFIER :



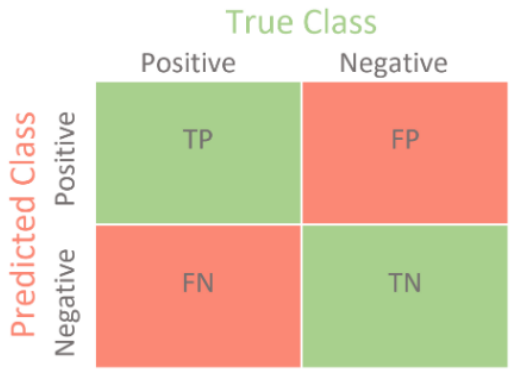
SUPPORT VECTOR MACHINE CLASSIFIER:



2.5 EVALUATION METRICS :

CONFUSION MATRIX:

A confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabelled as the other. The key to the confusion matrix is the number of correct and incorrect predictions are summarized with count values and broken down by each class not just the number of errors made.



ACCURACY:

The accuracy is calculated as:

$$\text{Accuracy} = TP+TN/\ TP+TN+FP+FN$$

Where,

• True Positive (TP) =Observation is positive, and is predicted to be positive.

• False Negative (FN) = Observation is positive, but is predicted negative.

• True Negative (TN) = Observation is negative, and is predicted to be negative.

• False Positive (FP) =Observation is negative, but is predicted positive

The obtained accuracy during training the data after feature selection using backward elimination was 86 % and during testing was 83%. The obtained accuracy during training the data after feature selection using REFCV method was 86 % and during testing was 85 %.

CHAPTER 3

RESULTS AND DISCUSSON

3.1 ADVANTAGES:

- Increased accuracy for effective heart. disease diagnosis.
- Handles roughest(enormous) amount of. data using random forest algorithm and. feature selection.
- Reduce the time complexity of doctors.
- Cost effective for patients.

APPLICATIONS:

Machine Learning can play an essential role in predicting presence/absence of Locomotor disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

FUTURE SCOPE :

Accuracy of the model can be improved by adding more data for training and testing and preferring the dataset which has least null values and not much differing in the values of the data. By adding the average values of the neighbour values would reduce the model from undergoing the underfitting.

This model can also be developed in many other ways like recurrent neural network(RNN),convolution neural network(CNN),deep residual network(DRN),combination of convolution and recurrent neural network(CRNN),support vector machines(SVM),Ensemble method.

3.2 RESULTS AND DISCUSSION:

RESULTS :

This extensive article highlights the importance of computer science, especially data science, to predict heart disease and prevent the seriousness of cardiovascular disease at an early stage. We note how existing datasets give a comprehensive picture by using each dataset variable to represent a biological indicator of cardiac disease. Next, we see how a machine learning approach can use these datasets to train various models to predict a target variable -- either detecting the presence of an ailment or classifying the intensity of risk of heart disease. We also walk through an end-to-end implementation of various models and compare their performance.

DISCUSSIONS :

The main purpose of designing this system is to predict the ten-year risk of future heart disease. We have used Logistic regression as a machine-learning algorithm to train our system and various feature selection algorithms like Backward elimination and Recursive feature elimination. These algorithms are discussed below in detail.

CHAPTER 4

CONCLUSION

CONCLUSION:

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. with 83% accuracy. The model used was Logistic Regression. Further for its enhancement, we can train on models and predict the types of cardiovascular diseases providing recommendations to the users, and also use more enhanced models.

REFERENCE :

[1] A. H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018.

[2] M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".

[3] K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available: https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machinelearning-36f00f3edb2c. [Accessed 2 March 2020].

[4] [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci#heart.csv.. [Accessed 05 December 2019].

[5] M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach".

APPENDIX :

CODE :

https://github.com/19pa1a04e7/Heart-Disease-Prediction