Forecasting Heart Problems By Implementing Machine Learning Algorithm

Mareedu Sai Sritha
Computer Science
University Of Central Missouri
7007429050

Nandini Ramshetty
Computer Science
University of Central Missouri
700734096

Nallakalva Rishitha Reddy
Computer Science
University Of Central Missouri
700742428

Kolanapaku Sai Sumanth
Computer Science
University of Central Missouri
700741157

Abstract—Gradually the cases of heart diseases are rising at a rapid rate it is important to be able to explain and anticipate such diseases beforehand. The correct prediction of heart disease can prevent life threats.The main aim of this study is to determine which patient is more likely to have heart disease based on a number of medical features.We developed a python based heart disease prediction system as it is most reliable and helps track and establish different types of health monitoring applications and to identify whether the person is likely to be diagnosed with a heart disease or not using the medical history of the person.

In this paper different machine learning algorithms such as KNN Classifier, Random Forest, Decision Tree, Support Vector Classifier are applied to predict and classify the patient with heart disease.Data analysis is required for this application, which is considered significant according to its accuracy rate over training data.The dataset available on Kaggle is used and the model is implemented using python. Different promising outcomes are accomplished and are approved utilizing exactness. The proposed model was likable and was able to predict of having a heart disease in a particular person by using Random Forest classifier which showed a higher accuracy in comparison to the other used classifiers. Moreover, We have used Tkinter standard python GUI library which takes patients data as inputs and produces the output.

Keywords—Supervised, Unsupervised, Reinforced, Regression, Cardiology, Tkinter, Interface.

I. Introduction

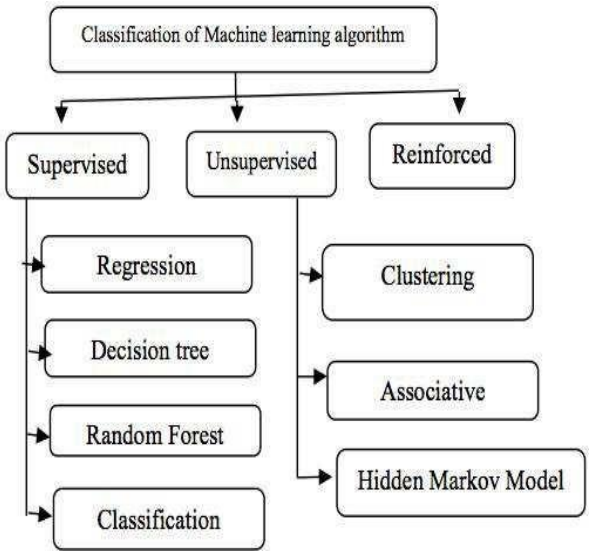
Heart is one of the most broad and imperative organ of human body so the consideration of heart is fundamental. Most of diseases are related to heart so the prediction about heart diseases is necessary and for this purpose comparative study needed in this field, today most of patient are died because 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.

ML is one of the effective innovation for the testing, which depends on preparing and testing. It is the part of Fake Intelligence(AI) which is one of wide area of realizing where machines imitating human capacities, AI is a particular part of computer based intelligence. Then again machines learning frameworks are prepared to figure out how to process and utilize information subsequently the mix of both innovation is likewise called as Machine Insight.

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 in this project we have used four algorithms which are decision tree, linear regression, k-neighbour, SVM. In this paper, we calculate the accuracy of four different machine learning approaches and on the basis of calculation we conclude that which one is best among them.

ML is one of productive innovation which depends on two conditions specifically testing and preparing for examplesystem take training directly from data and experience and in view of this preparing test ought to be applied on various kind of need according to the calculation required.

GIT HUB LINK : https://github.com/rishithareddy3339/ML-Final-Project



A. 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. Managed learning depends on "train me" idea. Supervised learning have following processes:

• Classification

• Random Forest

• Decision tree

• Regression

To perceive examples and measures likelihood of uninterruptable results, is peculiarity of relapse. Framework have capacity to distinguish numbers, their qualities and gathering feeling of numbers which implies width and level, and so forth. There are following administered AI calculations:

• Linear Regression

• Logistical Regression

• Support Vector Machines (SVM)

• Neural Networks

• Random Forest

• Gradient Boosted Trees

• Decision Trees

• Naive Bayes

B. 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 . Unaided learning depends on "independent " idea.

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 Solo learning expressed it as there are three distinct groups.

Unsupervised algorithms have following process:

• Dimensionality

• Clustering

There are following unsupervised machine learning algorithms:

• t-SNE

• k-means clustering

• PCA

C. Reinforcement

Supported learning is the specialist capacity to communicate with the climate and figure out the result. It depends on "hit and preliminary" idea. 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.

II. MOTIVATION

The main motivation of doing this project is to present a heart disease prediction model for the prediction of occurrence of heart disease. Furthermore, this research is aimed at identifying the best algorithm for identifying patients at risk for heart disease. This work is justified by performing a comparative study and analysis using four classification algorithms namely knn, Decision Tree, Support vector and Random Forest are used at different levels of evaluations. These are commonly used machine learning algorithms, but predicting heart disease is a vital task that requires the highest level of accuracy. Hence, the four

algorithms are evaluated at numerous levels and types of evaluation strategies.This will give specialists and clinical professionals to lay out a superior.

By analyzing the large quantities of health care data, machine learning can assist in making predictions and decisions. This undertaking expects to anticipate future Coronary illness by dissecting information of patients which arranges regardless of whether they have coronary illness utilizing ML calculation.

Although heart disease has different forms, there are certain risk factors that influence the likelihood of a person developing it. Our technique can be very well adapted to predict heart disease by collecting information from a variety of sources, categorizing it under appropriate headings, and analyzing it to extract the desired information.

III. OBJECTIVES

1. The significance of diagnosing heart disease in preliminary stage will minimize the risk of patients life.

2. Over here, machine learning will is used to diagnose,detect and forecast many problems in the medical industry.The main purpose of this project is to give disorders analyst a tool to identify cardiac problems at an preliminary stage.

3. Machine learning is used to discover patterns from the user data and then make predictions based on trained data. ML helps in analyzing the data and identifying trends.

4. The objective of this project is to check whether the patient is likely to be diagnosed with any cardiovascular heart diseases based on their medical attributes such as gender, age, chest pain, sugar level, etc.

IV. RELATED WORK

Heart is one of the core organ of human body, it play crucial role on blood pumping in human body which is as essential as the oxygen for human body so there is always need of protection of it, this is one of the big reasons for the researchers to work on this. There is always need of analysis of heart related things either diagnosis or prediction or you can say that protection of heart disease .A variety of fields contributed to this work, including artificial intelligence, machine learning, and data mining.

Performance of any algorithms depends on variance and biasness of dataset[4]. According to investigate on the AI for expectation of heart infections himanshu et al.[4] innocent bayes perform well with low change and high biasness as compare to high variance and low biasness which is knn. With low biasness and high change knn experiences the issue of over fitting this is the justification for why execution of knn get diminished. There are different benefit of utilizing low fluctuation and high biasness in light of the fact that as the dataset little it require less investment for preparing as well as testing od calculation yet there likewise a few burdens of utilizing little 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. Decision tree is one of the nonparametric machine learning algorithm but as we know it suffers from the problem over fitting but it cloud be solve by some over fitting removable techniques. Support vector machine is logarithmic and statics foundation calculation, it build a straight distinct n-layered hyper plan for the characterization of datasets.

The nature of heart is complex, there is need of carefully handling of it otherwise it cause death of the person. The severity of heart diseases is classified based on various methods like knn, decision tree, generic algorithm and naïve bayes [3]. 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.

Some of the researchers have worked on data mining for the prediction of heart diseases. 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.

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, here other different algorithms are naivy bayes, knn and decision tree.

Gavhane et al.[1] have dealt with the multi-facet perceptron model for the expectation of heart sicknesses in person and the exactness of the calculation utilizing computer aided design innovation. If the number of person using the prediction system for their diseases prediction then the awareness about the diseases is also going to increases and it make reduction in the death rate of heart patient.

Some researchers have work on one or two algorithm for predication diseases.Krishnan et al.[2] demonstrated that choice tree is more exact as contrast with the guileless bayes order calculation in their task.

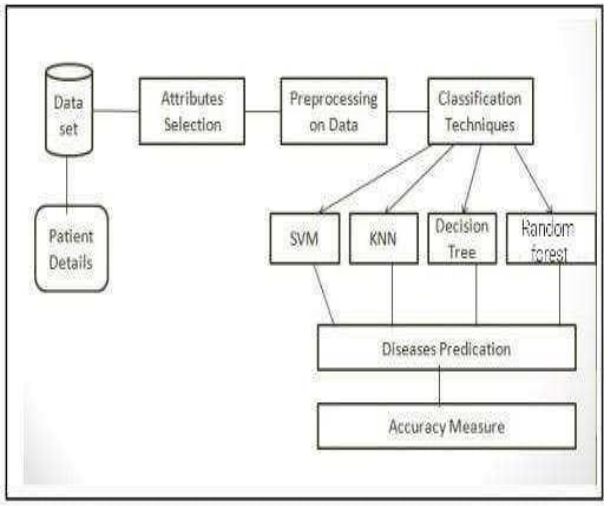
Machine learning algorithms are used for various type of diseases predication and many of the researchers have work on this like Kohali et al.[7] work on heart diseases prediction using logistic regression, diabetes prediction using support vector machine, breast cancer prediction using Adaboot classifier 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.

A study paper on heart infections predication have demonstrated that the old AI calculations doesn't perform

great exactness for the predication while hybridization perform great and give better precision for the predication[8]. But in our project we got 85% accuracy using one classifier than hybrid classifier.

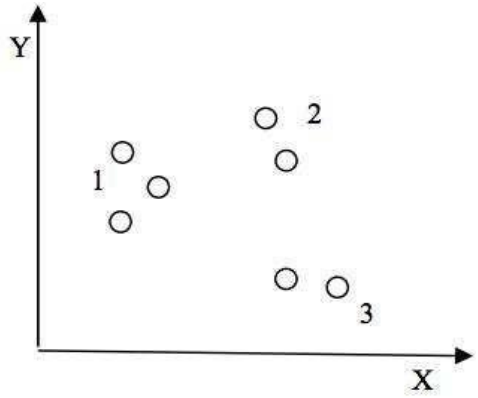
V. PROPOSED FRAMEWORK

Processing of system start with the data collection for this we uses the UCI repository dataset which is well verified by number of researchers and authority of the UCI .



1. KNN :

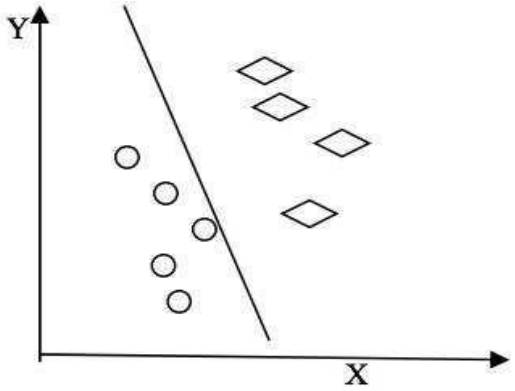
It work on the basis of distance between the location of data and on the basis of this distinct data are classified with each other. All the other group of data are called neighbor of each other and number of neighbor are decided by the user which play very crucial role in analysis of the dataset.



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

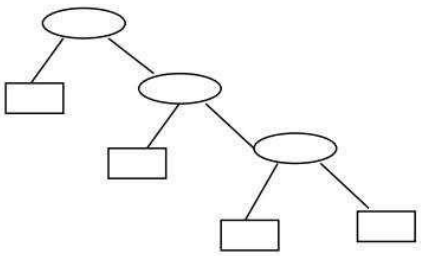
2. SVM :

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 plan decide the type of support vector such as example if a line is used as hyper plan then method is called linear support vector.



3. Decision Tree:

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



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

Entropy= -∑ Pij log Pij

In the above equation of entropy (1) Pij is probability of the node and according to it the entropy of each node is calculated. The node which have 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 are imbalanced then tree is create the over fitting problem which is not good for the calculation and this is one of reason Why decision trees are less accurate than linear regressions.

4. Random Forest:

RF is an ordinary bagging algorithm. Unlike conventional decision trees, RF trains each classifier using a randomly chosen subset of the dataset and a randomly chosen subset of the features.Each prepared classifier produces different expectation results for a similar info. Deciding in favor of the ouput of each prepared classifier, regularly utilizing the majority or the mean, prompts the last forecast outcome. As the elements of the calculation are haphazardly isolated, it will expand the variety of its classifiers and in this way upgrade the model's ability for speculation.

VI. DATA DESCRIPTION

1. Data Collection and Analysis :

First step for predication system is data collection and deciding about the training and testing dataset. In this project we have used 73% training dataset and 37% dataset used as testing dataset the system.

1. chest pain = 0 : typical angina

   1 : atypial angina

   2 : non angina

   3 : asymptomatic
2. blood pressure is in mm Hg.
3. cholestrol is in mg/dl
4. bloodsugar >120mg/dl [1=true; 0=false

   ecg = 0: normal

   1: having ST-T wave abnormality over here T = inversion

   ST = depression >0.05mv

   2:shows left ventricular
5. thalach is reaching max heart rate

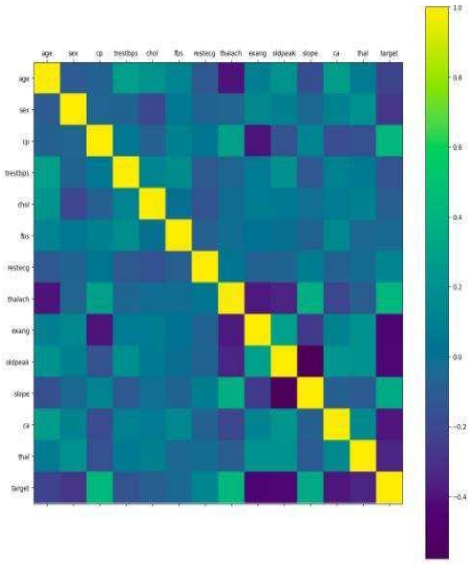
   exang is 1 = yes ; 0 = no
6. thal 3= normal;6 = fixed defect ; 7 = reversible defect

2. Attribute Selection :

Attribute of dataset are property of dataset which are used for system and for heart many attributes are like heart bit rate of person, gender of the person, age of the person and many more shown for predication system.

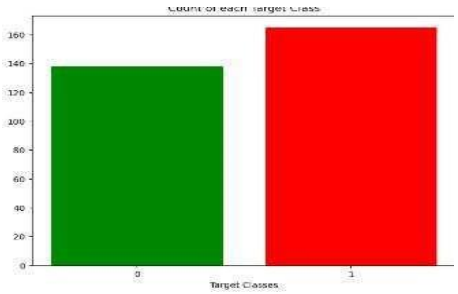
3. Preprocessing of data

Preprocessing required for accomplishing esteemed outcome from the AI calculations. For example Random forest algorithm does not support null values dataset and for this we have to manage null values from original raw data. For our project we have to convert some categorized value by dummy value means in the form of “0”and “1” correlation matrix.



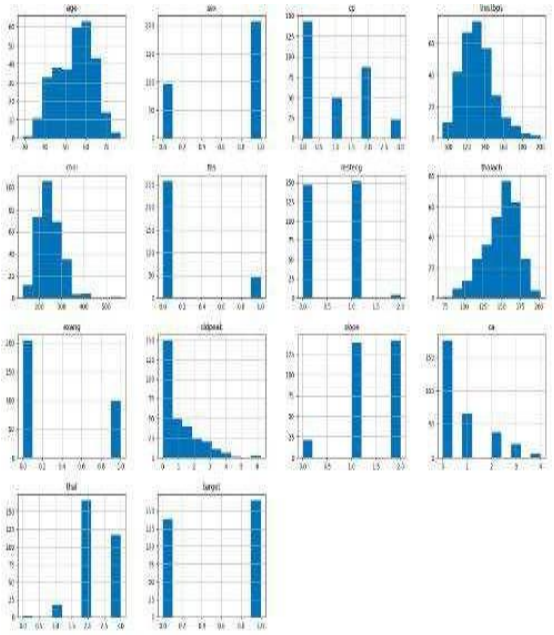
4. Data Balancing :

Data balancing is essential for accurate result because by data balancing graph we can see that both the target classes are equal. This addresses the objective classes where "0" addresses with heart infections patient and "1" addresses no heart illnesses pateints.



5. Histogram of attributes :

Histogram of traits shows the scope of dataset qualities and code which is utilized to make it. dataset.hist()



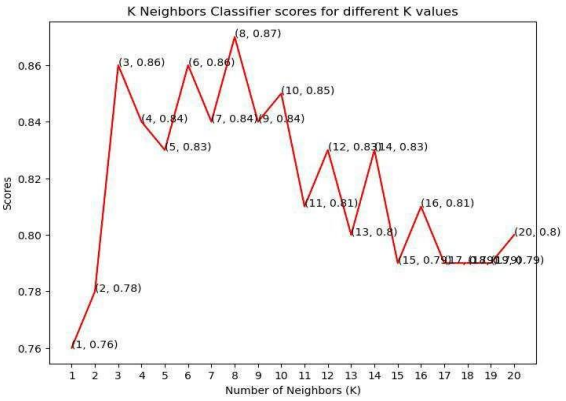
VII. RESULTS ANALYSIS

In our project we have used four machine learning algorithms those are KNN, SVM, Random forest and decision tree classifiers to find the accuracy.

1. KNN Classifier:

The classified score differs based on varied values of neighbors that we choose. Thus, We will plot a score graph for different K values and continuously check when we can get the best score.

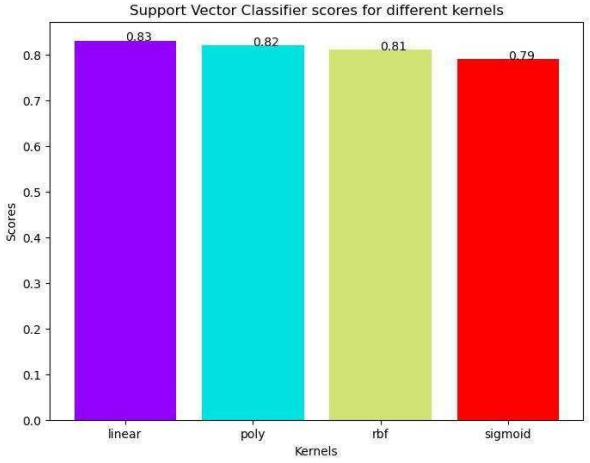
We have the scores for different neighbor values in array knn_scores. We now plot it and check for which value of K we get the best scores.



By the plot above, it is clear that the maximum score achieved was 0.87 for 8 neighbors.

2. SVM Classifier:

There are No.of kernels for Support Vector Classifier. We test some of them and check which has the best score. We now plot a bar plot for each kernel and see which got the best.



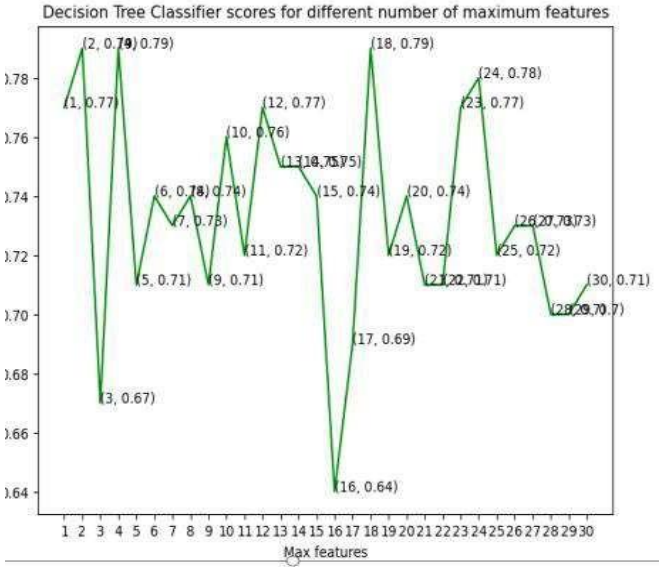
The rbf kernel performed the best, being slightly better than linear kernel. The score for Support Vector Classifier is 82.0% with linear kernel.

3 .Decision Tree Classifier:

Here, We use the Decision Tree Classifier to model the problem. We vary between a set of max_features and observes which returns the best accuracy. We picked the utmost number of features from 1 to 30 to split. Here we can observe the scores for each

case

Algorithm	Accuracy
KNN	82%
Support Vector	82%
Decision Tree	77%
Random Forest	85%

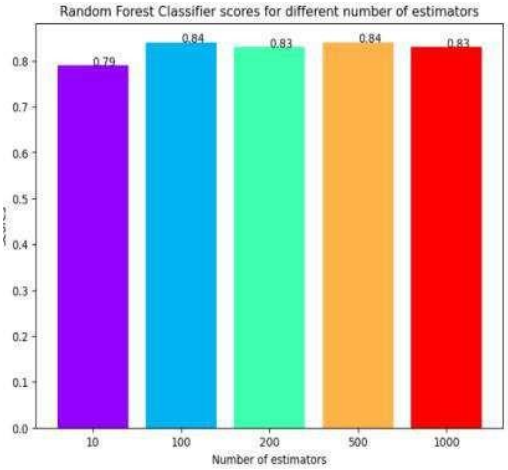


This ML model have attained the good accuracy at three values with maximum features, 2, 4 and 18

The score for Decision Tree Classifier is 77.0% with [2, 4, 18] maximum features.

4.Random Forest Classifier:

Now, We use the ensemble method, Random Forest Classifier, to create the model and vary the No.of estimators to see the effect.
The model is trained and the scores are noted. Now plotting a bar plot to compare the scores.
The highest score is achieved when the total estimators are 100 or 500.

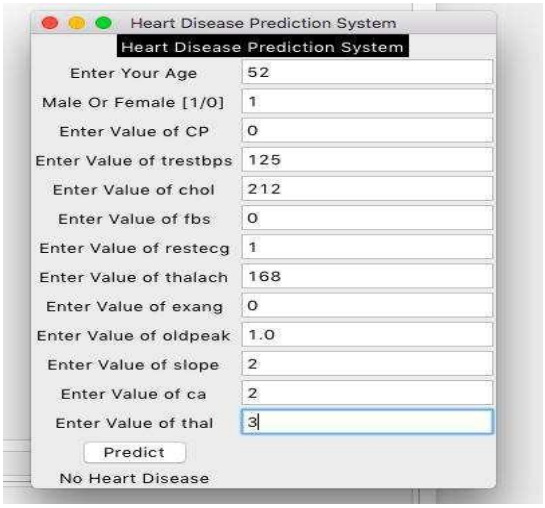


The score for Random Forest Classifier is 85.0% with [100, 500] estimators.

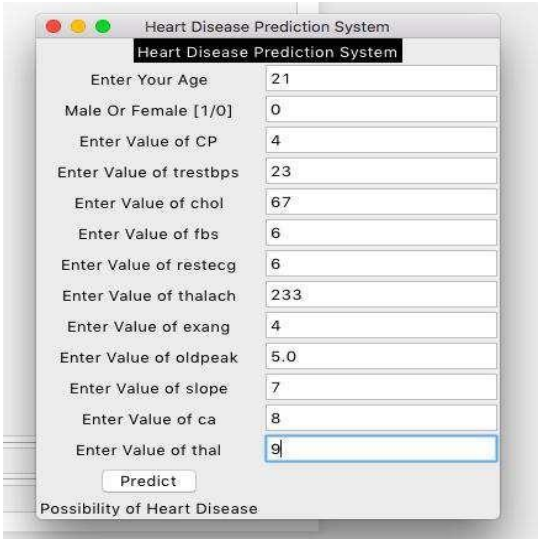
After performing the machine learning approach for testing and training we find that accuracy of the RANDOM FOREST is much efficient as compare to other algorithms. It is necessary to calculate accuracy based on the confusion matrix of each algorithm

Tkinter :

Finally, Inorder to check whether a set of values having a heart disease or not we have developed a interface with the usage of Tkinter library which is a standard python GUI .In this it takes the inputs as patients data and when we hit on the predict button it calculates and checks the range and shows whether he or she is having heart disease or not.



In this, We can observe the values and the prediction status which is having no heart disease.



In this, The patient is likely to get a heart disease with those values. To develop this block in GUI we have imported Tkinter library.

VIII. CONCLUSION

Implemented KNN Classifier, Suport vector classifier , Decision Tree and Random forest classifier to predict the accuracy. By differentiating the above four classifiers we came to know which model is having the best accuracy for heart disease prediction.We have extended this project with the usage of Tkinter python GUI library which takes the input values from users and displays the predicted value.

IX. REFERENCES

1. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.
2. V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.
3. H. Singh, T. Gupta and J. Sidhu, "Prediction of Heart Disease using Machine Learning Techniques," 2021 Sixth International Conference on Image Information Processing (ICIIP), 2021, pp. 164-169, doi: 10.1109/ICIIP53038.2021.9702625.
4. A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.
5. Senthilkumar Mohan Chandrasegar Thirumalai and Gautam Srivastava "Effective heart disease prediction using hybrid machine learning techniques" IEEE Access vol. 7 no. 2019 pp. 81542-81554.
6. A. Singh and R. Kumar "Heart Disease Prediction Using Machine Learning Algorithms". 2020 International Conference on E. "Ali Liaqat" An optimized stacked support vector machines based expert system for the effective prediction of heart failure" IEEE Access vol. 7 no. 2019 pp. 54007-54014 2010.
7. D. Baroud A. N. Hasan and T. Shongwe "A study towards implementing various artificial neural networks for signal classification and noise detection in PFDM/PLC Channels" 12th IEEE international symposium on communication systems networks and digital signal processing (CSNDSP) 2020.
8. Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019.
9. 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.
10. Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access 2019.
11. 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.
12. 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.
13. 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.
14. Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Diseases Prediction", 4th International Conference on Computing Communication And Automation(ICCCA), 2018.
15. M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.
16. S. Kumra, R. Saxena, and S. Mehta, "An Extensive Review on Swarm Robotics," pp. 140–145, 2009.

17. Hazra, A., Mandal, S., Gupta, A. and Mukherjee, " A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review" Advances in Computational Sciences and Technology , 2017.
18. Patel, J., Upadhyay, P. and Patel, "Heart Disease Prediction Using Machine learning and Data Mining Technique" Journals of Computer Science & Electronics , 2016.
19. Chavan Patil, A.B. and Sonawane, P."To Predict Heart Disease Risk and Medications Using Data Mining Techniques with an IoT Based Monitoring System for Post-Operative Heart Disease Patients" International Journal on Emerging Trends in Technology, 2017.
20. . Kirubha and S. M. Priya, "Survey on Data Mining Algorithms in Disease Prediction," vol. 38, no. 3, pp. 124–128, 2016.
21. M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.