

# A Study and Review on implementation of Machine Learning Techniques & Data Mining in the Domain of Heart Disease Prediction

<sup>1</sup>Nikita Shrivastava,<sup>2</sup>Shashank Girepunje,<sup>3</sup>Name of 3<sup>rd</sup> Author

<sup>1</sup>M. Tech Scholar,<sup>2</sup>Assistant Professor,<sup>3</sup>Head of Department,

<sup>1</sup>Computer Science and Engineering Department,

<sup>1</sup>Kalinga University, New Raipur.

**Abstract :** Heart disease, which is also known as cardiovascular disease, encloses different conditions that effect the heart and is the primary basis of death worldwide over the span of the last few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyses huge complex medical data, helping healthcare professionals to predict heart disease.

**IndexTerms - Heart Disease Prediction, Data Mining, Machine Learning Algorithms, Random Forest, Decision Tree, Support Vector Machine, Naïve Bais algorithm.**

## I. INTRODUCTION

Over the last decade, heart disease or cardiovascular remains the primary basis of death worldwide. Concerning the heart disease prediction numerous systems are being recommended which are being deployed by the means of various techniques and algorithms. Heart disease is one of the top leading causes of death accounting for 17.7 milliondeaths each year, 31% of all global deaths, as reported by World Health Organization 2017. Patients unhealthy habits such as tobacco use, unhealthy diet, physical inactivity and alcoholusage are the main reasons leading to many types of HD. Several clinical information andsymptoms are found to be related to HD including age, blood pressure, total cholesterol,diabetes, hyper tension.

Gaining quality service at affordable price remains the prime and challenging concern for the healthcare establishments. For offering quality services at par, there must be accurate diagnosis of the patients along with effective dosage of medicines. Low quality clinical diagnosis and treatment can yield in undesired and inadequate results.

Data mining refers to the extraction of required information from huge datasets in various fields such as the medical field, business field, and educational field. Machine learning is one of the most rapidly evolving domains of artificial intelligence. There have been numerous dedicated approaches based on datamining techniques proposed in recent years to help healthcare professionals in the diagnosisof Heart related issues. Heart disease prediction systems based on data mining techniques could assist doctors in giving accurately Heart diseases prediction making based on the clinical information data of patients. Datamining techniques which refers to mining the information, allow the extraction of hiddenknowledge and established the relationships between attributes inside the data, is the promising techniques for Heart attack prediction

## II. REVIEW OF LITERATURE

There have been numerous dedicated approaches based on datamining techniques proposed in recent years to help healthcare professionals in the diagnosisof Heart issues. Heart attack prediction systems based on data mining techniques could assist doctors in givingaccurately Heart attack prediction making based on the clinical information data of patients. Datamining techniques which refers to mining the information, allow the extraction of hiddenknowledge and established the relationships between attributes inside the data, is the promising techniques for Heart issues prediction

- **A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease**

This paper expects to explore and analyze the exactness of various information mining arrangement plans, utilizing Ensemble Machine Learning Techniques, for the forecast of coronary illness. The Cleveland informational index for heart infections, containing 303 cases, has been utilized as the fundamental data set for the preparation and testing of the created framework. 10-Fold Cross-Validation has been applied to build the measure of information, which would some way or another have been restricted. Various classifiers, in particular Decision Tree (DT), Naïve Bayes (NB), Multilayer Perceptron (MLP), K-Nearest Neighbor (K-NN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM), have been utilized. Besides, the outfit forecast of classifiers, stowing, boosting and stacking, has been applied to the dataset. The aftereffects of the examinations demonstrate that the SVM strategy utilizing the boostingtechnique outflanks the other previously mentioned strategies.

- **Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automated Optimization**

Diverse information mining and machine learning techniques are often used to research the gigantic quantity of facts to create greater commercial values in high-end organization systems. But, the development of technology has made statistics mining and machine learning possible on low-end systems, consisting of private computers or embedded structures. At the same time as researchers have proposed splendid work on the management designs and symptoms of different components of the system, most of the paintings are constructed upon the traits of the gadget, which might also alternate every now and then. This makes it not possible to optimize the device overall performance with static, or statically adaptive, machine designs. On this paintings, we propose to embed the supports of information mining and machine learning to the layout of working gadget, with a purpose to find out a new, automatized manner to adaptively optimize the device without the use of complex algorithms. To validate the proposed ideas, we pick the cache layout as a case look at, wherein the alternative of cached contents is automatically controlled by using a choice maker. The choice maker then replies on a facts miner, which analyzes the data collected through the gadget display. The efficacy of the considered case is proven by way of a series of experiments, wherein the effects are quite encouraging.

- **A Hybrid Machine Learning Approach for Prediction of Heart Diseases**

The point of this paper is to introduce a proficient strategy of foreseeing heart illnesses utilizing AI draws near. Subsequently we proposed a half breed approach for heart forecast utilizing Random woodland classifier and basic k-implies calculation AI procedures. The dataset is additionally assessed utilizing two other diverse AI calculations, to be specific, J48 tree classifier and Naive Bayes classifier and results are thought about. Results achieved through Random backwoods classifier and the relating disarray network shows vigor of the system.

- **A Survey on Predicting Heart Disease using Data Mining Techniques**

The point of this paper is to introduce a proficient strategy of foreseeing heart illnesses utilizing AI draws near. Subsequently we proposed a half breed approach for heart forecast utilizing Random woodland classifier and basic k-implies calculation AI procedures. The dataset is additionally assessed utilizing two other diverse AI calculations, to be specific, J48 tree classifier and Naive Bayes classifier and results are thought about. Results achieved through Random backwoods classifier and the relating disarray network shows vigor of the system.

- **Heart Disease Prediction System using Data Mining Techniques: A study**

Information Mining is the cycle of non-unimportant extraction of certain, already obscure and conceivably valuable data from information. An example is fascinating in the event that it is substantial for a given test information with some level of assurance, novel, possibly valuable and effectively comprehended by people. The immense measure of information produced for expectation of coronary illness is excessively intricate and voluminous to be prepared and broke down by customary strategies. Progressed Data Mining apparatuses conquer this issue by finding concealed examples and valuable data from mind boggling and voluminous information. Scientists explored writing on expectation of coronary illness utilizing information mining strategies and detailed that Neural Network procedure conquer any remaining methods with more significant levels of precision. Applying Data Mining procedures on medical care information can help in foreseeing the probability of patients getting coronary illness. This paper features the significant pretended by information mining instruments in investigating enormous volumes of medical services related information in expectation and analysis of infection.

- **Prediction of Heart Disease Using Machine Learning**

With the wild expansion in the heart stroke rates at adolescent ages, we need to set up a framework to have the option to recognize the indications of a heart stroke at a beginning phase and consequently forestall it. It is unfeasible for an average person to often go through exorbitant tests like the ECG and hence there should be a framework set up which is helpful and simultaneously solid, in anticipating the odds of a coronary illness. Hence we propose to foster an application which can anticipate the weakness of a coronary illness given fundamental side effects like age, sex, beat rate and so forth The AI calculation neural organizations has demonstrated to be the most exact and solid calculation and thus utilized in the proposed framework.

- **Design and Implementing Heart Disease Prediction Using Naives Bayesian**

The examination centers on coronary illness conclusion by thinking about past information and data. To accomplish this SHDP (Smart Heart Disease Prediction) is constructed through Navies Bayesian to anticipate hazard factors concerning coronary illness. The quick progression of innovation has prompted wonderful ascent in versatile wellbeing innovation that being one of the web application. The necessary information is collected in a normalized structure. For foreseeing the odds of coronary illness in a patient, the accompanying ascribes are being gotten from the clinical profiles, these include: age, BP, cholesterol, sex, glucose and so on The gathered qualities goes about as contribution for the Navies Bayesian arrangement for foreseeing coronary illness. The dataset used is parted into two areas, 80% dataset is used for preparing and rest 20% is used for testing. The proposed approach incorporates following stages: dataset assortment, client enrollment and login (Application based), grouping through Navies Bayesian, forecast and secure information move by utilizing AES (Advanced Encryption Standard). From that point result is created. The examination expounds and presents numerous information deliberation procedures by utilizing information mining techniques which are received for coronary illness expectation. The yield uncovers that the set up demonstrative framework viably helps with anticipating hazard factors concerning heart illnesses.

## Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems

Medical services associations are confronted with difficulties to give financially savvy and excellent patient consideration. The two directors and clinicians need to investigate an abundance of information accessible in the data sets of medical care data frameworks to find information and to settle on educated choices. This is basic specifically to improve the adequacy of illness treatment and counteractions. It happens to more significant if there should be an occurrence of coronary illness (HD) that is viewed as the essential explanation for death in grown-ups. Information mining fills in as an examination apparatus to find covered up connections and examples in HD clinical information. This paper surveys five models built of single and consolidated information mining strategies to help clinical choices in (HD) determination and forecast. The five frameworks give programmed design acknowledgment and endeavors to uncover connections among various boundaries and indications of HD. Every framework shows set of qualities and impediments as far as the kind of information it handles, exactness, simplicity of translation, dependability and speculation capacity. Helpless speculation capacity is as yet a significant open issue for information mining in medical services primarily as a result of the absence of information and cost of re-handling.

**Table 1: Comparative Analysis on deferent Machine learning Algorithms**

Random Forest	Decision Tree	Support Vector Machine	Logistic Regression	Naive Bayes Algorithm
It provides accurate predictions for many types of applications	Easy to interpret the decision rules	Support Vector Machine is a technique to make predictions, both in the case of classification and regression	Logistic regression model looks like an equation between independent variables with respect to its dependent variable.	Naive Bayes is a linear classifier
It can measure the importance of each feature with respect to the training data set.	Nonparametric so it is easy to incorporate a range of numeric or categorical data layers and there is no need to select uni modal training data.	SVM are in a class by Artificial Neural Network regarding functionality and condition problems can be solved.	Logistic regression is a parametric model, in which the model is defined by having parameters multiplied by independent variables to predict the dependent variable.	It tends to be faster when applied to big data. In comparison, Other Algorithms are usually slower for large amounts of data,
Pairwise proximity between samples can be measured by the training data set.	Robust with regard to outliers in training data.	Support Vector Machine is a selection method that compares the standard parameter set of discrete values, called the candidate set, and take the one that has the best classification accuracy.	Assumptions are made on response (or dependent) variable, with binomial or Bernoulli distribution.	Naive Bayes is highly accurate when applied to big data.

### III. COMPARATIVE CONCLUSION

Machine learning is scientific techniques where the computers learn how to solve a problem, without explicitly program them. Deep learning is currently leading the ML race powered by better algorithms, computation power and large data. Still ML classical algorithms have their strong position in the field. Naïve Bayes is a classification method based on Bayes theorem that derives the probability of the given feature vector being associated with a label. Naïve Bayes has a naive assumption of conditional independence for every feature, which means that the algorithm expects the features to be independent which not always is the case. Logistic regression is a linear classification method that learns the probability of a sample belonging to a certain class. Logistic regression tries to find the optimal decision boundary that best separates the classes. SVM can handle non-linear solutions whereas logistic regression can only handle linear solutions. Linear SVM handles outliers better, as it derives maximum margin solution. Hinge loss in SVM outperforms log loss in LR.

### REFERENCES

- [1] G. Sannino and G. De Pietro, "A smart context-aware mobile monitoring system for heart patients," in Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on. IEEE, 2011, pp. 655–695.
- [2] A. Adeli and M. Neshat, "A fuzzy expert system for heart disease diagnosis," in International Multi Conference of Engineers and Computer Scientists, vol. 1, 2010.
- [3] D. P. Mandic and J. Chambers, Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability, 2001.
- [4] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques, 2016
- [5] SeyedaminPouriye\*, Sara Vahid\*, Giovanna Sannino\*, Giuseppe De Pietroy, Hamid Arabnia\*, Juan Gutierrez "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017
- [6] Anjan Nikhil Repaka, Anjan Nikhil Repaka, Anjan Nikhil Repaka, "Design And Implementing Heart Disease Prediction Using NaivesBayesia" proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI