Predicting Heart Disease at Early Stages Using Machine Learning

Miss. Apeksha Barhate[1]*, Mr. Ravindra Pandit*[2]*, Dr. Umesh Pawar* [3]*, Mr. Ramesh Daund* [4]

Department Of Computer Engineering

Snd College of Engineering and Research Centre, Yeola, Nashik.

Abstract – Predicting and area of coronary sickness has persistently been an essential and testing task for clinical consideration trained professionals. Clinical facilities and various focuses are offering exorbitant medicines and exercises to treat heart contaminations. Subsequently, expecting coronary disease toward the starting stages will be useful to people all around the planet so they will take major actions already quitting any funny business. Coronary disease is a basic issue in later times; the major legitimization behind this sickness is the confirmation of alcohol, tobacco, and nonappearance of real action. All through the long haul, AI shows reasonable results in basically choosing and assumptions from the wide plan of data made by the prosperity care industry. A part of the controlled AI techniques used in this assumption for coronary disease are fake decision tree (DT), support vector machine (SVM), Naive Bayes (NB) and k closest neighbor estimation. Moreover, the shows of these estimations are summarized..

Keywords: Machine learning, supervised learning, health care services, heart disease.

INTRODUCTION

The heart is one of the key bits of the human body later the psyche. The fundamental limit of the heart is to siphoning blood to the whole body parts. Any issue that can provoke disturbing the handiness of the heart is called heart sickness. A couple of sorts of coronary disease are there in the world; Coronary Artery Disease (CAD), and cardiovascular breakdown (HF) are the most broadly perceived heart sicknesses that are accessible. The essential clarification for the coronary sickness is blockage then again lessening of the coronary veins [1] Coronary veins in like manner responsible for giving blood to the heart. PC supported plan is the primary wellspring of death in excess of 26 million people are encountering all over the planet, and it is growing 2 each consistently as a result of CAD 17.5 million passings happened overall in 2005[2]. In the creating scene, 2 for each of the general population all around the planet is moping from CAD, and 10 for every people are more settled than 65 years. Around 2 for every of the yearly clinical benefits spending plan spent just to treat CAD infection. USA government consumed 35 billion bucks for CAD in 2018 Different elements can raise the gamble of cardiovascular breakdown. Clinical researchers have arranged those variables into two distinct classes; one of them is risk factors that can't be changed, and another is risk factors that can be changed. Family ancestry, sex, age goes under risk factors that can't be changed. Elevated cholesterol, smoking, actual dormancy, high blood joy every one of these go under risk factors.

Coronary illness is a huge issue, so there is a requirement for finding or expectation of coronary illness there are a few strategies to analyze coronary illness among them Angiography is the moving strategy which is utilized by a large portion of the doctors across the world. Notwithstanding, there are a few disadvantages related with angiography procedure. It is a costly technique and doctors need to examine such countless elements to analyze a patient thus this cycle makes doctor work undeniably challenging, so these restrictions persuade to foster a harmless strategy for expectation of coronary illness.

SYSTEM ARCHITECTURE

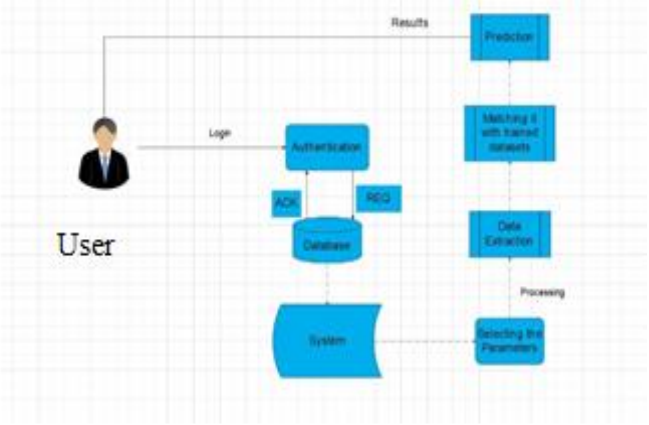


Figure: System Architecture

Clinical determination is considered as the need might arise to be done definitively and effectively. The computerization of the equivalent would be exceptionally gainful. Clinical choices are in many cases made in light of specialist's instinct and experience as opposed to on the information rich information concealed in the data set. This training prompts undesirable predispositions, mistakes and unreasonable clinical costs which influences the nature of administration gave to patients. Information mining can possibly produce an information rich climate which can serve to work on the nature of clinical choices essentially.

Irregular Forest is a famous classifier which is straightforward and simple to execute. It requires no space information or boundary setting and can deal with high layered information. The outcomes got from Random Forests are more straightforward to peruse and decipher. The drill through component to get to definite patients‟ profiles is just accessible in Random Forests.

Credulous Bayes is a measurable classifier which expects no reliance between credits. It endeavors to amplify the back likelihood in deciding the class. The benefit of utilizing gullible bayes is that one can work with the guileless Bayes model without utilizing any Bayesian strategies. Gullible Bayes classifiers have functions admirably in numerous complicated genuine circumstances.

LITERTURE SURVEY

Authors: A.M. Kavitha; G. Gnaneswar; R. Dinesh; Y. Rohith Sai; R. Sai Suraj Findings: Heart disease causes a significant mortality rate around the world, and it has become a health threat for many people. Early prediction of heart disease may save many lives. Detecting cardiovascular diseases like heart attacks, coronary artery diseases, etc. is a critical challenge for regular clinical data analysis. Machine learning (ML) can provide an effective solution for decision-making and accurate predictions. The medical industry is showing enormous development in using machine learning techniques. In the proposed work, a novel machine learning approach is proposed to predict heart disease. The proposed study used the Cleveland heart disease dataset, and data mining techniques such as regression and classification were used. Random Forest and Decision Tree machine learning techniques are used.The novel technique of the machine-learning model is designed. In implementation, 3 machine learning algorithms are used: 1. random forest, 2. decision tree, and 3. hybrid model (a hybrid of random forest and decision tree). Experimental results show an accuracy level of 88.7% through the heart disease prediction model with the hybrid model. The interface is designed to get the user’s input parameter to predict the heart disease, for which we used a hybrid model of decision trees and random forests. [1]

Authors: Akanksha Kumari; Ashok Kumar Mehta Findings: Heart disease is the leading cause of death and hospitalisation in the world. With the advancement of technology and the contribution of computer engineering, it is easy to detect heart disease, and thus treatment is fast and effective. Machine learning is becoming increasingly popular in the medical field for predicting disease.The authors attempted to predict heart disease using seven machine learning algorithms and to improve the accuracy of weakly performing algorithms using ensemble methods such as AdaBoost and the voting ensemble method in this paper.The performance of linear discriminate analysis is good among other algorithms; its mean value is approximately 0.847 and its mean absolute error is 0.185; the false acceptance rate is lowest among all i.e.; 0.33 and the false recognition rate is 0.076, accuracy is somehow coming 80% which is less if compared with Logistic Regression.[2].

Authors: A. Lakshmanarao; A. Srisaila; T.Srinivasa Ravi Kiran Findings: Cardiovascular diseases (heart-related diseases) are the reason for the deaths of 18 million people every year in the world. According to WHO,31% of the deaths worldwide are due to heart-related diseases. In this paper, we proposed a novel machine learning model for heart disease prediction. The proposed method was tested on two different datasets from Kaggle and UCI. We applied sampling techniques to the unbalanced dataset and feature selection techniques are used to find the best features. Later several classifier models were applied and achieved good accuracy with ensemble classifier. The experimentations on two datasets shown that the proposed model is effective for heart disease prediction. Python was used for all implementations.[3].

Authors: Sakshi Bhoyar, Nikki Wagholikar, and Kshitij Bakshi Findings: Stroke, heart failure, arrhythmia, and myocardial infarction are the most common cardiovascular diseases that record high mortality rates around the world. Because the available tests are prohibitively expensive, heart defects are not detected in their early stages.Thus, a fast, real-time, and reliable system that predicts the chances of a patient having heart disease in an optimised manner is required. In this research, a neural network model using a multilayer perceptron (MLP) is proposed for the prediction system. Experimental analysis resulted in an accuracy of 85.71% for the UCI Heart Disease dataset and 87.30% for the Cardiovascular Disease dataset. When compared to previous research, the increase in accuracy was approximately 12–13. A simple web application tool is also developed using Python programming to test the prediction system. This study aims to create a user-friendly tool for both medical professionals and the general public.[4]

OBJECTIVES

- Early detection: The primary objective is to develop or implement a system that can detect heart diseases at an early stage, ideally before the onset of severe symptoms or complications. Early detection enables timely intervention and treatment, leading to better outcomes and improved quality of life for individuals at risk
- Accuracy and reliability: The project aims to develop or utilize diagnostic tools, algorithms, or models that provide accurate and reliable results in detecting heart diseases. This involves minimizing false positives and false negatives to ensure that the detection system has a high level of precision and sensitivity.
- Accessibility and affordability: Making heart disease detection accessible and affordable is another objective. This could involve developing cost-effective diagnostic technologies.
- Public health impact: Ultimately, the objective of a heart disease detection project is to have a positive impact on public health. By improving detection rates, implementing preventive measures, and reducing the burden of heart disease, the project aims to promote cardiovascular health and decrease morbidity and mortality associated with heart diseases at a population level.Public health impact: Ultimately, the objective of a heart disease detection project is to have a positive impact on public health. By improving detection rates, implementing preventive measures, and reducing the burden of heart disease, the project aims to promote cardiovascular health and decrease morbidity and mortality associated with heart diseases at a population level.

CLASSIFICATION USING RANDOM FOREST ALGORITHM

Irregular Forest forms characterization or relapse models as a tree structure. It breaks down a dataset into increasingly small subsets while simultaneously a related Random Forest is steadily evolved. The eventual outcome is a tree with choice hubs and leaf hubs. A choice hub (e.g., Outlook) has at least two branches (e.g., Sunny, Overcast and Rainy). Leaf hub (e.g., Play) addresses a grouping or choice. The highest choice hub in a tree which compares to the best indicator called root hub. Arbitrary Forests can deal with both absolute and mathematical information.

A Random Forest is fabricated hierarchical from a root hub and includes parceling the information into subsets that contain occurrences with comparative qualities (homogenous). ID3 calculation utilizes entropy to compute the homogeneity of an example. On the off chance that the example is totally homogeneous the entropy is zero and assuming the example is a similarly separated it has entropy of one. To construct a Random Forest, we want to compute two kinds of entropy utilizing recurrence tables as follows:

a) Entropy using the frequency table of one attribute:

$$E(S)=\sum_{i=1}^{c}-p_i\log_2 p_i$$

Entropy using the frequency table of two attributes:

$$E(T,X)=\sum_{c\in X}P(c)E(c)$$

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a Random Forest is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

Step 1: Calculate entropy of the target.

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.

$$Gain(T,X)=Entropy(T)-Entropy(T,X)$$

Step 3: Choose attribute with the largest information gain as the decision node.

Step 4(a): A branch with entropy of 0 is a leaf node.

Step 4(b): A branch with entropy more than 0 needs further splitting.

Step 5: The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

A Random Forest can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one.

CLASSIFICATION USING AdaBoost CLASSIFIER

AdaBoost makes it conceivable to consolidate different "feeble classifiers" into a solitary classifier which is "areas of strength for designated." Decision trees with one level, or choice trees with just a single split, are the most well known calculation utilized with AdaBoost. Choice Stump is one more name for these trees. and is approach makes a model by relegating equivalent loads to every one of the data of interest. It then gives focuses that are mistakenly sorted with a higher weight. In the following model, all focuses with more prominent loads are given more significance. It will keep on preparing models till a lower mistake is gotten.

&e weight of the preparation set is utilized to begin the AdaBoost calculation. Allow us to think about preparing set (x1, y), . . . (xn, yn), in which every xi is in occurrence space X and each name yi is in assortment of marks Y, that is particularly like the assortment of $\{-1,+1\}$. Weight on preparing occasion I on the round t is doled out as DIt(i). Toward the beginning, a similar weight is utilized (DIt(i)) = 1/M, I = 1, . . ., M), where It is the emphasis number. then, weight of the misclassified case from the base learning calculation is then expanded in each round. The AdaBoost calculation's pseudocode is displayed beneath. What's more,

$$\alpha_{It}=\frac{1}{2}\mathrm{In}\left[\frac{P_{+1}-P_{-1}}{P_{-1}+P_{-1}}\right].$$

C_{It} is the normalization constant, αIt is used to allow the outcome to be generalized and to solve the problem of overfitting and noise sensitive situations. The real value of $\alpha_{It}h_{It}$ (x) is built using a class probability estimate (P).

CLASSIFICATION USING NAIVE BAYES CLASSIFIER

Bayes hypothesis gives an approach to computing the back likelihood, P (c|x), from P(c), P(x), and P (x|c). Gullible Bayes classifier expects that the impact of the worth of an indicator (x) on a given class (c) is free of the upsides of different indicators. This supposition that is called class contingent autonomy.

Likelihood Class Prior Probability

$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$$

Posterior Probability Predictor Prior Probability

$$P(c\mid X)=P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

- P (c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P (x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Thus, we can write:

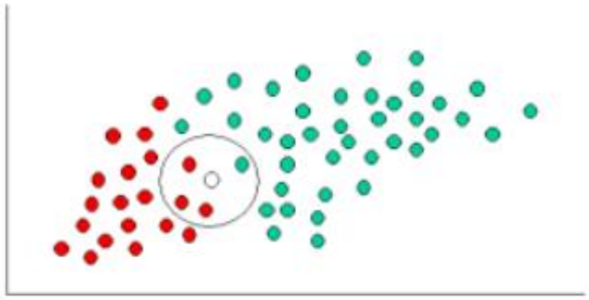
$$\text{Prior probability for GREEN} \propto \frac{\text{Number of GREEN objects}}{\text{Total number of objects}}$$

$$\text{Prior probability for RED} \propto \frac{\text{Number of RED objects}}{\text{Total number of objects}}$$

Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior probabilities for class membership are:

$$\text{Prior probability for GREEN} \propto \frac{40}{60}$$

$$\text{Prior probability for RED} \propto \frac{20}{60}$$



Having formulated our prior probability, we are now ready to classify a new object (WHITE circle). Since the objects are well clustered, it is reasonable to assume that the more GREEN (or RED) objects in the vicinity of X, the more likely that the new cases belong to that particular color. To measure this likelihood, we draw a circle around X which encompasses a number (to be chosen a priori) of points irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label. From this we calculate the likelihood:

$$\text{Likelihood of X given GREEN} \propto \frac{\text{Number of GREEN in the vicinity of X}}{\text{Total number of GREEN cases}}$$

$$\text{Likelihood of X given RED} \propto \frac{\text{Number of RED in the vicinity of X}}{\text{Total number of RED cases}}$$

Albeit the earlier probabilities show that X might have a place with GREEN (considering that there are two times as many GREEN contrasted with RED) the probability demonstrates in any case; that the class enrollment of X is RED (considering that there are more RED items nearby X than GREEN). In the Bayesian examination, the last grouping is created by consolidating the two wellsprings of data, i.e., the earlier and the probability, to shape a back likelihood utilizing the supposed Bayes' standard (named after Rev. Thomas Bayes 1702-1761).

$$\text{Posterior probability of X being GREEN} \propto$$
$$\text{Prior probability of GREEN} \times \text{Likelihood of X given GREEN}$$
$$=\frac{4}{6}\times\frac{1}{40}=\frac{1}{60}$$

$$\text{Posterior probability of X being RED} \propto$$
$$\text{Prior probability of RED} \times \text{Likelihood of X given RED}$$
$$=\frac{2}{6}\times\frac{3}{20}=\frac{1}{20}$$

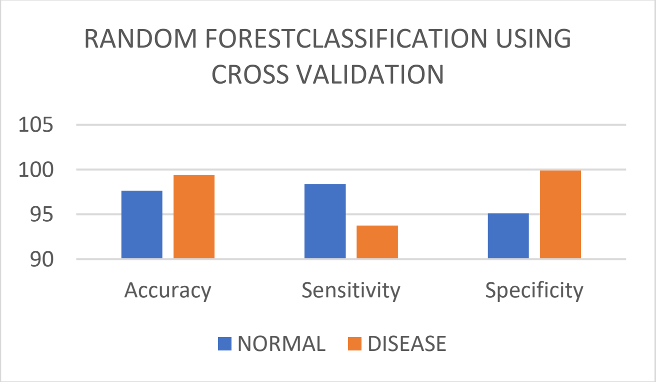
Finally, we classify X as RED since its class membership achieves the largest posterior probability. Naive Bayes can be modelled in several different ways including normal, lognormal, gamma and Poisson density functions:

$$p(x_k \mid C_j) = \begin{cases} \dfrac{1}{\sigma_{kj}\sqrt{2\pi}} \exp\left(\dfrac{-(x-\mu_{kj})^2}{2\sigma_{kj}}\right), \quad -\infty < x < \infty, -\infty < \mu_{kj} <, \sigma_{kj} > 0 & \text{Normal} \\ \mu_{kj} : \text{mean}, \sigma_{kj} : \text{standard devation} & \\ \dfrac{1}{x\sigma_{kj}(2\pi)^{1/2}} \exp\left\{\dfrac{-[\log(x/m_{kj})]^2}{2\sigma_{kj}^2}\right\}, \quad 0 < x < \infty, m_{kj} > 0, \sigma_{kj} > 0 & \text{Lognormal} \\ m_{kj} : \text{scale parameter}, \sigma_{kj} : \text{shape parameter} & \\ \dfrac{\left(\dfrac{x}{b_{kj}}\right)^{c_{kj}-1}}{b_{kj}\Gamma(c_{kj})} \exp\left(\dfrac{-x}{b_{kj}}\right), \quad 0 \le x < \infty, b_{kj} > 0, c_{kj} > 0 & \text{Gamma} \\ b_{kj} : \text{scale parameter}, c_{kj} : \text{shape parameter} & \\ \dfrac{\lambda_{kj} \exp(-\lambda_{kj})}{x!}, \quad 0 \le x < \infty, \lambda_{kj} > 0, x = 0,1,2,\ldots & \text{Poisson} \\ \lambda_{kj} : \text{mean} & \end{cases}$$

SUMMARY OF THE CLASSIFICATION ACCURACY

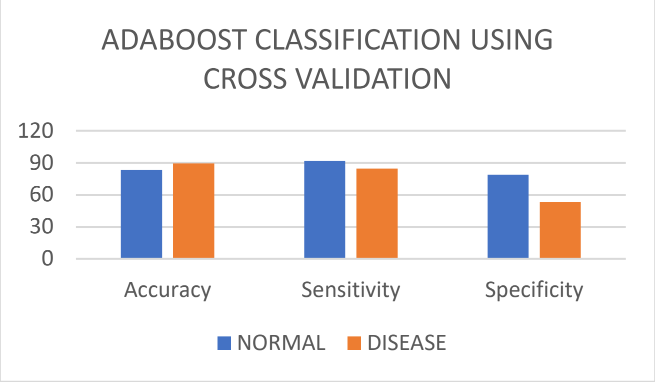
- **Random Forest Algorithm**

	NORMAL	DISEASE
Accuracy	97.6482	99.3885
Sensitivity	98.3686	93.7500
Specificity	95.1168	99.8974



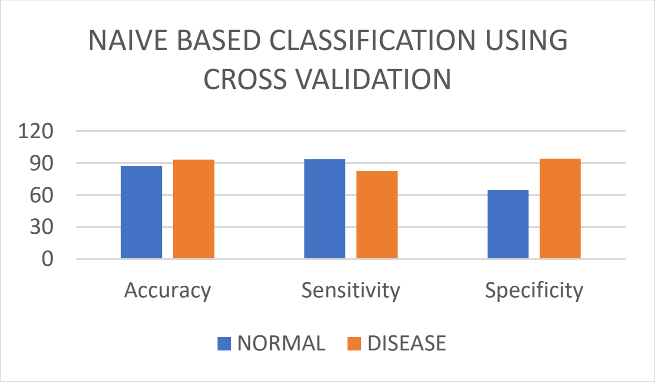
- **AdaBoost Algorithm**

	NORMAL	DISEASE
Accuracy	83.2501	89.3208
Sensitivity	91.76	84.3594
Specificity	78.7448	53.3807

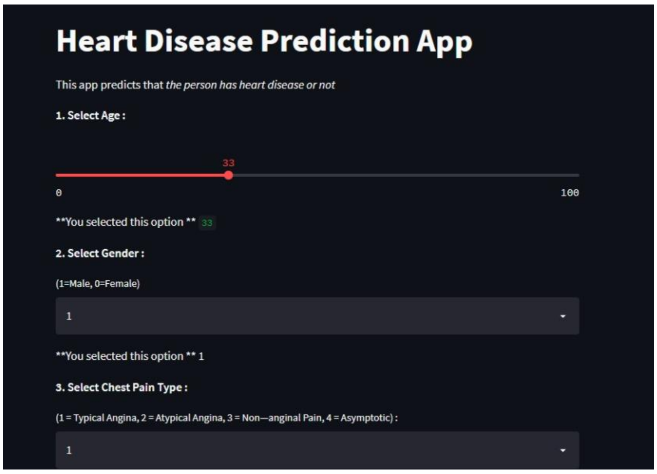


- **Naïve Based Algorithm**

	NORMAL	DISEASE
Accuracy	87.3001	93.3208
Sensitivity	93.7160	82.3864
Specificity	64.7558	94.3077



RESULTS



Heart Disease Prediction App

This app predicts that *the person has heart disease or not*

1. Select Age :

33

0 100

**You selected this option ** 33

2. Select Gender :

(1=Male, 0=Female)

1

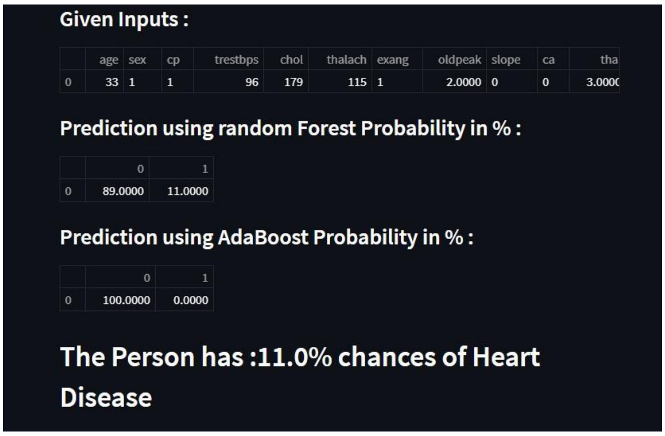
**You selected this option ** 1

3. Select Chest Pain Type :

(1 = Typical Angina, 2 = Atypical Angina, 3 = Non—anginal Pain, 4 = Asymptotic) :

1

Figure: Fill Details Page



Given Inputs :

	age	sex	cp	trestbps	chol	thalach	exang	oldpeak	slope	ca	tha
0	33	1	1	96	179	115	1	2.0000	0	0	3.000(

Prediction using random Forest Probability in % :

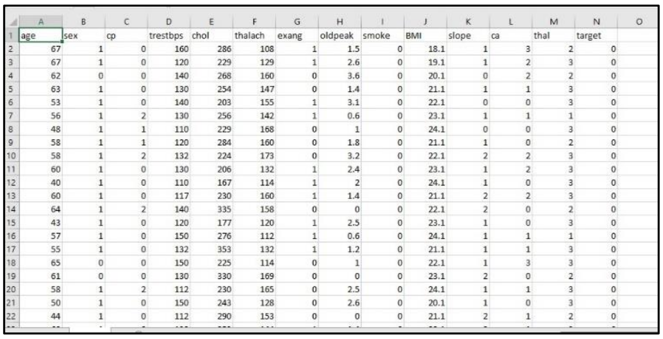
	0	1
0	89.0000	11.0000

Prediction using AdaBoost Probability in % :

	0	1
0	100.0000	0.0000

The Person has :11.0% chances of Heart Disease

Figure: Display Prediction



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	age	sex	cp	trestbps	chol	thalach	exang	oldpeak	smoke	BMI	slope	ca	thal	target	
2	67	1	0	160	286	108	1	1.5	0	18.1	1	3	2	0	
3	67	1	0	120	229	129	1	2.6	0	19.1	1	2	3	0	
4	62	0	0	140	268	160	0	3.6	0	20.1	0	2	2	0	
5	63	1	0	130	254	147	0	1.4	0	21.1	1	1	3	0	
6	53	1	0	140	203	155	1	3.1	0	22.1	0	0	3	0	
7	56	1	2	130	256	142	1	0.6	0	23.1	1	1	1	0	
8	48	1	1	110	229	168	0	1	0	24.1	0	0	3	0	
9	58	1	1	120	284	160	0	1.8	0	21.1	1	0	2	0	
10	58	1	2	132	224	173	0	3.2	0	22.1	2	2	3	0	
11	60	1	0	130	206	132	1	2.4	0	23.1	1	2	3	0	
12	40	1	0	110	167	114	1	2	0	24.1	1	0	3	0	
13	60	1	0	117	230	160	1	1.4	0	21.1	2	2	3	0	
14	64	1	2	140	335	158	0	0	0	22.1	2	0	2	0	
15	43	1	0	120	177	120	1	2.5	0	23.1	1	0	3	0	
16	57	1	0	150	276	112	1	0.6	0	24.1	1	1	1	0	
17	55	1	0	132	353	132	1	1.2	0	21.1	1	1	3	0	
18	65	0	0	150	225	114	0	1	0	22.1	1	3	3	0	
19	61	0	0	130	330	169	0	0	0	23.1	2	0	2	0	
20	58	1	2	112	230	165	0	2.5	0	24.1	1	1	3	0	
21	50	1	0	150	243	128	0	2.6	0	20.1	1	0	3	0	
22	44	1	0	112	290	153	0	0	0	21.1	2	1	2	0	

Figure: CSV File for Dataset Training

CONCLUSION

The general goal of our work is to foresee all the more precisely the presence of coronary illness. In this paper, three information mining order methods were applied to be specific Random Forest, AdaBoost and Naive Bayes. From results, it has been seen that Decision trees gives exact outcomes as contrast with Naive Bayes. This framework can be additionally extended. It can utilize a more prominent number of information sources. Different information mining strategies can likewise be utilized for predication e.g., Clustering, Time series, Association rules. The text mining can be utilized to mine immense measure of unstructured information accessible in medical care industry data set.

REFERENCES

[1] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," IEEE Access, vol. 7, pp. 54007–54014, 2019, doi: 10.1109/ACCESS.2019.2909969.

[2] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," IEEE Access, vol. 7, pp. 180235– 180243, 2019, doi: 10.1109/ACCESS.2019.2952107.

[3] M. Gjoreski, A. Gradisek, B. Budna, M. Gams, and G. Poglajen, "Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure from Heart Sounds," IEEE Access, vol. 8, pp. 20313–20324, 2020, doi: 10.1109/ACCESS.2020.2968900.

[4] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on χ2 Statistical Model and Optimally Configured Deep Neural Network," IEEE Access, vol. 7, pp. 34938–34945, 2019, doi: 10.1109/ACCESS.2019.2904800.

[5] M. R. Ahmed, S. M. Hasan Mahmud, M. A. Hossin, H. Jahan, and S. R. Haider Noori, "A cloud based four-tier architecture for early detection of heart disease with machine learning algorithms," 2018 IEEE 4th Int. Conf. Comput. Commun. ICCC 2018, pp. 1951–1955, 2018, doi: 10.1109/CompComm.2018.8781022.

[6] "types of heart disease." [Online]. Available: https://www.heartandstroke.ca/heart/what-is-heart-disease/typesof-heart-disease.

[7] J. Schmidhuber, "Deep Learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, 2015, doi: 10.1016/j.neunet.2014.09.003.

[8] N. H. Farhat, "Photonit neural networks and learning mathines the role of electron-trapping materials," IEEE Expert. Syst. their Appl., vol. 7, no. 5, pp. 63–72, 1992, doi: 10.1109/64.163674.

[9] A. K. M Sazzadur Rahman, M. Mehedi Hasan, S. Asaduzzaman, M. Asaduzzaman, and S. Akhter Hossain, "An analysis of computational intelligence

techniques for diabetes prediction Machine Learning View project An analysis of computational intelligence techniques for diabetes prediction," Int. J. Eng. &Technology, vol. 7, no. 4, pp. 6229–6232, 2018, doi: 10.14419/ijet.v7i4.28245.

[10] G. H. Tang, A. B. M. Rabie, and U. Hägg, "Indian hedgehog: A mechanotransduction mediator in condylar cartilage," J. Dent. Res., vol. 83, no. 5, pp. 434–438, 2004, doi: 10.1177/154405910408300516.