Smart Health Prediction Using Data Mining

K.Praveen[1], G.Prasanth[2], M.Thiruppathi[3]
[1, 2, 3] Dept of CSE
[1, 2, 3] Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract- *Heart disease is one of the most commonly found chronic disease that has become a mainstream health issue with the current lifestyle. It is essential to identify the symptoms and treat the disease at early stages. Data mining practices are used in number of applications. It is an exercise of determining a large amount of pre-existing database to produce new information. In health care system data mining renders a vital role to predict the illness with the given symptoms and classify the disease. The major reason of data mining in health care system is to evolve a new automated tool for determining and diffusing pertinent health care information. Here, the system is fed with various attributes. According to those attributes, the system compares the given symptoms with the actual dataset and predicts the relevant disease based on the user input. In this system, Random Forest model has been used for prediction. The goal is to develop a cost-effective and easily accessible healthcare system that can benefit the medical practitioners to combat the prolonged procedures of diagnosis and faster retrieval of results.*

Keywords- Heart Disease Prediction, Data Mining, Random Forest, Easily accessible.

I. INTRODUCTION

Smart Health Care Prediction using Data Mining is a new powerful technology which is of high interest in computer world. It is a sub field of computer science that uses already existing data in different database to transform it to new researches and result. The actual task is to extract data by automatic or semi-automatic means. The different parameters included in data mining include clustering, forecasting, path analysis and predictive analysis. With the growing researching the field of health informatics a lot of data is being produced. The analysis of such a large amount of data is very hard and requires excessive knowledge. Smart health care applies data mining techniques for health diagnosis. Data mining refers to extracting meaning full information from the different huge amount of dataset. It is the process of determining the unseen finding pattern and knowledge from the massive amount of data set. Data mining is significant research doings in the field of medical sciences since there is a requirement of well-organized methodologies for analysing, predict and detecting diseases. To detect and predict diseases Data mining applications are used for the management of healthcare, health information, patient care system, etc. It also plays a major role in analysing survivability of a disease. Data mining classification techniques play a vital role in healthcare domain by classifying the patient dataset. Data mining classification technique is used to analyses and predicts many diseases. The classification techniques like Feature selection methods, improve the performance accuracy of the algorithm by reducing the dimensionality of the feature and it can be grouped into a wrapper and filter method. Tendency for data mining application in healthcare today is great, because healthcare sector is rich with information, and data mining is becoming a necessity.

1.1 Background and motivation:

Healthcare organizations produce and collect large volumes of information on daily basis. Use of information technologies allows atomization of processes for extraction of data that help to get interesting knowledge and regularities, which means the elimination of manual tasks and easier extraction of data directly from electronic records, transferring onto secure electronic system of medical records which will save lives and reduce the cost of the healthcare services, as well and early discovery of contagious diseases with the advanced collection of data.

Raw data from healthcare organizations are voluminous and heterogeneous. They need to be collected and stored in the organized forms, and their integration enables forming of hospital information system. Healthcare data mining provides countless possibilities for hidden pattern investigation from these data sets. These patterns can be used by physicians to determine diagnoses, prognoses and treatments for patients in healthcare organizations. Data mining is the process of extracting hidden information from massive dataset, categorizing valid and unique patterns in data. There are many data mining techniques like clustering, classification, association analysis, regression etc. Heart disease has become a serious problem in the world in which the heart is damaged and it is the cause of improper function of heart. Life is dependent on competent functioning of heart, because heart is necessary part of our body. If function of heart is not suitable, it will affect the other body parts of human such as brain, kidney etc. Heart disease is a disease that affects on the function of heart. There are number of factors which increases risk of heart disease. At the present days, in the world heart disease is the main cause of deaths. The World Health Organization (WHO) has expected that 12 million deaths occur worldwide, every year due to the heart diseases. Prediction by using data mining techniques gives us accurate result of disease. It can answer complex queries for diagnosing the disease and thus help healthcare analysts and practitioners to make intelligent clinical decisions which conventional decision support systems cannot. There is vast potential for data mining applications in healthcare.

1.2 Problem statement:

Heart disease can be managed effectively with a combination of lifestyle change, medicine, and in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The prediction results can be used to prevent and thus reduce cot for surgical treatment and other expenses.

The goal is to develop a strong and accurate Heart disease prediction system that fulfils the above problems by predicting with few tests and attributes and deliver the result of presence of heart disease based on the prediction.

1.3 Objective of the project:

The main objective of this project is to develop a heart prediction system. The system can discover and extract knowledge associated with the diseases from a historical heart disease data set.

Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

To list out, the objectives are as follows:

- Provides new approach to concealed patterns in the data.
- Helps avoid human biasness.
- To implement Random Forest Classifier that classifies the presence of disease as per the input of the user.
- Reduce the costs involved in manual prediction.

II. REVIEW OF LITERATURE

[1] Harshitha M et al. proposed a system based on the Naïve Bayes algorithm. The methodology is to implement a smart health care prediction system that make use of data mining technique which include Naïve Bayes algorithm, this procedure is expressed as “Knowledge Discovery Process”, this process includes the steps: Data Selection, Data Pre-processing, and implementing the Naïve Bayes algorithm. They also concluded with the results as to examine essential patterns to predict and classify whether the symptoms of the patient indicate as heart disease or diabetes or neither of these. The results are displayed using R shiny which is an open source package in R to provide most powerful web application framework.

[2] Pradnya Suresh Joshi et al.proposed a prediction system using the Random Forest Classifier. This system supports an end user and online consultation. They proposed a framework that enables clients to get moment direction on their medical problems through an astute social intelligent health care system online. Also the system allows user to share their symptoms and issues. This system allows user/patients to share their symptoms and issues It then processes patients symptoms to check for various illnesses and based on input it predict the disease or disorder it feels user’s symptoms are associated with and also suggest the doctor to whom he or she can contact and also book an appointment.

[3] Manisha M S Pillai et al. proposed a prediction system for breast cancer using Logistic regression algorithm. It is more suited technique for prediction of breast cancer. They used two values 0 and 1, malignant is represented with a value of 1 and benign is represented with 0, as it is more efficient. User login into system using username and password, after login to the system, where the user can give clinical result data. They used the dataset taken from Wisconsin repository. Comparing the obtained data with user given data, they could predict the presence of breast cancer, accuracy of the system is high.

[4] N. Vijay Kumar et al.proposed a system that is based on the K-Nearest Neighbors (KNN) and the Decision Tree algorithms. They followed thee simple steps: Loading the Training Dataset. Pre-processing (remove unwanted values in Training Dataset), Map and Reduce (a programming model and an associated implementation for processing and generating big data sets with a parallel, distribution algorithm on cluster. Classify data usingclassificationalgorithms,KNNandDecisionTree.

[5] Konde T.R et al. proposed a Health Care System, which is a web based application and could be accessed throughout the specified department to handle the various processes involved in Smart health Prediction System where Patient can see various Doctor. Software also lists various expert Doctors available where user can search Doctor for their medical issue. If user’s symptoms do not exactly match any disease in the database, then it is shows the diseases user could probably have based on his/her symptoms.

[6] Pinky Saikia Dutta et al. proposed a prediction system, which is a web based application for Predicting diseases based on user input symptoms. It predicts diseases by mining data sets and provides remedial solutions for Effective Treatment. The core objective of their project is to develop a web application using data mining concept accompanied by JSP (Java Server pages) technology and MYSQL. The algorithm used for prediction is the Apriori Algorithm.

[7] J. Manikandan et al. proposed a system that allows users to get instant guidance on their health issues through an intelligent health care system online. The system is fed with various symptoms and the disease/illness associated with those systems. The system allows user to share their symptoms and issues. They use some intelligent data mining techniques to guess the most accurate illness that could be associated with patient's symptoms. If the system is not able to provide suitable results, it informs the user about the type of disease or disorder it feels user's symptoms are associated.

[8] N. Shabaz Ali et al. proposed a way to choose the best algorithm for health prediction using data mining. They ran through various algorithms like Support Vector Machine (SVM), Neural Networks, Logistic Regression, Discrimination analysis, Random forest, Linear Regression, Naïve Bayes, Nearest Neighbor, Decision tree, and came to a conclusion that Neural Networks, Random Forest and Decision Trees show higher accuracies than the other algorithms.

[9] Prof. Krishna Kumar Tripathi et al. proposed a system, in which hidden knowledge will be extracted from the historical data by preparing datasets by applying Naïve Byes algorithm. The main features of their system will be giving instant diagnosis on the user entered symptoms and getting tips for remaining fit. In the proposed system, they use the data mining method in which the symptoms entered by users are cross checked in the database and from that the frequent item sets are mined out of the existing datasets.

[10] G.Pooja reddy et al. proposed a specialist framework using Naïve Bayes and Decision Tree Algorithms, which is called Smart Health Prediction framework, which is utilized for improving the task of specialists. It involves fundamental parts, for example, quiet login, enter side effects in the System, and recommend medications, proposes an adjacent specialist. The application takes the contribution of different manifestations from the patient, does the examination of the entered side effects, and gives fitting sicknessexpectation.

III. PROPOSED SYSTEM

A Good GUI Desktop Application for accurate prediction of Heart Disease in patients. As only Python is going to be used, the entire system will be a light weight application that can seamlessly run even on low memory systems.

The user/medical practitioners have to enter the required details after taking those mentioned tests. The GUI can then handle the compilation of values and pass them to the Prediction module, so that the saved Random Forest model is loaded and the prediction takes place. There will be no need to re-train the model each time the user enters the data.

Finally, the result will be shown neatly in a separate window which intimates whether the patient is of High Risk or Low Risk.

The whole process will happen effortlessly whereby, the required tests have to be taken prior to the prediction using this tool.

IV. METHODOLOGY

4.1 Classification:

Classification comprises of two steps: - 1) Training and 2) Testing. Training builds a classification model on the basis of training data collected for generating classification rules. The IF-THEN prediction rule is popular in data mining; they signify facts at a high level of abstraction. The accuracy of classification model based on the degree to which classifying rules are true which is estimated by test data.

4.2 Prediction:

Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown. Prediction in data mining is to identify data points purely on the description of another related data value.

4.3 Algorithm:

-> Random forest classifier:

Random forest classifier, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the majority of votes becomes our model's prediction. Random forest fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and also control over-fitting.

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. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, 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. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

-> Advantages:

Overfitting is one of the critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't overfit the model. The another advantage is the classifier of Random Forest can handle missing values, and the Random Forest classifier can be modelled for categorical data.

4.4 Modules:

4.4.1 Main Module:

The main module consists of the declarations and imports of other modules, where the required dependencies are passed between them through inheritance. Basically, this module connects the other modules.

4.4.2 Random Forest Model Build Module:

The Random Forest Model Build Module is like the heart of the program, as it is in this module, the random forest model gets built by tuning its necessary parameters. Initially the dataset is imported from the 'heart.csv' file to a dataframe object with the help of the Pandas library, and this dataframe is then used for further processing. Later, the dataframe is then split into Train and Test sets with which the building of the random forest model takes place. The Test set is used to test the model that is built on the Train set.

Once the model is built successfully, it is then saved to local directory with name 'rf_model.pac' with the help of the Joblib package for Python. This local dump file of the trained and the tested random forest model will be used in the GUI part later for seamless and rapid result delivery, instead of training the model each time the data is entered into the GUI. The file 'rf_modelc.pac' is the compressed version of the original dump file, which is lighter than the original dump file.

This module acts as the major independent module and will not be called whenever the program runs to predict the result. As this module carries out the major task of saving the trained random forest model to be used later, this module also helps in reducing a great amount of time by not training the whole dataset again and again to predict every instances of data for prediction.

The saved model file is then imported by the Prediction module so that the collected data is used for prediction. This import of the random forest model file takes place with the help of the same Joblib library package, which helped to save the model.

4.4.3 Prediction Module:

As the name suggests, this model does the important job of importing the saved random forest model from the local directory. Also, this module has the prediction function, by which the collected data from the GUI module is passed, and the prediction takes place effortlessly. The result of the prediction is then returned by this function, so that it is passed to GUI for the result delivery.

4.4.4 GUI Module:

The GUI Module is one of the major part of this project, as it is with this module, the user interacts with the application to enter and select values that will be caught and sent for the prediction. This GUI module is entirely designed with the help of Tkinter library which has the components or the widgets for the required GUI design. The whole application is planned and neatly drawn out using the previously mentioned Tkinter library and gives the user a satisfactory UX.

The application is set to load initially for 3 seconds to provide a latency, which in-turn simply improves the usage. After the loading progresses, the widgets like the Entry, Buttons, Dropdowns, etc. are displayed.

The logo to the left of the title is a button specially assigned for displaying the "About" information of the application, on clicking it. Similarly, the "Info" to the right of the title is a button that is assigned to display the usage information of the application, on clicking it.

Apart from this miscellaneous buttons, the main body section of the GUI has the attributes that has to be filled by the user for prediction. The attributes include Age, Sex, Chest Pain Type, Resting Systolic Blood Pressure, Serum

Cholesterol, Fasting Blood sugar > 120mg/dl ?, Resting ECG results, Max Heart Rate, Exercise induced Angina ?, ECG – Old Peak, ECG – Slope, Cardiac Fluoroscopy – No. of Major Vessels, and Thallium Scintigraphy.

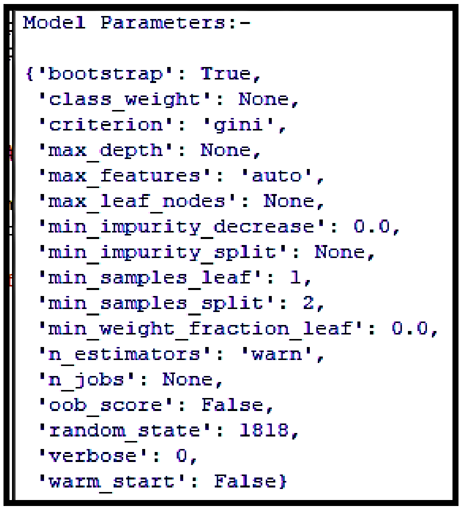
After the values are entered and set, the Submit button has a function attached to it, which when clicked, helps to collect the input data and process them into a valid reshaped array (done with the help of the Numpy library) to pass for the prediction.

This prediction is then passed to a Toplevel widget window which displays the result along with a small message which is either positive or negative with respect to the result. Also, if the values are entered incorrectly and needed to be changed, one can easily reset the entire form using the Reset button which is to the left of Submit.

The entire flow of the GUI is that it starts from the loading bar, input data by the user, passing the data, getting the data from the prediction and finally displaying the result in a separate window.

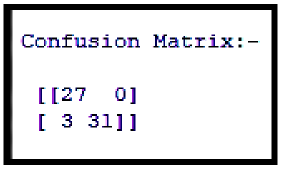
V. RESULTS

5.1 Model Parameters:



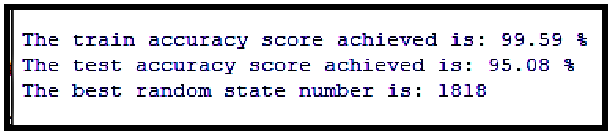
```
Model Parameters:-

{'bootstrap': True,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 'warn',
 'n_jobs': None,
 'oob_score': False,
 'random_state': 1818,
 'verbose': 0,
 'warm_start': False}
```

5.2 Confusion Matrix:



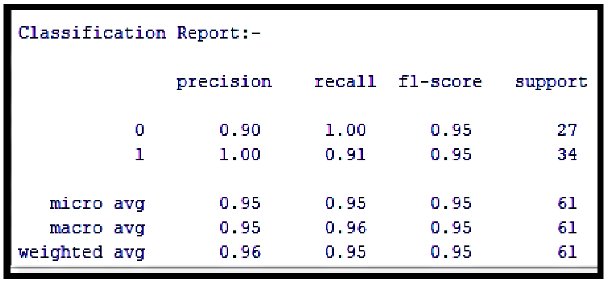
```
Confusion Matrix:-

 [[27  0]
 [ 3 31]]
```

5.3 Accuracy Scores:



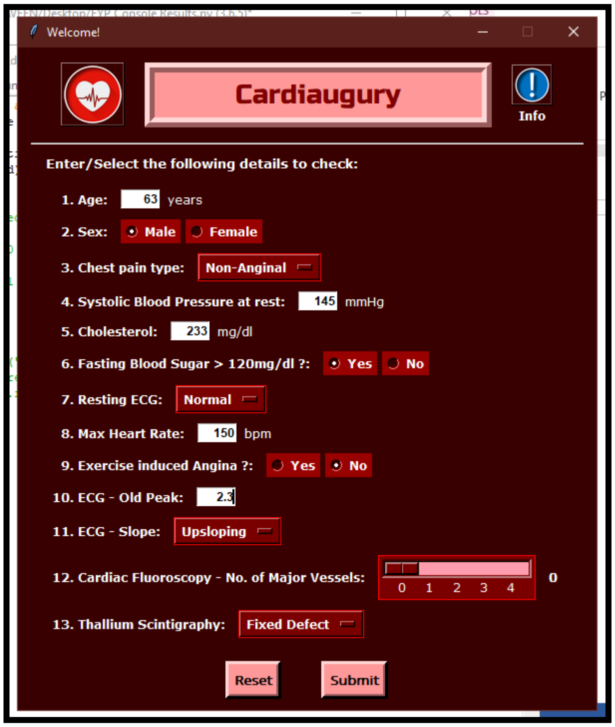
```
The train accuracy score achieved is: 99.59 %
The test accuracy score achieved is: 95.08 %
The best random state number is: 1818
```

5.4 Classification Report:



```
Classification Report:-

              precision    recall  f1-score   support

           0       0.90      1.00      0.95        27
           1       1.00      0.91      0.95        34

   micro avg       0.95      0.95      0.95        61
   macro avg       0.95      0.96      0.95        61
weighted avg       0.96      0.95      0.95        61
```

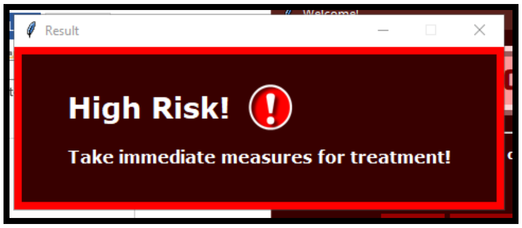
5.5 Data Entry in GUI:



Enter/Select the following details to check:

1. Age: 63 years
2. Sex: ● Male ○ Female
3. Chest pain type: Non-Anginal
4. Systolic Blood Pressure at rest: 145 mmHg
5. Cholesterol: 233 mg/dl
6. Fasting Blood Sugar > 120mg/dl ?: ● Yes ○ No
7. Resting ECG: Normal
8. Max Heart Rate: 150 bpm
9. Exercise induced Angina ?: ○ Yes ● No
10. ECG - Old Peak: 2.3
11. ECG - Slope: Upsloping
12. Cardiac Fluoroscopy - No. of Major Vessels: 0 1 2 3 4 0
13. Thallium Scintigraphy: Fixed Defect

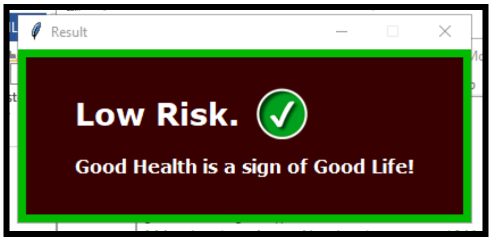
5.6 Result: High Risk:



High Risk!

Take immediate measures for treatment!

5.7 Result: Low Risk:



Low Risk. ✓

Good Health is a sign of Good Life!

VI. CONCLUSION

In this work, we proposed a prediction system for heart disease, which is based on the Random Forest classifier model of Machine Learning. Using Data Mining, the obtained data, which is in the format CSV, is searched for patterns in it, after cleaning ad preparing. With the prepared data, a simple Prediction model is built upon the Random Forest algorithm, and this model is then utilized for further testing and predictions. To achieve this system, a simple desktop application is to be built which has good GUI components for a satisfactory UX using Python as the base language. Also, the system will convey the predicted result back to the user through the GUI. We hope this system can reduce the overheads such as costs of manual predictions and enable any end-user to keep track of their health at the end of the day.

REFERENCES

[1] Harshitha M, Dr. B M Sagar, “Smart Health Care Implementation Using Naïve Bayes Algorithm”, International Journal of Innovative Research in Computer Science & Technology (IJIRCST), May 2019.

[2] Pradnya Suresh Joshi, Asst.Prof. Ashwini Gaikwad, “Smart Health Prediction System Using Data Mining”, Journal of Emerging Technologies and Innovative Research (JETIR), Dec. 2020.

[3] Manisha M S Pillai, Rahul Gopal, Roshitha Mariam Sunny, Revathy Chandran, Akhila Balachandran, “Smart Health Prediction System Using Python”, International Journal of Computer Sciences and Engineering (IJCSE), May 2019.

[4] N. Vijay Kumar, M. Udaya Prakash, “Smart Health Prediction using Data Mining with Effective Machine Learning”, International Journal of Scientific Engineering and Technology Research (IJSETR), Jan.-Dec. 2019.

[5] Konde T.R, Konde D.R, Khokrale P.V, Phulwade S.P, “Health Prediction System”, International Journal of Advance Engineering and Research Development (IJAERD), Feb. 2018.

[6] Pinky Saikia Dutta, Shrabani Medhi, Sunayana Dutta, Tridisha Das, Sweety Buragohain, “Smart Health Care Using Data Mining”, International Journal Of Current Engineering And Scientific Research (IJCESR), 2017.

[7] J. Manikandan, Mr. T. Muthusamy, Mrs. K. K. Kavitha, “Smart Health Predicting System Using Data Mining”, International Journal on Future Revolution in Computer Science & Communication Engineering (IJFRCSCE), Nov. 2018.

[8] N. Shabaz Ali, G. Divya, “Prediction of Diseases in Smart Health Care System using Machine Learning”, International Journal of Recent Technology and Engineering (IJRTE), Jan. 2020.

[9] Prof. Krishna Kumar Tripathi, Shubham Jawadwar, Siddhesh Murudkar, Prince Mishra, “A Smart Health Prediction Using Data Mining”, International Research Journal of Engineering and Technology (IRJET), Apr. 2018.

[10] G.Pooja reddy, M.Trinath basu, K.Vasanthi, K.Bala Sita Ramireddy, Ravi Kumar Tenali, “Smart E-Health Prediction System Using Data Mining”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Apr. 2018.