HEART DISEASE PREDICTION USING MACHINE LEARNING

Project report submitted in partial fulfilment of Requirement for the award of

Degree of

Bachelor of Science

By

MEHEK BANO

Under the Guidance of

RAGHAVA SWAMY

Lecturer in Computer Science

DEPARTMENT OF COMPUTER SCIENCE

D.R VS Krishna Govt. Degree College (Autonomous)

VISAKHAPATNAM

CERTIFICATE

This is to certify that the project report “HEART DISEASE PREDICTION “

Submitted by

MEHEK BANO

Record of the work carried out by her in partial fulfilment of the requirement for the award of the degree of bachelor of science (Computer science), as prescribed by the DR. V.S. Krishna Govt. Degree & P.G. college(A) in the Academic Year 2020-2023

GUIDE

RAGHAVA SWAMY

HEAD OF THE DEPARTMENT

ACKNOWLEDGEMENT

We take this opportunity to thank those who helped us immensely throughout this project and whose efforts we will always appreciate and remember.

We thank to RAGHAVA SWAMY for their valuable suggestions, their indefatigable efforts to review our work, inspiring us and helping us in all possible ways at any time.

We would also like to extend our sincere appreciation to all staff members of computer Department, without their help and co-operation the Project report would not have been a success.

Last but not the least we express our extreme gratitude to the Almighty, without whose blessing nothing is possible.

DECLARATION

I here by declare that this project report entitled “HEART DISEASE PREDICTION” is the result of original work done by me and to the best of my knowledge similar work has not been submitted previously to any other university or published any time before. This project is submitted on the partial fulfilment or the requirement for the award of degree of Bachelor of Science.

MEHEK BANO

CONTENTS

SI.NO **TITLE**

HEART DISEASE PREDICTION USING MACHINE LEARNING

Abstract

Day by day the cases of heart diseases are increasing at a rapid rate and it's very important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e., it should be performed precisely and efficiently. This project is mainly focused on which patient is more likely to have heart disease based on various medical attributes. Machine learning can play an essential role in predicting presence/absence of locomotors disorders, heart diseases and more. such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt to their diagnosis and dealing per patient basis. We work on predicting possible heart diseases in people using machine learning algorithms.

Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, accurate detection of heart diseases in all cases and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise.

In this study, a tentative design of a cloud-based heart disease prediction system had been proposed to detect impending heart disease using Machine learning techniques. For the accurate detection of the heart disease, an efficient machine learning technique should be used which had been derived from a distinctive analysis among several machine learning algorithms in a Java Based Open Access Data Mining Platform, WEKA.

The proposed algorithm was validated using two widely used open-access database, where 10-fold cross-validation is applied in order to analyze the performance of heart disease detection. An accuracy level of 97.53% accuracy was found from the SVM algorithm along with sensitivity and specificity of 97.50% and 94.94%respectively.

Moreover, to monitor the heart disease patient round-the-clock by his/her caretaker/doctor, a real-time patient monitoring system was developed and presented using Arduino, capable of sensing some real-time parameters such as body temperature, blood pressure, humidity, heartbeat. The developed system can transmit the recorded data to a central server which are updated every 10 seconds.

As a result, the doctors can visualize the patient's real-time sensor data by using the application and start live video streaming if instant medication is required. Another important feature of the proposed system was that as soon as any real-time parameter of the patient exceeds the threshold, the prescribed doctor is notified at once through GSM technology.

Keywords:

Data Mining, Machine Learning, It (Internet of Things), Patient Monitoring System, Heart Disease Detection and Prediction

Objective

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either it is 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 everyday 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.

- **Technology:** Machine learning using python

The importance and advantages of the application of Machine learning based heart disease detection and prediction system were discussed in several research findings. The application of artificial intelligence in disease detection system especially the cardiac disease system detection improves the performance of other existing widely used models like models provided by American College of Cardiology/American Heart Association (ACC/AHA) models in CVD detection and prediction.

The possibility and related matters of providing advanced services of a human health management system were analysed by Zhao, Wang, and Anahita, in 2011 and they had also given a research direction of medical technology on Iota . Many types of health-related sensors and technologies were analysed by them. They identified some issues which need to be solved. The home monitoring system and decision support system was schemed by Chanchiang and German in 2014.

This system contributed to home monitoring, diagnosis, medical prescriptions, medical treatment, rehabilitation and development of his patients with Parkinson's disease. Wireless Health Monitoring System (WHMS) has attracted considerable attention from the research community and industry over the last decade. Improvement of several Machine learning algorithms and classifier performances like weighted associative classifier were reported in the detection of cardiac abnormalities .

The elderly patients monitoring from indoor or outdoor locations had been presented by a real-time mobile healthcare system in A signal sensor and a smartphone were the primary components of the system. The bio-signal sensor data was transmitted to an intelligent server via GPRS/UMTS network for data collection. The system could perform in monitoring the mobility, vital signs, location, and condition of the elderly patient from a distant location. A fully functional wireless body area network (WBAN) system had been proposed in.

The designed system used medical bands to obtain physiological data from sensors. The author had chosen some medical bands in order to abate the interruption between the sensors and other existing devices. To increase the operating extent, the multi-hopping technique had been implemented and a medical gateway wireless board had been used in this regard

Chapter-1
Introduction

Introduction:

The most crucial part of our body is Heart. It's a muscular organ which placed directly behind and slightly left of breastbone. Heart Disease causes highest number of deaths globally, with approximately around 17.9 million people died from it every year which means around 31% of deaths are from the heart disease as per the WHO (World Health Organization). Heart disease are also called as Cardiovascular Disease which are group of complication of the blood vessels and heart which include cerebrovascular disease, rheumatic heart disease and some other heart conditions. Four out of five heart disease deaths are from the strokes and heart attacks.

Heart Disease is the most life-threatening disease in the world nowadays. Therefore, heart disease should be predicted at their early stage and healthy lifestyle are ways to prevent them. People who are at risk of heart disease may have symptoms like high blood pressure, obesity, cholesterol, diabetes, age, etc.

As there is recent improvement in medical health care is observed. The health care system has collected massive amount of data about heart disease and they have all those data and created datasets which consist of different medical parameter or features such as age, sex, blood pressure, cholesterol, chest type and so on, etc.

Datasets consist of around 13 to 15 different medical parameters. These datasets are now available for analysis and to extract crucial information from it. So, we can predict the heart disease at their early stage by applying machine learning algorithms on this massive amount of data to extract features (information/medical parameters) that we will extract from datasets. Various machine learning techniques like logistic regression, naïve bays, support vector machine, k nearest neighbor (ken), etc.

we can use for predicting heart disease means to classify whether a person is having cardiovascular disease or not, after applying them on the features extraction from datasets. Different algorithms will give different accuracy, so after comparing among them we can find the best algorithm which predicts heart disease with highest accuracy.

The main objective of our project is to enhance efficiency for predicting heart disease rate. The work proposed in this paper focus mainly on various data mining practices that are employed in heart disease prediction. Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to heart can cause distress in other parts of body. Any sort of disturbance to normal functioning of the heart can be classified as a Heart disease.

In today's contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension According to the World Health Organization more than 10 million die due to Heart diseases every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases.

1.1 EXISTING SYSTEM

The World Health Organization (WHO) has estimated that 12million deaths occur worldwide, every year due to the heart diseases. About 25% deaths in the age group of 25-69 year occur because of heart diseases. In urban areas, 32.8%. Deaths occur because of heart ailments, while this percentage in rural areas is 22.9.Over 80% of deaths in world are because of Heart disease. WHO estimated by 2030, almost 23.6 million. People will die due to Heart disease. The diagnosis of diseases is a significant and tedious task in medicine. Treatment of the said disease

is quite high and not affordable by most of the patients particularly in India.

1.2 PROPOSED SYSTEM

In this system, we are implementing effective heart attack prediction system using Machine-learning algorithm. We can give the input as in CSV file or manual entry to the system. After taking input, the algorithms apply on that input to algorithms. After accessing data set the operation is performed and effective heart attack level is produced. The proposed system will add some more parameters significant to heart attack with their weight, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

The heart is a kind of muscular organ which pumps blood into the body and is the central part of the body's cardiovascular system which also contains lungs. Cardiovascular system also comprises a network of blood vessels, for example, veins, arteries, and capillaries. These blood vessels deliver blood all over the body.

Abnormalities in normal blood flow from the heart cause several types of heart diseases which are commonly known as cardiovascular diseases (CVD). Heart diseases are the main reasons for death worldwide. According to the survey of the World Health Organization (WHO), 17.5 million total global deaths occur because of heart attacks and strokes. More than 75% of deaths from cardiovascular diseases occur mostly in middle-income and low-income countries. Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack.

Therefore, detection of cardiac abnormalities at the early stage and tools for the prediction of heart diseases can save a lot of life and help doctors to design an effective treatment plan which ultimately reduces the mortality rate due to cardiovascular diseases. Due to the development of advance healthcare systems, lots of patient data are nowadays available (i.e., Big Data in Electronic Health Record System) which can be used for designing predictive models for cardiovascular diseases.

Data mining or machine learning is a discovery method for analyzing big data from an assorted perspective and encapsulating it into useful information. “Data Mining is a non-trivial extraction of implicit, previously unknown and potentially useful information about data” [2]. Nowadays, a huge amount of data pertaining to disease diagnosis, patients etc. are generated by healthcare industries.

Data mining provides a number of techniques which discover hidden patterns or similarities from data. Therefore, in this paper, a machine learning algorithm is proposed for the implementation of a heart disease prediction system which was validated on two open access heart disease prediction datasets.

Another contribution of this paper is the presentation of a cardiac patient monitoring system using the concept of Internet of Things (Iota) with different physiological signal sensors and Arduino microcontroller.

Sensor networks are currently using the Internet of Things (Iota) technology to collect, analyze and passing of information from one node to another. Iota is a recently used rapidly expanding technology, where multiple sensors/data collectors can sense, share information and communicate over a private network, Internet Protocol (IP) or public networks.

The sensors collect the data after a specific time, analyze it and use it to initiate the required action, and provide an intelligent cloud-based network for analysis, planning and decision making. The products that are developed with Iota such as embedded technology, allow to exchange information among each other nodes or the Internet and it was assessed that about 8 to 50 billion devices will be connected by 2020

Chapter-2
Methodology

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this paper are K nearest neighbors (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease.

This paperwork includes examining the journals, published paper and the data of cardiovascular disease of the recent times. Methodology gives a framework for the proposed model. The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users.

The proposed methodology includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the pre-processing stage where we explore the data.

Data pre-processing deals with the missing values, cleaning of data and normalization depending on algorithms used. After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier.

Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective Heart Disease Prediction System (EHDPS) has been developed using different

classifiers. This model uses 13 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction. First of all, the collection of data and selection of relevant attributes are the initial steps in this study. After that, the relevant data is pre-processed into the required format. The given data is then separated into two categories: training and testing datasets.

Methodology:

In this study, an efficient machine earning algorithm was chosen from some available algorithms in a Java-based open access data mining platform (WEKA) to detect the presence or to decide the probability of having heart diseases form a large dataset. Then a continuous cardiac monitoring system design has been proposed using Arduino based microcontroller system. The step by step design approaches of the proposed system and the workflow of the complete system have been mentioned below:

➢ Collection and selection of different heart disease datasets in order to train various machine learning algorithms.

➢ Comparison of various data mining algorithm's accuracy and performance in predicting heart disease.

➢ Selection of the best algorithm from performance characteristics of the models to develop a cloud-based intelligent heart disease prediction mobile application.

➢ Storing of doctor and patient information following registration of patients and doctors separately through the application in a cloud-based server for analysis.

➢ A full-fledged tentative design of an android application with respective criteria has been shown in the Analysis and Results section.

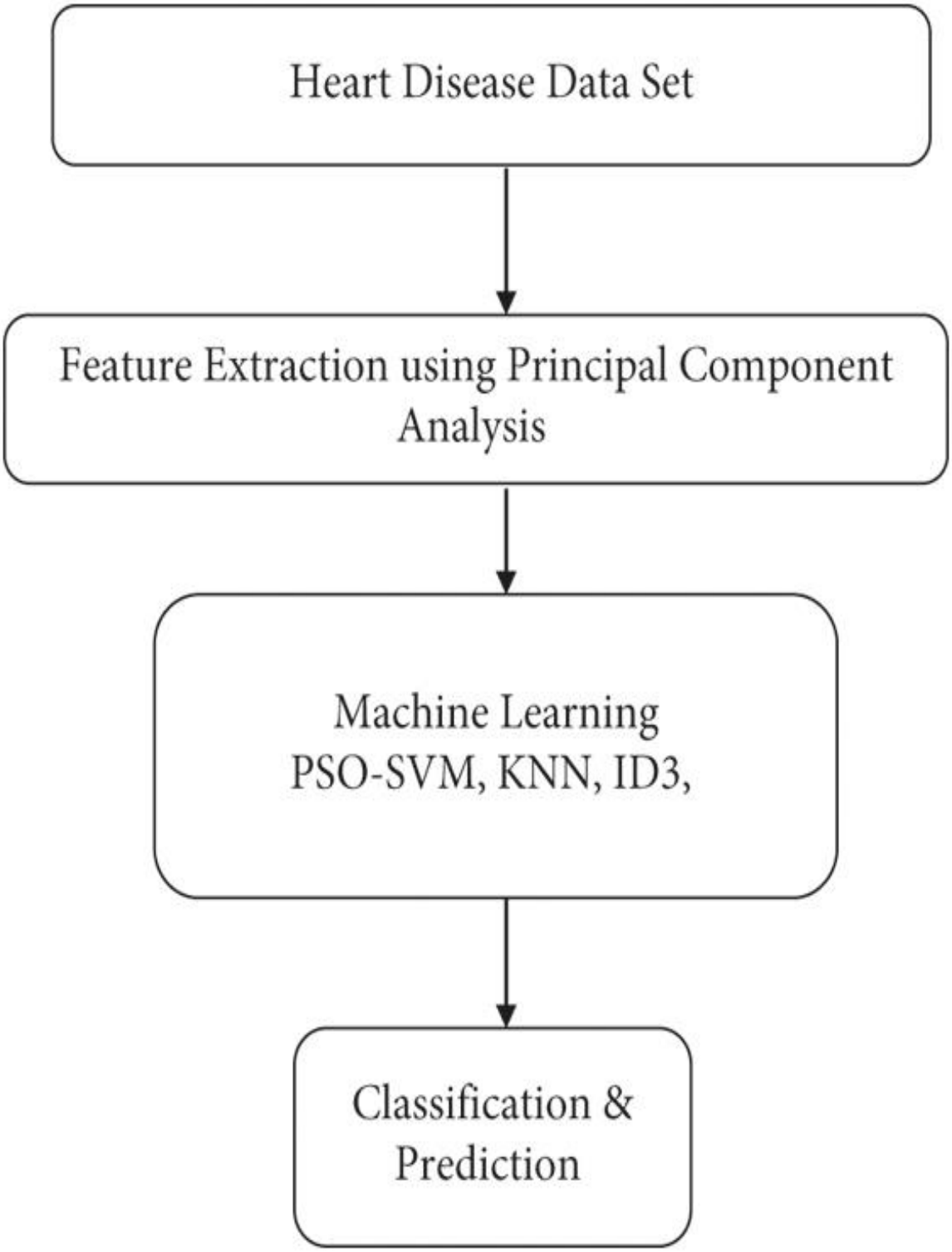
➢ On the application interfaces, the patient can input all parameters of heart disease manually and they can input sensor data like heartbeat using a specific button. Hence, they can predict whether they have heart disease or not.

➢ After that, the patient can send the result to any doctor registered in the application in a report format by typing doctor's ID in the search option. Furthermore, He/she can start a live video call with the doctor if any critical situation will arise.

➢ A primary wireless patient health monitoring system is developed using Arduino with temperature, humidity, and heartbeat sensors in order to collect real-time patient physiological data and to detect the critical situation of the patient.

➢ All the sensor real-time data are stored and updated on the cloud server database automatically after a specific time (every 10 sec) and if any unwanted value of the sensors is encountered then prescribed doctor from any location in the world is notified via mobile SMS using GSM module. In the case of ICU patients, an alarm system exists using a buzzer to alert the caretaker or nurses immediately.

➢ After receiving an alert message containing the current data of the sensors in his phone, the doctor will log in to the application with a view to checking the patient's previous physiological data and consulting appropriate medication via live video streaming


Heart Disease Data Set
Feature Extraction using Principal Component Analysis
Machine Learning
PSO-SVM, KNN, ID3,
Classification & Prediction

Chapter -3
Machine algorithms

Machine algorithms:

Machine Learning algorithms are the programs that can learn the hidden patterns from the data, predict the output, and improve the performance from experiences on their own. Different algorithms can be used in machine learning for different tasks, such as simple linear regression that can be used for prediction problems like stock market prediction, and the KNN algorithm can be used for classification problems**.**

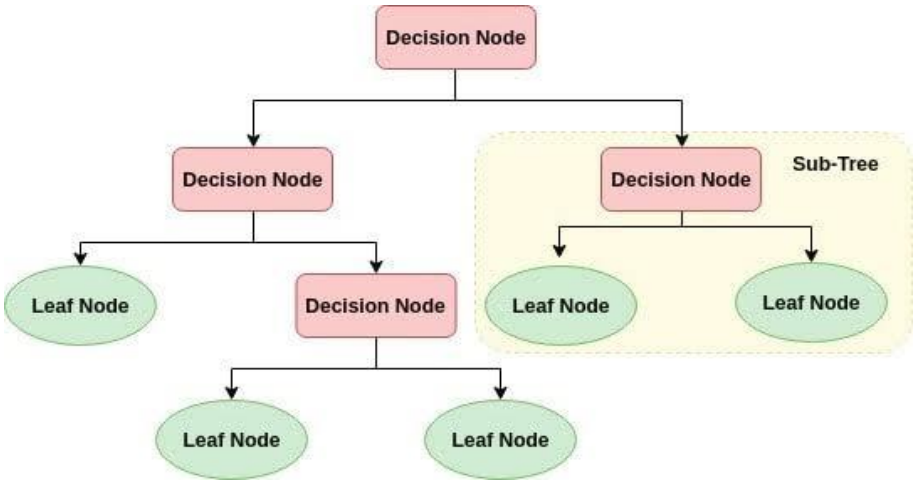
In this topic, we will see the overview of some popular and most commonly used machine learning algorithms along with their use cases and categories.

3.1 Decision tree algorithm:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

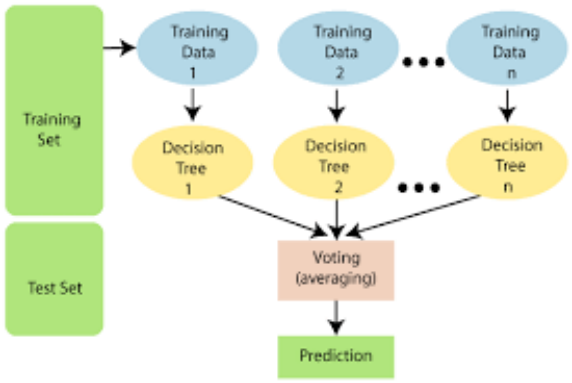


3.2 Random Forest Algorithm:

Random forest is the supervised learning algorithm that can be used for both classification and regression problems in machine learning. It is an ensemble learning technique that provides the predictions by combining the multiple classifiers and improve the performance of the model.

It contains multiple decision trees for subsets of the given dataset, and find the average to improve the predictive accuracy of the model. A random-forest should contain 64-128 trees. The greater number of trees leads to higher accuracy of the algorithm. To classify a new dataset or object, each tree gives the classification result and based on the majority votes, the algorithm predicts the final output. Random forest is a fast algorithm, and can efficiently deal with the missing & incorrect data.

Random Forest is a classifier that contains a number of decision trees on



various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

3.3 Naïve Bayes Algorithm:

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in *text classification* that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.** Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles**.** The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without

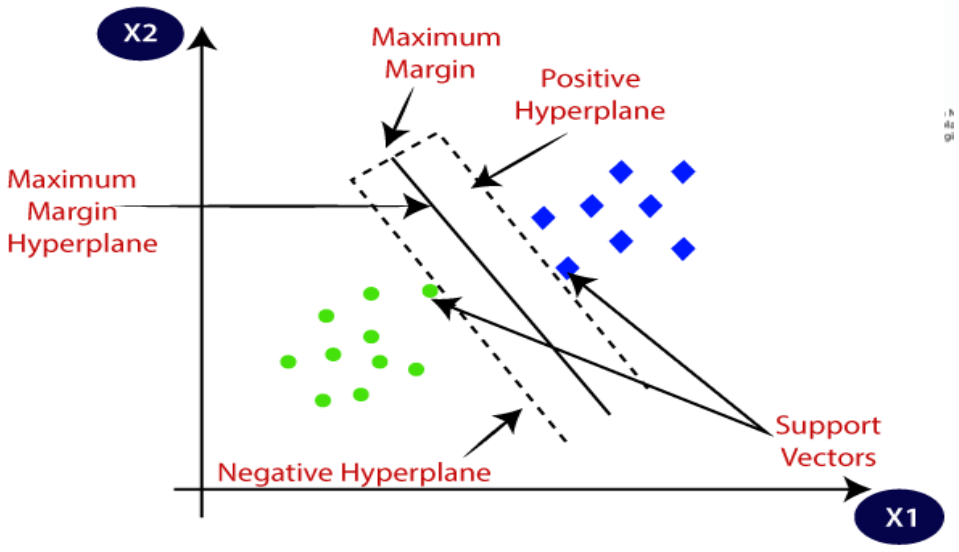
depending on each other. Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

3.4 Support Vector Machine Algorithm:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper plane.

SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper plane:



Chapter-4
Result analysis

Result analysis:

From these results we can see that although most of the researchers are using different algorithms such as SVC, Decision tree for the detection of patients diagnosed with Heart disease, KNN, Random Forest Classifier and Logistic regression yield a better result to out rule them.

The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by KNN and Logistic Regression are equal to 88.5% which is greater or almost equal to accuracies obtained from previous researches.

So, we summarize that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease.

Chapter-5
coding

CODING:

After an extensive introduction, we can finally perform heart disease detection in Python using a hands-on tutorial that implements several machine learning algorithms, primary exploratory data analysis, and inbuilt data analysis techniques for feature importance. To test out how different models perform on the task, we will use the UCI Heart Disease dataset having 14 columns and over 300 samples. You can find the dataset on.

Heart Disease Prediction UCI Dataset and download the CSV file.

```
1   import jumpy as no
2   import pandas as pd
3   import matplotlib. pilot as plot
4
5   from sclera. Model selection import train test split, Rando mazed Search CV
6   from sclera. pre-processing import Standardisable
7   from sclera. neighbours import Neighbour's Classifier
8   from sclera. sum import SVC
9   from sclera. tree import Decision Tree Classifier
10  from sclera. ensemble import
```

Random Forest Classifier, Gradient Boosting Classifier

First, we must load Python's necessary libraries crucial for data science. We use Sci-kit Learn to implement the k neighbours' classifier, SVM, Random Forest, and Decision Tree.

Implementations from Sci-kit learn are well optimized with the option for tuning. It also lets you perform a Grid Search to find

the best set of hyper parameters for a model.

The search is performed using module in SC learn.Prediction

of Heart Disease using Machine Learning Algorithms

Next, we load the dataset CSV file using Pandas. The Pandas package efficiently loads and access large datasets directly from an external file and performs data processing, feature engineering, and data analysis. In our case, we will use the reeds function to load our file.

Heart Disease Prediction using Machine Learning Python

```
1 dataset = pd.read_csv('/content/heart.csv')
2 dataset.info()
```

The last line in the code above will list the columns, datatype, and the number of values, as shown below.

```
# Create another figure plt.figure(fig size=(10, 8))

# Scatter with positive examples plt.scatter(df.age[df.target==1], df.thalach[df.target==1], c="salmon")

# Scatter with negative examples plt.scatter(df.age[df.target==0], df.thalach[df.target==0], c="lightblue")

# Add some helpful info
```

```
plt.title("Heart Disease in function of Age and Max Heart Rate") plt.xlabel("Age") plt.ylabel("Max Heart Rate")

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


 from sklearn.metrics import accuracy_score, confusion_matrix, classification_report


def print_score(clf, X_train, y_train, X_test, y_test, train=True):    if train:

      pred = clf.predict(X_train)

      clf_report = pd.DataFrame(classification_report(y_train, pred, output_dict=True))       print("Train

Result:\n================================================")

      print(f"Accuracy Score: {accuracy_score(y_train, pred) * 100:.2f}%") print("_____") print(f"CLASSIFICATION REPORT:\n{clf_report}") print("_____") print(confusion Matrix: \n {confusion_matrix(y_train, pred)}\n")


  elif train==False:          pred = clf.predict(X_test)          clf_report = pd.DataFrame(classification report(test, pred,


output_dict=True))       print("Test

Result:\n================================================") print(f"Accuracy Score: {accuracy_score(y_test, pred) * 100:.2f}%") print("_____") print(f"CLASSIFICATION REPORT:\n{clf_report}") print("_____")
```

print(f"Confusion Matrix: \n {confusion_matrix(y_test, pred)}\n") import numpy as np import pandas as pd from scipy.stats import mode import matplotlib.pyplot as plt import seaborn as sns

from sklearn.preprocessing import LabelEncoder

from sklearn.model_selection import train_test_split, cross_val_score from sklearn.svm import SVC

from sklearn.naive_bayes import GaussianNB from sklearn.ensemble import RandomForestClassifier from sklearn.metrics import accuracy_score, confusion_matrix

%matplotlib inline

Chapter-6
screenshot

SCREENSHOT:

Predict the presence of heart disease

Patient I’d

Patient Name

Age

Sex

Chest Pain

Test blood pressure in mm Hg

Serum cholesterol in mg/do

Fasting blood sugar greater than 120 mg/dl Resting electrocardiographic results

Maximum Heart rate achieved

The slope of the peak exercise ST segment

Chapter-7
output

OUTPUT:

RangeIndex: 303 entries, 0 to 302 Data columns (total 14 columns):

#	columns	Non-Null Count	Dtype
1	age	303 non-null	int64
2	sex	303 non-null	int64
3	cp	303 non-null	int64
4	trestbps	303 non-null	int64
5	chol	303 non-null	int64
6	fbs	303 non-null	int64
7	restecg	303 non-null	int64
8	thalach	303 non -null	int64
9	exang	303 non-null	int64
10	oldpeak	303 non-null	float64
11	slope	303 non-null	int64
12	ca	303 non-null	int64
13	thal	303 non-null	int64
14	target	303 non-null	int64

dtypes: float64(1), int64(13)

Heart Disease Prediction using Machine Learning Python Before moving, we must perform exploratory data analysis (EDA) on the dataset we just loaded. Looking at the description of the columns on Kaggle, we can infer that some of them are categorical variables. The basics of

feature engineering and data science tell us that such columns need to be encoded to avoid unintentional bias. For example, as shown below, columns like chest pain (cp), thal, and ca need to be one-hot encoded in addition to others.


```
1       print(np.unique(dataset[[‘cp']].values))
2       print(np.unique(dataset[[‘thal']].values))
3       print(np.unique(dataset[[‘ca']].values))
```

[0 1 2 3]

[0 1 2 3]

[0 1 2 3 4]

Moreover, the columns of age, cholesterol (chol), Rest BP (trestbps), thalach, and oldpeak need to be normalized.

Train: (3936, 132), (3936,)

Test: (984, 132), (984,)

==========

age : [63 37 41 56 57 44 52 54 48 49 64 58 50 66 43 69 59 42 61 40 71 51 65 53

46 45 39 47 62 34 35 29 55 60 67 68 74 76 70 38 77]

========== sex : [1 0]

========== cp : [3 2 1 0]

==========

trestbps : [145 130 120 140 172 150 110 135 160 105 125 142 155 104 138 128 108 134

122 115 118 100 124 94 112 102 152 101 132 148 178 129 180 136 126 106

156 170 146 117 200 165 174 192 144 123 154 114 164]

==========

chol : [233 250 204 236 354 192 294 263 199 168 239 275 266 211 283 219 340 226

247 234 243 302 212 175 417 197 198 177 273 213 304 232 269 360 308 245

208 264 321 325 235 257 216 256 231 141 252 201 222 260 182 303 265 309

186 203 183 220 209 258 227 261 221 205 240 318 298 564 277 214 248 255

207 223 288 160 394 315 246 244 270 195 196 254 126 313 262 215 193 271

268 267 210 295 306 178 242 180 228 149 278 253 342 157 286 229 284 224

206 167 230 335 276 353 225

330 290 172 305 188 282 185 326 274 164 307

249 341 407 217 174 281 289 322 299 300 293 184 409 259 200 327 237 218

319 166 311 169 187 176 241 131]

========== fbs : [1 0]

==========

restecg : [0 1 2]

==========

thalach : [150 187 172 178 163 148 153 173 162 174 160 139 171 144 158 114 151 161

179 137 157 123 152 168 140 188 125 170 165 142 180 143 182 156 115 149

146 175 186 185 159 130 190 132 147 154 202 166 164 184 122 169 138 111

145 194 131 133 155 167 192 121 96 126 105 181 116 108 129 120 112 128 109 113 99 177 141 136 97 127 103 124 88 195 106 95 117 71 118 134 90]

==========

exang : [0 1]

==========

oldpeak : [2.3 3.5 1.4 0.8 0.6 0.4 1.3 0. 0.5 1.6 1.2 0.2 1.8 1. 2.6 1.5 3. 2.4

0.1 1.9 4.2 1.1 2. 0.7 0.3 0.9 3.6 3.1 3.2 2.5 2.2 2.8 3.4 6.2 4. 5.6

2.9 2.1 3.8 4.4]

==========

slope : [0 2 1]

==========

ca : [0 2 1 3 4]

==========

thal : [1 2 3 0]

========== target : [1 0]

Chapter-8
Testing algorithms

Testing algorithms:

After running the 3 algorithms on the 4 data sets (600, 800, 1000 and 1200lines), we test the accuracy of each algorithm on the different data sets. Table 1 illustrates how we calculated the accuracy from the confusion matrix (we took the case of the last data set which contains 1200 lines), on the 3 algorithms. Now we can display the accuracy of each algorithm on the 4 data sets, in

Table 1. The confusion matrix with accuracy

Neural network dataset(1200lines)		SVM dataset (1200 lines)		KNN dataset (1200 lines)	
Sick	Not sick	Sick	Not sick	Sick	Not sick
94	8	90	11	84	18
6	92	9	90	11	87
93%		90%		85.5%	

Table 1

Fig. 1 shows a comparison between the 3 algorithms in terms of accuracy and stability against changes in the data sets.

Choosing the best algorithm After analyzing the results found previously. We find that neural networks is the best algorithm in our study, it is always stable in its results, and gives the best accuracy

Table 2 The accuracy of different algorithms

Algorithm	Data set (600 lines)	Data set (800 lines)	Data set (1000 lines)	Data set (1200 lines)
Neural network	91%	92%	92%	93%
SVM	89%	89%	90%	88%
KNN	85%	85%	86%	85%

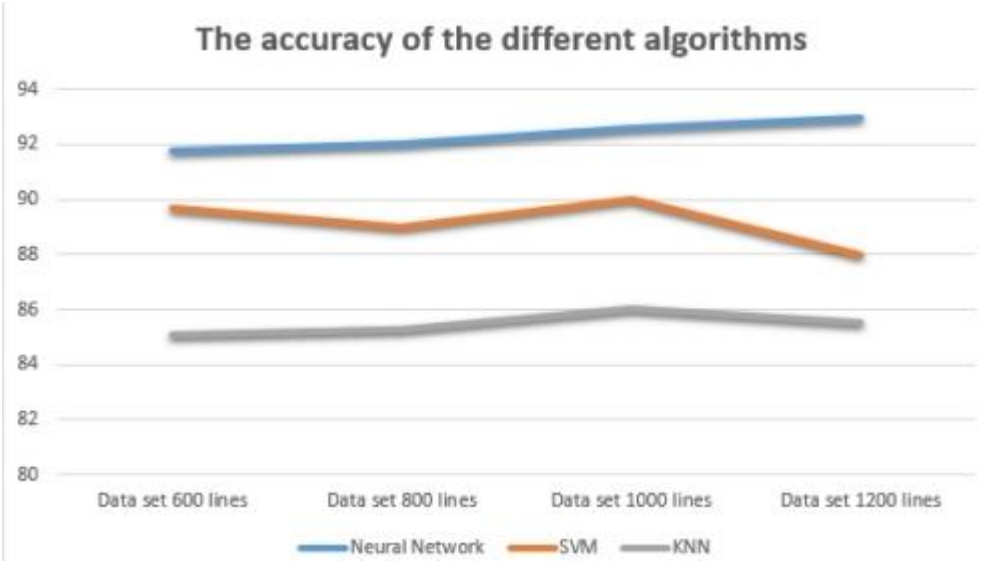


Fig. 1 The accuracy of different algorithms

Chapter -9
Conclusion

Conclusion:

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases

effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases.

The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society.

This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the seven different machine learning algorithms used to measure the performance are SVM, Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting applied on the data set.

Chapter -10
Reference

Reference:

[1] Sony J, Ansari U, Sharma D & Sony S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

[2] Dungaree C S & Pate S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shined R, Arjun S, Patel P & Wag mare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qajar U & Jived M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (I-Society 2014) (pp. 259-64). IEEE.