Miss. D. Gowdhami[1], Mr. S. Sivasaravanababu[2], R. Aravindhan[3], D. David Ezra[4]

Heart health prediction and classification using random forest and SVM algorithm in ML

Miss. D. Gowdhami[1], Mr. S. Sivasaravanababu[2], R. Aravindhan[3], D. David Ezra[4]

Abstract

Technology is helping us a lot to improve the health sector in both pre and post-diagnostics. We wanted to leverage the AI/ML technology in the area of cardiac health prediction as we found out that cardiac-related health issues lead to fatal circumstances frequently. Our idea is to make an attempt to predict the heart health condition using some vital information collected from the patient. We have collected about 500 patient's sample health data using which we wanted to build a prediction model using AI/ML. Our focus is more on implementing a prediction model as a technology Vs accuracy of the prediction as we are not the subject matter expert in the field of cardiac health. However, we are planning to talk to some medical care experts/doctors to understand some basics about cardiac health and its vital parameters to assess the condition, etc so that our project reflects reliability to some extent for the scope of our submission but later it can be extended to reflect the complete reliability as we gain more knowledge and feed more data to the model.

Keywords- ML, Hearth Health Prediction

INTRODUCTION

A major challenge faced by health care organizations (hospitals, medical centers) is that the provision of treatments that are qualitatively effective. Poor and ignorant clinical decisions can result in disastrous consequences and are unacceptable. Hospitals must also minimize the price of clinical tests. They will achieve these results by employing support systems. Most hospitals today employ some style of hospital information systems to manage their health care or patient data. These systems typically generate huge amounts of knowledge which take the shape of numbers, text, charts, and pictures. Unfortunately, these data are rarely wont to support clinical higher cognitive process. We turn data into useful information that may enable health care practitioners to form intelligent clinical decisions. This can be the most motivation for this research.

[1]Asst. Prof, Department of Electronics and Communication Vel Tech High Tech Dr.RR Dr.SREngineering College, Avadi
[2]Asst. Prof, Department of Electronics and Communication Vel Tech High Tech Dr.RR Dr.SREngineering College, Avadi
[3]Department of Electronics and Communication Vel Tech High Tech Dr.RR Dr.SREngineering College, Avadi
[4]Department of Electronics and Communication Vel Tech High Tech Dr.RR Dr.SREngineering College, Avadi

To design a Health Prediction System for medical data classification and early disease prediction by using supervised learning. The system effectively identifies the disease and its subtypes, the subtype which is referred to as the level of class such as normal and mild or extreme. Using combinatorial methods from data mining decision making has been simplified and our proposed algorithm achieved the highest accuracy.

Machine learning is a part of data science wherein computer systems are made to learn from different data sets based on patterns generated from the datasets. It enables a machine to learn on its own based on some useful algorithms specifically developed for it. Machine learning is being heavily used in all sectors including manufacturing, healthcare, Research, and Development, etc. In this work, we are predicting the heart disease occurrence in a patient based on some important characteristics which are best suited based on our data set that we have collected.

The circulatory system is one of the needed life processes. And the heart is one of the vital human organs, it plays a crucial role in the circulatory system. The main function of the heart is to pump blood through the blood vessels of the circulatory system. If the functioning of the heart is improper then it will affect health adversely and can even lead to death.

In section II, we discuss what are all algorithms prevalent in the market and what all symptoms do they measure. Our main aim behind developing the system is to make it user-friendly so that regular monitoring of the patient is made possible. Thus we aim to get maximum accuracy with minimum input data from the user

Section III. discusses our system design and expounds on its workflow and structure. The algorithm we have used is explained in greater detail.

Related Works

Predicting heart diseases has been one of the favored research topics. Various algorithms were utilized by different researchers to try and do the identical by collecting data from different areas.

Many researchers are have tried to unravel this by using Deep Neural Network and also the different algorithm has been used in ML-like(Decision Tree, k-nearest neighbors) and most of the researchers have only handle data and its accuracy level and they do not have any front end to communicate with the user

There are many approaches within the literature for detecting artery illness. within the present work, a choice tree classification algorithm has been to assess the events related to cardiac disease. Indecision tree ID3, C4.5 algorithms are accustomed perform a number of the guts Attack Prediction projects get an accuracy level of just about 81% in their project they use to gather plenty of information from the user for prediction

The major disadvantage of what we analyze from previous work are Detection isn't possible at an earlier stage and a few others are the practice to use various collected data is time-consuming .and formal logic isn't always accurate, therefore the results are perceived as a supported assumption, so it should not be widely accepted. These systems haven't got the potential of

Miss. D. Gowdhami[1], Mr. S. Sivasaravanababu[2], R. Aravindhan[3], D. David Ezra[4]

machine learning yet as neural network type pattern recognition. to beat all this disadvantage we came out with some new ideas

Proposed system

We proposed a system that is proved to be an aid for the physicians with the diagnosis. To design a Health Prediction System for medical data classification and early disease prediction by using supervised learning. The system effectively identifies the disease and its subtypes, the subtype which is referred to as the level of class such as normal and mild or extreme. Using combinatorial methods from data mining decision making has been simplified and our proposed algorithm achieved the highest accuracy. we use a Support Vector Machine, random forest algorithm, and Logistic regression

SVM

By using a Support Vector Machine we can solve both classification and regression type of problems. The aim of using the SVM algorithm is to make the best line between the boundaries that can segregate n-dimensional space. By doing this we can put the data point in the right category. By doing this we will get the best decision boundary known as a hyperplane. For creating a hyperplane extreme vectors are chosen by SVM These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine.

The primary advantage of using SVM is it has a high performance and high accuracy rate and SVM has very flexible and is widely in various domains with high rates of success and SVM is relatively memory efficient. and when the number of dimensions is greater than the number of samples it will be very efficient.

SVM is as you see in Fig 3.1 the centerline is the margin which is used to split the data and other to the line is support vectors and the centerline is known as separating hyper plane
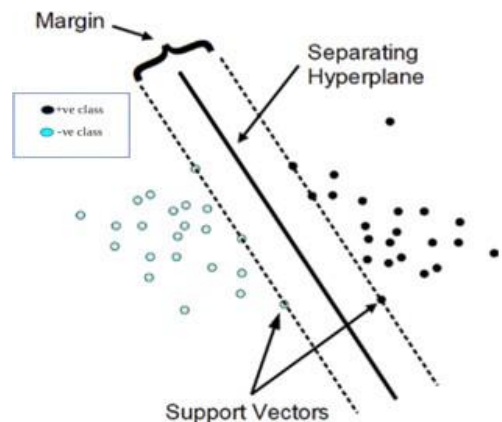

Fig 3.1 support vectors
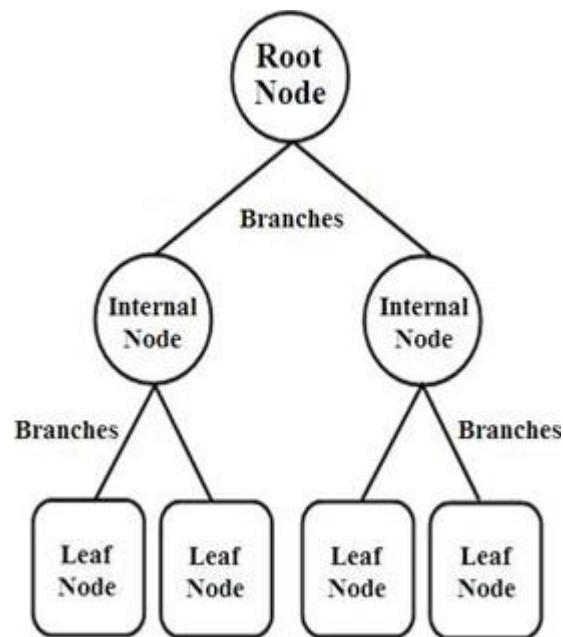
Random Forest

Random forest is a supervised learning algorithm. Building an ensemble of the decision tree is referred to as a forest, generally, these decision trees are trained with a bagging method. The

purpose of the bagging method is that when we do the combination of learning models the overall result will be increased.

First when we see the decision tree is a chart that people use to conclude a course of the show a statistical probability. At the end of Each branch of the decision, a tree represents a possible decision. The major advantage of using a decision tree is it is simple to understand, interpret, visualize. and Can also handle multi-output problems. The below image is a simple exp of decision tree

When we see Fig 3.2 the first is the root node which is the starting node and the second is the internal node and the connecting is called branches and the last node is the leaf node which is helping to make a decision

Fig 3.2 Decision Tree

The combination of more decision trees is called a random forest. The more tree in the forest the more robust the result the below fig3.3 is the example for random forest

Miss. D. Gowdhami[1], Mr. S. Sivasaravanababu[2], R. Aravindhan[3], D. David Ezra[4]
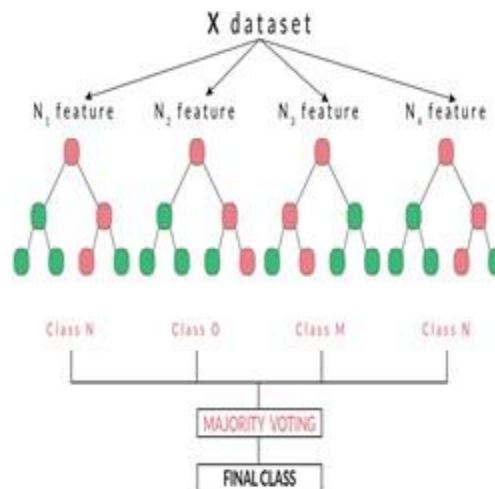
Fig 3.3 Random Forest

In this study, a prototype of a system that comprises a binary classification model to predict the heart attack risk factor of an individual based on his/her medical data is proposed. The system is well equipped with a comprehensive UI and UX that is easy to use and understand. The classification follows supervised learning wherein the dataset used was obtained from the reliable open-source platform. The dataset represents a binary classification problem. The dataset comprises 14 attributes in total, out of which 13 are predictor variables and one feature is a binary response variable. Therefore, the dataset represents a binary classification problem. In the end, an easy-to-use web interface was developed so that the system can be used by persons with little or no technical knowledge, thereby completely abstracting the core functionality and implementation details of the system
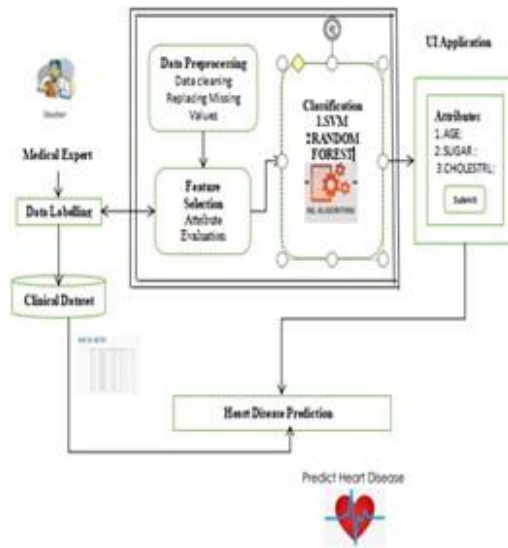

Fig 3.4 Architecture diagram

IMPLEMENTATION

Data Collection
In this process, by using the proposed method the collected data is analyzed. the dataset is employed in our research for identifying the performance of proposed methods.

Age: displays the age of the user

Sex: using the binary format displaying 1 as male and 0 as female the gender of the individual is displayed during this format

Chest-pain type: displays the kind of chest pain experienced by the user using the subsequent **format :**
1 = typical angina 2 = atypical angina
3 = non — anginal pain 4 = asymptotic

Resting Blood Pressure: mmHg (unit) this is often the format that we use to displays the resting

4535

pressure value of a personal

Serum Cholesterol: mg/dl (unit) during this format we use to displays the serum cholesterol
Fasting Blood Sugar: 120mg/dl is that the standard fasting blood glucose value which is able to be compared with a private
If fasting blood glucose > 120mg/dl then : 1 (true) else : 0 (false)

Resting ECG:displays   resting electrocardiographic results
0 = normal
1 = having ST-T wave abnormality 2 = left ventricular hypertrophy

Max rate achieved: this can be used to display the max vital sign achieved by a private.
chest pain during Exercise-induced : 1 = yes 0 = no
This data is collected from the project from a reliable open-source platform

Data Preprocessing

In this project, we collected data from different sources so the data will be in raw format which is not good for the analysis. For we use a technique to convert the raw data into a clean data set. we use it to check whether it has any missing data if any missing data is found we use it to add duplicate values if we leave the missing data this will affect the final result For the purpose of data preprocessing data needs to go through several steps that are. First, it will start with data cleaning where the data will be cleaned by filling the missing data in the data set resolving the inconsistencies in the data. and it will go to data integration and data transformation and finally, it will go to data reduction Here the raw data in the heart disease data is cleaned and the metadata is appending to it by removing the things which are converted to the integer. So, the data is easy to train. Hear all the data. during this pre-processing, we first load the metadata into this, then this metadata is going to be attached to the information and replace the converted data with metadata. This data will be moved to train and test data. train test split which is in scikit-learn which will take care of splitting data based on the weight given in the code. We took 0.2 and 0.8 that is 20 percent as test data and 80 percent as train data Here SVM is used for these processes.

Classification

For classification, the algorithm was used from the scikit-learn package. Multiple data mining tools are encapsulated by scikit-learn. As we see before training data and test data set were split. The training data was converted into a NumPy array and an algorithm was applied. The test data is fed into the model accuracy and other parameters were calculated. The classification method makes use of the Random Forest algorithm.
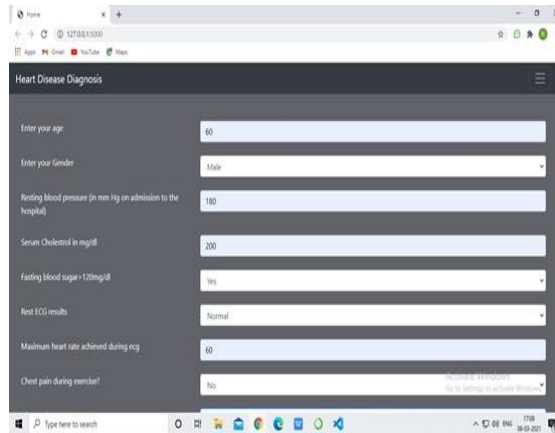
User Interface / User Experience

An interactive web interface was developed to access the classifier and check the risk factor. The backend consists of a CGI script written in Python. It accepts the medical form data as input to the classification model trained and predicts the risk factor of the individual The sequence of events taking place while using

the system The predictor receives the form data from the web interface and inputs the same to the classification model. Once the model makes a prediction, it is displayed to the user and simultaneously updated on the local database.

Result

Based on the values entered by the user on the web page the input goes to the ML module and finally, it will predict whether the patient is having Heart disease or not. The prediction is classified into 5 type thy are stage 1 or 2 or 3 or 4 .and 0 which is no disease and overall we obtained more than 97% Fig 7.2 shows the output of the given input data
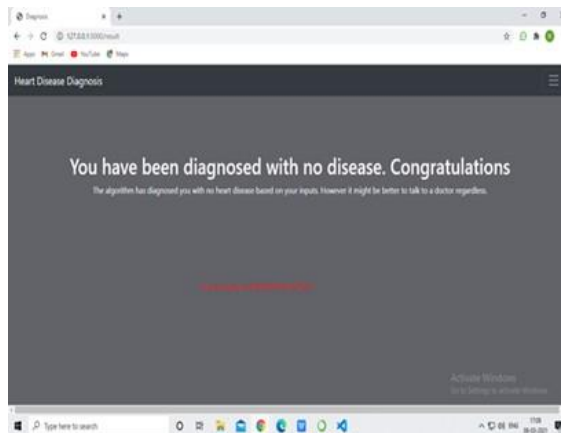
Fig. 7.1 shows the web page where the user needs to enter the values for(Age, enter the value of gender, enter the value of cholesterol &amp, blood pressure).

Fig. 7.1 Web Page Input

Fig. 7.2 Web Page Output

As you see in Fig 7.2 shows the output wether The user is diagnosed with a disease or not if it is yes then it will show what stage it is

CONCLUSION

Nowadays heart diseases became quite common thanks to the unhealthy lifestyle we've got and ignored the first symptoms of heart diseases may find yourself in death. To avoid unnecessary things and to understand more about our health condition this can be very helpful to watch ourselves. During this covid time, taking care of health is important. within the above paper, we've studied various classification algorithms which will be used for the

classification of cardiopathy databases also we've seen different techniques that may be used for classification and also the accuracy obtained by them was low. This investigation tells us about dissimilar technologies utilized in dissimilar papers with a dissimilar count of attributes with different accuracy reckoning on the tools designed for execution. we've obtained quite 97% it's high compared to other papers within the future, we've planned to merge with a sensible device which will automatically enter the worth and therefore the result is directly given to the user and there are lots of smart devices which doesn't have this feature we attending to implement it

REFERENCE

1. Prediction of Heart Disease Using Machine Learning, Aditi Gavhane; Gouthami Kokkula;
2. Isha Pandya; Prof. Kailas Devadkar
3. Blood Viscosity based Heart Disease Risk Prediction Model in Edge/Fog Computing, R Latha ; P Vetrivelan
4. Predictive Analysis of Rapid Spread of Heart Disease with Data Mining, Radhanath Patra; Bonomali Khuntia
5. Newfangled Approach for Early Detection and Prevention of Ischemic Heart Disease using Data Mining, Dhara B. Mehta; Nirali C. Varnagar,
6. Real-time machine learning for early detection of heart disease using big data approach Abderrahmane Ed-Daoudy; Khalil Maalmi
7. Automated Prediction of Heart Disease Patients using Sparse Discriminant Analysis,
8. K. M. Zubair Hasan; Shourob Datta; Md Zahid Hasan; Nusrat Zahan
9. [7].Heart Disease Prediction using Evolutionary Rule Learning, Aakash Chauhan; Aditya Jain; Purushottam Sharma; Vikas Deep An Investigation of Heart Disease PredictionTechniques, G. Shanmugasundaram; V. Malar Selvam; R. Saravanan; S. Balaji Non-Laboratory-Based Risk Factors for Automated Heart Disease Detection, H. Mai; T. T. Pham; D. N. Nguyen; E. Dutkiewicz
10. Heart Disease Prediction System using CART-C, Priyal Chotwani; Asmita Tiwari; Vikas Deep; Purushottam Sharma Sampathkumar, A., S. Murugan, Ahmed A. Elngar, Lalit Garg, R. Kanmani, and A. Christy Jeba Malar. "A Novel Scheme for an IoT-Based Weather Monitoring System Using a Wireless Sensor Network." In Integration of WSN and IoT for Smart Cities, pp. 181-191. Springer, Cham, 2020.
11. Sampathkumar, A., R. Maheswar, Pon Harshavardhanan, S. Murugan, P. Jayarajan, and V. Sivasankaran. "Majority Voting based Hybrid Ensemble Classification Approach for Predicting Parking Availability in Smart City based on IoT." In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-8. IEEE, 2020

Miss. D. Gowdhami[1], Mr. S. Sivasaravanababu[2], R. Aravindhan[3], D. David Ezra[4]

12. UshaKiruthika,**S. Kanaga Suba Raja,** C.J. Raman ,V.Balaji. (2020) 'A Novel Fraud Detection Scheme for Credit Card Usage Employing Random Forest Algorithm Combined with Feedback Mechanism', IEEE Second International Conference on Power, Energy, Control and Transmission Systems (ICPECTS2020), Sairam Engineering College, Chennai ,Tamilnadu, India. **(Scopus Indexed)**

13. UshaKiruthika,**S. Kanaga Suba Raja,**V.Balaji ,C.J. Raman, (2020) 'E-Agriculture for Direct Marketing of Food Crops using Chatbots', IEEE Second International Conference on Power, Energy, Control and Transmission Systems (ICPECTS2020), Sairam Engineering College, Chennai Tamilnadu, India. **(Scopus Indexed)**

14. Abukharis, S., Alzubi, J. A., Alzubi, O. A., Alamri, S., & O'Farrell, T. (2016). Packet error rate performance of IEEE802. 11g under bluetooth interface. *arXiv preprint arXiv:1602.05556.*