**HEART DISEASE PREDICTION USING MACHINE LEARNING**

Jaidi Akshitha Reddy Jaidi Akanksha Reddy Varsha Reddy Baddam

[axj17770@ucmo.edu](mailto:axj17770@ucmo.edu) [axj17760@ucmo.edu](mailto:axj17760@ucmo.edu) vxb99720@ucmo.edu

**Abstract:** In recent times, heart disease prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart disease. Day by day the cases of heart diseases are increasing at a rapid rate and it’s very Important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e., it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc. So, a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances as heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This paper makes use of heart disease dataset available in UCI machine learning repository. The proposed work predicts the chances of heart disease and classifies patient's risk level by implementing different data mining techniques such as Naive Bayes, Decision Tree, Logistic Regression and Random Forest. Thus, this paper presents a comparative study by analyzing the performance of different machine learning algorithms. The trial results verify that Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

Keywords - optimization, Prediction model, classification, accuracy, training data set, testing, algorithm.

**Introduction**

Diagnosing heart disease early is crucial for saving lives and preventing further damage, as it is one of the most prevalent diseases today. Medical professionals can use this machine learning prediction to better assess their patients' conditions. Thus, the key to discovering, comprehending, and alleviating heart-related ailments is the application of Machine learning. In this method, a machine uses preexisting data to learn the patterns (through algorithms) and then returns the findings. Data mining-related algorithms abound, and they can be used to better recognize the medical data. Combining many classifiers can boost performance. The primary goals of this research are (1) to increase accuracy and (2) to demonstrate the significance of early disease prediction. The dataset is utilized to make predictions using supervised machine learning algorithms like KNN and Naive bayes, and the results are analyzed.

Heart-related disorders, often known as cardiovascular diseases, have been the leading cause of death worldwide during the past ten years. According to the World Health Organization, more than 17.9 million people die annually from cardiovascular illness, with coronary heart disease as well as cerebral stroke accounting for 80 percent of these fatalities. Numerous deaths are common in both low- and middle-income economies (Seckeler 2011). Heart disease is caused by several factors, including personal and occupational behaviors and genetic predisposition (Gaziano 2010). Cigarettes, excessive drinking and espresso consumption, stress, inactivity, and medical conditions such as obesity, asthma, high blood cholesterol, and a history of heart disease are all risk factors for heart disease. The ability to identify cardiac illness rapidly, accurately, and precisely is essential for initiating preventative measures to prevent death.

Knowledge discovery through data mining is the most popular method of knowledge extraction. Machine learning has several potential applications for issue solving. Algorithms for data mining and machine learning are becoming important in the healthcare industry. In the health care industry, there is a wealth of data. Unfortunately, most of it was wasted. Coronary artery disease ranks among the leading causes of death around the world. Heart disease is the leading cause of death in the Western world. Predictions of heart illness can be made using a variety of statistical methods, including the Decision Tree, J48 algorithm, Logistic model tree approach, Random Forest algorithm, Naive Bayes, KNN, Support Vector Machine, and Nearest Neighbor. Including more attributes raises the forecast level's accuracy. The purpose of this research is to perform a predictive analysis on heart disease using machine learning algorithms and to compare the various machine learning algorithms utilized to decide which are the most effective and practical.

The healthcare industry has an abundance of data collected from patients. Several machine learning methods will need to be used to analyze these data. Practitioners reviewed the data to determine the best course of treatment.

**MOTIVATION**

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. Early detection of heart disease is crucial in saving the life of people. he chances of survival are greater when emergency treatment begins quickly. Good data-driven systems for predicting heart disease can improve the entire research and prevention process, making sure that more people can live healthy lives. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms, k Nearest Neighbor, Support Vector Machine for predicting heart disease.

.

**CONTRIBUTION** **AND** **OBJECTIVES**

This work aims in developing the best accuracy and performance among Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, K nearest Neighbor etc. By using several cardiovascular system parameters such as age, blood pressure, ECG results, sex, and blood sugar, it is possible to measure the possibility of getting affected by heart disease. For deriving the algorithm with the best accuracy in the detection and prediction of heart disease, a comparative analysis of chosen machine learning algorithms has been shown. This algorithm takes the medical parameters such as age, blood pressure, heartbeat, sex, ECG results, blood sugar etc. as input and shows the probability of getting affected by heart disease as output.

**RELATED WORK AND IMPLEMENTATION**

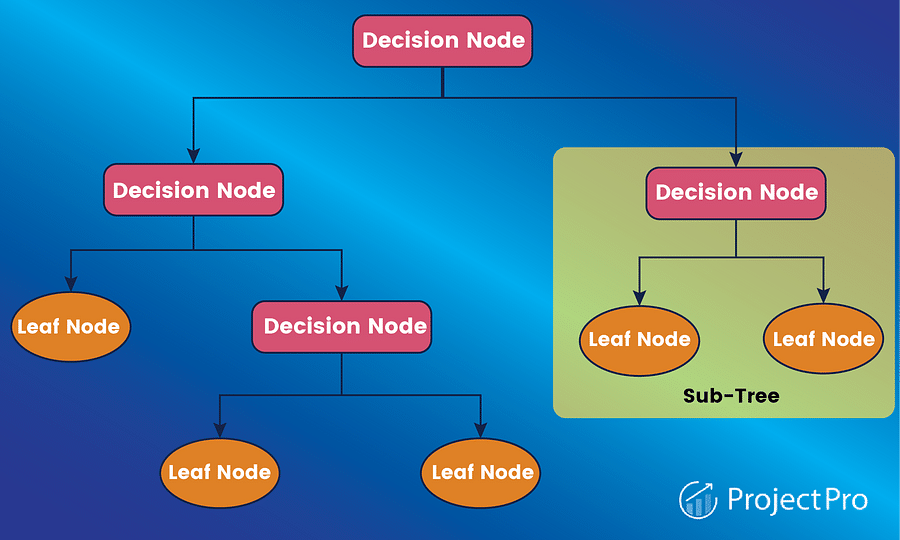
**DECISION TREE ALGORITHM**

Decision tree classifiers are highlighted as the best option for predicting heartbeats in WEKA in this paper. As a decision-making aid, a decision tree uses a tree-like graph or model of potential actions and the consequences (such as costs, benefits, and probabilities) that might result from doing those actions. This is one representation of a program that operates solely on if-then statements. These methods make use of supervised learning for classification and regression, although they are not parametric. The goal is to build a model that can reliably predict a target variable's value by deducing simple decision rules from those properties. For this reason, decision trees can take either numerical or categorized data as input. In medicine, decision trees are used to decide the relative importance of multiple qualities and the ensuing course of action.

An Urgent Plea Using the values of the input (independent) characteristics, a tree can be used to learn a classification function. The presence of the dependent property and the ability to count classes (values) establish that this is, in fact, a supervised classification problem. A tree's precision may vary with its degree of intricacy. Numerous metrics, including but not limited to the number of nodes, the number of leaves, the tree's depth, and the number of features used in tree development, are commonly employed to assess tree complexity. An effective method for controlling a tree's growth is pruning.

The J48 algorithm, the logistic model tree algorithm, and the Random Forest decision tree algorithm are contrasted in this study as potential decision tree approaches. The proposed approach reduces the number of false positives by cutting out unnecessary information and includes a confidence factor and fast features. It has been proven that decision tree performance can be greatly improved by reducing the number of errors trimmed from the data. The best performance in recognizing patients with heart disease is then compared between the three decision tree algorithms. Training => Algorithm => Model => Testing => Evaluation

**Fig 1.0 Decision Tree Algorithm**



**Advantages**:

* Understandable prediction rules are created from the training data.
* Builds the fastest tree.
* Builds a short tree.

**Disadvantages:**

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

**RANDOM FOREST ALGORITHM**

As a form of supervised classification, the random forest algorithm is one example. This program, as its name implies, generates a forest full of trees. Most people would agree that a forest appears healthier when there are more trees there. Similarly, the random forest classifier's accuracy improves as the number of trees in the forest grows. Random Forest can be implemented using one of three distinct approaches, including Forest-RI (Random Input choice), Forest-RC (Random Blend), and a hybrid of the two.

Few other conventional models can match the random forest algorithm's adaptability and durability when it comes to classification jobs using tabular data. The random forest classifier is frequently used for forecasting loan risk, fraud detection, and even cardiovascular illness because of its ease of use and adaptability.

Random forest classifier makes use of the ensemble learning theorem to maximize training by combining the output of several decision trees. The purpose is to maximize the predicted accuracy of the dataset by utilizing several subsets in various permutations. As a first stage, construct various decision trees, then optimize them, combine and contrast them. Then, the trees are used for prediction, and the results are ensembled to produce an overall forecast.

The Random Forest procedure has some desirable qualities, for example

* It is not difficult to utilize, basic, and effortlessly parallelized.
* It doesn't "t oblige models or parameters to choose aside from the number of indicators to pick at arbitrary at every node.
  + It runs effectively on extensive databases; it is moderately strong to anomalies and commotion.
  + It can deal with a considerable number of information variables without variable deletion; it evaluates what variables are important in classification.
  + It has a successful system for assessing missing information and keeps up accuracy when a vast extent of the data is missing. It has methods for adjusting errors in class populace unequal data sets.

**Advantages**

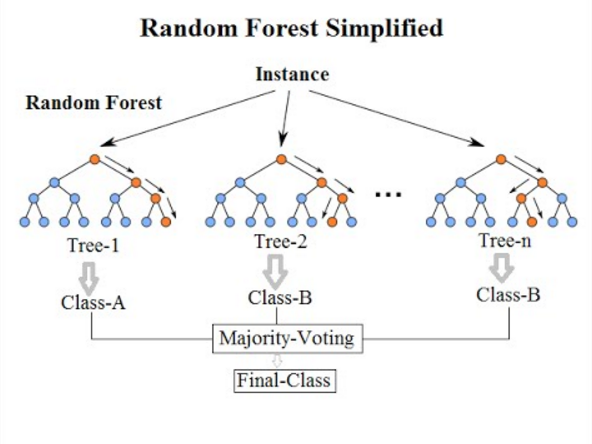
The same random forest algorithm or the random forest classifier can use for both classification and the regression task.

* Random forest classifier will handle the missing values.
* A random forest classifier won "t overfit the model when we have more trees in the forest.

**Disadvantages**:

* Quite slow to create predictions once trained. More accurate ensembles require more trees, which means using the modem becomes slower.
* Results of learning are incomprehensible. Compared to a single decision tree, or a set of rules, they don't give you much insight.

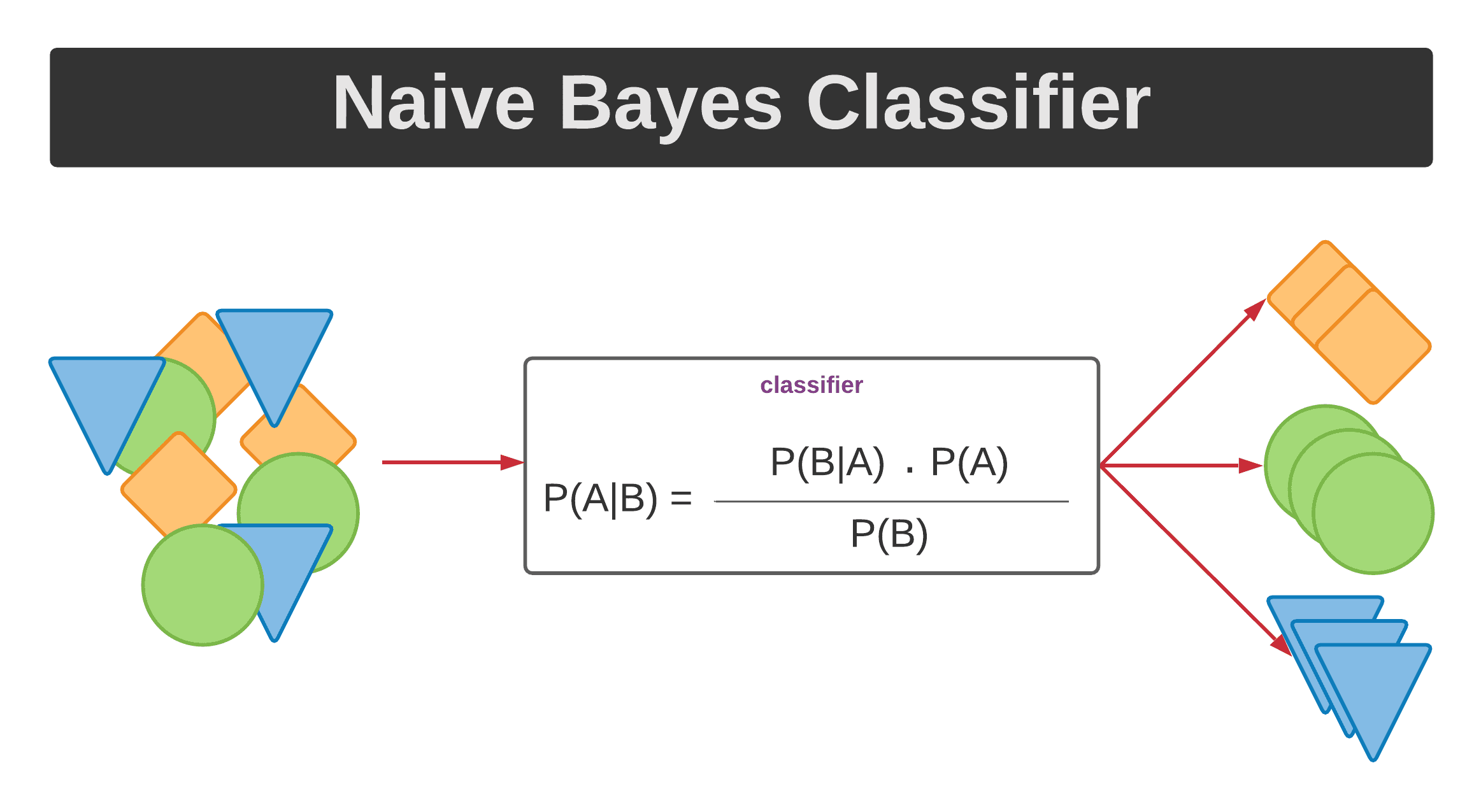
**Fig 2.0 Random Forest Algorithm**



**NAÏVE BAYES CLASSIFIER**

This classifier has received a great deal of interest since it is an extremely efficient probabilistic representation. Using the training data, this classifier determines the probability of each attribute Ai given the class label C. The class with the highest posterior probability is predicted after assessing the likelihood of C given specific instances of A1 according to the Bayes rule. The goal of classification is to use a set of predictors, or attributes, to reliably predict the value of a discrete class variable. Particularly, the Naive Bayes classifier is a Bayesian network where the class has no parents, and each point just has the course as a parent. Despite its apparent lack of complexity, the naive Bayesian (NB) algorithm has proven to be superior to more established approaches on a wide variety of real-world datasets. This is because NB learns linearly by combining classifier predictions using ensemble mechanisms like bagging and boosting, and it outperforms them. Predictive accuracy, however, decreases when features that are redundant or not typically distributed are considered.

**Fig 3.0 Implementing Naïve Bayes Classification Using Python**

To begin with, the Naive Bayes classifier is a controlled operation. Bayes' theorem provides a simple framework for a sorting algorithm. Strong (Naive) freedom to ascribe indicates this. The Bayes theorem is a tool for determining the likelihood of events. Each predictor exists in isolation from every other predictor. Every one of the characteristics increases the possibility of optimizing it all by itself. To put it another way, it will struggle when using the Naive Bayes model but will work fine when applying Bayesian techniques. In many intricate real-world situations, Naive Bayes classifiers are put to good use.

P(X/Y) is the likelihood function, P(X) is the class prior probability, P(Y) is the predictor prior probability, and P(Y/X) is the probability. Naive Bayes is a non-linear, sophisticated approach for classifying data that is intuitive, simple to configure, and potent. However, there is a lack of accuracy because it is built on inference and class conditional freedom.

**Advantages:**

1) Easy to implement.

2) Requires a small amount of training data to estimate the parameters. 3) Good results were obtained in most of the cases

**Disadvantages:**

1) Assumption: class conditional independence, therefore loss of accuracy.

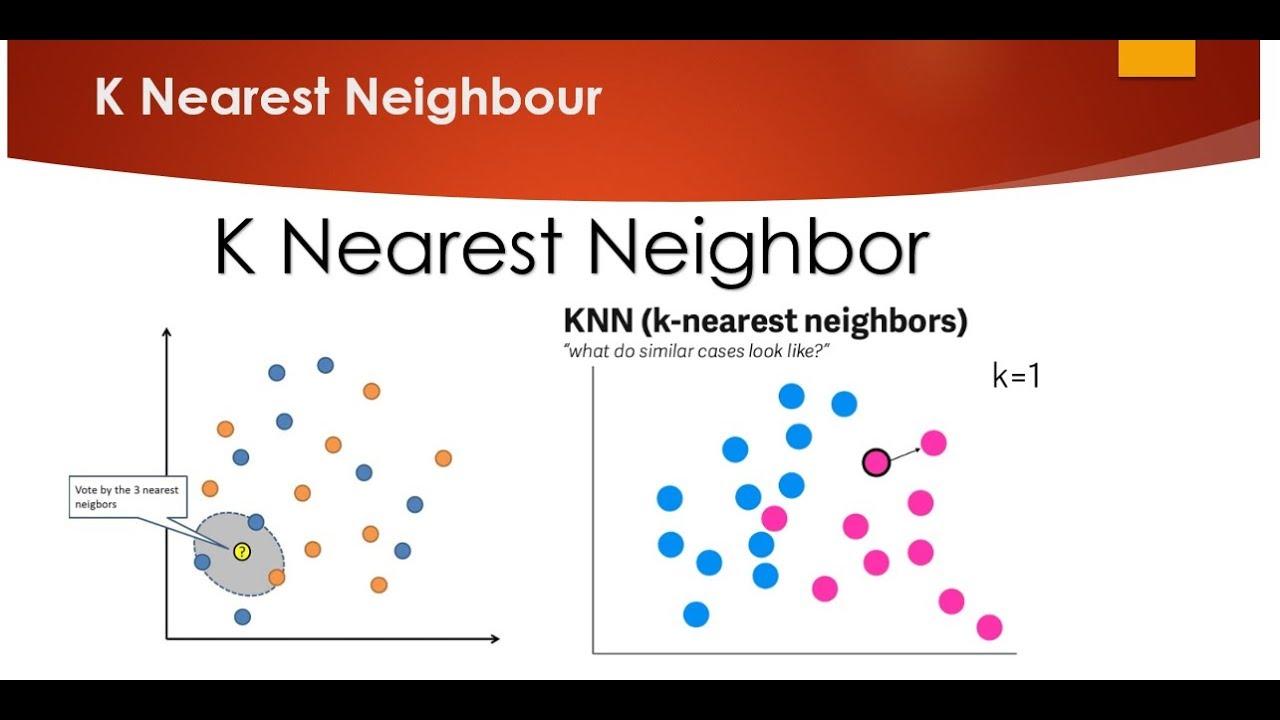
2) Practically, dependencies exist among variables.

3) Naïve Bayesian Classifier cannot model dependencies among these.

**K-NEAREST NEIGHBOUR**

The K-nearest neighbors' strategy is a supervised classification technique. It sorts things according to their closeness to one another. This is an illustration of case-based learning. The Euclidean distance is utilized to evaluate the separation between a characteristic and its neighbors. It uses a collection of named points to decide how to identify a new point. The findings are classified into 9 categories based on their commonalities, and K-NN may be used to fill in missing data values. Once the missing values have been filled in, a number of prediction algorithms are used to the data set. Using several rounds of these algorithms, accuracy can be enhanced. The K-NN technique is simple to build and does not require a plan or other assumptions. This method has applications in sorting, regression, and searching. Although K-NN is the simplest method, it is influenced by irrelevant and noisy functions.

**Figure 4.0 K-Nearest Neighbour Easily Explained With Implementation**

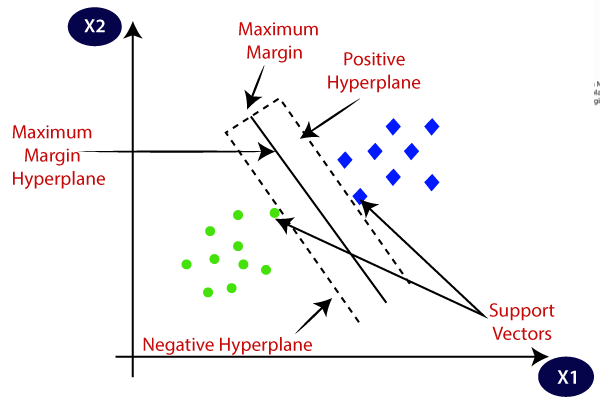


**SUPPORT VECTOR MACHINE**

The family of support vector machines includes both non-linear and linear variants. A supervised classifier, or support vector machine (SVM) is one that is given training data from a human. SVMs often employ a training and testing dataset. In a perfect society, divisions between social classes would be as simple as drawing a line. In such a circumstance, there exists a clear demarcation between the two categories. In this case, nevertheless, the data set can be visually divided into several distinct parts by means of several bars. The "dividing line" can be any one of these options. For the best line, we look over the training set and find the points that are closest to each category. This problem can be simplified by solving its equivalent minimization problem. Support vectors are the points of data that lie along the maximal margin lines. In non-normally distributed datasets, a line or higher-order function cannot be used to partition classes. The data in real-world datasets is typically flawed and noisy.

Unfortunately, it's not hard to build a model that over-fits the training data. When a model does not adequately capture noise and random fluctuations, over fitting occurs. Thus, the model is not universal and yields greater error rates in general, regardless of the database used. A simple model is less likely to be overfit. As the complexity of a model grows, it must be flexible enough to be applied in a wide range of situations while maintaining its essential simplicity. One strategy for achieving this goal is to facilitate model error. To prevent over-fitting, a support vector machine can produce mistakes. The likelihood of mistakes is hoped to be reduced. Support vector machine (SVM) based classifiers are popular. They've recently been a major focus of academic inquiry. Its widespread acceptance is attributable to its high empirical success rate. SVM classifiers have surpassed naive Bayes in popularity in recent years.

**Figure 5.0 Support Vector Machine Algorithm**



**DATA DESCRPTION**

The dataset used for this research purpose was the Public Health Dataset and it is dating from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The “target” field refers to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease. The first four rows and all the dataset features are shown in Table 1 without any preprocessing. Now the attributes which are used in this research purpose are described as follows and for what they are used or resemble:

* Age—age of patient in years,
* sex—(1 = male; 0 = female).
* Cp—chest pain type.
* Trestbps—resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if you have a normal blood pressure reading, it is fine, but if it is a little higher than it should be, you should try to lower it. Make healthy changes to your lifestyle).
* Chol—serum cholesterol shows the number of triglycerides present. Triglycerides are another lipid that can be measured in the blood. It should be less than 170 mg/dL (may differ in different Labs).
* Fbs—fasting blood sugar larger than 120 mg/dl (1 true). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
* Restecg—resting electrocardiographic results.
* Thalach—maximum heart rate achieved. The maximum heart rate is 220 minus your age.
* Exang—exercise-induced angina (1 yes). Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease.
* Oldpeak—ST depression induced by exercise relative to rest.
* Slope—the slope of the peak exercise ST segment**.**
* Ca—number of major vessels (0–3) colored by fluoroscopy.
* Thal—no explanation provided, but probably thalassemia (3 normal; 6 fixed defects; 7 reversible defects).
* Target (T)—no disease = 0 and disease = 1, (angiographic disease status).

**RESULTS**

In the end, an experiment was run to assess the performance of different classifiers in making predictions. Based on their qualitative performance, four popular classifiers were selected for the experiment. Furthermore, a dataset close to our hearts was selected from the machine learning repository at UCI. The best results can be obtained with a naive base classifier. The classification performance of four machine learning algorithms is compared by applying classifiers to the same data and analyzing the rates of misclassification and correct classification. Naive base classifier is superior to SVM, DT, and KNN in terms of accuracy. Their performance is very comparable, with just a slight difference, as shown by the quantitative data produced by computer simulations.

**ANALYSIS**

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.

**REFERENCES**

[1] Gaziano, T. A., Bitton, A., Anand, S., Abrahams-Gessel, S., & Murphy, A. (2010). Growing epidemic of coronary heart disease in low-and middle-income countries. *Current problems in cardiology*, *35*(2), 72-115.

[2] Iftikhar, S., Fatima, K., Rehman, A., Almazyad, A. S., & Saba, T. (2017). An evolution based hybrid approach for heart diseases classification and associated risk factors identification. *Biomedical Research*, *28*(8), 3451-3455.

[3] Jin, H., & Jung, Y. G. (2016). Performance Comparison of Decision Trees of J48 and Reduced-Error Pruning. *International journal of advanced smart convergence*, *5*(1), 30-33.

[4] Kumar, P. M., & Gandhi, U. D. (2018). A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. *Computers & Electrical Engineering*, *65*, 222-235.

[5] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

[6] Seckeler, M. D., & Hoke, T. R. (2011). The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. *Clinical epidemiology*, *3*, 67.

[7] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

[8] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[9] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[10] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[11] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072 9

[12] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[13] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[14] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Mutli-Conference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

[15] Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

[16] Soni Jyoti. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-8.

[16] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-60). IEEE.

[17] Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).

[18] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H (2016). Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92.

[19] Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device -kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.

[20] Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T (2011). U.S. Patent No. 8,014,863. Washington, DC: U.S. Patent and Trademark Office.

[21] Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.

[22] Buechler K F & McPherson P H (1999). U.S. Patent No. 5,947,124. Washington, DC: U.S. Patent and Trademark Office.

[23] Takci H (2018). Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences, 26(1), 1-10.

[24] Worthen W J, Evans S M, Winter S C & Balding D (2002). U.S. Patent No. 6,432, 124. Washington, DC: U.S. Patent and Trademark Office.

[25] Acharya U R, Fujita H, Oh S L, Hagiwara Y, Tan J H & Adam M (2017). Application of deep