International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

HEART DISEASE PREDICTION

A.Vindhya Sree[1], CH. Neha[2], K. Hima Bindu[3],Sumera[4].
Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana, India

Abstract: Heart, an important organ in human body regulates the blood flow and circulation. Heart related diseases are the one of the major reasons for high mortality in world. In the last few decades these are emerged as the most life-threatening diseases. So, there is a need for a reliable, accurate and feasible approach to diagnosis such diseases in time for proper treatment and medication. Machine Learning algorithms offers an opportunity to improve accuracy by exploiting the frequency of diseases related to heart based on the recorded data. In this project prediction of heart diseases from the dataset using Machine Learning Algorithms such as Logistic Regression, SVM, Naïve _Bayes, Decision Tree, Random Forest, KNN classification algorithms has been done. The results verify that Random Forest algorithm has achieved the highest accuracy of 89.7% compared to other Machine learning algorithms implemented.

Keywords: Heart diseases data set, supervised, unsupervised, reinforced, Logistic Regression, SVM, Naïve _Bayes, Decision Tree, Random Forest, KNN classification algorithms. Python programming, jupytor Notebook, confusion matrix.

1.INTRODUCTION

Heart is one of the most extensive and vital organ of human body so the care of heart is essential. Most of diseases are related to heart so the prediction about heart diseases is necessary and for this purpose comparative study needed in this field, today most of patient are died because their diseases are recognized at last stage due to lack of accuracy of instrument so there is need to know about the more efficient algorithms for diseases prediction. Machine Learning is one of the efficient technology for the testing, which is based on training and testing. It is the branch of Artificial Intelligence(AI) which is one of broad area of learning where machines emulating human abilities, machine learning is a specific branch of AI. On the other hand machines learning systems are trained to learn how to process and make use of data hence the combination of both technology is also called as Machine Intelligence. As the definition of machine learning, it learns from the natural phenomenon, natural things so in this project we uses the biological parameter as testing data such as cholesterol, Blood pressure, sex, age, etc. and on the basis of these, comparison is done in the terms of accuracy of algorithms such as Naïve Bayes, Logistic Regression, SVM, Decision Tree, Random Forest, KNN classification algorithms. In this project, we calculate the accuracy of six different machine learning approaches and on the basis of calculation we conclude that which one is best among them.

Machine Learning

Machine Learning is one of efficient technology which is based on two terms namely testing and training i.e. system take training directly from data and experience and based on this training test should be applied on different type of need as per the algorithm required.

There are three type of machine learning algorithms:

Supervised Learning:

Supervised learning can be define as learning with the proper guide or you can say that learning in the present of teacher. We have a training dataset which act as the teacher for prediction on the given dataset that is for testing a data there are always a training dataset. Supervised learning is based on "train me" concept. Supervised learning have following processes:

- Classification
- Random Forest
- Decision tree
- Regression

To recognize patterns and measures probability of uninterruptable outcomes, is phenomenon of regression. System have ability to identify numbers, their values and grouping sense of numbers which means width and height, etc.

There are following supervised machine learning algorithms:

- Linear Regression
- Logistical Regression
- Support Vector Machines (SVM)
- Neural Networks
- Random Forest
- Gradient Boosted Trees
- Decision Trees
- Naive Bayes

Unsupervised Learning:

Unsupervised learning can be define as the learning without a guidance which in Unsupervised learning there are no teacher are guiding. In Unsupervised learning when a dataset is given it automatically work on the dataset and find the pattern and relationship between them and according to the created relationships, when new data is given it classify them and store in one of them relation. Unsupervised learning is based on "self sufficient " concept. For example suppose there are combination fruits mango, banana and apple and when Unsupervised learning is applied it classify them in three different clusters on the basis if there relation with each other and when a new data is given it automatically send it to one of the cluster . Supervisor learning say there are mango, banana and apple but Unsupervised learning said it as there are three different clusters. Unsupervised algorithms have following process:

- Dimensionality
- Clustering
- There are following unsupervised machine learning algorithms:
- k-means clustering

Reinforcement:

Reinforced learning is the agent ability to interact with the environment and find out the outcome. It is based on "hit and trial" concept. In reinforced learning each agent is awarded with positive and negative points and on the basis of positive points reinforced learning give the dataset output that is on the basis of positive awards it trained and on the basis of this training perform the testing on datasets

About Jupytor Notebook:

Jupiter notebook is used as the simulation tool and it is comfortable for python programming projects. Jupytor notebook contains rich text elements and code also, which are figures, equations, links and many more. Because of the mix of rich text elements and code, these documents are perfect location to bring together an analysis description, and its results, as well as, they can execute data analysis in real time. Jupyter Notebook is an open-source, web-based interactive graphics, maps, plots, visualizations, and narrative text.

1.1About Project

The work proposed in this project focuses mainly on various data mining practices that are implemented 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 the heart can cause distress in other parts of the 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 World Health

Organization more than 10 million die due to heart disease every single year around the world. A healthy lifestyle and earliest prediction are only ways to prevent heart related diseases.The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis. Even if heart disease is found as the prime source of death in world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of prediction of that disease.The proposed work makes an attempt to predict these heart diseases at an early stage to avoid disastrous consequences. Data Mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Machine learning (ML) which is a sub field of data mining handles large scale well formatted datasets effectively. In medical field, Machine learning can be used for diagnosis, detection, prediction of various diseases. The main goal of this project is to provide a tool for doctors to predict heart disease at an early stage. This in turn will help to provide effective treatment to patient's and advise several consequences. ML plays a very important role to predict the hidden discrete patterns and thereby analyze the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis. This project presents performance analysis of various ML techniques such as Naïve Bayes, Logistic Regression, SVM, Decision Tree, Random Forest, KNN classification algorithms.

1.2 Objectives of the Project

- To develop machine learning mode of predict future possibility of heart disease by implementing some classification algorithm such as Naïve bayes, Logistic Regression, SVM, Decision Trees, Random Forest, KNN algorithms.
- To determine significant risk factors based on medical dataset which may lead to heart disease.
- To analyse feature section methods and understand their working principles.

1.3 Scope of the Project

The scope of this project is to predict the heart disease using different classification algorithms. Machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of death cases can be minimized by the awareness about the diseases. Therefore, predicting the disease before becoming infected decreases the risk of death.

1.4 Advantages

- Increased acuuracy for effective heart duseas diagnosis.
- Handles enormous amount of data.
- Reduce the time complexity of doctors.
- Cost effective for patients.

1.5 Disadvantages

- Prediction of cardiovascular disease results is not that much accurate.
- Data mining techniques does not help to provide effective decision making.

1.6 Applications

This heart disease prediction model is very useful in health care sectors, especially for predicting the heart diseases, clinicians and institutions can provide better and improvised outcomes for the patients through scalable and dynamic applications.

1.7 Hardware and Software Requirements:

1. Hardware Requirements:
   Processor: Intel i5 core
   Ram: 8GB
   Hard Disk:120 GB

Monitor: 15" LED
Input Devices: Keyboard, Mouse

2. Software Requirements:
Operating System: Windows 8/8.1/10
Coding Language: Python
Special Tools: Jupyter Notebook.
Packages:

1. NUMPY: NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.
It is the fundamental package for scientific computing with Python.
It contains various features including these important ones:
•A powerful N-dimensional array object
•Sophisticated (broadcasting) functions
•Tools for integrating C/C++ and Fortran code
•Useful linear algebra, Fourier transform, and random number capabilities.
NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

2. PANDAS: Pandas is the most popular python library that is used for data analysis. It provides highly optimized performance with back-end source code is purely written in C or PYTHON.

3. SCIKIT LEARN: Scikit-learn is the most useful library for machine learning in Python. The Sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction. sklearn is used to build machine learning models.It should not be used for reading the data, manipulating and summarizing it. There are better libraries for that (e.g. NumPy, Pandas etc.)

4. MATPLOTLIB: Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multiplatform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.
One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Fig 1.7.1: Software and Package Requirements

2. LITERATURE SURVEY

Santhana Krishnan J and Geetha S[1] proved that decision tree is more accurate as compare to the naïve bayes classification algorithm in their project.

Gavhane et al.[2] have worked on the multi layer perceptron model for the prediction of heart diseases in human being and the accuracy of the algorithm using CAD technology.

Mohan et al.[3] define how you can combine two different approaches to make a single approach called hybrid approach which have the accuracy 88.4% which is more than of all other.

Himanshu et al.[4] define naive bayes perform well with low variance and high biasness as compare to high variance and low biasness which is knn.

Kumar et al.[5] have worked on various machine learning and data mining algorithms and analysis of these algorithms are trained by UCI machine learning dataset which have 303 samples with 14 input feature and found svm is best among them.

Kaur et al.[6] have worked on this and define how the interesting pattern and knowledge are derived from the large dataset. They perform accuracy comparison on various machine learning and data mining approaches for finding which one is best among then and get the result on the favor of svm.

Kohli et al.[7] work on heart diseases prediction using logistic regression, diabetes prediction using support vector machine, breast cancer prediction using Adaboot classifier and concluded that the logistic regression give the accuracy of 87.1%, support vector machine give the accuracy of 85.71%, Adaboot classifier give the accuracy up to 98.57% which good for predication point of view.

2.1 Existing System

In existing system, the naive bayes perform well with low variance and high biasness as compare to high variance and low biasness which is knn. With low biasness and high variance knn suffers from the problem of over fitting this is the reason why performance of knn get decreased. There are various advantage of using low variance and high biasness because as the dataset small it take less time for training as well as testing algorithm but there also some disadvantages of using small size of dataset. When the dataset size get increasing the asymptotic errors are get introduced and low biasness, low variance based algorithms play well in this type of cases. The dataset used was the Heart Disease dataset which is a combination of 4 different database, but only the UCI Cleveland dataset was used. This database consists of total 76 attributes, but all refer to using a subset of only 14 features. Therefore, we used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis.

2.1 Proposed System

Heart Disease can be managed effectively with a combination of lifestyle changes, medicine and in some cases, surgery. In this project we want to predict the heart disease from dataset using machine learning algorithms such as Naïve bayes, Logistic Regression, SVM, Decision Trees, Random Forest and KNN algorithms. The prediction results can be used to prevent and thus reduce cost for surgical treatment and other expenses.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

TABLE.1 Attributes of the Dataset

S. No.	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numaric
2	Sex	Gender of patient(male-0 female-1)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl, values from 126 to 564)	Numerical
6	Fbs	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	Resting	Resting electrocardiographics result (0 to 1)	Nominal
8	Thali	Maximum heart rate achieved(71 to 202)	Numerical
9	Exang	Exercise included agina(1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Ca	Number of major vessels (0-3)	Numerical
13	Thal	3-normal	Nominal
14	Targets	1 or 0	Nominal

3. PROPOSED ARCHITECTURE

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

3.1 BLOCK DIAGRAM

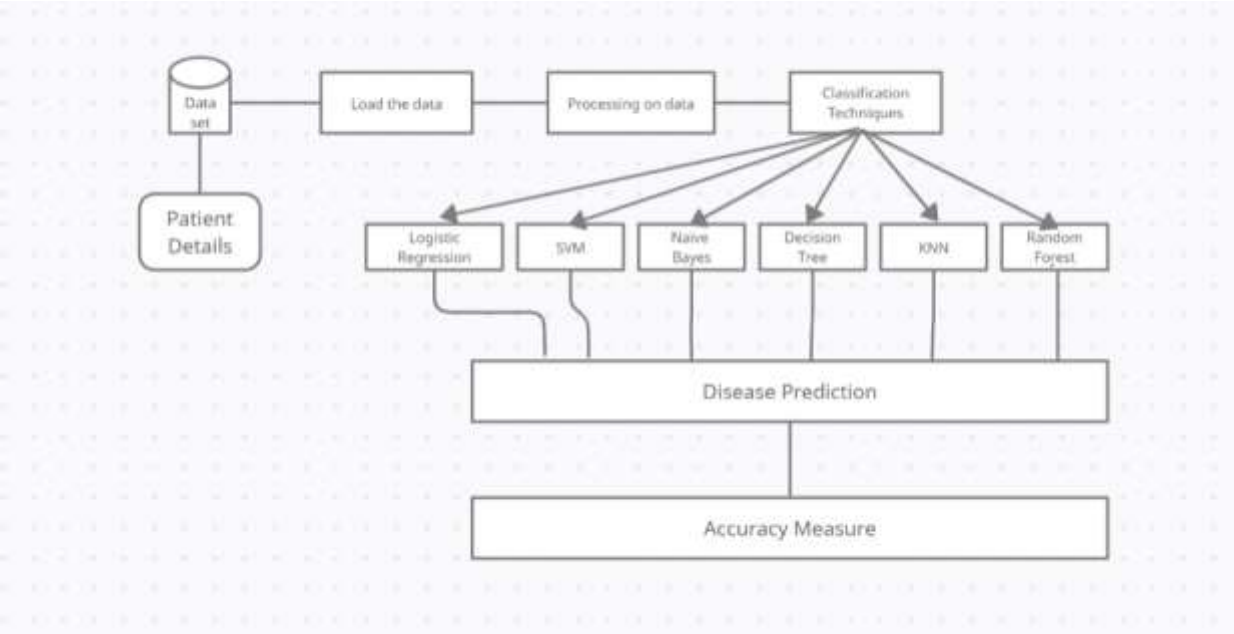
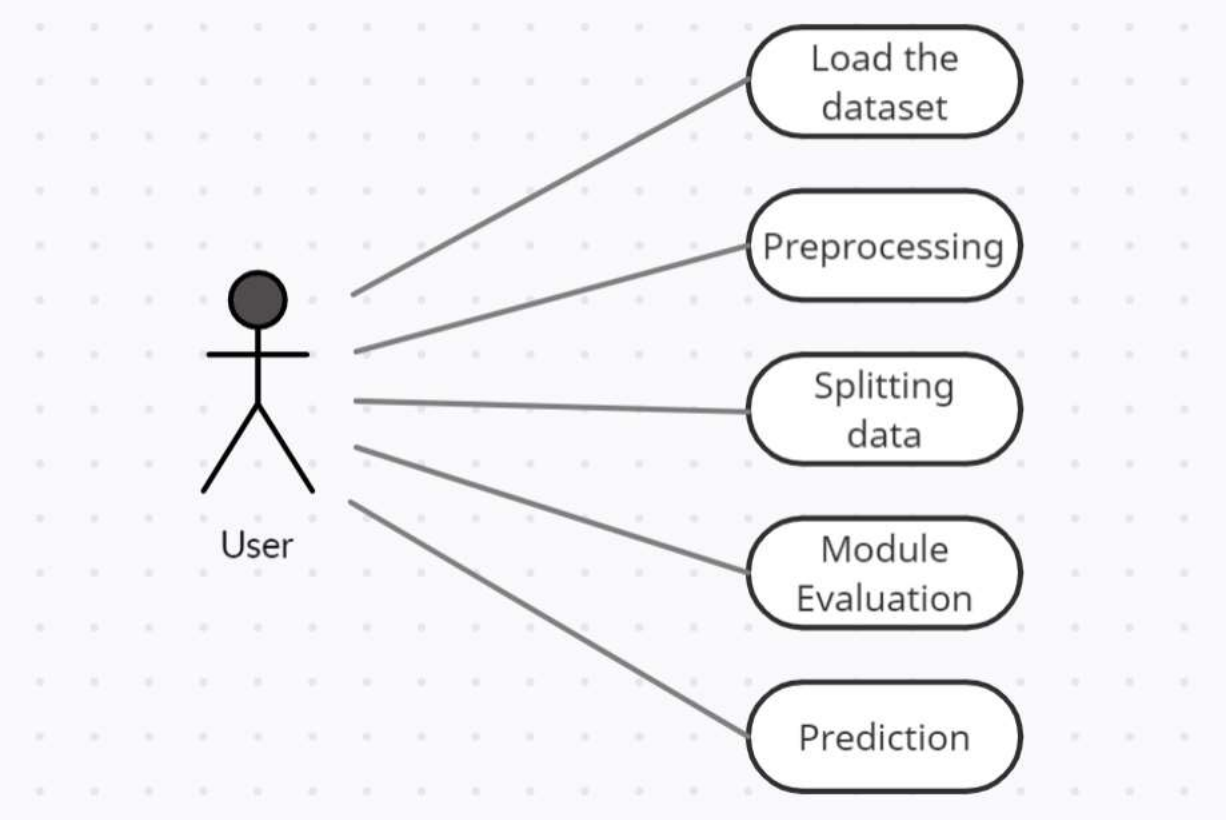


Fig3.1: Architecture diagram

Use-Case Diagram:



International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Fig3.2: Use case diagram

Class Diagram

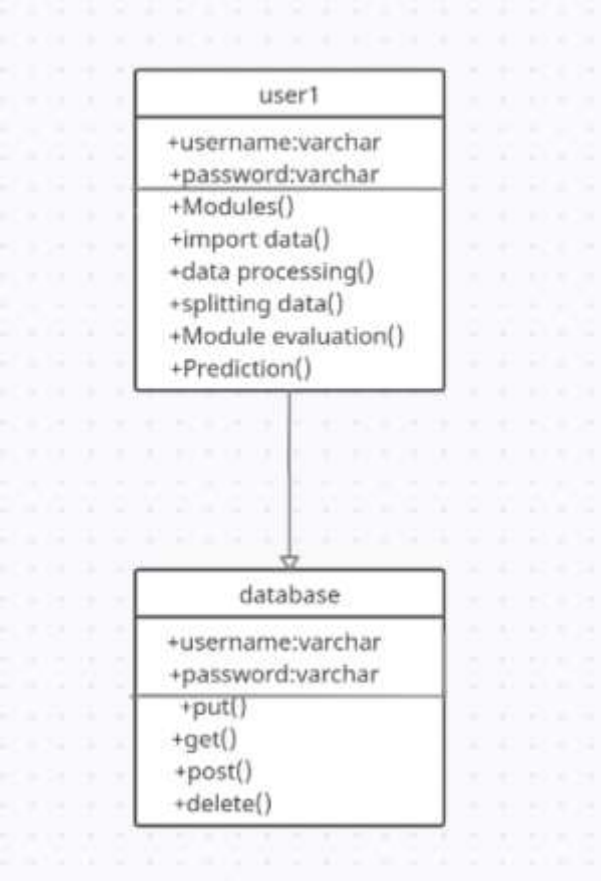


Fig3.3: Class diagram

Activity diagram

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Fig3.4: Activity diagram

Sequence diagram

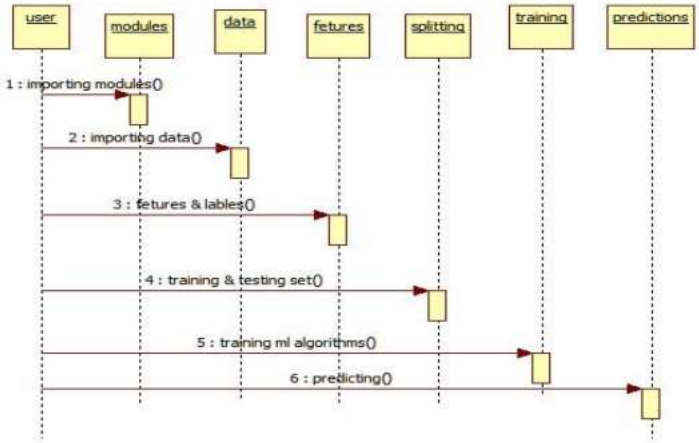


Fig3.5: Sequence diagram

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

4. IMPLEMENTATION

Environmental Setup:

- We need to install and setup the IDE
- After installing we need to set the path in environmental variables
- The process for installing is as below

Steps Installing Anaconda:

1.Downloads and install Anaconda from https://repo.anaconda.com/archive/Anaconda32021.05Windows-x86_64.exe.
2.After opening link u can see this download option
3.Click on the download option.
4.After downloading start installation.
5.Select the default options when prompted during the installation of Anaconda.
6.Ensure that the path to the folder where Anaconda is installed is added to your computer/system.

Start Jupyter Notebook:

1.Open anaconda navigator and the screen which is similar to below appears.

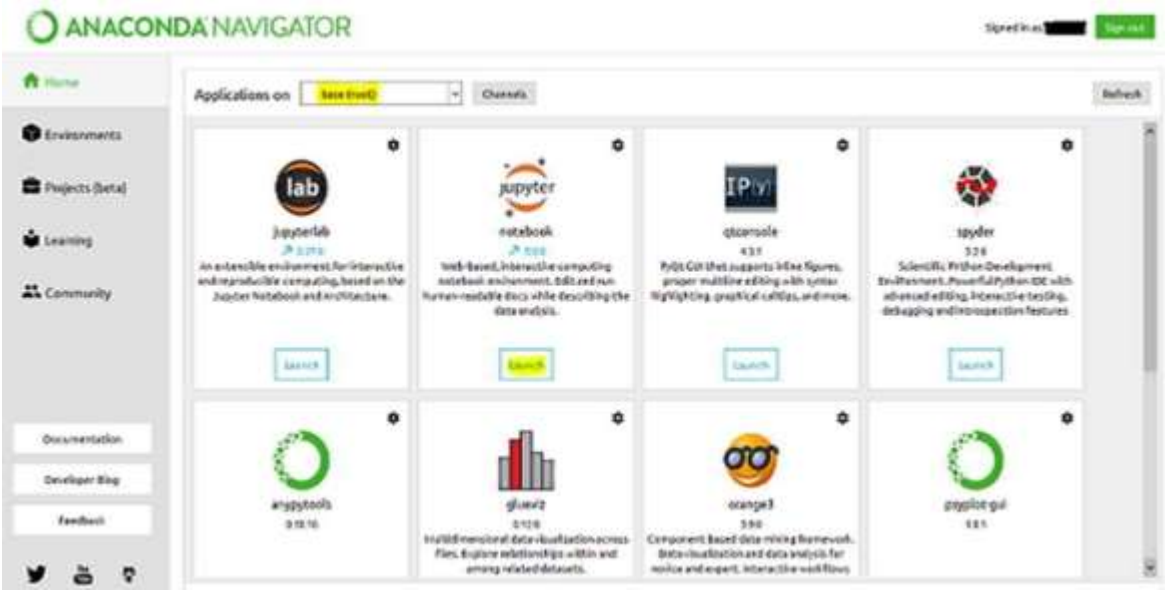


Fig4: Jupyter Notebook Screen

2.Open anaconda prompt to open jupyter notebook.

3.Now open jupyter new kernel.

4.Install required packages.

4.1 Algorithm

1.Data Collection and Processing.

- Loading the csv file
- Print first 5 rows of the dataset.
- Separating the Features and Target.

2.Splitting the dataset into training and testing data.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

3.Model Training.

Logistic Regression

Logistic Regression is one of the most popular Machine Learning algorithm, which comes under the Supervised Learning technique. It is based on the relationship between independent variable and dependent variables. It predicts the output of the categorical dependent variable.

SVM

Support Vector Machine or SVM is one of the Supervised Learning algorithms, which is used for both classification and regression problems. However primarily it is used as a classification algorithm. It is one category of machine learning technique which work on the concept of hyperplan means it classify the data by creating hyper plan between them. Training sample dataset is (Yi, Xi) where i=1,2, 3……. n and Xi is the ith vector, Yi is the target vector. Number of hyper plans decide the type of support vector such as example if a line is used as hyper plan, then method is called linear support vector.

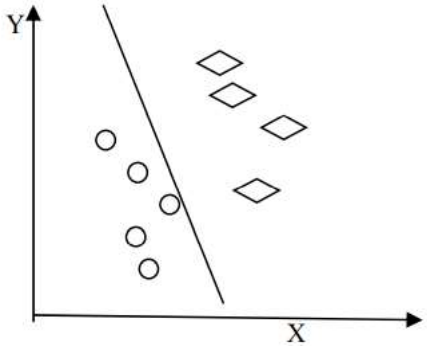


Fig4.2.1: SVM

Naive Bayes:

Naïve Bayes algorithm is one of the supervised learning algorithms, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that include a high-dimensional training dataset. It is a Probabilistic classifier, which means it predicts on basis of the probability of an object. 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.

3.4Decision Trees:

On the other hand, decision tree is the graphical representation of the data, and it is also the kind of supervised machine learning algorithms.

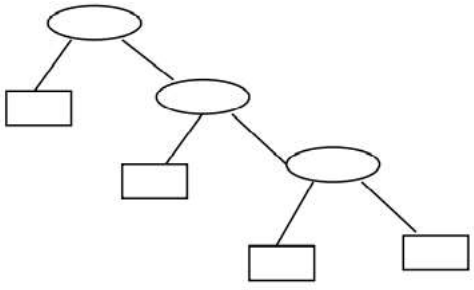


Fig4.4.1: Decision tree

For the tree construction we use entropy of the data attributes and on the basis of attribute root and other nodes are drawn.

$$\text{Entropy} = -\sum Pij \log Pij$$

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

In the above equation of entropy Pij is probability of the node and according to it the entropy of each node is calculated. The node which has highest entropy calculation is selected as the root node and this process is repeated until all the nodes of the tree are calculated or until the tree constructed. When the number of nodes is imbalanced then tree is creating the over fitting problem which is not good for the calculation.

KNN*:*

It works based on distance between the location of data and on the basis of this distinct data are classified with each other. All the other group of data are called neighbor of each other, and number of neighbors are decided by the user which play very crucial role in analysis of the dataset.

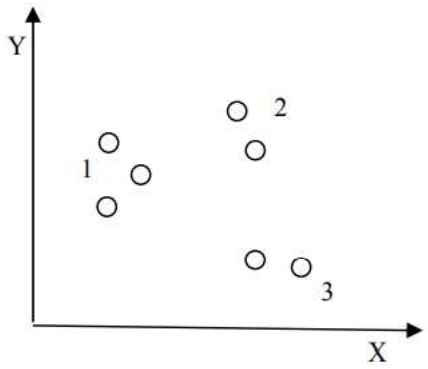


Fig4.5.1: KNN where k=3

In the above Fig. k=3 shows that there are three neighbor that means three different type of data are there. Each cluster represented in two-dimensional space whose coordinates are represented as (Xi, Yi) where Xi is the x-axis, Y represent y-axis and i= 1,2, 3…. n.

Random Forest:

Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

In simple, random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

ACCURACY CALCULATIONS*:*

Accuracy of the algorithms depends on four values namely true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

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

The numerical value of TP, FP, TN, FN defines as:

TP= Number of people with heart diseases.

TN= Number of people with heart diseases and no heart diseases.

FP= Number of people with no heart diseases.

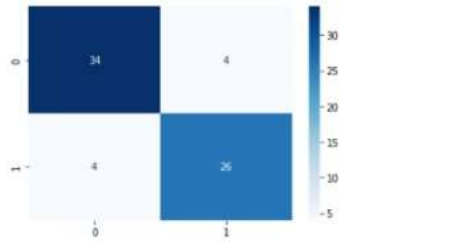
FN= Number of people with no heart diseases and with heart diseases.

5. RESULT

After performing the machine learning approach for testing and training we find that accuracy of the Random Forest is much efficient as compare to other algorithms. Accuracy should be calculated with the support of confusion matrix of each algorithm as shown in Fig.3.4.1 and Fig.3.5.1 here number of counts of TP, TN, FP, FN are given and using the equation of accuracy, value has been calculated and it is concluded that Random Forest is best among them with 89.7% accuracy and the comparison is shown in TABLE.2.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

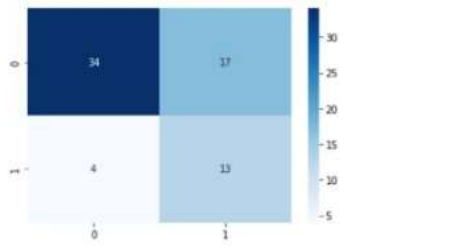
1.Logistic Regresion:



	precision	recall	f1-score	support
Absence	0.89	0.89	0.89	38
Presence	0.87	0.87	0.87	30
accuracy			0.88	68
macro avg	0.88	0.88	0.88	68
weighted avg	0.88	0.88	0.88	68

Fig5.1.1: Logistic Regression Confusion Matrix

2.SVM

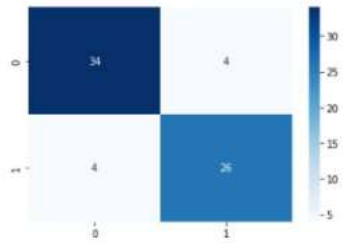


	precision	recall	f1-score	support
Absence	0.89	0.67	0.76	51
Presence	0.43	0.76	0.55	17
accuracy			0.69	68
macro avg	0.66	0.72	0.66	68
weighted avg	0.78	0.69	0.71	68

Fig5.2.1: SVM Confusion Matrix

3.Naive bayes

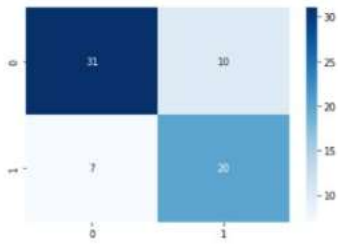
International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org



	precision	recall	f1-score	support
Absence	0.89	0.89	0.89	38
Presence	0.87	0.87	0.87	30
accuracy			0.88	68
macro avg	0.88	0.88	0.88	68
weighted avg	0.88	0.88	0.88	68

Fig5.3.1: Naïve bayes Confusion Matrix

4.Decision Trees:



	precision	recall	f1-score	support
Absence	0.82	0.76	0.78	41
Presence	0.67	0.74	0.70	27
accuracy			0.75	68
macro avg	0.74	0.75	0.74	68
weighted avg	0.76	0.75	0.75	68

Fig5.4.1: Decision Trees Confusion Matrix

5.KNN:

	precision	recall	f1-score	support
Absence	0.79	0.70	0.74	43
Presence	0.57	0.68	0.62	25
accuracy			0.69	68
macro avg	0.68	0.69	0.68	68
weighted avg	0.71	0.69	0.70	68

Fig5.6.1: KNN Confusion Matrix

6.Random Forest:

	precision	recall	f1-score	support
Absence	0.95	0.88	0.91	41
Presence	0.83	0.93	0.88	27
accuracy			0.90	68
macro avg	0.89	0.90	0.89	68
weighted avg	0.90	0.90	0.90	68

Fig5.6.1: Random Forest Confusion Matrix

TABLE.2 Accuracy comparison

Algorithm	**Accuracy**
Logistic Regression	88%
SVM	69%
Naïve Bayes	88%
Decision Tree	75%
KNN	69%
Random Forest	90%

6. CONCLUSION

This system is capable of providing most of the essential features required to predict the heart disease. Heart is one of the essential and vital organs of human body and prediction about heart diseases is also important concern for the human beings so that the accuracy for algorithm is one of parameter for analysis of performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose. When we perform the analysis of algorithms on the basis of dataset whose attributes are shown in TABLE.1 and on the basis of confusion matrix, we find Random Forest is best one.The Random Forest algorithm will perform better than compared to other algorithms.

7. FUTURE SCOPE

For the Future Scope more machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases.

8. REFERENCES

1. [1] Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019.

2. [2] Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, "Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018.

3. [3] Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access 2019.

4. [4] Himanshu Sharma and M A Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.

5. [5] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.

6. [6] Amandeep Kaur and Jyoti Arora,"Heart Diseases Prediction using Data Mining Techniques: A survey" International Journal of Advanced Research in Computer Science , IJARCS 2015-2019.

7. [7] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Diseases Prediction", 4th International Conference on Computing Communication And Automation(ICCCA), 2018.