Mini review Article

A model for Coronary Heart Disease Prediction using Data Mining Classification Techniques

ABSTRACT

Nowadays the guts malady is one amongst the foremost causes of death within the world. Thus it's early prediction and diagnosing is vital in medical field, which might facilitate in on time treatment, decreasing health prices and decreasing death caused by it. The treatment value the disease isn't cheap by most of the patients and Clinical choices are usually raised supported by doctors" intuition and skill instead of on the knowledge-rich information hidden within the stored data. The model for prediction of heart disease using a classification techniques in data mining reduce medical errors, decreases unwanted exercise variation, enhance patient well-being and improves patient results. The model has been developed to support decision making in heart disease prediction based on data mining techniques. The experiments were performed using the model, based on the three techniques, and their accuracy in prediction noted. The decision tree, naïve Bayes, KNN (K-Nearest Neighbors) and WEKA API (Waikato Environment for Knowledge Analysis-application programming interface) were the various data mining methods that were used. The model predicts the likelihood of getting a heart disease using more input medical attributes. 13 attributes that is: blood pressure, sex, age, cholesterol, blood sugar among other factors such as genetic factors, sedentary behavior, socio-economic status and race has been use **to predict the likelihood of** patient getting a Heart disease until now. This study research added two more attributes that is: Obesity and Smoking.740 Record sets with medical attributes was obtained from a publicly available database for heart disease from machine learning repository with the help of the datasets, and the patterns significant to the heart attack prediction was extracted and divided into two data sets, one was used for training which consisted of 296 records & another for testing consisted of 444 records, and the fraction of accuracy of every data mining classification that was applied was used as standard for performance measure. The performance was compared by calculating the confusion matrix that assists to find the precision recall and accuracy. High performance and accuracy was provided by the complete system model. Comparison between the proposed techniques and the existing one in the prediction capability was presented. The model system assists clinicians in survival rate prediction of an individual patient and future medication is planned for. Consequently, the families, relatives, and their patients can plan for treatment preferences and plan for their budget consequently.

Keywords: WEKA API; Decision Tree; Naïve Bayes; KNN, Cardiovascular disease, KDD.

1. INTRODUCTION

The Heart is a strong organ, situated close to the middle of the chest; it is duty is pumping blood to different parts of the body and together with system of vessels and blood from the human body's cardiovascular framework; Interferences to this dissemination of blood can result in serious medical issue including death [5]. People have been affected by dangerous sicknesses all through the past. The system for prediction can assist to lower the dangers of the disease. Prediction is done dependent on the present data fed to the framework model Using WEKA API which is open source information mining programming in Java. The model is being created dependent on three distinct information mining strategy that is Nave

Bayes, KNN, decision tree with WEKA API. The input dataset is analyzed using different classification algorithms and comparison is done for accuracy.

Nowadays an immense measure of information is gathered and kept in a daily basis. There is a significant need to break down this information yet with no scientific device, this appears to be unimaginable. This has prompted the improvement of Knowledge Discovery in Databases (KDD) which changes the low dimension information to a top state learning. KDD comprises of different procedures at various advances and Data mining is one of those procedures. Information mining is the way toward finding fascinating learning from huge measure of information kept in databases, information stockrooms or other data vaults. The fundamental point of information mining procedure is to separate data from a dataset and change it into a reasonable structure so as to help basic conclusions [45]. A tremendous measure of information is accessible in healthcare industry however the mining of this information is poor. In this way, the investigation of the medicinal services information is a must. Information Discovery in databases is getting to be famous research instrument for open human services information. In this study, we will do the exhibition investigation of various information mining grouping strategies on medicinal services information from the Cleveland, Hungary, Switzerland and the VA Long Beach Clinics Foundation, medical records department. This work will help discovering the best information mining arrangement method as far as precision on the specific dataset. The examined characterization systems are K-closest neighbor (KNN), Naive Bayes, Decision tree. The exhibition of these procedures is estimated dependent on their exactness. This investigation will assist the future scientists with getting proficient outcomes in the wake of realizing best information mining grouping method for specific dataset.

Information Mining is the nontrivial procedure of recognizing substantial, novel, conceivably valuable and at last reasonable example in information with the wide utilization of databases and the touchy development in their sizes. Information mining refers to removing or "mining" learning from a lot of information. Information digging is the quest for the connections and worldwide examples that exist in enormous databases however are tucked away among a lot of information [17]. The fundamental procedure of Knowledge Discovery is the change of information into learning so as to help in making judgments is known as information mining. Information Discovery procedure comprises of an iterative grouping of information cleaning, information coordination, information determination, information mining design acknowledgment and learning introduction. Information digging is the quest for the connections and worldwide examples that exist in enormous databases bramble are tucked away among a lot of information.

Many hospitals have put in databases systems to manage their clinical data or patient data. These data systems generally generate giant amounts of information which may be in any format like numbers, text, charts and pictures however sadly, this info that contains made information isn't used for clinical deciding. There’s abundant data keep in repositories that may be used effectively to support deciding in attention. Data processing techniques is wide utilized in medical field for extracting information from info. In data processing call tree may be a technique that is employed extensively. Call trees are non-parametric supervised learning technique used for classification.

The most aim is to form a model that predicts the worth of a target variable by learning straightforward call rules inferred from the info options. The structure of the choice tree is within the type of tree and leaf nodes. Decision trees are most typically utilized in research, principally in call analysis. Blessings are that they're straightforward to know and interpret. They’re strong, performed well with giant datasets, able to handle each of the numerical and categorical information.

By providing economical treatments, it will facilitate to scale back prices of treatment. Mistreatment data processing techniques it takes less time for the prediction of the un-wellness with a lot of accuracy. The most necessary step a company will absorb terms of information mining is to require advantage of the opportunities afforded by it. Collect information and place it to smart use with data processing, and you’ll before long begin reaping the advantages that’s ; more cash by Learning that varieties of merchandise customers have purchased and maximize that insight to individualize expertise, increase client loyalty, and boost client time period price. Improve stigmatization and promoting through Get feedback and use data processing to spot what's operating and what isn't with branding and marketing. contour reach by creating all of your outreach a lot of timely and relevant with data processing, faucet into new markets by Use different databases to spot potential customers and conduct relevant reach, Learn from the past by comparison current information to past data to search out trends to stay in mind once creating business choices.

Data mining has become more and more necessary, particularly in recent years, once nearly all industries and sectors everywhere the planet face issues on information explosion. All of unforeseen, there's just too abundant information, and this fast rise within the quantity of information demands a corresponding increase in the amount of knowledge and knowledge. Thus, there's a requirement to quickly, expeditiously and effectively method all that information into usable data and data processing offers the answer. In fact, you'll say that data processing is that the resolution. You'll realize data processing to be most frequently used or applied in organizations or businesses that maintain fairly giant to large databases. The sheer size of their databases and also the quantity of knowledge contained among them need over a little live of organization and analysis that is wherever data processing comes in. Through data processing, users are able to investigate information from multiple views in their analysis. It'll additionally build it easier to categorize the knowledge processed and establish relevant patterns, relationships or correlations among the assorted fields of the data. Therefore, we are able to deduce that data processing involves tasks of a descriptive and prognosticative nature. Descriptive, as a result of it involves the identification of patterns, relationships and correlations among giant amounts of information, and prognosticative, as a result of its application utilizes variables that are accustomed predict their future or unknown values. The use of information mining (DM) model allows machine intelligence in nosology processes.

DM is that the machine intelligence-based process of extracting important data from the set of huge quantity of information. DM may be a speedily growing field in a very big selection of health science applications. Applicable DM-based classification techniques and sensible cardiovascular disease prediction systems will lead toward quality health care in terms of accuracy and low economical health care services. The most motivation behind digitization of health information and utilization of sentimental computing tools is to lower the value of health care and cut back the quantity of preventable errors. Among numerous DM techniques, like agglomeration, association rule classification and regression, the classification is one among the foremost necessary techniques used for categorization of information patterns. In DM, essentially the classification-based machine learning algorithms are accustomed predict membership perform for labeling CVD information instances. Classification will be information analysis technique that extracts labels describing necessary data categories. The classifier's model is portrayed as classification rules, call trees or mathematical formulae, and it's termed as supervised learning. The model is employed for classifying future or unknown objects. The classification algorithmic program predicts un-wellness categorical class (eg, negative and positive) and build classifier model supported the coaching set. If the accuracy of the model is suitable, the model may be applied to categorize information tuples whose class labels are unknown. The classification contains 2 basic steps of learning and classification. In learning, coaching information is analyzed by classification algorithmic program and classifier's model is made. Within the classification section, check information are utilized to estimate the accuracy of the classification model. A healthy range of researchers are applying numerous algorithms and techniques like classification, clustering, multivariate analysis, artificial neural networks (ANNs), call trees, genetic algorithmic program (GA), KNN strategies, single DM model and hybrid and ensemble approaches to help health care professionals with improved accuracy within the identification of cardiovascular disease. During this study, the analysis quest of however the burden of artery un-wellness may be considerably reduced through soft machine strategies is explored. The final drawback statement of this study is to develop approach-based classifier's model that may be applied to CVD information sets to boost model prediction's outcomes for higher prediction accuracy and responsibility. Additionally to the current, the study presents example of intelligent cardiovascular disease prediction system supported associate degree approach with totally different classifiers, namely, Naïve theorem and KNN. The planned prediction system is computer program primarily based, having the power of scaling and enlargement as per user's additional demand.

The figure beneath illustrates Steps of the Knowledge Discovery in Databases process on the most proficient method to separate learning from information with regards to enormous databases Fayyad et.al [14].

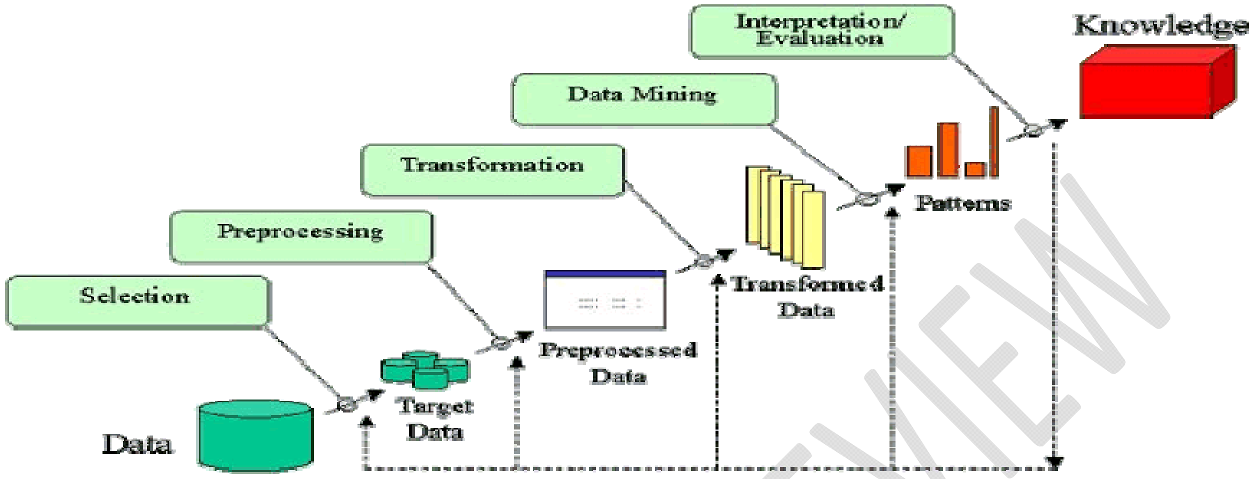


Figure1.0 Steps of Knowledge Discovery in Databases process by Fayyad et.al [14]

Various health industry information systems are structured to help patient charging, stock organization and making some simple calculation. A couple of health sectors utilize decision model systems yet are, as it were, limited. They can address simple inquiries like "What is the ordinary time of patients who have coronary disease? “What number of therapeutic techniques had achieved crisis facility stays longer than 10 days?", "Recognize the female patients who are single, more than 30 years old, and who have been treated for coronary sickness." However they can't respond to complex inquiries like "Given patient records, foresee the probability of patients getting a coronary disease." Clinical decisions are as often as possible made subject to experts' impulse and experience rather than on the learning rich data concealed in the database.

This preparation prompts bothersome tendencies, botches and super helpful costs which impacts the idea of care provided for patients. The proposed structure that coordinates the clinical decision help with PC based patient records could reduce therapeutic errors, overhaul tolerant security, decrease bothersome practice assortment, and improve getting result. This suggestion is promising as data modeling and analysis tool like data mining can make a learning rich condition which can help to in a general sense improve the idea of clinical decisions.

In this fast moving world people need to continue with an extravagant life so they work like a machine to win some portion of money and continue with a pleasant life appropriately in this race they disregard to manage themselves, because of this there sustenance affinities change in their entire lifestyle change, in this sort of lifestyle they are logically stressed they have heartbeat, sugar at a young age and they don't give enough rest for themselves and eat what they get and they even don't overemphasize the idea of the sustenance whenever cleared out the go for their own special prescription in light of all these little indiscretion it prompts a significant threat that is the coronary disease [7]. On account of this people go to therapeutic administrations experts but the prediction made by them is not 100% definite [25].

Quality facility proposes diagnosing patients precisely and controlling medications that are convincing. Poor clinical decisions can incite tragic outcomes which are along these lines unsatisfactory. Medicinal centers ought to in like manner limit the cost of clinical tests. They can achieve these results by using fitting PC based information or decision support system.

The treatment cost of heart disease is not affordable by most of the patients, and the Clinical decisions are often made based on doctors’ intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed model for Heart Disease Prediction using Data Mining Classification Techniques reduces medical errors, enhances patient safety, decrease unwanted practice variation, reduce treatment cost and improves patient outcome. This suggestion is promising as data modeling and analysis tools have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions [32].

2. LITERATURE REVIEW

This part goes for investigating the different information mining methods presented as of late for coronary illness expectation. The man-made brainpower methods centering K-closest neighbor (KNN), Naive Bayes and Decision tree will be presented. Recently distributed papers in displaying survival will be talked about and the recommendations for another strategy are introduced

2.1 Theoretical and Empirical Review

Various information mining systems have been utilized in the analysis of cardiovascular disease (CVD) over various Heart illness datasets. A few papers utilize just a single method for conclusion of coronary illness and different scientists utilize more than one information mining technique for the finding of coronary illness.

In [23,27] Jyoti et.al presented three classifiers Decision Tree, Naïve Bayes and Classification by methods for gathering to break down the proximity of coronary sickness in patients. Request by methods for bundling: Clustering is the route toward social occasion relative segments. This framework may be used as a preprocessing adventure before urging the data to the portraying model. Preliminaries were driven with WEKA 3.6.0 gadget Enlightening list of 909 records with 13 particular properties. All properties were made supreme and anomalies were made due with straightforwardness. To update the desire for classifiers, innate request was joined. Observations show that the Decision Tree data mining technique beats other two data mining methods in the wake of intertwining feature subset assurance yet with high model improvement time.

[27] Nidhi et.al discernments revealed that the Neural Networks with 15 characteristics improved in examination with other data mining frameworks [27]. The investigation concentrate assumed that Decision Tree technique showed better execution with the help of innate figuring's using included subset assurance. This examination work furthermore proposed a model of Intelligent Heart Disease Prediction structure using data mining frameworks explicitly Decision Tree, Naïve Bayes and Neural Network. An aggregate of 909 records were obtained from the Cleveland Heart Disease database. The results declared in the investigation work guarded the better execution of Decision Tree methodology with 99.6% accuracy using 15 qualities. In any case, Decision tree technique in mix with inherited estimation the introduction declared was 99.2% using 06 qualities.

In [8,9] Chaitrali et.al exhibited that Artificial Neural Network outmaneuvers other data mining methodology, for instance, Decision Tree and Naïve Bayes. In this investigation work, Heart disorder desire system was made using 15 characteristics [8,9]. The investigation work included two extra properties weight and smoking for capable finish of coronary sickness in making convincing coronary disease desire system.

[31] Researchers in year 2013 showed Hybrid Intelligent Techniques for the figure of coronary ailment. Some Heart Disease gathering system was researched in this examination and shut with legitimization noteworthiness of data mining in coronary sickness end and course of action. Neural Network with separated getting ready is helpful for sickness conjecture in starting time and the extraordinary execution of the structure can be gotten by preprocessed and institutionalized dataset. The game plan precision can be improved by decline in features.

[47] Vikas et.al, in their investigation work used three standard data mining figuring's CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) removed from a decision tree or rule based classifier to develop the conjecture models using a greater dataset. Discernment showed that presentation of CART computation was better when differentiated and other two course of action procedures.

V. Manikandan et.al in [46] recommended that association standard mining is used to remove the thing set relations. The data game plan relied upon MAFIA counts which achieved better precision. The data was surveyed using entropy based cross endorsement and bundle strategies and the results were considered. MAFIA (Maximal Frequent Item set Algorithm) used a dataset with 19 characteristics and the goal of the examination work was to have exceedingly definite audit estimations with bigger measures of precision.

Beant et.al in [6] circulated an investigation paper in IJRITCC "Review on Heart Disease using Data Mining Techniques". The maker referenced created by gigantic number of experts and investigated diverse data mining strategies reliant on execution and accuracy.

Methaila et.al [3] in their examination work focused on using different counts and mixes of a couple of target qualities for amazing heart ambush figure using data mining. Decision Tree has beated with 99.62% precision by using 15 characteristics. Moreover the exactness of the Decision Tree and Bayesian Classification further improves in the wake of applying inherited computation to diminish the genuine data size to get the perfect subset of value satisfactory for coronary disease estimate.

The experts [19] proposed a model for desire for coronary ailment using J48, Bayes Net, and Naïve Bayes, Simple CART and REPTREE Algorithms using understanding educational accumulation from Medical Practitioners.

Appraisal of the disorder matrix showed that J48, REPTREE and SIMPLE CART exhibit a figure model of 89 cases with a peril factor positive for heart attacks. The strategies immovably prescribed that data mining counts can foresee a class for judgments.

B.Venkatalakshmi et.al [5] played out an examination on coronary disease finding using data mining methodology Naïve bayes and Decision Tree techniques. Different sessions of examinations were coordinated with the proportional datasets in WEKA 3.6.0 contraption. Instructive gathering of 294 records with 13 attributes was used and the results revealed that the Naïve Bayes beat the Decision tree frameworks.

The synopsis of looked into writing alongside the quantity of properties utilized for the forecast of Cardiovascular Disease (CVD) is given in table beneath

Table 1.0: Table shows different data mining techniques used in the diagnosis of Heart disease.

Author/Researcher	Data Mining Technique used	Year	Number of Attributes Selected
Jyoti Sonia, et.al.	Naïve Bayes, Decision Tree, KNN	2011	13
K.Srinivas et.al.	Naïve Bayes, knn and D.L.	2011	14

Nidhi Bhatla et.al.	Naïve Bayes, Decision Tree, Neural Network	2012	15 and 13
Chaitrali S.Dangare & Sulabha S.Apte	Naïve Bayes, Decision Tree, Neural Network	2012	13 and 15
Abhishek Taneja	Naïve Bayes,J48 unpruned tree, Neural Network	2013	15 and 8
R. Chitra et. al.	Hybrid Intelligent Techniques	2013	15
Vikas Chaurasia, et.al.	CART,ID3,Decision Table	2013	Not mentioned
V. Manikandan et al.	K-Mean based on MAFIA, K-Mean based on MAFIA with ID3, K-Mean based on MAFIA with ID3 and C4.5	2013	19
Beant Kaur & Williamjeet	Papers Reviewed	2014	Nil
Aditya Methaila et. al.	Decision Tree, Naive Bayes, Neural Network ,Genetic Algorithm	2014	15 and 16
Hlaudi Daniel Masethe, Mosima Anna Masethe	J48,REPTREE,Naïve Bayes, Bayes net, Simple CART	2014	15
B.Venkatalakshmi and M.V Shivsankar	Decision Tree and Naïve Bayes	2014	13

2.2 Artificial Intelligence Techniques in Heart Disease Prediction

Information mining has been generally connected in the therapeutic field as this give enormous measure of information. Different scientists had connected the various information mining procedures on social insurance information [11]. connected 5 arrangement calculations for example choice tree, fake neural system, strategic relapse, Bayesian systems and credulous Bayes and stacking-sacking technique for structure arrangement models and thought about the precision of the plain and outfit model to foresee whether a patient will return to a medicinal services Center or not. From results, the best order model relies upon informational collection for example ANN (Artificial neural networks) in 3M informational index, choice tree in 6M and strategic relapse in 12M informational collection [23, 26] contrasted the information mining and conventional insights and expresses a few focal points of mechanized information framework. This paper gives an outline of how information mining is utilized in social insurance and medication. Patil Dipti [29] decides if an individual is fit or unfit dependent on authentic and constant information utilizing grouping calculations that is K-means and D-stream are connected. The presentation and precision of D-stream calculation is more than K-implies [4] utilized choice tree to construct an arrangement model for anticipating representative's exhibition. To manufacture a characterization model CRISP-DM was received.

PC reproduction demonstrates that the strategic relapse, neural system model and troupe model delivered best generally speaking grouping precision. Koç et al [24] connected ANN and strategic relapse to anticipate if the customer will buy in a term store or not subsequent to promoting effort. ANN orders 84.4% information accurately while calculated relapse characterizes 83.63% information effectively however LR takes 54 seconds and ANN takes 11 seconds to run. Along these lines, with more information and higher dimensional element space, utilizing ANN will be progressively productive. Fartash et.al [13] contrasted the different order calculations with anticipate the transmission capacity use design in various time interims among various gatherings of clients in the system correlation of various characterization calculations including. Choice Tree and Naïve Bayesian utilizing Orange is finished. The Decision Tree calculation accomplished 97% exactness and effectiveness in foreseeing the required data transfer capacity inside the system. Sakshi and Prof. Sunil Khare [35] gave a total examination of various information mining characterization procedures that incorporates choice tree, Bayesian systems, k-closest neighbor classifier and fake neural system.

Clinical databases have gathered enormous amounts of data about patients and their ailments. The term Heart illness includes the assorted sicknesses that influence the heart. Coronary illness is the real reason for setbacks on the planet. The term Heart illness includes the assorted ailments that influence the heart. Coronary illness kills one individual at regular intervals in the United States [48]

2.3 Data Mining Review

Notwithstanding the way that data burrowing has been around for more than two decades, its potential is simply being recognized now. Data mining solidifies quantifiable examination, AI and database advancement to think hid models and associations from gigantic databases Fayyad portrays data mining as "a method of nontrivial extraction of saw, in advance darken and possibly profitable information from the data set away in a database" [44] describes it as "a method of assurance, examination and showing of colossal measures of data to discover regularities or relations that are at first cloud with the purpose of getting clear and accommodating results for the owner of database" [17]

Data mining uses two systems: oversaw and unsupervised learning. In oversaw learning, a planning set is used to learn model parameters however in unsupervised adjusting no arrangement set is used (e.g., k means grouping is unsupervised) [28]. Each datum mining methodology fills another need dependent upon the exhibiting objective. The two most ordinary showing goals are gathering and figure. Game plan models predict full scale names (discrete, unordered) while estimate models envision steady regarded limits Decision Trees and Neural Networks use portrayal counts while Regression, Association Rules and Clustering use desire figuring's [10].Decision Tree figuring's consolidate CART (Classification and Regression Tree), ID3 (Iterative Dichotomized [10] and C4.5. These computations shift in selection of parts, when to keep a center point from part, and undertaking of class to a non-split center [11] CART uses Gini rundown to check the dirtying impact of a package or set of getting ready tuples [17].It can manage high dimensional unmitigated data.

Decision Trees can moreover manage constant data (as in backslide) yet they ought to be changed over to straight out data. Gullible Bayes or Bayes' Rule is the explanation behind a few, AI and data mining methods [42] .The standard (estimation) is used to make models with insightful capacities. It gives better methodologies for researching and getting data. It gains from the "evidence" by figuring the association between the goal (i.e., subordinate) and other (i.e., independent components. Neural Networks includes three layers: input, concealed and yield units (factors). Relationship between data units and concealed and yield units rely upon centrality of the doled out worth (weight) of that particular data unit. The higher the weight the more huge it is. Neural Network computations use Linear and Sigmoid trade limits. Neural Networks are sensible for setting up a ton of data with few wellsprings of information. It is used when various systems are unacceptable.

3. RESEARCH DESIGN

Methodology provides a framework for endeavor the projected DM modeling. The methodology may be a system comprising steps that remodel information into recognized data patterns to extract information for users. The DM methodology framework breaks down the mining method of vast knowledge into phases. It shows associate degree unvaried DM method for implementing machine learning strategies on the vast knowledge set taken for application. The projected methodology includes steps, stated because the preprocessing stage wherever the thoroughgoing exploration of the information is disbursed. It'll account for handling missing values, equalization knowledge and normalizing attributes counting on algorithms used. Once pre-processing of information is performed, prognostic modeling of the information is disbursed victimization classification models and ensemble approach. Finally, prescriptive modeling is undertaken, wherever the prognostic model is evaluated in terms of performance and accuracy victimization varied performance metrics. The figure below shows a framework break down of the unvaried data mining process of vast knowledge into phases

The advantage of this methodology of use is that: it provides High performance by an entire system model as compared to different techniques, Additional features and functions can be easily added even as late as the testing phase, offers a transparent and concrete approach and it's straightforward to use and access

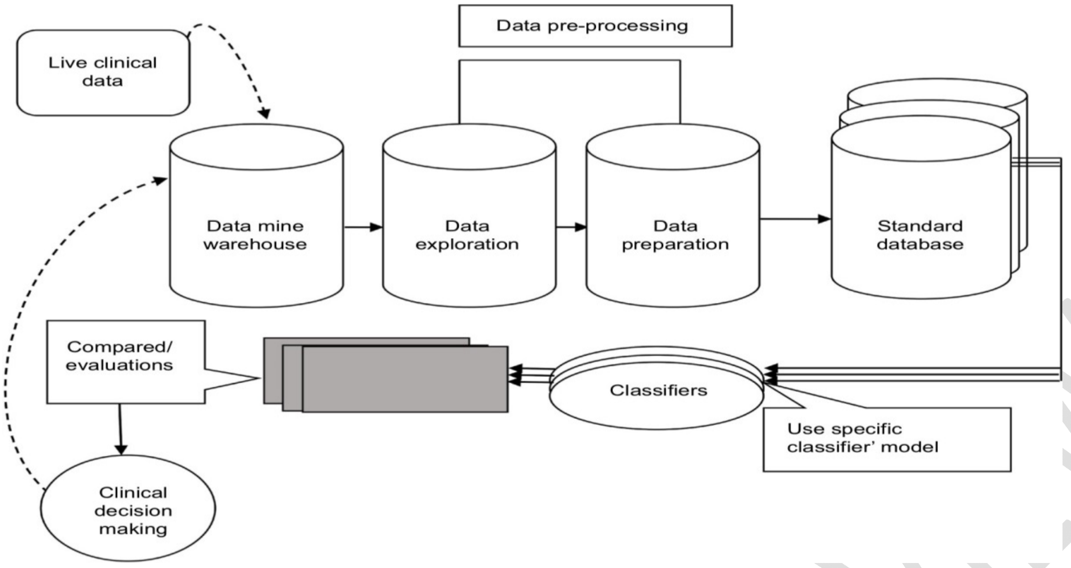


Figure 2.0: Methodology for mining heart disease data.

In this examination, three information digging procedures for prescient information mining assignment were utilized, that incorporates Decision tree, K-NN and Naïve Bayes. These strategies were utilized for producing learning to settle on it valuable for basic leadership. Every strategy delivered various outcomes to arrange the locale into centered or non-centered states involving the accessible factors in dataset .The experimentation was performed utilizing WEKA programming interface.

3.1. Proposed Model

The proposed engineering of coronary illness forecast framework is given below. The figure illustrates the frame work of the coronary heart disease prediction model steps activities.

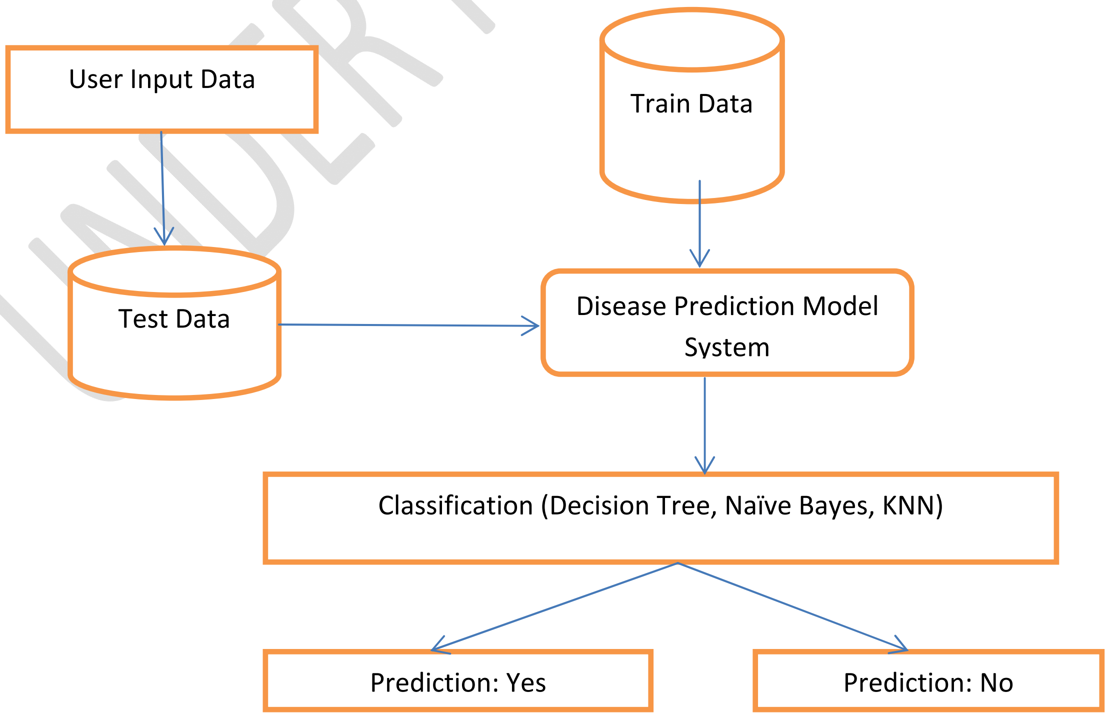

User Input Data
Train Data
Test Data
Disease Prediction Model System
Classification (Decision Tree, Naïve Bayes, KNN)
Prediction: Yes
Prediction: No

Figure 3.0: The System Model

It comprises of preparing dataset and client contribution as the test dataset. Weka information mining apparatus with programming interface was utilized to actualize the coronary illness forecast framework. The source code of Weka is in java. The framework is planned with java swing and use Weka programming interface to call the various techniques for Weka. The segments utilized are cases, various classifiers and strategies for assessment. Administered learning strategy is utilized here. A directed learning calculation examinations the preparation information and derives a capacity from the named preparing set. It tends to be utilized for mapping new models. The preparation information got from ucl repository coronary illness database is the preparation model. This preparation information comprise of the class name and its comparing esteem. Credulous Bayes, KNN and choice tree classifiers are administered learning calculations. They gain from the given preparing models. At the point when another case with same characteristics as in preparing information with various qualities other than those in the preparation model comes, these calculations effectively characterize the new case dependent on the speculation made from the preparation set. Gullible Bayes, KNN and choice tree classifiers are order the new perception into two classifications based on preparing dataset. The preparation dataset is in the ARFF group. The preparation set comprises of 296 traits including the class characteristic. Coronary illness forecast framework acknowledges contribution from the client through a graphical UI. Every one of the traits required for grouping is gotten from a content field. The graphical UI is fabricated utilizing swing. The following procedure is to move the client information acquired from graphical UI into a record of CSV (Comma isolated Value) augmentation. At that point the CSV record is changed over into ARFF document. Weka programming interface give local strategies to changing over from CSV to ARFF. The changed over client information is treated as test information. The test informational index will contain every one of the characteristics of preparing dataset. In the event that the client did not enter a property estimation a '?' will be relegated at the estimation of that comparing trait. Weka will deal with this missing worth. This test information is kept running on Naive Bayes, KNN and choice tree calculation. These calculations order the occasions got from the client and foresee the opportunity to have coronary illness. Net beans IDE is utilized to code in Java.

3.1.1 Decision Tree

A call tree could be a decision support tool that uses a tree-like model of selections and their doable consequences, as well as happening outcomes, resource prices, and utility. It’s a way to show AN algorithmic program that solely contains conditional management statements. Decisions trees are ordinarily utilized in research, specifically in call analysis, to assist determine a technique possibly to succeed in a goal, however also are preferred tools in machine learning. Classification is that the method of building a model of categories from a collection of records that contains category labels. Decision Tree algorithmic program is to seek out the method the attributes-vector behaves for variety of instances. Additionally on the bases of the coaching instances the categories for the freshly generated instances are being found. This algorithmic program generates the principles for the prediction of the target variable. With the assistance of tree classification algorithmic program the vital distribution of the information is well comprehendible [50]. J48 is AN extension of ID3. The extra options of J48 are accounting for missing values, call trees pruning, continuous attribute worth ranges, derivation of rules, etc. within the wood hen data processing tool, J48 is AN open supply Java implementation of the C4.5 algorithmic program. The wood hen tool provides variety of choices related to tree pruning. Just in case of potential over fitting pruning is used as a tool for précising. In different algorithms the classification is performed recursively until each single leaf is pure, that's the classification of the information ought to be as excellent as doable. This algorithmic program it generates the principles from that specific identity of that knowledge is generated. The target is more and more generalization of a call tree till it gains equilibrium of flexibility and accuracy.

Advantages of J48

a. Whereas building a tree, J48 ignores the missing values i.e. the worth for that item are often foretold supported what's better-known regarding the attribute values for the opposite records.

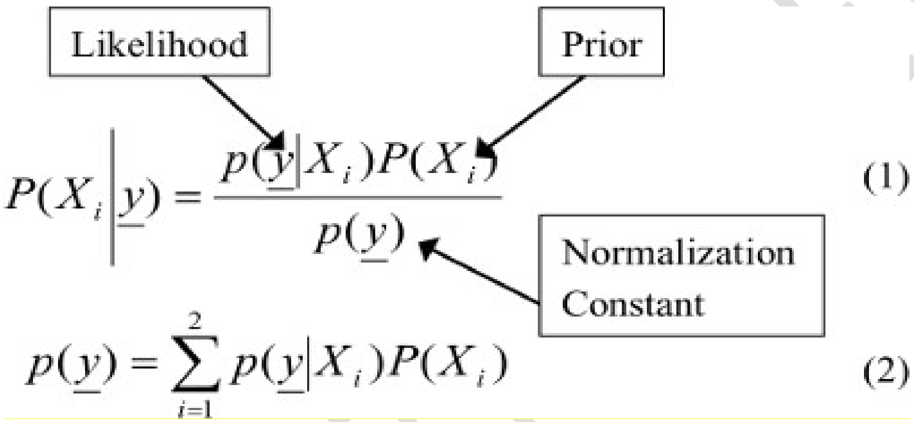
b. Just in case of potential over fitting pruning are often used as a tool for précising.

3.1.2 Naïve Bayes

This technique depends on probabilistic information. The gullible Bayes principle yields probabilities for the anticipated class of every individual from the arrangement of test example. Gullible Bayes depends on administered learning. The objective is to foresee the class of the experiments with class data that is given in the preparation information.

The quality "Analysis" is distinguished as the anticipated characteristic with worth "1" for patients with coronary illness and worth "0" for patients with no coronary illness. "Quiet Id" is utilized as the key; the rest are info traits. It is expected that issues, for example, missing information, conflicting information, and copy information have all been settled.

It is a probabilistic classifier supported Bayes' theorem such by the previous possibilities of its root nodes. The mathematician theorem is given in Equation one and social control constant is given in Equation a pair of. It proves to be associate best formula in terms of diminution of generalized error. It will handle statistical-based machine learning for feature vectors Y= [Y1, Y2….Yl]T and assign the label for feature vector supported supreme probable among on the market categories . It means feature "y" belongs to Xi category, once posterior likelihood P(X1/Y) is most i.e Y=X,: P(X1/Y)Max. The Bayesian classification downside is also developed by a-posterior possibilities that assign the category label ωi to sample X specified P(X1/Y) is supreme. The Bayesian classification downside is also developed by a-posterior possibilities that assign the category label ωi to sample X specified P(X1/Y) is supreme.



$$P(X_i|\underline{y}) = \frac{p(\underline{y}|X_i)P(X_i)}{p(\underline{y})} \qquad (1)$$

$$p(\underline{y}) = \sum_{i=1}^{2} p(\underline{y}|X_i)P(X_i) \qquad (2)$$

Application of Bayes' rule with the mutual exclusivity in diseases and also the conditional independence in findings is understood because of Naïve theorem Approach. It's a probabilistic classifier supported Bayes' theorem with robust independence assumptions between the options. Naïve theorem classifier despite its simplicity, it astonishingly performs well and infrequently outperforms in advanced classification. Straightforward Naïve theorem will be enforced by plugging within the following main Bayes formula

P(X1,X2,…,Xn|Y)=P(X1|Y)P(X2|Y)…P(Xn|Y) (3)

The abovementioned Naïve theorem network produces a mathematical model, that is employed for modeling the sophisticated relations of random variables of un-wellness attributes and call outcome. The formula uses the formula to calculate chance with regard to un-wellness condition attributes worth and call attribute value supported by previous information, the formula classifies the choice attribute into labels allotted, and thus the conditional support is computed for every variable attribute [51].

The Advantage of this formula is, it needs solely a tiny low quantity of coaching information for estimating the parameters essential for classification, straightforward to implement and sensible results obtained in most of the cases

Implementation of Bayesian Classification

The Naïve Bayes Classifier strategy is especially fit when the dimensionality of the sources of info is high. In spite of its effortlessness, Naive Bayes can frequently outflank increasingly advanced grouping techniques. Gullible Bayes model recognizes the attributes of patients with coronary illness. It demonstrates the likelihood of each information trait for the anticipated state.

Why favored Naive Bayes calculation

Credulous Bayes or Bayes' Rule is the reason for some, AI and information mining techniques. The standard (calculation) is utilized to make models with prescient abilities. It gives better approaches for investigating and getting information.

Why preferred naive Bayes implementation:

a. At the point when the information is high.
b. At the point when the properties are free of one another.
c. When we need increasingly proficient yield, when contrasted with different strategies yield

Bayes Rule

A restrictive likelihood is the probability of some end, C, given some proof/perception, E, where a reliance relationship exists among C and E.

This likelihood is meant as P(C |E) where

$$P\ (C/E) = P\ (E/C)\ P\ (C)/p\ (E)$$

3.1.3 K-NN – K-Nearest Neighbors

K-NN is a kind of occasion based learning or apathetic realizing, where the capacity is just approximated locally and all calculation is conceded until characterization. K-NN arrangement, the yield is class participation. An article is ordered by a dominant part vote of its neighbors, with the item being doled out to the class most basic among its k closest neighbors (k is a positive whole number, normally little). In the event that k = 1, at that point the item is just appointed to the class of that solitary closest neighbor. K-Nearest Neighbors have been used in statistical estimation and pattern recognition i.e

If K=1, select the nearest neighbor

•If K>1, for classification select the most frequent neighbor, for regression calculate the average of K neighbors

X	Y	Distance
Attribute 1	Attribute 1	0
Attribute 1	Attribute2	1

$X=Y \Rightarrow D=0$

$X=\neq Y \Rightarrow D=1$

The advantage of this technique is: K-NN is pretty intuitive and easy: K-NN formula is extremely simple to grasp and equally straightforward to implement. To classify the new information K-NN formula reads through whole information set to search out K-nearest neighbors. This algorithm needs solely a tiny low quantity of coaching data for estimating the parameters essential for classification

3.2 Experiments Data Set

The information set for this analysis was taken from UCI data repository [49]. Information accessed from the UCI Machine Learning Repository is freely obtainable. This info contains seventy six attributes, and when neglecting redundant and unsuitable attributes, fifteen attributes were hand-picked. Below is that the list of fifteen attributes and their temporary description. Specially, the Cleveland, Hungarian, Switzerland and therefore the VA urban center databases are employed by several researchers and located to be appropriate for developing a mining model, attributable to lesser missing values and outliers. The information were cleansed and preprocessed before they were submitted to the planned algorithmic rule for coaching and testing. The 740 record sets were the valid instances for supervised machine-learning model building. The below shows the chosen vital risk factors from databases and their corresponding values Predictable attribute

1. Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (has heart disease))

Key attribute

Patient Id – Patient's identification number

Input attributes (Description of attributes)

1. Age in Year
2. Sex (value 1: Male; value 0: Female)
3. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non angina pain; value 4: asymptomatic)
4. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
5. Restecg – resting electrographic results (value 0: normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
6. Exang - exercise induced angina (value 1: yes; value 0: no)
7. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping)
8. CA – number of major vessels colored by floursopy (value 0-3)
9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10. Trest Blood Pressure (mm Hg on admission to the hospital)
11. Serum Cholesterol (mg/dl)
12. Thalach – maximum heart rate achieved
13. Old peak – ST depression induced by exercise
14. Smoking – (value 1: past; value 2: current; value 3: never)
15. Obesity – (value 1: yes; value 0: no)Execution of Bayesian Classification

Attribute choice or feature sub-setting technique was applied for any reduction of information to form patterns easier and comprehendible, however found negligible effects on performance measures of the model engaged during this study. Visible of the above, all the thirteen attributes were taken into the thought for developing a classifier's model and getting CVD prognostic outcome. The info mining approach was used for evaluating the classification algorithms engaged and the DM tool was accustomed to build the model. In these experiments, 10-fold cross-validations were utilized to partition the info set into coaching and check sets; this fulfills the necessity of model training and testing purpose

3.3 Data Source

The publicly available heart disease database from Cleveland, Hungary, Switzerland and the VA Long Beach Clinical databases [49] have aggregated enormous amounts of data about patients and their ailments. The term Heart infection includes the assorted illnesses that influence the heart. Coronary illness is the real reason for setbacks on the planet. Coronary illness kills one individual at regular intervals in the United States. Coronary illness, Cardiomyopathy and Cardiovascular infection are a few classifications of heart ailments. The expression "cardiovascular malady" incorporates a wide scope of conditions that influence the heart and the veins and the way where blood is siphoned and coursed through the body. Cardiovascular ailment (CVD) results in extreme disease, incapacity, and passing.

740 Record sets with therapeutic qualities will be gotten from a freely accessible database for coronary illness from AI archive will be utilized, that is Cleveland, Hungary, Switzerland and the VA Long Beach Heart Disease databases [49] with the assistance of the datasets, and the examples noteworthy to the heart assault forecast are separated.

3.4. Processing and Analysis

The record sets were split into 2 datasets: coaching dataset and testing dataset. A complete 740 record sets with fifteen medical attributes were obtained from the guts illness info. The records were split into 2 datasets like coaching dataset (296 record sets) and testing dataset (444 record sets). To avoid bias, the records for every set were hand-picked willy-nilly in a very quantitative relation of 1 to 1.5. In machine learning, a coaching set consists of associate degree input vector and a solution vector, and is employed along with a supervised learning methodology to coach the information (e.g. decision tree, KNN or a Naive Thomas Bayes classifier) employed by associate degree in AI machine. In a very dataset a coaching set is enforced to make up a model, whereas a check (or validation) set is to validate the model designed. Knowledge points within the coaching set are excluded from the check (validation) set. When a model has been processed by victimization the coaching set checks the model by creating predictions against the test set as a result of the information within the testing set already contained in the celebrated values for the attribute to predict.

The table below shows the description of dataset selected for this work. The total record sets divided into two with 13 and 15 attributes respectively.

Dataset	No. Of Attributes		Instances	Classes
Health Services Data	A	B	740	2
	13	15		

Table 2.0 Dataset Description

The model was developed and the first 13 input attributes were used then two more other attributes which are **obesity and smoking** were added, as these attributes are considered as important attributes for heart disease.

Also the deaths due to heart disease in many countries occur due to: work overload, mental stress and many other problems, these are the other factor attributes we had considered in observing the prediction change.

Most of the research papers referred upon have used 13 input attributes for prediction of Heart disease, to get more appropriate results two more important attributes were added that is obesity and smoking.

Healthcare industry is generally “information rich”, but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making- that’s why we looked for more other attributes which contribute to the heart disease

4. EXPERIMENTS AND RESULTS

The exhibition survey of a model for Heart Disease Prediction, utilizing Decision Trees, Naive Bayes, and KNN displaying strategies were assessed concerning AI calculations. The targets of the trials were: To break down the exhibition for the coronary illness expectation procedures, and portray how to improve their forecast power, Efficient and precise in coronary illness forecast; To examine the centrality of symptomatic highlights that best depict coronary illness information utilizing information mining strategies. The Experiments demonstrated that the proposed technique gives the exact conclusion of coronary illness than the current strategies

4.1 Experimental Setup

This exploration utilized classifiers given by Weka. The informational indexes were utilized as contribution to three AI calculations; Naive Bayes (NB), K-Nearest Neighbor (KNN) and Decision Trees (DT). The investigations began with 13 info properties and then15 information traits esteems. Investigation results were then exhibited in tables, broke down and deciphered as definite

4.2 Experimental Results and Analysis

The test results and investigation accomplished for this examination was spoken to as in the tables beneath. The exploration system has been clarified in the past area. For the tests, different information mining grouping strategies were connected on the dataset. In this investigation, WEKA AI apparatus for information mining was utilized to achieve the goals. The level of precision rate and mistake rate of information mining Classification methods was utilized as the estimation parameters for investigation. These parameters recommend that the classifier having a higher exactness rate and lower estimation of blunder rate arrange the dataset in very amended way and the other way around. In this examination, the information was right off the bat isolated into preparing information and testing information. The preparation set was utilized to build the classifier and test set utilized for approval. In this examination, the level of dataset utilized for preparing and testing information were 40% and 60% individually. At that point, the 10 overlay cross approval technique was connected to create the classifiers utilizing recently referenced AI apparatuses. At last the outcomes were archived as far as precision rate and mistake rates.

The table beneath Displays the results for classification techniques applied on health facility services data in WEKA Considering accuracy and error rates as performance measure the classification techniques with highest accuracy are obtained for health facility Services data in given different techniques used.

Table 3.0 Results Using WEKA API

Technique Used	Accuracy Rate		Error Rate	
	13 Attributes	15 Attributes	13 Attributes	15 Attributes
Naive Bayes	90.76	94.59	9.24	5.41
Decision Tree	97.07	99.77	2.93	0.23
KNN	79.28	82.43	20.72	17.57

The graph below displays the performance analysis of classification techniques for 15 attributes using WEKA tool. The best classifier for this particular data set will then be chosen.

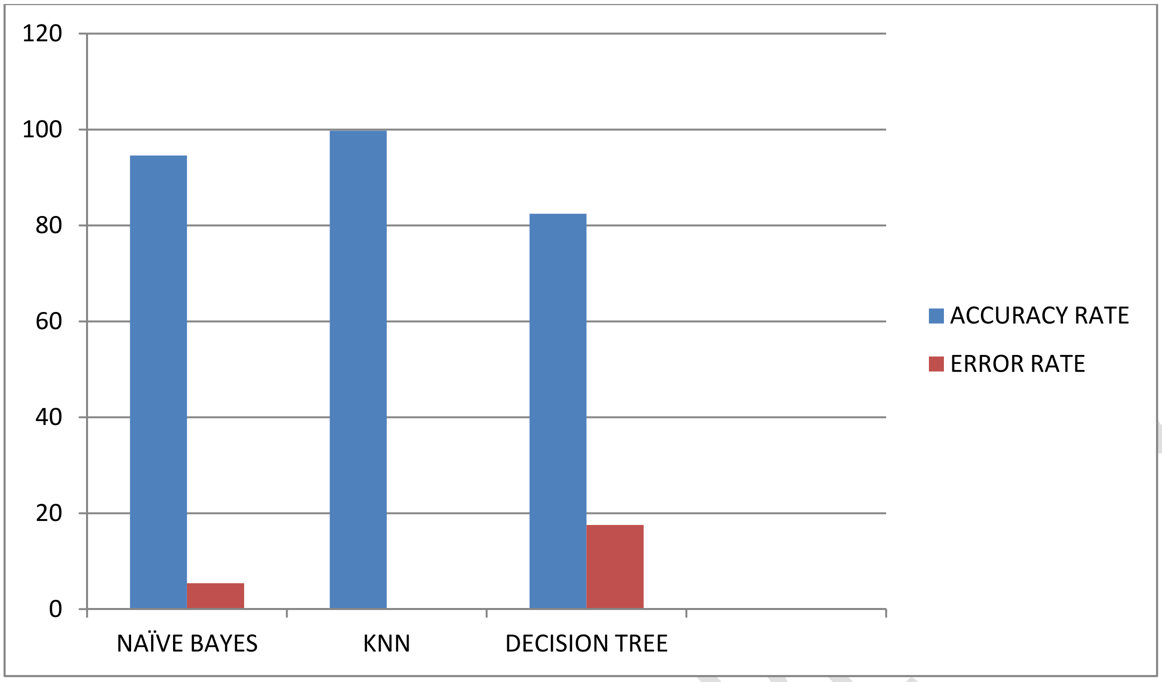


Fig 4.0 Performance analysis of classification techniques using WEKA API

4.3. Results

The dataset comprised of all 740 Record sets in Heart illness database. The records were then divide into two, one utilized for preparing comprises of 296 records and another for testing comprises of 444 records. The information mining apparatus Weka 3.6.6 was utilized for trial. At first dataset contained a few fields, in which some incentive in the records was absent. These were recognized and supplanted with most fitting qualities utilizing Replace Missing Values channel from Weka 3.6.6. The Replace Missing Values channel checks all records and replaces missing qualities with mean mode technique. This procedure is known as Data Pre-Processing. After pre-handling the information, information mining order procedures, for example, KNN, Decision Trees, and Naive Bayes were connected. A disarray lattice is acquired to figure the exactness of arrangement. A perplexity grid demonstrates what number of occurrences has been doled out to each class. In our analysis we have two classes, and in this manner we have a 2x2 perplexity network

Class A= YES (Has coronary illness)

Class B = (No coronary illness)

Table 4.0 a Disarray Network

	A(Has heart disease)	B(Has no heart disease)
A(has heart disease)	TP	FN
B(has no heart disease)	FP	TN

TP (True Positive): It indicates the quantity of records named genuine while they were in reality evident. FN (False Negative): It signifies the quantity of records delegated false while they were in reality evident. FP (False Positive): It indicates the quantity of records named genuine while they were in reality false. TN (True Negative): It means the quantity of records named false while they were in reality false. Results got with 13 properties are determined beneath

Table 3.3 Confusion Networks Got For Three Arrangement Techniques with 13 Qualities

Confusion matrix for Naive Bayes:

	A	B
A	182	13
B	28	221

Confusion matrix for Decision Trees:

	A	B
A	205	6
B	7	226

Confusion matrix for KNN:

	A	B
A	160	30
B	62	192

Results obtained by adding two more attributes i.e. total 15 attributes are specified below.

Table 3.4 Confusion matrixes obtained for three classification methods with 15 attributes

Confusion matrix for Naive Bayes:

	A	B
A	187	11
B	13	233

Confusion matrix for Decision Trees:

	A	B
A	168	0
B	1	275

Confusion matrix for KNN

	A	B
A	153	36
B	42	213

Table 5.0 shows accuracy for different classification methods with 13 input attributes and 15 input attributes values.

Classification Techniques	Accuracy with	
	13 Attributes	15 Attributes
Naive Bayes	90.76	94.59
Decision Tree	97.07	99.77
KNN	79.28	82.43

The accuracy of each of the method is plotted on a graph as below:

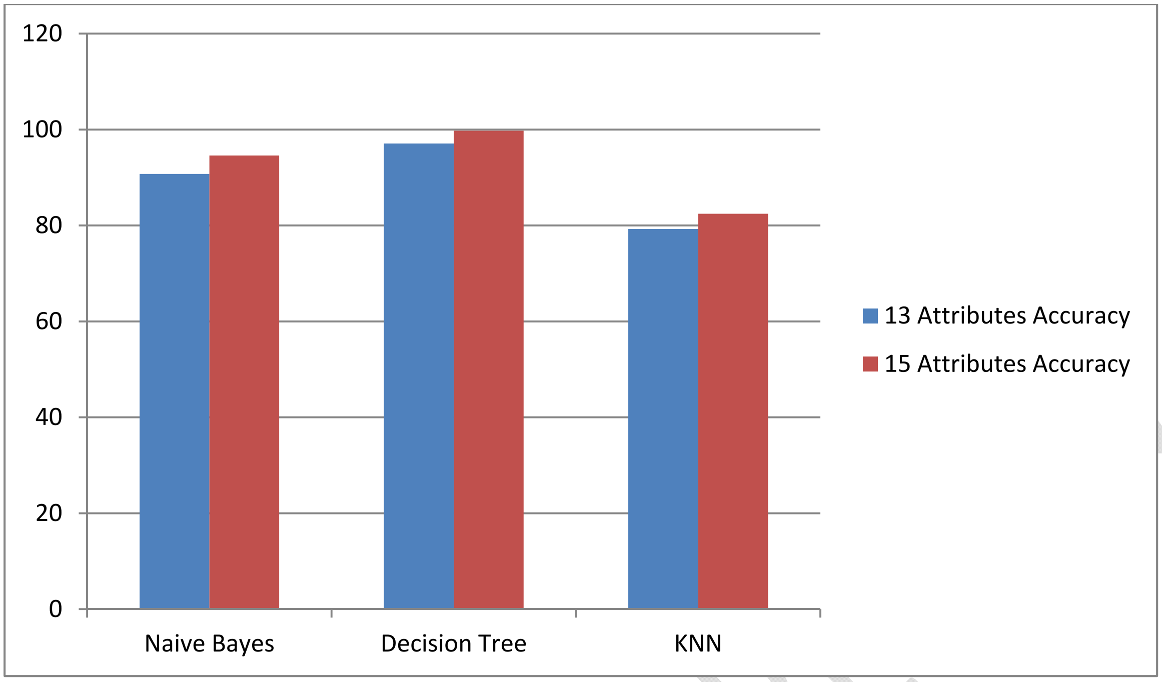


Figure 5.0: Graphical representation of accuracy for each method.

The performance and accuracy of every experiment are evaluated through performance measures like true positive rate, precision, F-measure, receiver in operation characteristic (ROC) space, letter statistics and root mean sq. (RMS) error. Identical measures are used for comparative analysis of enforced algorithms. Once the experiments, subsequent step is to match algorithms employed in these experiments for lightness the most effective one in terms of un-wellness prediction chance and classifier's accuracy. Having a glance at the results, it becomes apparent that the goal to supply AN ensemble classifier for early diagnostic screening with needed level of accuracy is triple-crown. A correlation between accuracy and therefore the quantity of attributes employed in the creation of the classifier was found. In general, additional attributes offer larger accuracy as visualized by results. With relation to mythical creature space as performance live, AN optimal/perfect classifier can score one on this take a look at, therefore this will build our results trying less dimmed with results quite 0.9 mythical creature for all classifiers. The comparative performance outline of enforced algorithms is given in table above. In general, the results of all the enforced rules are far better by all algorithms with specially the choice tree algorithm leading in accuracy and prediction chance. The accuracy of enforced algorithms on the given heart condition knowledge set is given within the table given above, and therefore the lowest accuracy is 84.43% for KNN analysis and therefore the highest accuracy is 98.17% for the choice rule supported on the top mentioned results and comparisons with relation to the chosen performance measures, the naïve Bayes and decision tree performed well and every rule with quite 94 prediction chance increased responsibility of the prediction system. Additional stress is given to pick out the algorithms having high true positive rate, as being the core live for early designation of heart condition

.

5. CONCLUSION AND FUTURE WORK

5.1 Knowledge Contributions

This research that proposed the use of a model for Heart Disease Prediction using Data Mining Classification Techniques provided a set of contributions that can be summarized while considering different points of view. On the more theoretical and modeling side, heart disease model for prediction analysis was proposed.

On the implementation side, this research improved results on accuracy with increase in number of attributes. This is supported by the high levels of classification accuracy exhibited when data sets that were used showed that there is increase in classification accuracy as the number of the attributes used for testing increased.

5.2 Conclusion

This approach-based paradigm for cardiopathy prediction model has been projected as a system whereas utilizing Naïve Bayesian, decision tree and KNN classifiers. The projected system is GUI-based, easy, scalable, reliable and expandable system, that has been enforced on the maori hen platform. The projected operating model can even facilitate in reducing treatment prices by providing Initial medical specialty in time. The model can even serve the aim of coaching tool for medical students and can be a soft diagnostic tool obtainable for MD and heart specialist. General physicians will utilize this tool for initial diagnosing of cardio patients. Various information mining characterization procedures were connected on the particular dataset, the order procedure inside the framework model is performed with traits like age, sex, heart beat rate, cholesterol level and so on. The expectation is then made dependent on this arrangement results. Here the AI ability of the PC framework can be stretched out into the medicinal field. The proposed framework model is best for lessening the blunder event during the illness expectation. In this examination the exactness and precision of three unique classifiers are estimated, the outcome demonstrates choice tree arrangement has high precision and less mistake rate, Naïve Bayer characterization strategy creates preferred outcome over KNN grouping. This investigation can assist scientists with getting productive outcomes in the wake of knowing the best order strategy for this specific dataset. The general target of the examination was to foresee all the more precisely the nearness of coronary illness. In this exploration, more information characteristics weight and smoking were utilized to get progressively precise outcomes.

5.3 Future Work

Heart Disease Prediction using Data Mining Classification Techniques can be used largely in hospital based sectors for disease prediction, However, there is need for more research to be done on contextual knowledge being incorporated as part of feature selection and model creation for specific domains where precise context, which does not depend on attributes needs to be used in learning and prediction is required also. There is need to experiment the prediction models with real live testing of heart disease. This research can also be enhanced by experiment with more attributes in training and testing data sets. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health's diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

REFERENCES

1. Abdullah H. Wahbeh, “A Comparison Study between Data Mining Tools over some Classification Methods” (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, vol. 3, no. 2, p 18-26*,* 2012.
2. Abhishek Taneja, Heart Disease Prediction System Using Data Mining Techniques; Oriental Journal of computer science & Technology ISSN: 0974-6471 December2013.
3. Aditya Methaila, Early Heart Disease Prediction Using Data Mining Techniques; CCSEIT, DMDB, ICBB, MoWiN, AIAP pp. 53–59, 2014.
4. Al-Radaideh “Using data mining techniques to build a classification model for predicting employee’s performance”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, pp.60-71, 2014.
5. B. Venkatalakshmi and M. Shivsankar, “Heart disease diagnosis using predictive data mining,” International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, no. 3, pp. 1873–1877, 2014.
6. Beant Kaur and Williamjeet Singh.,” Review on Heart Disease Prediction System using Data Mining Techniques”, IJRITCC , pp. 56-72, October 2014.
7. Blake, C.L.,Mertz, C.J.“UCI Machine Learning Databases”
8. Chaitrali S. Dangare, Sulabha S. Apte, —Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques; International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012
9. Chaitrali S.Danagre, Sulabha S.Apte, Ph.D, Improved Studyof Heart Disease Prediction Systemusing Data mining Classification Techniques,IJCA,June 2012.
10. Charly, K.: “Data Mining for the Enterprise”, 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 2014.
11. Choi Keunho et al. “Classification and Sequential Pattern Analysis for Improving Managerial Efficiency and Providing Better Medical Service in Public Healthcare Centers” health inform res, pp.67-76, June 2014
12. D.K, “Classification of women health disease (Fibroid) using decision tree algorithm”, International Journal of Computer Applications in Engineering Science Vol.2, Issue 3, pp.84, September2016,]
13. Fartash. Haghanikhameneh “A Comparison Study between Data Mining Algorithms over Classification Techniques in Squid Dataset” International Journal of Artificial Intelligence, Autumn (October) 2015, Vol. 9, pp66-68.
14. Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining,* AAAI Press / The MIT Press, Menlo Park, CA, 2014, pp. 1-34.
15. Garchchopogh et al, “Application of decision tree algorithm for data mining in healthcare operations: A case study”, International Journal of Computer Applications Vol 52 – No. 6, August 2014, pp.567-280.
16. Global Atlas on Cardiovascular Disease Prevention and Control (PDF). World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. pp. 3–18. ISBN 978-92-4-156437-3 September 2011.
17. Han, J. and Kamber, M. (2014). Data Mining: Concepts and Techniques. fourth Edition, Morgan Kaufmann Publishers, San Francisco Vol. 16, No. 3, 2013, pp. 291-296
18. Hearty “Analysis of meal patterns with the use of supervised data mining techniques-Artificial Neural Network and Decision Tree”, The American Journal of Clinical Nutrition Vol. 18, No. 3, 2013, pp. 192-190, 2015
19. Hlaudi Daniel Masethe, Mosima Anna Masethe-prediction of Heart Disease using Classification Algorithms; Proceedings of the World Congress on Engineering and Computer Science 2014.
20. Ho, T. J.: Data Mining and Data Warehousing, Prentice Hall, 2016, pp.66-69.
21. Huang, Li, Su, Watts, & Chen, 2007; Ishibuchi, Kuwajima, Nojima, 2007; Karabatak & Ince, 2009; Shin et al., 2010; Wang & Hoy, 2015, pp256-267.
22. Jabbar et al “Heart disease prediction system using associative classification and Genetic Algorithm”, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2015

23. Jyoti Soni et.al. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction; International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.
24. Koç et al, “A comparative study of artificial neural network and logistic regression for classification of marketing campaign results”, Mathematical and Computational Applications, Vol. 18, No. 3, 2013, pp. 392-398
25. Mrs.G.Subbalakshmi, “Decision Support in Heart Disease Prediction System using Naive Bayes”, Indian Journal of Computer Science and Engineering. Vol. 3, No. 5, May 2014, pp.227-227-238.
26. Nakul Soni, Chirag Gandhi, “Application of data mining to health care”, International Journal of Computer Science and its Applications Volume 36– No.10,vol.5 June 2014,
27. Nidhi Bhatla, Kiran Jyoti, “ An Analysis of Heart Disease Prediction using Different Data Mining Techniques” International Journal of Engineering and Technology Vol.1 issue 8 2012, pp.234-241..
28. Obenshain, M.K: “Application of Data Mining Techniques to Healthcare Data”, Infection Control and Hospital Epidemiology, 25(8), 690–695, 2014
29. Patil Dipti “An adaptive parameter for data mining approach for healthcare applications” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 1, 2014, pp.66.70..
30. Pushpalata Pujari “ Classification and comparative study of data mining classifiers with feature selection on binomial data set” Journal of Global Research in Computer Science, Vol. 3, No. 5, May 2016, pp.675-682.
31. R. Chitra, Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques; Ictact Journal On Soft Computing, July 2013,volume: 03, Issue: 04, pp781-785.
32. R.Wu, W.Peters, M.W.Morgan, “The Next Generation Clinical Decision Support: Linking Evidence to Best Practice”, *Journal of Healthcare Information Management*. 16(4), pp. 50 55, 2016.
33. S.Asha Rani and Dr.S.Hari Ganesh, “ A comparative study of classification algorithm on blood transfusion” International Journal of Advancements in Research & Technology, Volume 3, Issue 6, June-2014, pp.56-63.
34. Saichanma et al. "The Observation Report of Red Blood Cell Morphology in Thailand Teenager by Using Data Mining Technique." Advances in hematology, 2014 pp.104-109.
35. Sakshi and Prof.Sunil Khare “A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining” International Journal on Recent and Innovation Trends in Computing and Communication Vol. 3 Issue: 8,pp.5142 – 5147
36. Sayad AT, Halkarnikar PP. Diagnosis of heart disease using neural network approach. Int J Adv Sci Eng Technol. 2014;2:88–92.
37. Setiawan, *et al*,” A Comparative Study of Imputation Methods to Predict Missing Attribute Values in Coronary Heart Disease Data Set**”,** Journal in Department of Electrical and Electronic Engineering,Vol.21, PP. 266–269, 2008
38. Shadab Adam Pattekari and Asma Parveen, prediction system for heart disease using naïve bayes, International Journal of Advanced Computer and Mathematical Sciences, 2012, pp.476-484.
39. Shanthi Mendis; Pekka Puska; Bo Norrving; World Health Organization (2011).
40. Shelly Gupta et al. “Performance Analysis of Various Data Mining Classification Techniques on Healthcare Data” International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011,pp.877-892.
41. Sundar et al. “Performance analysis of classification data mining techniques over heart disease database”, [IJESAT] International Journal of Engineering Science and Advanced Technology, Volume-2, Issue-3,pp. 470 – 478,2013.
42. Tang, Z. H., MacLennan, J.: Data Mining with SQL Server 2005, Indianapolis: Wiley, 2015, pp.445-450.
43. Tariq O. Fadl Elsid and Mergani. A. Eltahir “An Empirical Study of the Applications of Classification Techniques in Students Database” Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 4, Issue 10(Part - 6), pp.01-10, October 2014
44. Thuraisingham, B.: “A Primer for Understanding and Applying Data Mining”, IT Professional, 28-31, 2015
45. Umadevi, D.Sundar, Dr.P.Alli, “A Study on Stock Market Analysis for Stock Selection – Naïve Investors’ Perspective using Data Mining Technique”, International Journal of Computer Applications (0975 – 8887), Vol 34– No.3,2011.

46. V. Manikandan and S. Latha, “Predicting the Analysis of Heart Disease Symptoms Using Medical Data Mining Methods “International Journal of Advanced Computer Theory and Engineering”, Vol. 2, Issue. 2,pp.236-240, 2013.
47. Vikas Chaurasia, et al, Early Prediction of Heart Diseases Using Data Mining Techniques; Caribbean Journal of Science and Technology ISSN 0799-3757, Vol.1,208-217, 2013.
48. World Health Organization; Cardiovascular Diseases (CVDs) Fact Sheet Reviewed June 2016
49. Cleveland, Hungary, Switzerland, & VA Long Beach Database: http://archive.ics.uci.edu/ml/datasets/Heart+Disease
50. Nadali, A; Kakhky, E.N.; Nosratabadi, H.E., "Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system," Electronics Computer Technology (ICECT), 2011 3rd International Conference on , vol.6, no., pp.161,165, 8-10 April 2011
51. Dangare CS, Apte SS. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *Int J Comput Appl*. 2012;47(10):44–48**.**