***Research on***

**PREDICTION OF HEART DISEASES USING MACHINE LEARNING ALGORITHMS**

**BY**

**SIVA SANKAR REDDYVARI**

(700693139)

Big Data Analytics & Information Technology

**TABLE OF CONTENTS**

Page

**ABSTRACT …………………………………………………………………………………………. 3**

**INTRODUCTION ……………………………………………………………………………………. 3**

**ABOUT HEART DISEASE ………………………………………………………………………… 4**

**METHODOLOGY …………………………………………………………………………………… 6**

**MACHINE LEARNING ALGORITHMS …………………………………………………………... 6**

**Decision tree ……………………………………………………………………………….. 6**

**Super Vector Machines …………………………………………………………………... 7**

**Naive Bayes Classifier ……………………………………………………………………. 8**

**Logistic Model Tree Classifier …………………………………………………………. 10**

**DEEP LEARNING FOR PREDICTION …………………………………………………………. 10**

**COMPARE MAJOR MACHINE LEARNING ALGORITHMS ………………………………… 12**

**CONCLUSION ………………………………………………………………………………………12**

**ABSTRACT:**

Heart disease is one of the most reason for death of individuals within the world. About 47% of all passings are caused by heart maladies. Agreeing to later overview by WHO association 17.5 million individuals dead each year. It'll increment to 75 million within the year 2030. This disease assaults a individual so immediately that it barely gets any time to urge treated with. So diagnosing patients correctly on opportune premise is the foremost challenging assignment for the medical fraternity. Destitute clinical choices may conclusion to understanding death and which cannot be managed by the hospital because it loses its notoriety.

Medical professionals working within the field of heart disease have their possess impediment, they can anticipate chance of heart assault up to 67% precision, with the current plague situation specialists require a back framework for more precise expectation of heart illness. Machine learning algorithm and deep learning offer new possibilities for predicting heart attack accurately. Precision of the forecast level is high when utilizing more number of attributes. Our point is to perform predictive analysis utilizing these information mining, machine learning algorithms on heart diseases and analyze the different mining, Machine Learning algorithms utilized and conclude which procedures are successful and proficient.

**1. INTRODUCTION**

One of the major challenges in heart disease is correct discovery and finding nearness of it interior a human. Early methods have not been so much proficient in finding it indeed medical specialist are not so much effective enough in predicating the heart disease. There are various medical instruments available within the advertise for predicting heart illness there are two major issues in them, the first one is that they are exceptionally much costly and moment one is that they are not proficiently able to calculate the chance of heart disease in human.

Agreeing to most recent survey conducted by WHO, the therapeutic proficient able to correctly anticipated as it were 67% of heart disease so there's a vast scope of inquire about in zone of predicating heart infection in human. There is a necessity for medical practitioners to predict cardiopathy before they occur in their patients. The options that increase the likelihood of heart attacks are smoking, lack of physical exercises, high pressure, high sterol, unhealthy diet, harmful use of alcohol, and high sugar levels. Cardio Vascular illness (CVD) incorporates coronary heart, neural structure (Stroke), hypertensive heart, inborn heart, peripheral artery, rheumatic heart, inflammatory cardiopathy. Cardio Vascular Disease (CVD) has become a leading cause of death due to changes in life styles in developing countries. CVD is projected to be one of the world's largest death killers. Data mining plays an important role in predicting diseases in the health sectors.

Machine Learning is one such instrument which is broadly utilized in distinctive spaces since it doesn’t require different algorithm for diverse dataset. Reprogrammable capacities of machine learning bring a parcel of quality and opens new doors of openings. Because the prediction of this disease involves a lot of parameters and technicality. Machine learning could be a better choice to achieve predictive accuracy. Algorithms such as Naive Bayes, Decision Tree, KNN, are used to predict the risk of heart disease in each algorithm, as Naive Bayes used the probability of predicting heart disease, though choice decision tree is utilized to provide classified report for the heart disease, This predication framework for heart disease helps doctors to foresee heart disease within the early arrange of disease resulting in sparing millions of life.

It can therefore be more likely to predict that patients will be diagnosed with heart disease by implementing a heart disease prediction system using data mining techniques and performing some sort of data mining on different heart disease attributes.

**2. ABOUT HEART DISEASE**

The term "heart disease" is regularly utilized traded with the term "cardiovascular illness." Cardiovascular illness for the most part alludes to conditions that include limited or blocked blood vessels that can lead to a heart assault, chest torment (angina) or stroke. Other heart conditions, such as those that influence your heart's muscle, valves or beat, moreover are considered shapes of heart illness.

**2.i. Symptoms:**

Men are more likely to experience chest pain; Women are more likely to develop other symptoms along with chest discomfort such as shortness of breath, nausea, and extreme tiredness.

**Here are some:**

* Chest pain, chest snugness, chest weight and chest discomfort (angina)
* Shortness of breath
* Pain, numbness, weakness or coldness in your legs or arms if the blood vessels in those parts of your body are narrowed
* Pain in the neck, jaw, throat, upper abdomen or back

**2.ii. TYPES OF HEART DISEASES:**

**Heart Arrhythmia**

A heart arrhythmia, is an unusual pulse. Your heart may beat as well rapidly, as well gradually or irregularly. Heart arrhythmia indications can incorporate:

* Fluttering in your chest, Hustling pulse(tachycardia), moderate pulse (bradycardia), Chest torment or discomfort, Shortness of breath, Light headedness, Discombobulation, Fainting (syncope) or close

**Heart Failure**

With heart failure, your heart doesn't pump blood as well because it ought to meet your body's needs. It is as a rule caused by coronary course infection, but it can also happen since you've got thyroid illness, high blood pressure, heart muscle disease (cardiomyopathy), or certain other conditions.

**Heart valve disease**

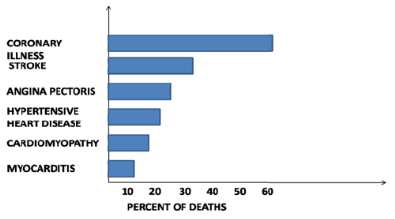
Heart valve disease occurs when not working properly one or more of the four valves in the heart. Heart valves help ensure that the blood that is pumped through the heart continues to flow. Cardiac valve disease makes it difficult. Examples are stenosis, prolapse of the mitral valve

**Cardiomyopathy (Heart Muscle Disease)**

Cardiomyopathy is a heart muscle disease, or myocardial disease. It becomes stretched, thickened, or stubborn. Your heart may become too weak for a good pump. The disease has many possible causes, including genetic heart conditions, reactions to certain drugs or toxins (such as alcohol), and virus infections. Chemotherapy sometimes leads to cardiomyopathy. Doctors are often unable to find the exact cause.

**Congenital Heart Disease**

Congenital heart disease is a type of birth defect that causes heart problems at birth and occurs in about 1 in 100 live births.



**3. METHODOLOGY**

The prediction methodology's goal is to design a model that can infer predicted class characteristics from combining other data. In this research, the task of data mining is to develop class prediction models based on selected attributes. The following algorithms are used in the research: **Decision tree, Support vector machine, Naïve Bayes, and Logistic model tree algorithm** for classifying and developing a model for diagnosing heart attacks in medical practitioners patient data set.

**4. MACHINE LEARNING ALGORITHMS FOR HEART DISEASE PREDICATION**

**4.i. DECISION TREE**

A decision tree is a decision support device that employments a tree-like chart or demonstrate of decisions and their conceivable results, counting chance occasion results, asset costs, and utility. fundamental component of decision tree includes root, nodes, and branching decision. A set of rules can be concluded from the tree which can be utilized to classify the obscure information record to its output value. The goal is to create a model that predicts the value of a target variable by learning simple rules of decision based on data characteristics. Both numerical data and categorical data can be handled by decision trees.

Decision trees determine order in various attributes for medical purposes and decisions are made based on the attribute. There are few approaches for building tree such as ID3, CYT, and J48 has utilized the approaches to classify the dataset utilizing J48, essentially have compared decision tree with classification yield of other calculation. Decision tree is utilized in those region of the restorative science where various parameters included in classification of information set. Since decision tree among all machine learning algorithms is the most compressive approach. These clearly reflect important characteristics in the data set. In heart disease where there are numbers of parameters that affect patients such as blood pressure, blood sugar, age, sex, genetic and other factors.

Doctor can clearly identify the most effective feature among all the parameters by seeing the decision tree. They can also produce the most influential feature in the population mass. Decision tree is based on entropy, and gaining information clearly indicates the value of the dataset.

**Benefits:**

* From the training data, understandable prediction rules are created.
* Builds the tree that is fastest.
* Building a short tree. Only enough attributes need to be tested until all data are classified.

**Drawbacks:**

* Data may be over-fitted or over-classified.
* Only one attribute is tested for decision-making at a time.

Due to the decision tree spilt data set aligned to the axis, overfitting occurred means that a lot of nodes are needed to spread data. This problem is solved by J48 explained in the greedy method leads to less optimal tree. J48's additional features include missing values, pruning decision trees, continuous value ranges of attributes, rule derivation, etc. It generates the rules from this algorithm from which specific identity of that data is generated. The aim is to generalize a decision tree gradually until it achieves a balance of flexibility and accuracy

**4.ii. SUPPORT VECTOR MACHINES:**

Vector support machines are available in various forms, both linear and non-linear. A vector supporting machine is a supervised classifier. By finding the hyper plane that maximizes the margin between two classes, an SVM performs classification. Support vectors are the vectors that define the hyper plane. Steps for hyperplane calculation.

* Establish data on training
* Set parameter SVM
* Train the SVM
* SVM classified region
* Support vector

Building an optimal classification decision plan requires minimizing the error function. The shape of the error function becomes the basis for further classification of these algorithms in linear, polynomial, sigmoid and radial SVM categories.

So installing simple phrases the philosophy of SVM is to obtain an optimum hyper plane for facts points that are linearly separable. support vectors actually check with the records factors that are closest to the demarcating floor which can be subsequently intricate to categorise. The metric that alludes to the optimality of a hyper plane is the margin across the hyper aircraft. So the trouble transitions into that of an optimization one. As mounted the maximum margin classifier learnt and derived from the education statistics could lead us to most suitable hyper plane. this is finished by means of remodelling the maximal margin classifier because the inner product (sum of multiplication of pair values) of given statistics factors in preference to the facts factors.

The general function of the kernel could be defined as follows:

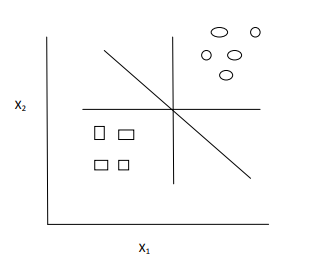
𝑓(𝑥) = 𝐵0 + ∑(𝑎𝑖 × (𝑥, 𝑥𝑖 ))

Here x is the new input vector and from training data it is necessary to estimate the coefficients B0 and ai.

**Advantages:**

* Firstly regularisation parameters which avoid hassle of over fitting which one of the most important demanding situations is in choice tree.
* Kernel tree is used to keep away from the professional expertise through the expertise of kernel.
* SVM is an effective technique as it make use of convex optimisation problem (COP) which imply it has doesn’t neighbourhood minima.
* error rated is tested which provide a more guide after misclassification of dataset.

All the above functions might be beneficial for scientific diagnose dataset that is ensuing in constructing extra efficient predication machine for the health practitioner. It doesn’t imply it has all correct aspect .coin has usually facet on the opposite side it has best feature removal of over fitting trouble is quite touchy and it need optimizing parameter flaw in optimisation might also bring about error and may cause over becoming.

Fig: SVM Classifier

**4.iii. NAÏVE BAYES CLASSIFIER**

Instead of a single classifier, it is a combination of multiple classifiers all working on the principle of independent features of Naïve Bayes. Each feature is assumed to be independent and autonomous, individually contributing to the probability of the training data point belonging to a specific class. A NaïveBayesian output is definitely not difficult to run, with no clogged repetitive parameter estimation that makes it particularly supportive of broad datasets despite its effortlessness, the Naïve Bayesian classifier generally does its job shockingly well and is widely used in view of the fact that it often overlaps high-order techniques that are complex.

The Naïve Bayes treats each variable as independent, which helps to predict even if the variables are not related to each other. This classifier is a powerful probabilistic representation and has been given considerable attention to its classification use.

Classification is then done by means of applying Bayes rule to compute the chance of C given the particular instances of A1. An and then predicting the elegance with the best posterior chance. The goal of class is to efficaciously expect the fee of a designated discrete class variable given a vector of predictors or attributes.

In specific, the Naive Bayes classifier is a Bayesian community in which the elegance has no mother and father and each attribute has the class as its sole figure. although the naive Bayesian (NB) set of rules is easy, it may be very effective in many actual international datasets because it can deliver higher predictive accuracy. but, whilst attributes are redundant and now not usually disbursed, the predictive accuracy is decreased.

**Benefits:**

* clean to put in force
* requires a small amount of schooling information to estimate the parameters
* proper outcomes acquired in maximum of the cases

**Drawbacks:**

* Assumption: magnificence conditional independence, therefore loss of accuracy.
* almost, dependencies exist amongst variables.
* Dependencies among those can't be modelled with the aid of Naïve Bayesian Classifier.

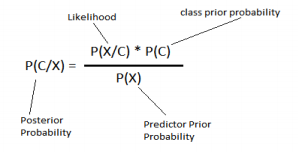


Fig: Naïve Bayes process

* 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.

**4.iv. LOGISTIC MODEL TREE CLASSIFIER:**

Logistic Model Tree is the classifier for the construction of logistic model trees, consisting of a decision tree structure with the leaves ' logistic regression function. The algorithm can monitor target variables, numerical and nominal attributes and missing qualities in parallel and multi-class. The tested attributes are associated with each inner node, as in the decision tree. The attributes with k values, the node has k infant nodes for nominal attributes and the instances are sorted according to the value of the attribute. The node has two child nodes for the numeric attributes and compares the test value attributes to a threshold.

Ideally, we want to adapt our algorithm to the dataset concerned: for small datasets where a simple linear model offers the best bias-variance trade. The "tree" logistical model should consist of a single logistical regression model, i.e. it should be tapped back to the root. A more elaborate tree structure is appropriate for other datasets. Building a logistic model tree by developing a standard classification tree, building logistic regression models for all nodes, pruning a percentage of sub-trees using a pruning model and in some way combining logistic models in a single model. Cross-validation is used by the pruning plan to achieve more steady pruning results. Despite the fact that this expanded the multifaceted computational nature, it produced littler trees and, for the most part, more accurate trees.

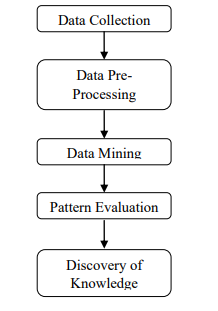
These ideas lead to the following algorithm for the development of logistic model trees:

The development of tree begins with the construction of a logistic model at the root using the algorithm LogitBoost. The amount of is solved using 10 fold cross-validation. In this process the information is part of the preparation and test set 10 times, for each preparation set LogitBoost is precipitated to a maximum number of cycles and the lapse rates on the test set are logged for each cycle and summarized over the distinctive folds. The amount of stresses with the least number of blunders is used to prepare the LogiBoost algorithm for all the information. This gives the model of logistic regression at the tree base.

Like other tree impelling systems, LMT does not require parameter tuning. LMT produces a solitary tree with double parts of numerical properties, multi-route parts of ostensible properties and logistic regression models at the leaves, and the algorithm ensures that only applicable attributes are incorporated in the last.

**4.v. DEEP LEARNING FOR PREDICATION IN HEART DISEASE:**

Deep learning can be defined as a machine learning subfield based on learning at multiple levels of representation and abstraction; each level contains multiple processing units for multiple processing of input and output layers. Deep learning work on the principle of characteristic hierarchy in which higher level hierarchy is formed by lower level characteristics of composition. Deep getting to know carry renaissance to the neural network version foremost paintings is going in the subject of in its implementation via stacked confined Boltzmann device and automobile encoder-decoder technique. This strategy awe inquires about with their performance in field of picture preparing and layer shrewd pre training procedures other zones of its application include Natural dialect processing, acoustic handling. RNN is consider to be best suited for consecutive include and sequential information their exist different strategy working on these two form LSTM was proposed by Hochreiter and Schmidhuber, the performance in the field of sequence-based tasks is quite impressive. A worldly based heart malady expectation has been wiped out paper where author utilized GRE to achieve the tall precision



*Fig: Deep Learning flowchart:*

From the flow chart there are five modules it has their own specific operation, the objective is to display the over stream chart in most common way. Data collection is a phase in which the data set is collected from the standard repository followed by the pre-processing phase, which includes noise reduction functionality and selection of features. subsequent step is core for deep learning as it enforces the primary algorithmic approach adapted for manipulation of data set. Algorithms can range from deep belief to recurring neural networks. The performance analysis of the above data mining technique was the main module because the basic comparison of the above adapted method was illustrated. Our desire goal, which includes percentage or probability of occurrence of instances, will be achieved in the last discovery of the knowledge module. The likelihood of a heart attack in the patient is in our case.

**5.** **Compares major algorithms based on different parameters of MACHINE LEARNING:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Technique | Outlier | Online learning | Over/under  Fitting | Parametric | Accuracy | Execution  Technique |
| SVM | SVM can  handle properly | It takes less time  than ANN | Better performance  than fitting and  fitting | Its not parametric  Model | Superior to  other  parametric  models | Depending on the dataset used, the  NLP operation is  generally rather  slow |
| Naïve Bayes | It is less  pruned to  outlier | This can perform on  Online | It is not affected  by fitting and under  fitting | Yes | High with a  limited set of  data | Low with limited  Set of data |
| Decision  Tree | Outlier does  now not play  vital position in  interoperation of dataset by  way of  decision tree | Doesn’t support | It is affected by  fitting and under  fitting | Non-parametric  Model | Depends on  dataset,  technique.  This have high  accuracy than  SVM | Need less time  than Any other  parametric  Model. |

**Conclusion:**

Heart attack is a crucial human society health problem. This paper summarizes state-of - the-art techniques and methods available to prevent this disease. It Provides a deep insight into the classification of heart disease machine learning techniques. Deep learning an emerging artificial intelligence field has shown some promising results with high accuracy in other fields of medical diagnosis. The part of classifier is significant in healthcare industry so that the comes about can be utilized for foreseeing the treatment which can be given to patients. Machine learning procedures altogether progresses precision of cardiovascular hazard expectation through which patients can be recognized amid an early organize of malady and can be benefitted by preventive treatment.

When we see at Naïve Bayes comes about we are getting more precise comes about with probabilities of all other possibilities but due to direction to as it were one arrangement decision trees may miss lead. At long last ready to say by this try that Naïve Bayes is more precise in the event that the input information is cleaned and well kept up indeed in spite of the fact that ID3 (J48) can clean it self it cannot donate precise comes about each time, and in this same way Naïve Bayes moreover will not deliver exact results about each time we have to be consider results about of diverse algorithms and by all its comes about in case a forecast is made it'll be exact. But able to utilize Naïve Bayes consider factors as person ready to utilize combination of algorithms like Naïve Bayes.

By analysing the test results, it is concluded thatJ48 tree method turned out to be best classifier for heart infection expectation since it contains more exactness and slightest add up to time to construct. Highest accuracy belongs to the J48 algorithm with reduced pruning of errors followed by LMT, whereas applying reduced pruning of errors to J48 results in higher performance without pruning, it results in lower Execution. The finest algorithm J48 based on UCI information has the most noteworthy exactness

In conclusion, as distinguished through the writing audit, we accept as it were a minimal victory is accomplished within the creation of prescient demonstrate for heart disease patients and thus there's a need for combinational and more complex models to extend the exactness of anticipating the early onset of heart malady.

**REFERENCES**

[1]. Prerana T H M1, Shivaprakash N C2, Swetha N3 "Prediction of Heart Disease Using Machine Learning Algorithms-Naïve Bayes" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99 ©IJSE Available at www.ijse.org ISSN: 2347-2200

[2]. A.T.Azar and S.M.El-Metwally,"Decision tree classifiers for automated medical diagnosis," Neural Comput. Appl., vol. 23, no. 7–8, pp.2387–2403, Dec.2013. 10Y.C.T.Bo Jin, "Support vector machines with genetic fuzzy feature transformation for biomedical data classification.," Inf Sci, vol.177, no.2,pp.476–489,2007

[3]. Elkan C."Naive Bayesian Learning, Technical Report CS97-557",Department of Computer Science and Engineering, University of California, San Diego, USA, 1997

[4]. T.M.Lakshmi, A.Martin, R. M. Begum, and V. P. Venkatesan, “An Analysis on Performance of Decision Tree Algorithms using Student’s Qualitative Data,” Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 5, pp. 18–27, 2013.

[4]. P.Sharma and A. P. R. Bhartiya, “Implementation of Decision Tree Algorithm to Analysis the Performance,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 1, no. 10, pp. 861–864, 2012.

[5]. D.K.Srivastava and L.Bhambhu, “Data classification using support vector machine,” J. Theor. Appl. Inf. Technol., 2009

[6]. J. Schmidhuber, “Deep Learning in neural networks: An overview,” 2015.

[7]. Yuming Hua, Junhai Guo, and Hua Zhao,“ deep learning,” Proc. 2015 Int. Conf. Intell. Comput. Internet Things, pp. 1–4, 2015.

[8]. J.R. Quinlan, “Induction of Decision Trees,” Expert System, vol. 1, no. 1, pp. 81–106, 2007

[9]. D.S. Medhekar, M.P. Bote, and S. D. Deshmukh, “Heart Disease Prediction System using Naive Bayes,” International Journal of Enhanced Research in Science Technology and Engineering, vol. 2, no. 3, Elsevier 2013.

[10]. Y. Zhang et al., “Studies on application of Support Vector Machine in diagnose of coronary heart disease,” Electromagnetic Field Problems and Applications 2012 Sixth International Conference (ICEF), Dalian, IEEE 2012.

[11]. S. A. Pattekari and A. Parveen, “PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES,” International journal of Advanced Computer and Mathematical Sciences, vol. 3, no. 3, pp. 290–294, 2012

[12]. Prerana T H M1, Shivaprakash N C2, Swetha N3 "Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes" International Journal of Science and Engineering Volume 3,Number 2 – 2015 PP: 90-99 ©IJSE Available at www.ijse.org ISSN: 2347-2200