Prediction Of Heart Diseases Using Machine Learning

Guide- Dr Pavan Kumar

Team Members

Manav Nair - 19BOE10007

Shirish Waghmode - 19MIM10026

Vartika Pandey - 19MIM10053

Ashwini Darade - 19BCE10311

Ananya Saxena - 19MIM10018

Anant Kumar Pandey - 19MIM10013

Prince -19MIM10007

Ayush Joshi - 19BCE10294

INTRODUCTION

- Cardiovascular diseases kill approximately 17 million people globally every year, and they mainly exhibit as myocardial infarctions and heart failures. Heart failure (HF) occurs when the heart cannot pump enough blood to meet the needs of the body.

- A study shows that from 1990 to 2016 the death rate due to heart diseases have increased around 34 percent from 155.7 to 209.1 deaths per one lakh population in India.

- Machine learning, in particular, can predict patients' survival from their data and can individuate the most important features among those included in their medical records.

MOTIVATION

- **Healthcare organisations are required to provide quality service at affordable costs,Qualityservice implies diagnosing patients correctly and administeringtreatments that are effective.Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests.**

- **They can achieve these results by employing appropriate computer-based information and/ordecision support systems. Most hospitals today employ some sort of hospital informationsystems to manage their healthcare or patient data. Unfortunately, these data are rarely usedto support clinical decision making. There is a wealth of hidden information in this data that islargely untapped. This raises an important question: "*How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions*?"**

- **This is the main motivation for this research.**

OBJECTIVE

- **The main objective of this research is to develop a heart prediction system.The systemcan discover and extract hidden knowledge associatedwith diseases from a historicalheart data set.**
- **Heart disease prediction system aims to exploit data mining techniques on medical data sets to assist in the prediction of heart diseases.**

SPECIFIC OBJECTIVE

- Provides a new approach to concealed patterns in the data.
- Helps avoid human biases.
- To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.
- Reduce the cost of medical tests.

TOPIC OF WORK

- The goal of our heart disease prediction project is to determine if a patient should diagnosed with heart disease or not, which is abinary outcome, so: Positive result = 1 the patient will be diagnosed with heart disease.

- The Health Prediction system is an end user support and onlineconsultation project.The system allows user to share their symptoms and issues. It then processes users symptoms to check for various illness that could be associated with it.

- For the disease prediction, we use K-Nearest Neighbor (KNN) and Convolutional neural network (CNN) machine learning algorithm fora ccurate prediction of disease. For disease prediction required disease symptoms dataset.

EXISTING WORK

- **The algorithm to diagnose HF in a non-acute setting is the following. First the probability of HF based on prior clinical history of the patient, the presenting symptoms, physical examination, and resting ECG is estimated. The process of diagnosis of HF can be:**

  (i) less time consuming

  (ii) supported, and

  (iii) performed with the same accuracy by the applications of machine learning techniques on the available data.

continuation;

Machine learning techniques have been applied to classify HF subtypes. This discovery has the potential to impact on clinical practice, becoming a new supporting tool for physicians when predicting if a heart failure patient will survive or not. Indeed, medical doctors aiming at understanding if a patient will survive after heart failure may focus mainly on serum creatinine and ejection fraction. Machine learning applied to medical records, in particular, can be an effective tool both to predict the survival of each patient having heart failure symptoms, and to detect the most important clinical features (or risk factors) that may lead to heart failure.

CONCLUSION

This project provides the deep insight into machine learning techniques for classification of heart diseases. The role of classifier is crucial in healthcare industry so that the results can be used for predicting the treatment which can be provided to patients. The existing techniques are studied and compared for finding the efficient and accurate systems. Machine learning techniques significantly improves accuracy of cardiovascular risk prediction through which patients can be identified during an early stage of disease and can be benefitted by preventive treatment. It can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases

CONTINUATION WITH ZEROTH REVIEW PPT

CONTINUATION

The project involved analysis of the heart disease patient dataset with proper data processing. Then, different models were trained and and predictions are made with different algorithms KNN, Decision Tree, Random Forest,SVM,Logistic Regression etc This is the jupyter notebook code and dataset I've used for my Kaggle kernel 'Binary Classification with Sklearn and Keras'

We've used a variety of Machine Learning algorithms, implemented in Python, to predict the presence of heart disease in a patient. This is a classification problem, with input features as a variety of parameters, and the target variable as a binary variable, predicting whether heart disease is present or not.

Machine Learning algorithms used:

- Logistic Regression (Scikit-learn)
- Naive Bayes (Scikit-learn)
- Support Vector Machine (Linear) (Scikit-learn)
- K-Nearest Neighbours (Scikit-learn)
- Decision Tree (Scikit-learn)
- Random Forest (Scikit-learn)
- Accuracy achieved: 95%

Literature review

1. Several approaches have been performed on this popular dataset. The accuracy obtained by all the approaches is more with time computations.
2. Different levels of accuracy have been attained using various data mining techniques.
3. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification.
4. The research concludes that accuracy obtained by Decision Tree was highest.
5. Researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success.

MODULE REVIEW

MODULE NUMBER	MODULE NAME
1	Data fetching from UCI Preprocessing of data
2	Normalization/scaling
3	Feature extraction
4	Feature selection
5	Splitting dataset into training and testing phase
6	Applying model on X train and Y train

CONTINUATION:

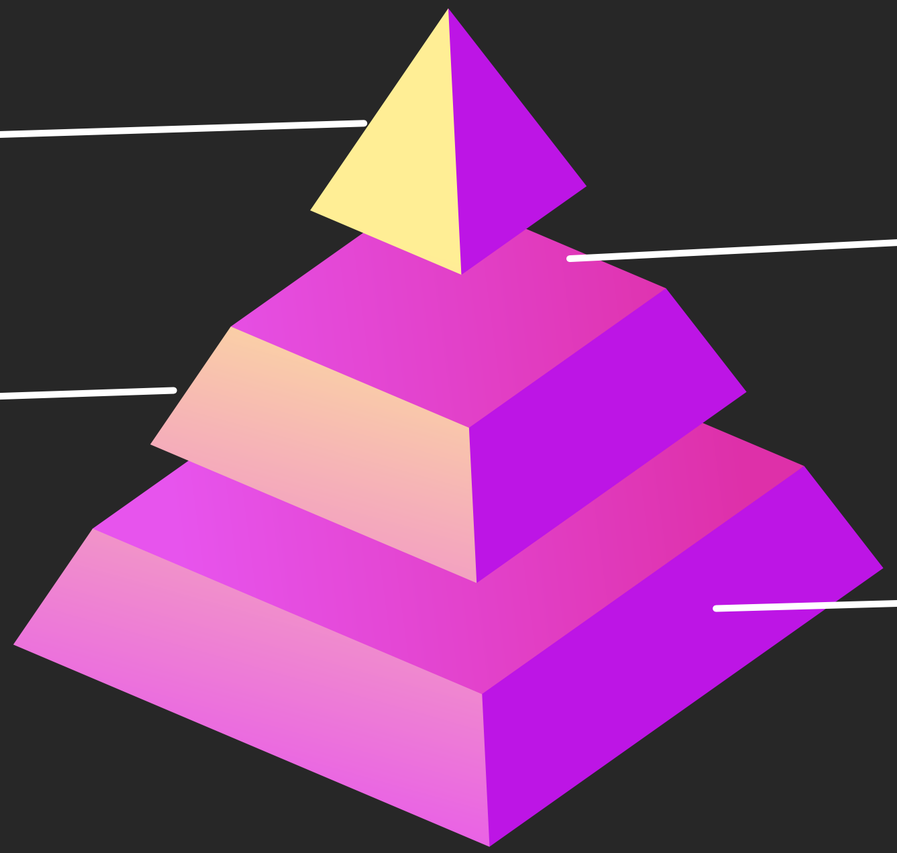
MODULE NUMBER	MODULE NAME
7	Calculate the prediction labels
8	Calculate the confusion matrix

MODULE WORK FLOW EXPLANATION

1. **Data analysis this is nothing but collecting, measuring and anlyzing accurate insights for research.**

2. Data processing

Before starting the application of machine learning algorithm we prepare the data to be implemented and this phase is achieved by feature selection



3. Applying classifiers and regression algorithm

In this we apply the selected algorithm (SVm, KNN, neural networks) and analyze

4. Testing and comparing each algo result

In this we test the accuracy of each algorithm on the different data sets.and the illustrates how we calculate the accuracy from the confusion matrix..

MODULE WORK FLOW EXPLANATION

1. Data Fetching from UCI-

Go to the UCI ML repository to retrieve the data. Click on the Data Set Description link. This opens a page of valuable information about the data set, including source material, publications that use the data, column names, and more.

2.Preprocessing of data - A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.It involves below steps:

Getting the dataset - To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset.

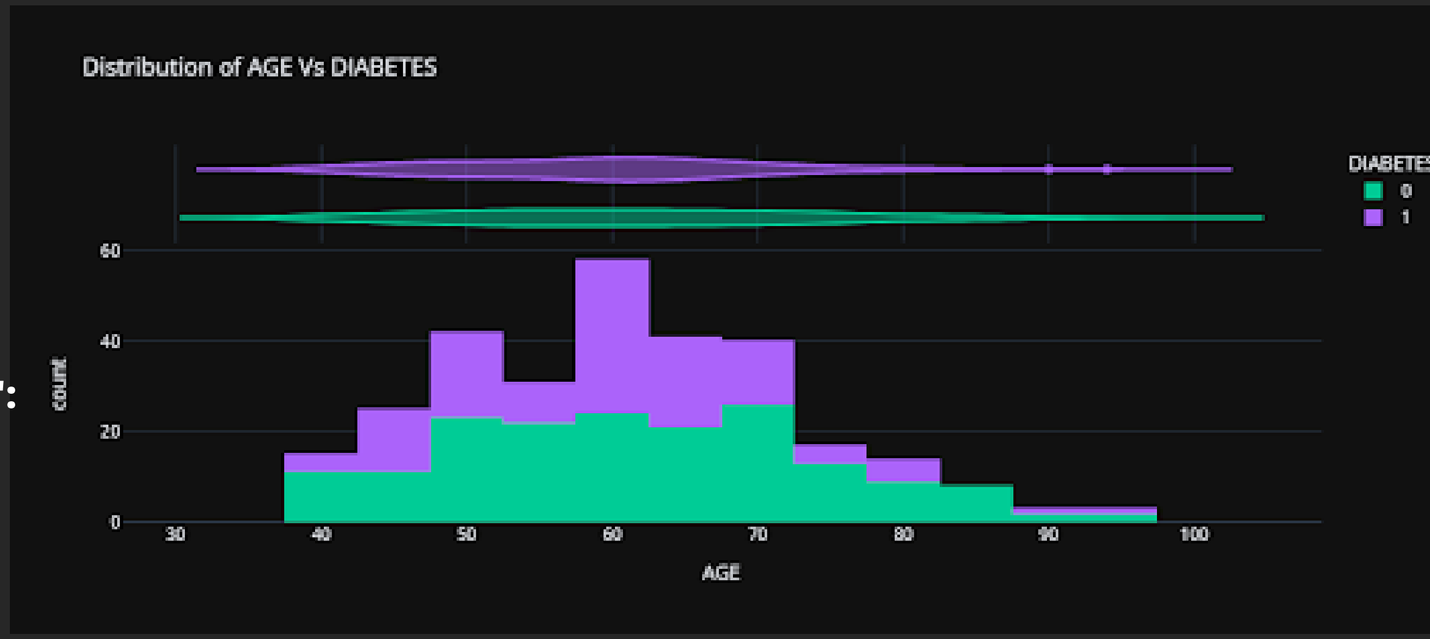
Importing libraries - In order to perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data preprocessing, which are:

- Numpy: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python. It also supports to add large, multidimensional arrays and matrices. So, in Python, we can import it as:
- Matplotlib: The second library is matplotlib, which is a Python 2D plotting library, and with this library, we need to import a sub-library pyplot. This library is used to plot any type of charts in Python for the code. It will be imported as below
- Pandas: The last library is the Pandas library, which is one of the most famous Python libraries and used for importing and managing the datasets. It is an open-source data manipulation and analysis library:

- **Importing datasets -I**n machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model.
- **Finding Missing Data -**The next step of data preprocessing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.
- **Encoding Categorical Data -**Categorical data is data which has some categories such as, in our dataset; there are two categorical variable, **Country**, and **Purchased**.
- Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers.
- **Splitting dataset into training and test set -**If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:
- **Feature scaling -** Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominate the other variable.

IMPLEMENTATION AND CODING

```
import plotly.express as px
fig = px.histogram(dataset, x="age", color="diabetes",
marginal="violin",hover_data=dataset.columns,
                title ="Distribution of AGE Vs DIABETES",
                labels={"diabetes": "DIABETES", "age": "AGE"},
                template="plotly_dark",
                color_discrete_map={"0": "RebeccaPurple", "1":
"MediumPurple"})
fig.show()
```



1. LOGISTIC REGRESSION

INPUT

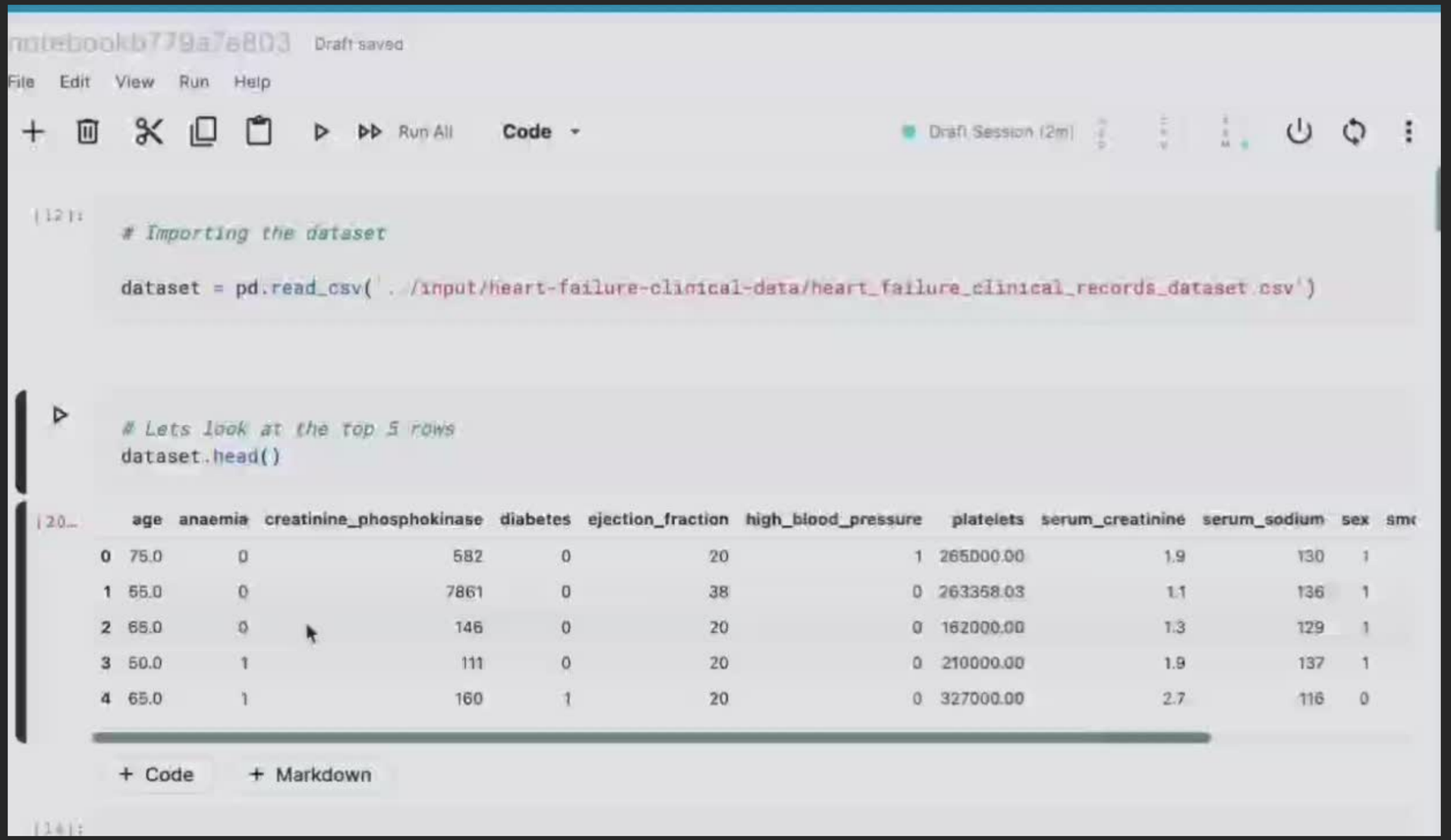
```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression()
classifier.fit(x_train, y_train)

y_pred = classifier.predict(x_test)
mylist = []
from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, y_pred)
ac = accuracy_score(y_test, y_pred)
mylist.append(ac)
print(cm)
print(ac)
```

OUTPUT

```
[[40 3]
[ 4 13]]
0.8833333333333333
```

DEMO VIDEO



	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smo
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	

Contiuation of 2nd Review

FINAL PRESENTATION

Introduction

- Cardiovascular diseases are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age.

- Heart disease remains the #1 cause of death worldwide in the latest annual Statistical Update from the American Heart Association

- Experts predict the global burden of cardiovascular disease will grow exponentially over the next few years as the long-term effects of the current COVID-19 pandemic evolve.

Objective

- Provides a new approach to concealed patterns in the data.
- Helps avoid human biases.
- To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.
- Reduce the cost of medical tests.

- The main objective of this research is to develop a heart prediction system.The system can discover and extract hidden knowledge associatedwith diseases from a historicalheart data set.
- Heart disease prediction system aims to exploit data mining techniques on medical data sets to assist in the prediction of heart diseases.

EXISTING WORK

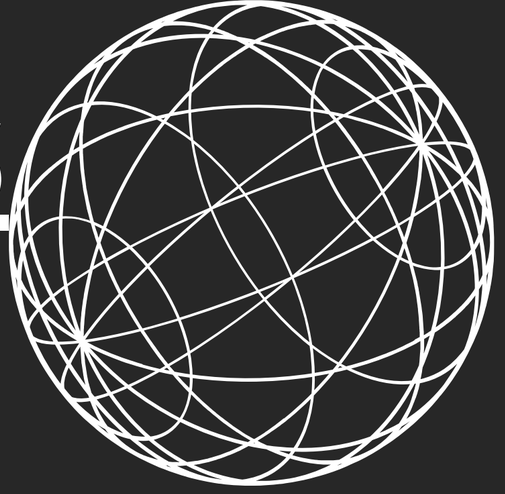
Most of the researchers used Framingham machine learning databases to predict heart disease in their proposed work.Researchers in their experimental study achieved a different level of accuracy by using different machine learning algorithms. Further, existing works with reference to heart disease prediction with their limitations are described as follows.A. Golande et al. explores various machine learning methods that can be used to diagnose heart disease. In their proposed work, authors implemented DT, KNN, and K-Meansal algorithms for heart disease as a classification problem. The experimental results conclude that the accuracy obtained by the DT was significantly higher than was anticipated by the combination of various techniques and parameter correction. An important limitation of their proposed work was they did not use the data processing to clean up the data set.F. S. Alotaibi et al. created a machine learning model and compared different machine learning algorithms .In their proposed work they used a quick miner tool to test the performance of different machine learning algorithms. In their proposed work the accuracy of the DT, LR, Naive Bayes and SVM classification algorithms were compared. The decision of the drug algorithm is very high.An important limitation of their approach was author did not use the data processing process to clean up the data set. T. R. Prince, et al. conducted research on machine learning algorithms divided into categories for predicting heart disease. The classification strategies used by the author are Naive Bayes, KNN, Decision tree, Neural network. Classification classifiers were analyzed by taking different numbers of symbols . The key limitations of the proposed work were they did not perform many experimentation work.

Chen et al. in their proposed work compared the accuracy of different algorithms such as SVM, neural networks, Bayesian segregation, tree decision, and Logistic retrospective considerations . In their tests, they found that SVM had a really high accuracy of 90.5%, neural networks 88.9%, Bayesian 82.2%, Decision tree 77.9% and logistic regression 73.9%. Shoumanetal et al. statistically significant risk factors for age, blood pressure, cholesterol, smoking, total cholesterol, diabetes, high blood pressure, genetics, obesity, metabolic syndrome . The same paper also listed the Cleveland Heart Disease Database is a standard database for cardiovascular research as it is widely accepted. Detra noe tal used logistic regression to obtain 77% accuracy of prediction. Further, in short, the existing works with reference to heart disease prediction are described in Table 1. Table 1 highlights the proposed methodology and limitations of the existing work.

Table1: Summary of Existing Works on CHDD

Writer Name	Proposed methodology	Limitations
Detrano et al.[6]	Logistic regression (77%)	There is only very less percentage of accuracy i.e., 77%
Cheung [9]	C4.5(81.11%) Naive Bayes (81.48%) BNNF (80.96%)	Author has performed naïve bayes algorithm acquired the better accuracy, but the drawback of this algorithm is the assumption of independent predictor features.
Tu et al. [10]	J4.8 Decision Tree (78.9%) Bagging Algorithm (81.41%)	Bagging is certainly a star classifier when we need to fight against the variance, create a more stable and robust model that can be run parallel. The only limitation is that it is computationally expensive since it requires a high number of estimators.

Project Modules



Module-1 Data Preprocessing

1) Data Cleansing :- Data cleaning is the process of fixing or removing incorrect, corrupt incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multi data sources, there are many opportunities for data to be duplicated or mislabeled.

2)Replacing Missing Values:-The problem of missing value is quite common in many real- datasets. Missing value can bias the results of the machine learning models and/or reduce accuracy of the model.

3)Handling Outliners-One of the most important steps as part of data preprocessing is detect and treating the outliers as they can negatively affect the statistical analysis and the train process of a machine learning algorithm resulting in lower accuracy

Module-2 Feature Selection

In this stage, experiments are conducted with and without feature selection to assess the effect of feature selection. Feature selection helps to construct a more accurate model by eliminating or underrepresenting the less relevant features, minimizing training time and enhancing learning performance . The behavior of several feature selection approaches under the three major categories (filter, wrapper, and embedded) is assessed in this experiment.

- **Feature Selection Using Filter Methods:** Filter-based selection techniques utilize statistical methods to determine the dependence or association among independent features (input attributes) and the dependent feature (target attribute).This proposed work mainly uses four filter-based feature selection algorithms, namely, Chi-square, mutual information, ReliefF, and ANOVA F method.

- **Feature Selection Using Wrapper Methods:** This method mainly utilizes a search strategy to evaluate the variable subsets of independent features by feeding the chosen learning algorithm and then assessing the performance of a learning algorithm. The searching method can identify various types of techniques and in this experiment mainly used four types of techniques are: forward feature selection, backward feature elimination, recursive feature elimination, and exhaustive feature selection
- **Feature Selection Using Embedded Methods:** Embedded techniques include integrating the feature selection process into the machine learning algorithm's development. This is a hybrid technique that combines the filter and wrapper methods. Here, the algorithms include their feature selection strategy. These techniques need less computation than wrapper methods. The design of embedded feature selection approaches is algorithm dependent. This uses Lasso Regression and Ridge Regression methods.

Module-3 Classification

SVM(Support Vector Machine): Among various data mining methods, SVM is well known for its discriminative power for classification, especially in the cases where sample sizes are small and a large number of features (variables) are involved (i.e., high-dimensional space). Guyon also showed that SVM performs much better than correlation based techniques in a gene selection problem for cancer classification with the feature selection method, recursive feature elimination (RFE). In addition, in comparing SVM with logistic regression in predicting diabetes and pre-diabetes, demonstrated that SVM can be a promising tool.

XGBoost Classifier-XGBoost classifier is a Machine learning algorithm that is applied for structured and tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. XGBoost is an extreme gradient boost algorithm. And that means it’s a big Machine learning algorithm with lots of parts. XGBoost works with large, complicated datasets. XGBoost is an ensemble modelling technique.

K-Nearest Algorithm-

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.
It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

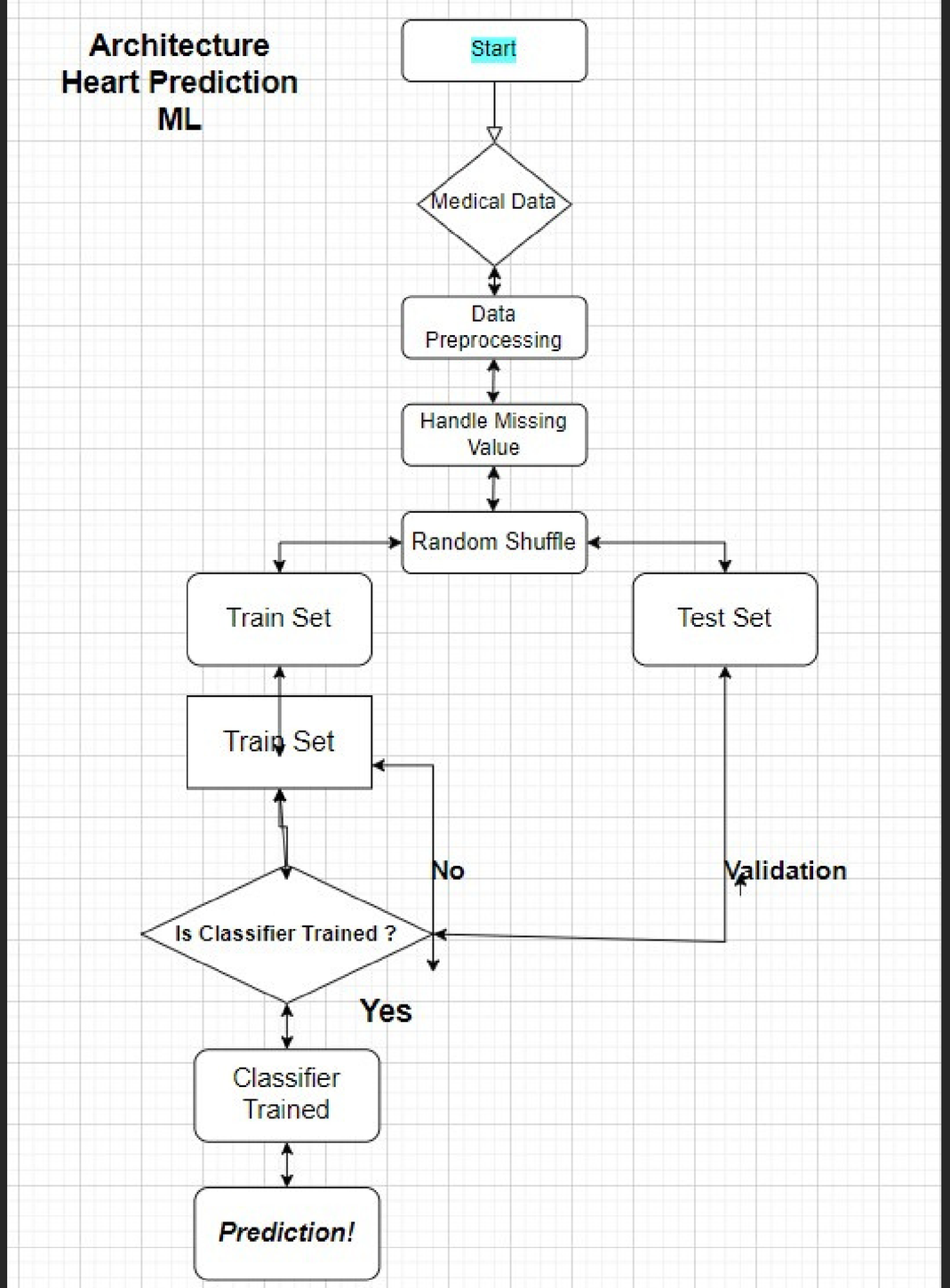
Logistic Regression-

Logistic Regression is a “Supervised machine learning” algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. That means Logistic regression is usually used for Binary classification problems.

Random Forest Classifier- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "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." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Artificial neural network-Artificial neural network is used for complex and difficult tasks. Neural network is data mining tool used for classification & clustering. A neural network is typically a collection of neuron-like processing units with weighted connections between the units. There are two modes in artificial neural networks

Architecture

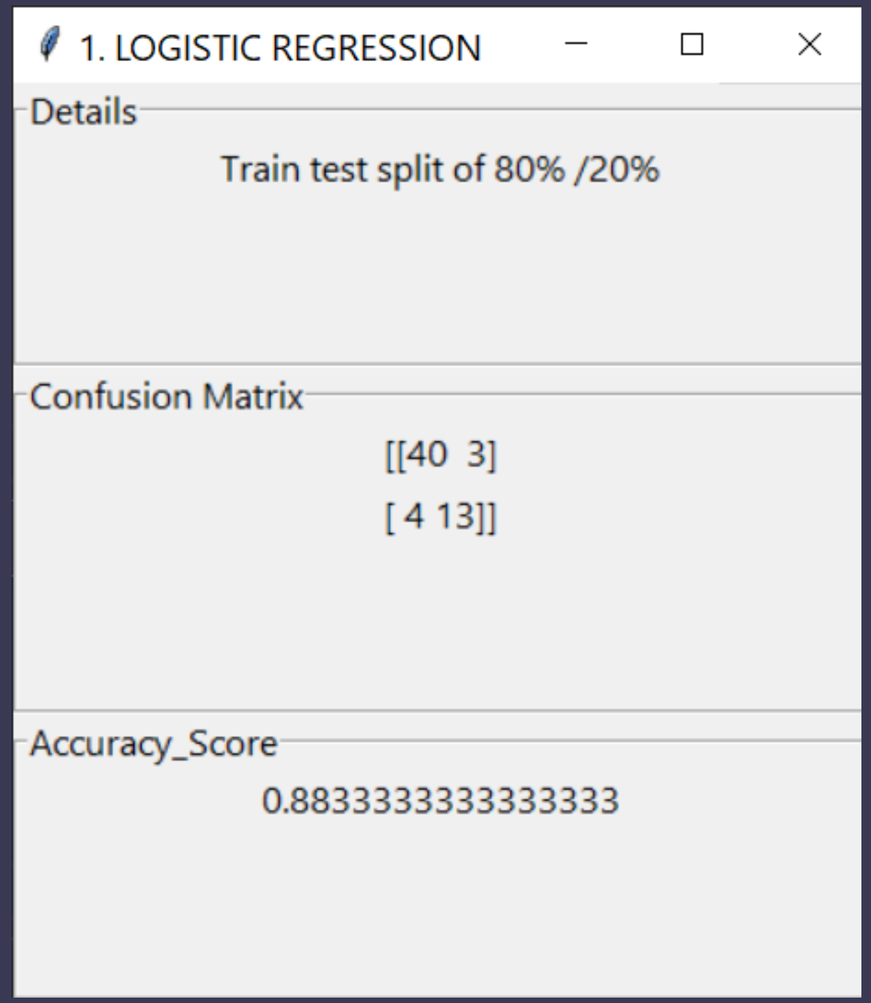


Outcome

Once data mining ,data cleaning are done then we can move towards output

1. LOGISTIC REGRESSION

When the latter occurs, a binary logistic regression model is an appropriate method to present the relationship between the disease's measurements and its risk factors.

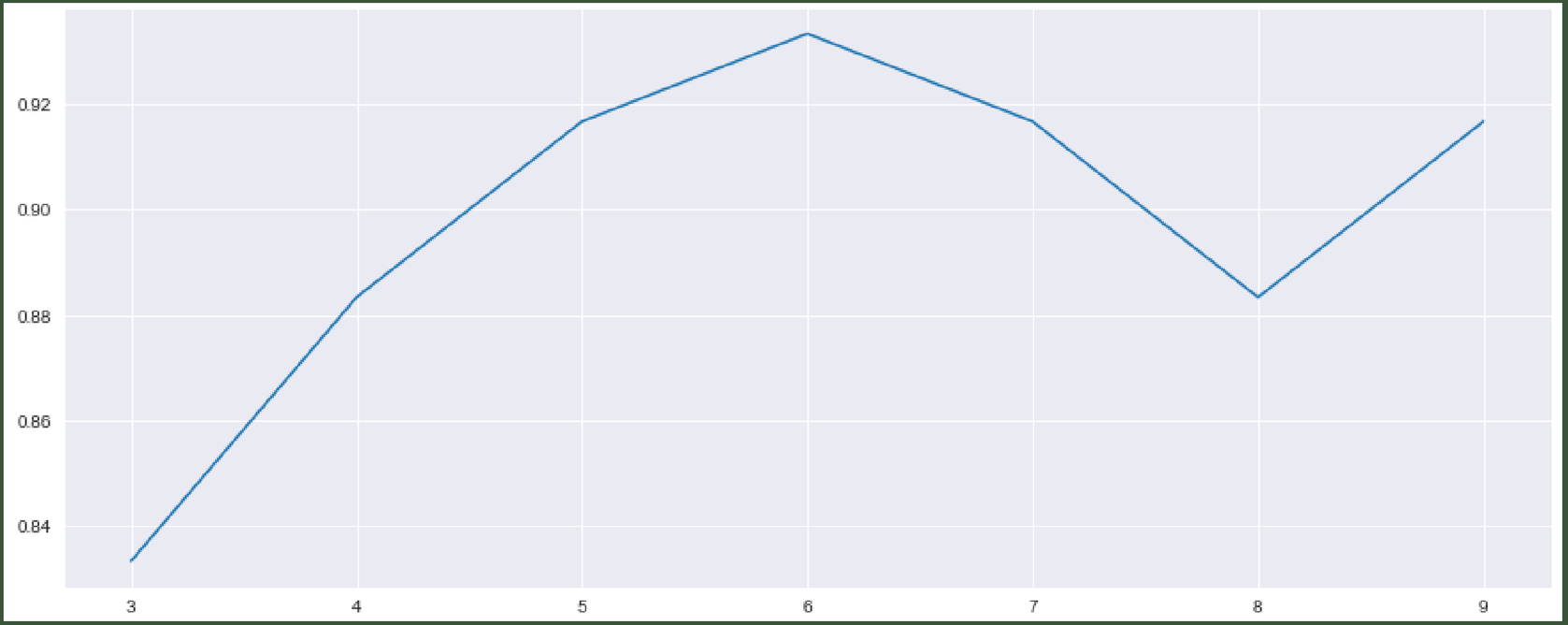


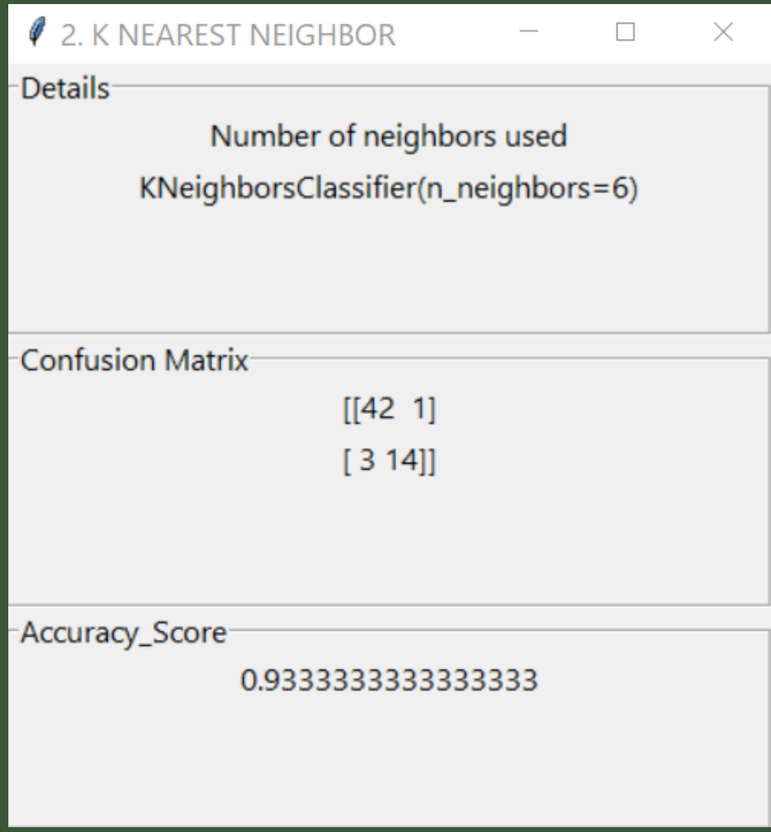
~Prince

Outcome

2. K NEAREST NEIGHBOR

KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closet to the test data.



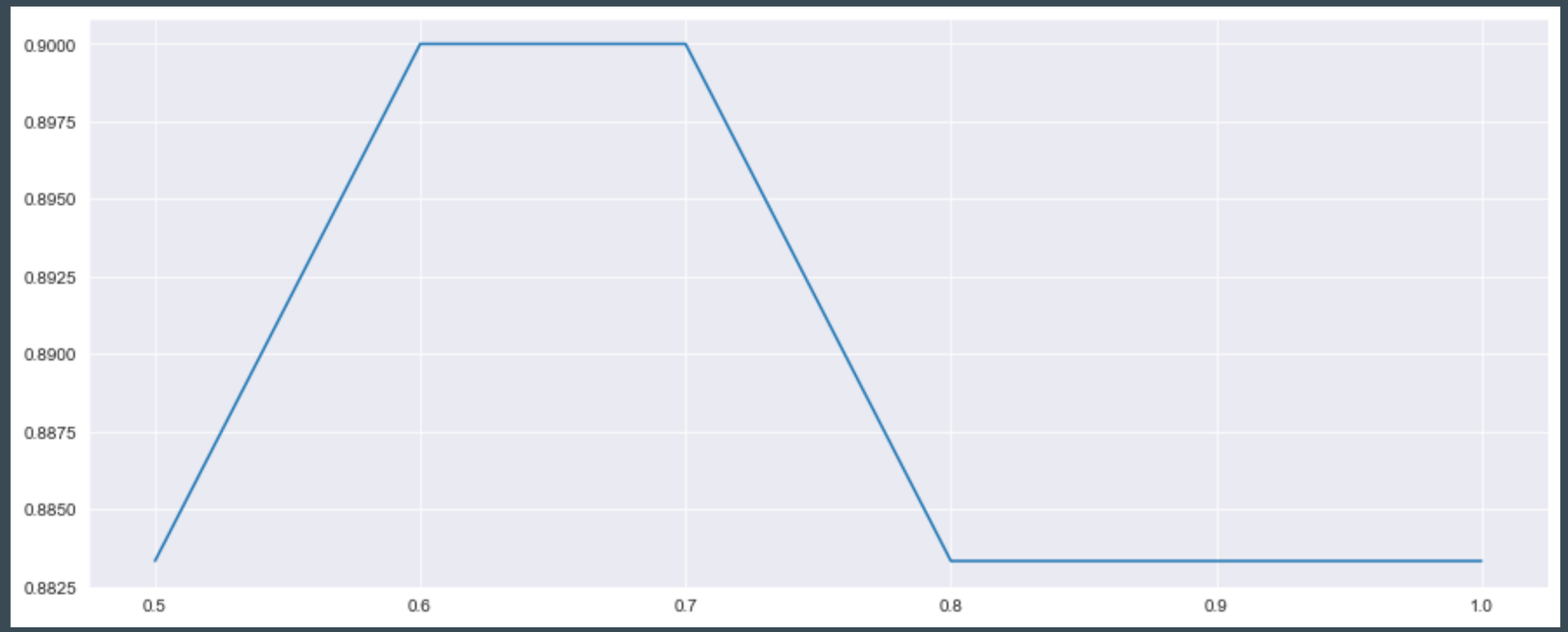


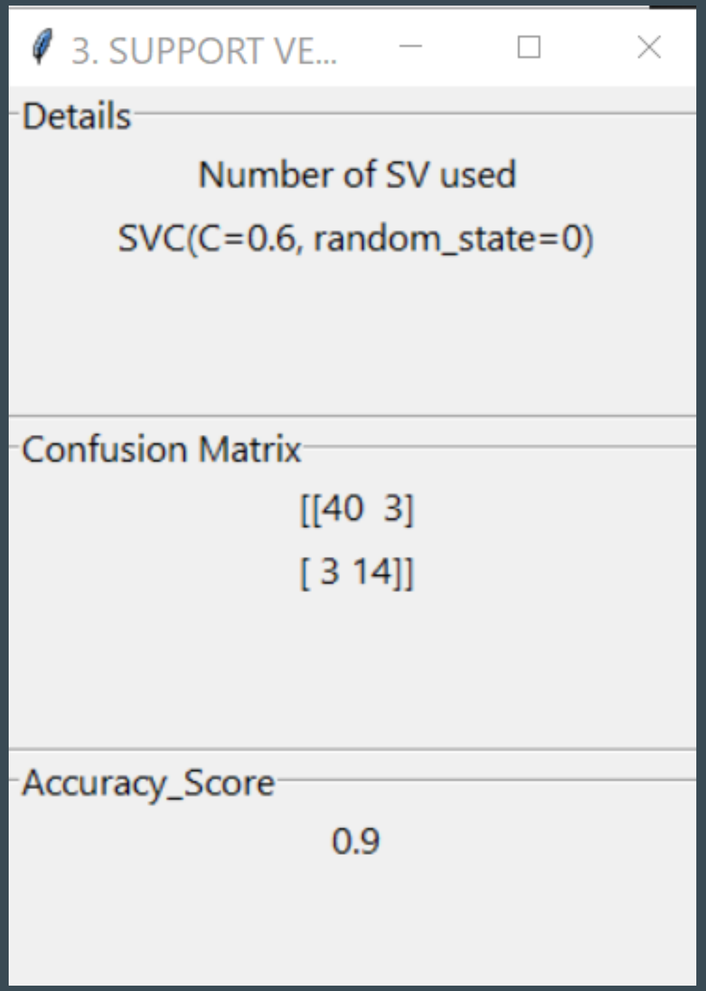
~Prince

Outcome

3. SUPPORT VECTOR CLASSIFIER

Support Vector Classifier is an extension of the Maximal Margin Classifier. It is less sensitive to individual data. Since it allows certain data to be misclassified, it's also known as the " Soft Margin Classifier". It creates a budget under which the misclassification allowance is granted.



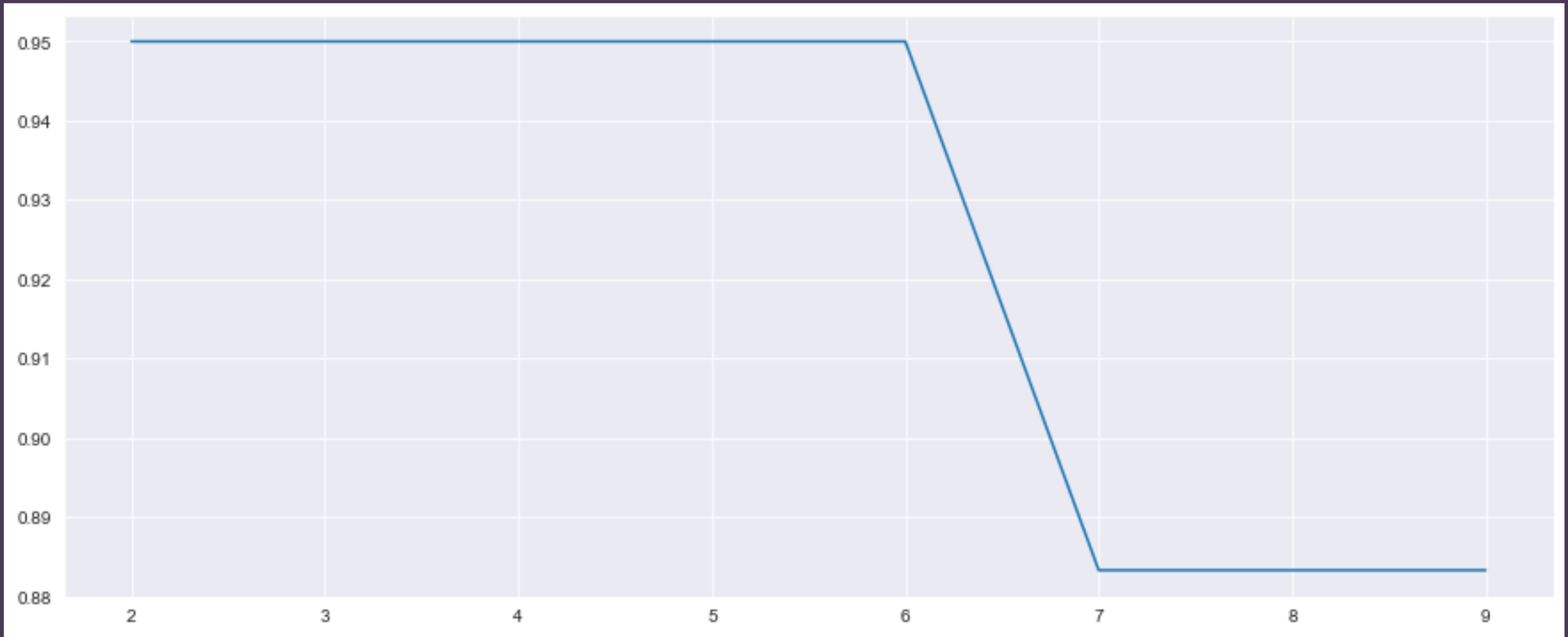


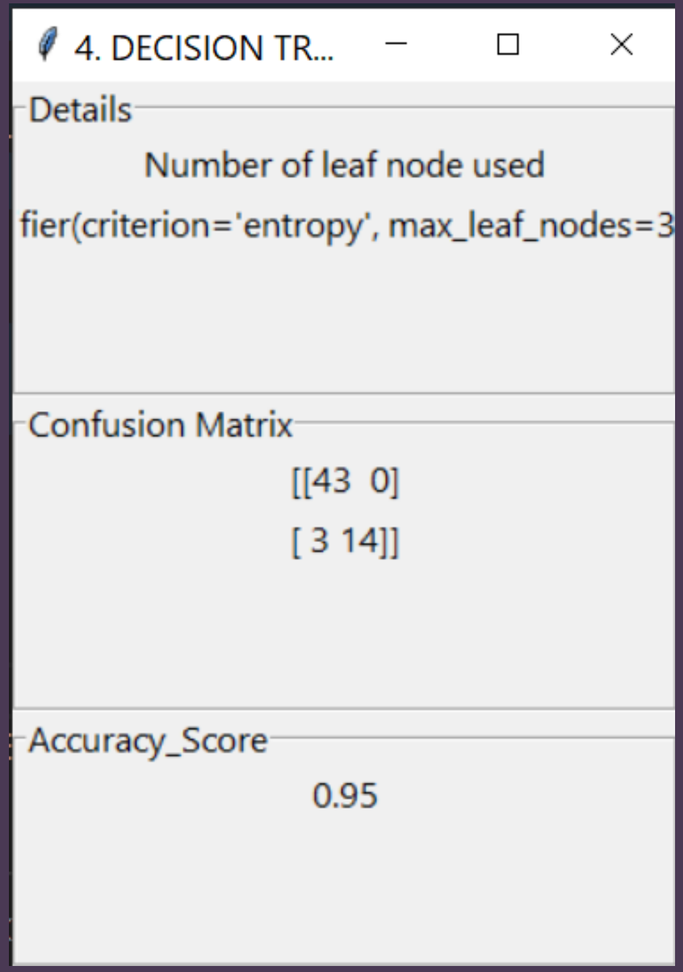
~Prince

Outcome

4. DECISION TREE CLASSIFIER

use prelabelled data in order to train an algorithm that can be used to make a prediction. Decision trees can also be used for regression problems



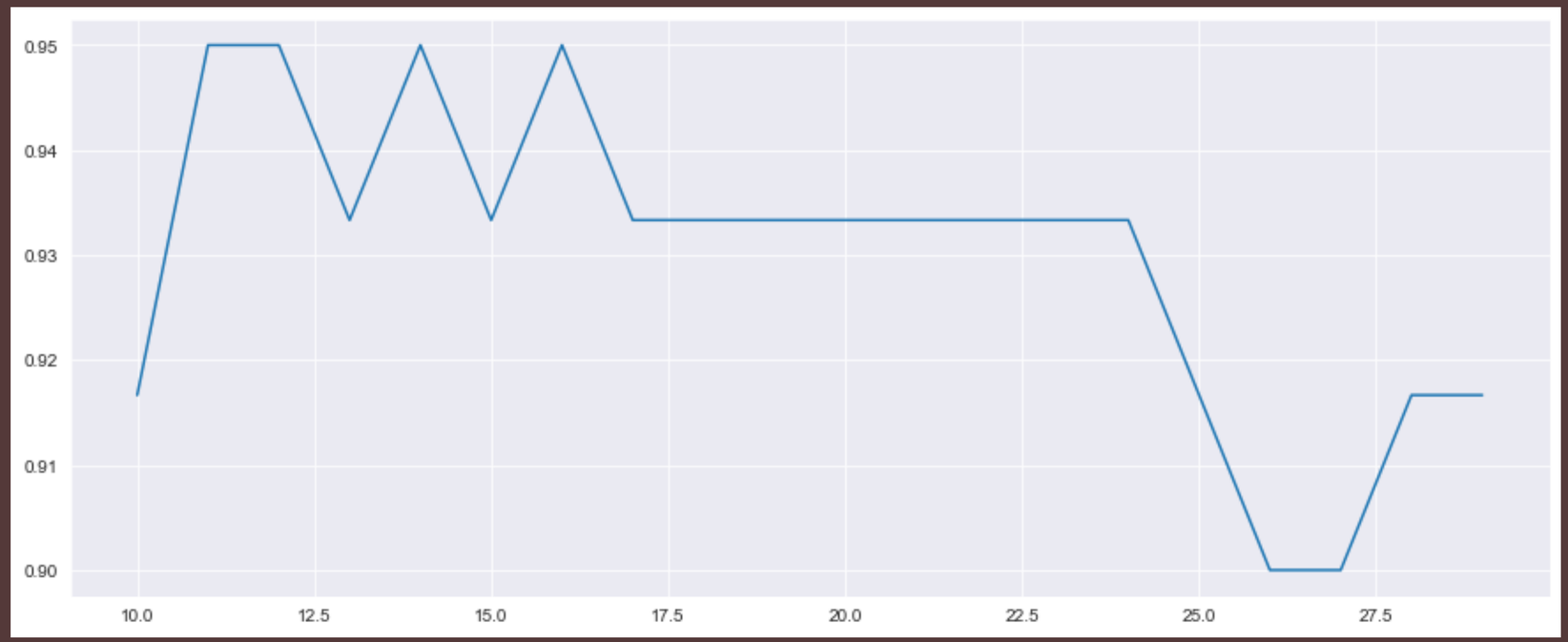


~Prince

Outcome

5.RANDOM FOREST CLASSIFCATION

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems.random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting



~Prince

Outcome

6. ANN (5 layers used)

ANNs are also named as “artificial neural systems,” or “parallel distributed processing systems,” or “connectionist systems.” ANN acquires a large collection of units that are interconnected in some pattern to allow communication between the units. These units, also referred to as nodes or neurons, are simple processors which operate in parallel.

Confusion Matrix
[[40 3]
[2 15]]
Accuracy
0.9166666666666666

~Prince

How our Project serve to the community ?

Not only for heart disease: if we have dataset of any disease we can predict the intensity of that disease through our application.

Easy to use: Our application can be used even by non-technical person,they can do it in two steps by selecting the database and running the algorithm by just one click.

High rates of heart disease: Heart diseases have become more and more frequent among people including our country (India). Therefore, predicting the disease before becoming infected decreases the risk of death. This prediction is an area that is widely researched.

Reference

1. https://ieeexplore.ieee.org/abstract/document/8474922/
2. https://ieeexplore.ieee.org/document/8741465
3. http://sersc.org/journals/index.php/IJAST/article/download/5545/3446/

4.https://www.geeksforgeeks.org/machine-learning/

5.https://medium.com/swlh/top-20-websites-for-machine-learning-and-data-science-d0b113130068

6. Simple guide to confusion matrix terminology (dataschool.io)

9. https://www.kaggle.com/code/faressayah/predicting-heart-disease-using-machine-learning/notebook

10 Decision Tree - GeeksforGeeksdecision-tree/?msclkid=b58b685bc08e11eca1d4289ca07eab76.

Responsibilty

Problem Statement : Manav Nair

Motivation of Project-Anant kumar Pandey

Research and Analysis - Ashwini Darade

Project Modules :-

- **Project Prototype-1** Vartika Pandey
- **Pseudocode -2** Ananya Saxena
- **Design & Architecture -3** Shirsh Waghmode

Implementation and Results- Prince & Ayush Joshi

Thank you!