Heart Disease Prediction Using Machine Learning Algorithm

Mihir Raj Singh
School of Computer Science and Engineering,
Galgotias University
Grater Noida, India
mihirsingh270@gmail.com

Aniket Sharma
School of Computer Science and Engineering,
Galgotias University
Grater Noida, India
aniket18121999@gmail.com

Deepanshu Singh
School of Computer Science and Engineering,
Galgotias University
Grater Noida, India
arsuhragsingh27@gmail.com

Abstract—Heart disease, additionally called cardiovascular disease, includes any range of conditions involving the human heart or blood vessels. It's the leading reason for death round the world: artery sickness, stroke, hypotensions, arrhythmia, and angina area unit every type of heart condition. Scientists estimate that ninety percent of all heart condition is preventable, for the most part through fashion modifications together with a healthy diet, regular exercise, the turning away of tobacco, and limiting alcohol intake. the danger factors of heart condition for developing heart condition area unit various and well-know. Major risk factors for heart condition embody the following: Age, Gender, Tobacco usage, One’s level of physical inactivity, Obesity, cardiovascular disease, Diabetes, High cholesterol. A case history of heart condition. variant folks worldwide struggle to regulate the danger factors that cause upset, several others stay unaware that they’re at high risk, an oversized range of heart attacks and strokes is prevented by dominant major risk factors through fashion interventions and drug treatment wherever necessary.

Keywords—Machine Learning, Data Set, Classifier, Data Processing

I. INTRODUCTION

A many individuals in this nation are experiencing cardiovascular or heart disease. In emanant countries, the norm of medical care benefits actually should be improved. Like in Indonesia, they don't have a lot of clinical professionals where for each 1000 inhabitants just 0.36 specialists are available. Subsequently, many exploration and advancements in medical care administration enhancements are flourishing, for instance: the use of machine-to-machine (M2M) innovation in tolerant observing. With its concentrated turns of events, M2M innovation are becoming greatly used in different fields, including medical care. While many investigators are cleared out clinical discipline related with CVD, Machine learning procedures are used in medical care determination moreover. AI might be a course of extricating or investigating enormous size of information to figure it out. AI is that the method involved with tracking down beforehand obscure examples and patterns in data sets and utilizing that data to make perceptive models the planet Health Statistics 2012 report edifies the very reality that one out of three grown-ups worldwide has raised crucial sign – a condition that causes around a large portion of all passings from stroke and heart condition. Heart condition, additionally alluded to as confusion (CVD), encases assortment of conditions that impact the guts – not simply heart assaults. Counting India heart condition was the primary clarification for losses inside the various nations. Heart condition kills one individual in at regular intervals inside us. Coronary heart condition, Cardiomyopathy and confusion are a few classes of heart diseases. The expression "cardiovascular disease" incorporates a decent scope of conditions that influence the guts and subsequently the change the veins and the way during which blood is siphoned and coursed through the body. Determination is confounded and significant assignment that should be executed precisely and effectively. The determination is normally made, upheld specialist's insight and information. This outcomes in undesirable outcomes and unreasonable clinical expenses of medicines gave to patients. Consequently, a programmed determination framework would be really helpful. We have attempted to introduce the itemized learn about the two-AI calculation. Conclusion of HD is generally finished by the investigation of the clinical record of the patient, actual assessment report and examination of concerned side effects by a doctor. Additionally, it is costly and computationally hard to examine. In this manner, to foster a painless finding framework upheld classifiers of AI to determine these issues. AI is currently an arising field because of the expanding measure of information. AI makes it conceivable to aggregate information from a colossal measure of information, which is incredibly weighty for man and in some cases outlandish. For great outcomes appropriate an AI model is vital. In view of an AI procedure called arrangement, 70% of the information is regulated or prepared and 30% is tried as a piece of this text. Clearly, a good AI model might be a model that not exclusively performs well on data seen all through preparing (else an AI model may just get familiar with the preparation information), but also on inconspicuous data. To guage all classifiers on data and notice that they get, on the normal, half of the cases right. Also, adequate cross approval strategies and execution examination measurements region unit indispensable essential for a model once model is train and mind dataset. Arrangement is accomplished by trademark the closest neighbor to sort out the class of an example. K-closest neighbor is one among the foremost widely used classification problem [M+03]. It is the most straightforward classifier.

II. LITERATURE REVIEW

These years, Machine Learning has become in style in several fields of business, because of its purpose to
convert massive to become valuable data samples of machine learning use square measure mentioned below:

- A mercantile establishment arrange their merchandise by looking its client shopping for behavior and product association data.
- Analysis of internet browsing pattern to optimize a web site style.
- Analysis for monetary fraud detection.
- Analysis for a particular sickness designation.

There square measure several in style machine learning algorithms, particularly classification techniques, that every of them has superiority and weakness moreover, 3 of them are: Random Forest, call Tree, and k-Nearest Neighbor (KNN). Random forest may be a versatile, straightforward to use machine learning rule that produces, even while not hyper-parameter standardization, an excellent result most of the time it's conjointly one among the foremost used algorithms, attributable to its simplicity and variety (it is used for each classification and regression tasks). Random forest may be a supervised learning rule. The "forest" it builds, is Associate in Nursing ensemble of call trees, sometimes trained with the "bagging" methodology. The final plan of the sacking methodology is that a mixture of learning models will increase the result. Meanwhile, call Tree algorithms square measure one among the highest in data processing world moreover, because of its speed in coaching part and clear modeling call Tree works by classifying the trained information that create a tree. This tree is made in coaching part to examine however correct the classifier for check information. Then the check information is classified mistreatment the tree. The other rule here mentioned is k-Nearest Neighbor (KNN). This may be the easy and basic classification technique that often employed in several studies, particularly once there's solely few or no data concerning the info distribution. it's a non-constant quantity rule, implies that KNN doesn't create presumptions concerning distribution of knowledge employed in analysis. It fits in sensible environments, wherever often real information doesn’t follow theoretical statistics like distribution KNN does not make generalization which implies that KNN maintains all training data. KNN conjointly referred to as the lazy rule, or it solely use fast training part. Euclidean distance is typically utilised in KNN classifier for the calculation of the similarity between training and test data. It is calculated with formula below:

\[
d_{xy}(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

For clinical reason there are many investigates about utilizing machine learning, particularly in heart disease, for instance Chaurasia's examination: "Early Prediction of Heart Diseases Using Data Mining Techniques. Truck, ID3, and Decision Table were the 3 information mining algorithm he contrasted and utilized with foresee whether or not a patient will have heart disease. A dataset from California University, Irvine (UCI) is taken to do the examination, it involved 10 out of 76 boundaries in this review, they are: (1) Sex, (2) Age, (3) Chest Pain(cp), (4) Blood Pressure(trestbps), (5) Fasting Blood Pressure(fbs), (6) Cholestrol(chol), (7) Maximum Heart Rate Achieved(thalach), (8) Resting Electrocardiographic Results (restecg), (9) Slope, and (10) Exercise Induced Angina (exang). The featured outcomes were the exactness of the algorithms: ID3 = 72.9% CART = 83.5% and DT = 82.5%. Jarad et al with one more related review: "Wise Heart Disease Prediction System with MongoDB". They utilized data set programming called Naive Bayes with MongoDB, KNN algorithm and Decision List to anticipate patient's heart disease. They likewise utilized example dataset from UCI with 13 out of 76 boundaries accessible: (1) Sex, (2) Age, (3) Blood Pressure(trestbps), (4) Chest Pain(cp) (5) Fasting Blood Pressure(fbs), (6) Cholestrol(chol), (7) Maximum Heart Rate Achieved(thalach), (8) Resting Electrocardiographic Results (restecg), (9) Oldpeak, (10) Exercise Induced Angina (exang), (11) Slope, (12) Thalium Stress Result(tha), and (13) Number of major vessels(ca). The aftereffect of this review gives precision of algorithms utilized: Decision List 52%, Naive Bayes 52.33%, and KNN with 45.67% accuracy. Study done by Karthiga et al is the last model: "Heart Disease Analysis System Using Data Mining Techniques". They work was to anticipate heart disease also, with MAF algorithm and K-implies grouping. They use UCI dataset with 11 boundaries picked: (1) Sex, (2) Age, (3) Slope, (4) Fasting Blood Pressure(fbs), (5) Fasting, (6) Painloc, (7) Cholestrol(chol), (8) Thalium Stress Result(tha), (9) Blood Pressure(trestbps), (10) Maximum Heart Rate Achieved(thalach), and (11) Exercise Induced Angina (exang). They delivered significant outcomes: 89% to 74% exactness with various procedures proposed.

The creator Carlos Ordonez [1] has taken a review and learned with regards to the affiliation rule mining with the train and actually look at build on dataset for the expectation of heart condition. The inconvenience about Association decide mining is that it delivers uncommonly sizable number of decides that the majority of the square measure restoratively digestive. By and large, additionally affiliation runs square measure mined in general informational index while not approval on an independent example. To disentangle this, the creator has taken an alternate algorithmic program that utilizations search requirements to downsized the amount of rules. The algorithmic program initially looks for affiliation rules on a preparation set and finally approves them on an independent really look at set. Search imperatives and check set approval extensively downsized the amount of affiliation rules and turn out an assortment of rules with high prognostic exactness. These standards address important clinical information.

The creator Asha Rajkumar [2] dealt with diagnosing of heart condition abuse grouping upheld managed machine learning. An instrument named Tanagra is utilized to group the data, 10-overlap cross approval is utilized to measure the data and in this manner the outcomes were taken and analyzed. Tanagra might be a free information handling bundle for instructional exercise and examination capacities. Numerous information handling methodologies from educational information examination, applied arithmetic learning, machine learning and data space. The dataset is
parted into two sections, 80% is utilized for preparing and 20% for testing.

The creator Sairabi H. Mujawar [3] utilized Naïve Bayes and k-implies. Distinguishing proof of cardiovascular disease could be a convoluted errand and requires pleasant abilities. The dataset he utilized is given by Cleveland cardiovascular disease information. The objective worth with ‘1’ demonstrates the presence of cardiovascular disease and a worth ‘0’ shows the shortfall of cardiovascular disease. Changed k-implies chips away at each absolute and combinative information that we tend to experience here. Exploitation two introductory centroids we tend to obtain two farthest groups. It at long last offers a suitable assortment of groups. Gullible Bayes makes a model with prophetic capacities. The outcomes show a 93% accuracy for the right expectation of a cardiovascular disease and 89% exactness in that a patient doesn’t experiencing cardiovascular disease.

The creator S. Suganya et al. [4] anticipated cardiopathy exploitation fluffy truck recipe. It was acquainted inside the deliberate information with dispose of the vulnerability in information. To demonstrate conservative with significance various classifiers of consistent methods Minimum distance CART classifier was utilized. The heart disease dataset is at first sequestered into credits that expansion (heart condition | cardiopathy | cardiovascular disease) hazard. Then, at that point, fluffy enrollment work is applied to dispose of vulnerability and in the long run ID-3 algorithm is run recursively through the non-leaf branches till all the data are characterized. Java language is utilized in this undertaking. The creator Ashwini Shetty An et al. [5] for foreseeing cardiovascular disease the methodology was totally unique with a digit various information handling procedure. Creator's investigation work examinations the hereditary rule and neural organization for the forecast of cardiovascular diseases. Here, the neural organization utilizes 10 secret layers, 13 info layers and a couple of result layers. Characteristic layers were the info layers (explicitly age, resting beat rate, circulatory strain, blood glucose and others for example a sum of 13 credits). Back engendering rule of Levenberg-Marquardt is utilized for preparing and testing. To execute this framework enhancement tool stash is utilized. 'design' work is utilized with neural organization any place each weight lies between -2 to 2. Wellness work which getting utilized inside the hereditary decide is that the Mean sq. Blunder (MSE). For the change of loads Genetic rule is utilized. For everyone dependent on MSE, wellness work will be determined. Later the choice is done, change and hybrid in hereditary decide replaces the chromosome that is having lower adaption with the higher qualities. By improving the response that compares to interconnecting loads and edge of neural organization Fitter strings are acquired. The framework calculates accuracy of MATLAB.

By using WEKA pre-processing is finished. The hybrid system of neural network and genetic rule works far better than the performance of neural network alone is concluded. The author G Purusothaman et al. [6] have studied and compared totally different classification techniques for the prediction of cardiovascular disease. He applied Decision tree, artificial neural network and Naïve Bayes these three algorithms rather than applying only one, the authors interested in the working of hybrid models i.e. models which more than one classification technique. The performances of each models show promising results like decision tree, artificial neural network and Naïve Bayes are 76%, 85% and 69% severally. But the hybrid model purely dominated by giving 96% accuracy. Therefore, it is better for using hybrid models as it leads to a better result as a classifier for the prediction of heart diseases with sensible accuracy.

III. RELATED WORK

KNN, Random Forest are a portion of the calculation that is utilized broadly in these kinds of ventures and creators lean toward these calculations alongside others for their examination. An enormous number of works is being completed for discovering the proficient technique for the determination of different infection Our work depends on the push to foresee the conclusion with decreased number of properties so it can offer more towards the coronary illness issue utilizing characterization. Like said KNN and Random Forest were utilized. We utilized 6 credits rather than 13 and were prepared to accomplish a comparable exhibition.

IV. IMPLEMENTATION OF MODEL

Data Set: Data that is collected as a set is called dataset. It consists of a different attribute that a model uses to get trained. Dataset is the most important part for a machine learning projects. The dataset that is used in this project is taken from Kaggle. Here are some attributes that is there in the dataset.

- age - age in years
- sex - (0 = female; 1 = male)
- cptr - chest pain type
  - 0: Typical angina
  - 1: Atypical angina
  - 2: Non-anginal pain
  - 3: Asymptomatic
- restbps - resting blood pressure anything above 130-140 is typically cause for concern
- chlrl - serum cholesterol
- fbs - (0 = false; 1 = true)
  - '>126' mg/dL signals diabetes
- restecg - resting electrocardiographic results
  - 0: Nothing to note
  - 1: ST-T Wave abnormality
    - can range from mild symptoms to severe problems
  - 2: Possible or definite left ventricular hypertrophy
    - Enlarged heart's main pumping chamber
- thalach - maximum heart rate achieved
- exang - exercise induced angina (1 = yes; 0 = no)
- oldpeak - ST depression induced by exercise relative to rest looks at stress of heart during exercise unhealthy heart will stress more
- slope - the slope of the peak exercise ST segment
• 0: Upsloping: better heart rate with exercise (uncommon)
• 1: Flatsloping: minimal change (typical healthy heart)
• 2: Downsloping: signs of unhealthy heart
  - ca – no. of major vessels (0-3) colored by flouroscopy
  - thali - thalium stress result
  - 1.3: normal
  - 6: fixed defect
  - 7: reversible defect: no proper blood movement when exercising
  - target - have disease or not (1=yes, 0=no) (= the predicted attribute)

This makes our model ready.
\[ d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_k - p_k)^2} \]

\[ = \sqrt{\sum_{i=1}^{k} (q_i - p_i)^2} \]

**Random Forest**

We can understand the working of Random Forest algorithm with the assistant of following steps –

- ‘K’ attributes/features are selected in a random order from the wholesome ‘m’ attributes/features with the condition that the ratio between m and k is very high.
- Within the K attributes, a certain node D is picked with the logic of the best split point.
- The node is then split to further divisions of daughter nodes using the same logic of best split.
- Repeat above steps to the point where ‘L’ nodes are represented.
- Build the Random Forest by cycling above all steps for a total of ‘N’ number of times, that on completion will generate N trees.

**Logistic Regression**

Logistic Regression was utilized in the organic sciences in mid-20th century. It was then utilized in numerous sociology applications. Logistic Regression is utilized when the ward variable(target) is straight out.

For example,

To anticipate whether an email is spam (1) or (0)

Regardless of whether the tumor is harmful (1) or not (0)

The calculation that utilized are KNN and Random Forest. K-closest neighbors (KNN) calculation could likewise be such a directed milliliter calculation which could be utilized for each grouping additionally as relapse prognostic issues. In any case, it’s basically utilized for arrangement prognostic issues in business. Irregular backwoods might be a managed learning algorithmic principle that is utilized for every order additionally as relapse. In any case, yet it’s basically utilized for arrangement issues. Also, we realize that a timberland is framed by trees and extra trees proposes that extra solid woods. Likewise, the irregular timberland algorithmic principle makes call trees on information tests at that point gives the forecast from everything about and finally chooses the least complex goal by getting the vote.

**K-nearest neighbor (KNN)**

‘Feature similarity’ is used by K-nearest neighbors (KNN) algorithm. It is considered as a lazy algorithm that classifies data sets based on their similarity with neighbors.

- A number K of the neighbors is selected on initial phase.
- After the estimation of K, the Euclidean distance of number of neighbors (K) is calculated.
- Euclidean distance calculation is the factor came handy while taking K nearest neighbors

In simple words you can say that a random forest is an algorithm which uses different decision trees. Then from the decision tree it takes the results from each of them and gives the best solution by voting.

458
Consider a situation where we need to arrange if an email is spam. In the event that we utilize straight relapse for this issue, there is a requirement for setting up a limit dependent on which arrangement should be possible. Say if the genuine class is dangerous, anticipated consistent worth 0.4 and the edge esteem is 0.5, the information point will be named not harmful which can prompt genuine result progressively.

From this model, it tends to be construed that linear regression isn't appropriate for characterization issue. Straight relapse is unbounded, and this brings logistic regression into picture. Their worth stringently goes from 0 to 1.

Fig 3. Dataset provided by "Kaggle"

Where:
0 = patient suffering from heart disease
1 = patient not suffering from heart disease

V. EXPERIMENTAL RESULTS

Fig. 6. Count plot showing dataset is balanced

Fig. 7. Graph between number of neighbor and score
we are equally indebted to galgotia’s university for providing us such competitive task which enhance our professional as well as communication skills and make us learn about teamwork. i am ineffably indebted to runumi devi mam for conscientious guidance and encouragement to accomplish this assignment. finally, i would also like to thank all our friends for their cooperation and interest, which was necessary for completing our project report.

VII. CONCLUSION

The conclusion is that, machine learning algorithm can play a vital part in predicting whether the person is having heart disease or not. The algorithm is very promising and show accurate results. But still a lot of work can also be done to handle the high dimensional data and overfitting.

REFERENCES

[1] Carlos Ordonez, 2006, “Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction”, IEEE Transactions on Information Technology in Biomedicine (TITB), pp. 334-343, vol. 10, no. 2.

[2] Asha Rajkumar, and Mrs G. Sophia Reena, 2010, “Diagnosis of Heart Disease using Data Mining Algorithms”, Global Journal of Computer Science and Technology, Vol. 10, Issue 10, pp.38-43, September.

[3] Sairabi H. Mujawar, and P. R. Devale, October 2015, “Prediction of Heart Disease using Modified k-means and by using Naive Bayes”, International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 10, pp. 10265-10273.

[4] S. Suganya, and P.Tamije Selvy, January 2016, “A Proficient Heart Disease Prediction Method using Fuzzy-Cart Algorithm”, International Journal of Scientific Engineering and Applied Science (IJSEAS), Vol. 2, Issue1, ISSN: 2395-3470.

[5] Ashwini Shetty A, and Chandra Naik, May 2016, “Different Data Mining Approaches for Predicting Heart Disease”, International Journal of Innovative Research in Science,Engineering and Technology(An ISO 3297: 2007 Certified Organization), Vol. 5, Special Issue 9, pp. 277-281.

[6] G. Purusothaman, and P. Krishnakumar, June 2015, “A Survey of Data Mining Techniques on Risk Prediction: Heart Disease”, Indian Journal of Science and Technology, Vol. 8(12), DOI:10.17485/ijst/2015/v8i12/58385, pp. 1-5.

[7] Kaggle website for providing dataset

[8] Machine Learning with Python KNN introduction, TutorialsPoint

[9] “Machine Learning with Python Random Forest introduction, Javapoint