ISSN 2319 – 1953
International Journal of Scientific Research in Computer Science Applications and Management Studies

Constructive Heart Disease Prophecy with Hybrid Machine Learning Strategy

T R Mani Chigurupati[# 1], SK Mahaboob Basha[*2], E Sunil[@3]

[#1, *2, @3]*Assistant Professor, CSE Department, Holymary Institute of Technology and Science, Keesara, Medchal, Hyderabad, Telangana.*

[1]tulasiratnamani.chigurupati@gmail.com,[2]bashashaik1617@gmail.com,[3]e.sunil1820@gmail.com

***Abstract:* Coronary illness is one of the hugest reasons for mortality on the planet today. Expectation of cardiovascular infection is a basic test in the region of clinical information examination. AI (ML) has been demonstrated to be powerful in helping with settling on choices and forecasts from the enormous amount of information delivered by the human services industry. We have likewise observed ML strategies being utilized in ongoing improvements in various regions of the Internet of Things (IoT). Different investigations give just a brief look into anticipating coronary illness with ML methods. In this paper, we propose a novel strategy that targets finding noteworthy highlights by applying AI strategies bringing about improving the precision in the expectation of cardiovascular malady. The expectation model is presented with various blends of highlights and a few known arrangement procedures. We produce an upgraded exhibition level with a precision level of 75.7% through the forecast model for coronary illness with the mixture irregular woods with a straight model.**

I. Introduction

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. In this paper, we propose a novel method that aims finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 100% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naïve Byes (NB). The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation.

II. Literature Survey

A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier:

Coronary Heart Disease (CHD) is a common form of disease affecting the heart and an important cause for premature death. From the point of view of medical sciences, data mining is involved in discovering various sorts of metabolic syndromes. Classification techniques in data mining play a significant role in prediction and data exploration. Classification technique such as Decision Trees has been used in predicting the accuracy and events related to CHD. In this paper, a Data mining model has been developed using Random Forest classifier to improve the prediction accuracy and to investigate various events related to CHD. This model can help the medical practitioners for predicting CHD with its various events and how it might be related with different segments of the population. The events investigated are Angina, Acute Myocardial Infarction (AMI), Percutaneous Coronary Intervention (PCI), and Coronary Artery Bypass Graft surgery (CABG). Experimental results have shown that classification using Random Forest Classification algorithm can be successful used in predicting the events and risk factors related to CHD.

Using PSO Algorithm for Producing Best Rules in Diagnosis of Heart Disease:

Heart disease is still a growing global health issue. In the health care system, limiting human experience and expertise in manual diagnosis leads to inaccurate diagnosis, and the information about various illnesses is either inadequate or lacking in accuracy as they are collected from various types of medical equipment. Novel ensemble method for the prediction of response to fluvoxamine treatment of obsessive–compulsive disorder: About 30% of obsessive–compulsive disorder (OCD) patients exhibit an inadequate response to pharmacotherapy. The detection of clinical variables associated with treatment response may result in achievement of remission in shorter period, preventing illness development and reducing socioeconomic costs.

Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques:

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. In this paper, we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

III. Objective of the Project

3.1. Existing System:

This was on different Machine Learning Algorithms like Decision Trees and Random Forest. We predicted the accuracy by using above two algorithms in Existing system.

Disadvantages:

The accuracy is very low when we use Decision Trees. Comparatively with other algorithms Decision Trees produce less accurate scores in prediction in this kind of scenarios.

3.2. Proposed System:

In this study, we have used an R studio rattle to perform heart disease classification of the Cleveland UCI repository. It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. ML process starts from a preprocessing data phase followed by feature selection based on DT entropy, classification of modeling performance evaluation, and the results with improved accuracy. The feature selection and modeling keep on repeating for various combinations of attributes. Table 1 shows the UCI dataset detailed information with attributes used. Table 2 shows the data type and range of values. The performance of each model generated based on 13 features and ML techniques used for each interaction and performance is recorded. Section A summarizes the data pre-processing, Section B discusses the feature selection using entropy, Section C explains the classification with ML techniques and Section D presented for the performance of the results.

Advantages:

The results obtained by the Logistic Regression Algorithm are best compared to any other Algorithms. The Accuracy obtained was almost equal to cent percent which proves using of Logistic algorithm gives best results.

3.3. Modules Description Supervised Classification (Training Dataset):

The data has been divided into two parts i.e., training and testing data in the 70:30 ratios. Learning algorithms have been applied on the training data and based on the learning, predictions are made on the test data set.

Supervised Classification (Test Dataset): The test dataset is 30% of the total data. Supervised learning algorithms have been applied on the test data and the output obtained is compared with the actual output.

Pandas: pandas are an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

NumPy: NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array objec0t, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

IV. System Analysis

4.1 Feasibility Report

Preliminary investigation examine project feasibility, the likelihood the system will be useful to the organization. The main objective of the feasibility study is to test the Technical, Operational and Economical feasibility for adding new modules and debugging old running system. All system is feasible if they are unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation:
Technical Feasibility
Economical Feasibility
Operation Feasibility

4.1.1. Technical Feasibility:

In the feasibility study first step is that the organization or company has to decide that what technologies are suitable to develop by considering existing system. Here in this application used the technologies like Visual Studio 2008 and SQL Server 2005. These are free software that would be downloaded from web. Visual Studio 2008 –it is tool or technology.

4.1.2. Operational Feasibility:

Not only must an application make economic and technical sense, it must also make operational sense. Issues to consider when determining the operational feasibility of a project

4.1.3. Economic Feasibility:

It refers to the benefits or Outcomes we are deriving from the product as compared to the total cost we are spending for developing the product. If the benefits are more or less the same as the older system, then it is not feasible to develop the product. In the present system, the development of new product greatly enhances the accuracy of the system and cuts short the delay in the processing this application. The errors can be greatly reduced and at the same time providing a great level of security. Here we don't need any additional equipment except memory of required capacity. No need for spending money on client for maintenance because the database used is web enabled database

V. SYSTEM REQUIREMENT SPECIFICATIONS.

A Software Requirements Specification (SRS)–a requirements specification for a software system – is a complete description of the behavior of a system to be developed. It includes a set of use cases that describe all the interactions the users will have with the software. In addition to use cases, the SRS also contains non-functional requirements. Non-functional requirements are requirements which impose constraints on the design or implementation (such as performance engineering requirements, quality standards, or design constraints).

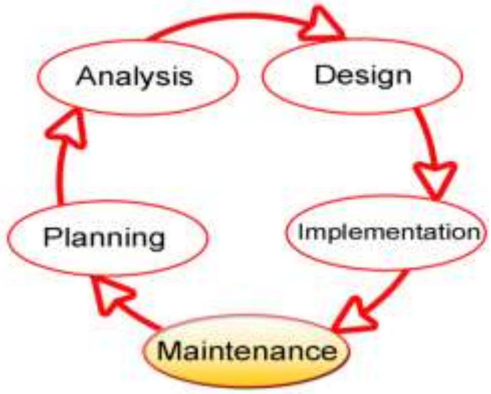
System requirements specification:

A structured collection of information that embodies the requirements of a system. a business analyst, sometimes titled system analyst, is responsible for analyzing the business needs of their clients and stakeholders to help identify business problems and propose solutions. Within the systems development life cycle domain, typically performs a liaison function between the business side of an enterprise and the information technology department or external service providers. Projects are subject to three sorts of requirements: Business requirements describe in business terms what must be delivered or accomplished to provide value. Product requirements describe properties of a system or product (which could be one of several ways to accomplish a set of business requirements.) Process requirements describe activities performed by the developing organization. For instance, process requirements could specify specific methodologies that must be followed, and constraints that the organization must obey. Product and process requirements are closely linked. Process requirements often specify the activities that will be performed to satisfy a product requirement. For example, a maximum development cost requirement (a process requirement) may be imposed to help achieve a maximum sales price requirement (a product requirement); a requirement that the product be maintainable (a Product requirement) often is addressed by imposing requirements to follow particular development styles

5.1. Non-Functional Requirements

Secure access of confidential data (user's details). SSL can be used.24 X 7 availability. Better component design to get better performance at peak time Flexible service based architecture will be highly desirable for future extension

5.2. SDLC Methodologies: SDLC MODEL:



The Software Development Lifecycle (SDLC) for small to medium database application development efforts.

This project uses iterative development lifecycle, where components of the application are developed through a series of tight iteration. The first iteration focus on very basic functionality, with subsequent iterations adding new functionality to the previous work and or correcting errors identified for the components in production.

VI. SYSTEM DESIGN

6.1 UML Diagrams

The Unified Modeling Language (UML) is used to specify, visualize, modify, construct and document the artifacts of an object-oriented software intensive system under development. UML offers a standard way to visualize a system's architectural blueprints, including elements such as:

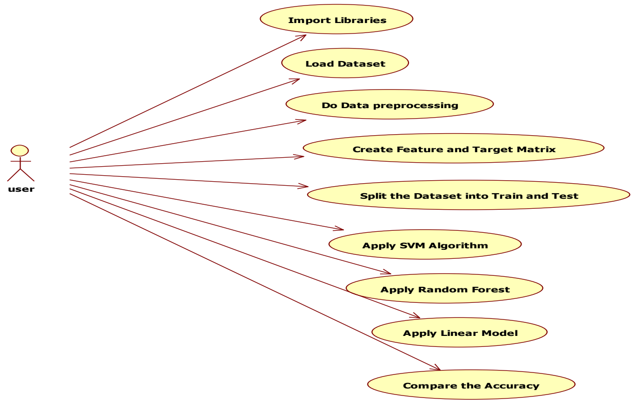
Actors
Business processes
(logical) components
Activities
programming language statements

Database schemas, and
Reusable software Components.

UML combines best techniques from data modeling (entity relationship diagrams), business modeling (work flows), object modeling, and component modeling. It can be used with all processes, throughout the software development life cycle, and across different implementation technologies. UML has synthesized the notations of the Booch method, the Object-modeling technique (OMT) and Object-oriented software engineering (OOSE) by fusing them into a single, common and widely usable modeling language. UML aims to be a standard modeling language which can model concurrent and distributed systems.

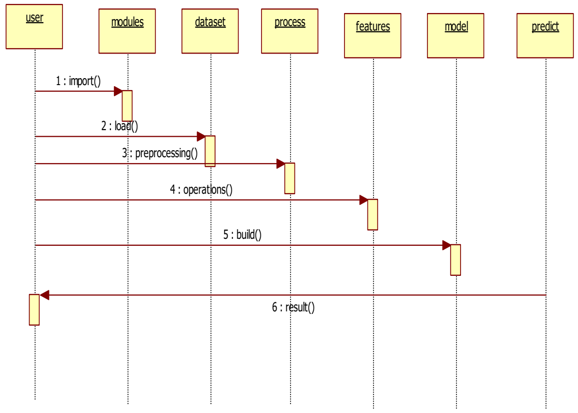
6.1.1 Use Case Diagram:

Over View Use Case Diagram: Server:



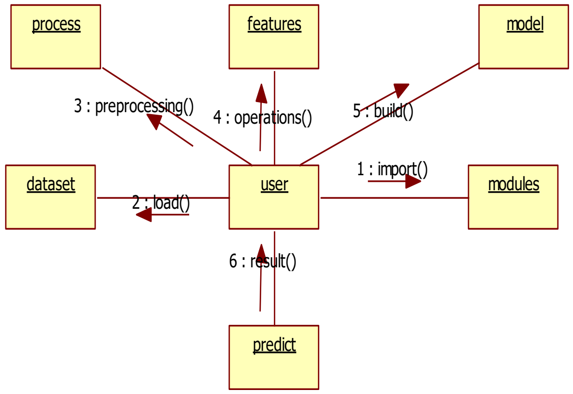
6.1.2 Sequence diagram:

Sequence Diagrams Represent the objects Participating the interaction horizontally and time vertically. A Use Case is a kind of behavioral classifier that represents a declaration of an offered behavior.



Each use case specifies some behavior, possibly including variants that the subject can perform in collaboration with one or more actors. Use cases define the offered behavior of the subject without reference to its internal structure. These behaviors, involving interactions between the actor and the subject, may result in changes to the state of the subject and communications with its environment. A use case can include possible variations of its basic behavior, including exceptional behavior and error handling.

6.1.3 Collaboration Diagram:



VII. Input and Output Design

7.1 Input and Output

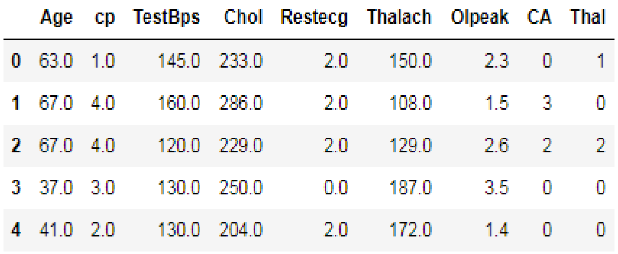
The following some are the projects inputs and outputs.

Inputs:

Importing the all required packages like NumPy, pandas, matplotlib, scikit – learn and required machine learning algorithms packages. Setting the dimension of visualization graph. Downloading and importing the dataset and convert to data frame.

Outputs:

Preprocessing the importing data frame for imputing nulls with the related information.



	Age	cp	TestBps	Chol	Restecg	Thalach	Olpeak	CA	Thal
0	63.0	1.0	145.0	233.0	2.0	150.0	2.3	0	1
1	67.0	4.0	160.0	286.0	2.0	108.0	1.5	3	0
2	67.0	4.0	120.0	229.0	2.0	129.0	2.6	2	2
3	37.0	3.0	130.0	250.0	0.0	187.0	3.5	0	0
4	41.0	2.0	130.0	204.0	2.0	172.0	1.4	0	0

All are displaying cleaned outputs. After applying machine learning algorithms it will give good results and visualization plots

VIII. Conclusion

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world data sets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.

References

[1] A. S. Abdullah and R. R. Rajalaxmi, ``A data mining model for predictingthe coronary heart disease using random forest classi_er," in Proc. Int.Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012,pp. 22_25.

[2] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar,``Using PSO algorithm for producing best rules in diagnosis of heartdisease," in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306_311.

[3] N. Al-milli, ``Backpropogation neural network for prediction of heartdisease," J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131_135, 2013.

[4] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya,and R. Deepika, ``Analysis of neural networks based heart disease predictionsystem," in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk,Poland, Jul. 2018, pp. 233_239.

[5] P. K. Anooj, ``Clinical decision support system: Risk level prediction ofheart disease using weighted fuzzy rules," J. King Saud Univ.-Comput. Inf.Sci., vol. 24, no. 1, pp. 27_40, Jan. 2012. doi: 10.1016/j.jksuci.2011.09.002.

[6] L. Baccour, ``Amended fused TOPSIS-VIKOR for classication(ATOVIC) applied to pp. 1011_1014.

[7] C.-A. Cheng and H.-W. Chiu, ``An articial neural network model forthe evaluation of carotid artery stenting prognosis using a national-widedatabase," in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.(EMBC), Jul. 2017, pp. 2566_2569.

[8] H. A. Esfahani and M. Ghazanfari, ``Cardiovascular disease detectionusing a new ensemble classi_er," in Proc. IEEE 4th Int. Conf. Knowl.-Based Eng. Innov. (KBEI), Dec. 2017,

[9] F. Dammak, L. Baccour, and A. M. Alimi, ``The impact of criterion weightstechniques in TOPSIS method of multi-criteria decision making in crispand intuitionistic fuzzy domains," in Proc. IEEE Int. Conf. Fuzzy Syst.(FUZZ-IEEE), vol. 9, Aug. 2015, pp. 1_8.

[10] R. Das, I. Turkoglu, and A. Sengur, ``Effective diagnosis of heart diseasethrough neural networks ensembles," Expert Syst. Appl., vol. 36, no. 4,pp. 7675_7680, Ma 2009.doi:10.10.16/j.eswa.2008.09.013.