International Journal of Scientific Research & Engineering Trends
Volume 8, Issue 1, Jan-Feb-2022, ISSN (Online): 2395-566X

A Survey on Data Mining Techniques for Analysis and Prediction of Heart Diseases

Research Scholar Sugendran G, Asst. Prof. Dr.S.Sujatha
Department of omputer Science
G.R.DamodaranCollege of Science
Coimbatore,India

Abstract- A classification is one of the most important and beneficial techniques. whenever interacting with a huge amount of data, order methods are highly beneficial. to forecast all of the class names, an arrangement is being used. recently accessible data is commonly represented by a class name in grouping models. segmentation is the very first step towards developing a model which describes and identifies data. analysis refers to the process of describing a model that shows and identifies data categories or concepts. description techniques can deal with both mathematical and obvious attributions. building fast and accurate classifiers for large datasets is a crucial task in data mining and data breaches. grouping predicts understandable class names and classifies data based on the preparation set. classification is a two-stage process. we propose a study of several data mining arrangement strategies such as decision tree, k-nearest neighbour, support vector machines, naive bayesian classifiers, neural networks, and random forest in this article.

Keywords- decision tree, k-nearest neighbour, support vector machine, bayesien classifiers, neural networks, random forest

I. INTRODUCTION

Analysing data and the cycle to obtain insightful data are instances of information digging. There are several machine learning tools available that may be used to analyse various types of data. Machine learning utilizes a range of applications to dissect relevant information, including decision making, market analysis, production control, and customer support, as well as logical discoveries and training frameworks. [1]. This approach is applied to comparative groupings not same type of data is alluded to grouping. The groups are formed by researching comparable examples of the data. While classifying attributes with identical utility and in populace pick up understanding into structures can be acquired in science for inferring plant and creature scientific categorizations. Similar dwellings and terrains region can be discovered in a city through using bunching in topography.

Data bunching may be used to describe all reports available on the Web in order to discover new hypotheses. The single information bunching arrangement technique creates groups and things since those in various bunches are unmistakable and those in the same bunch are basically the same. In data mining, group inquiry is regarded as a standard subject that is used for data exchange. The information items are organised into a grouping of distinct classes called as a group. [2].Objects that are divided into separate classes are more exceptional, and objects within a class have a strong similarity to one another. Perceptive analysis is the process of separating from existing data sets in order to decide on designs and predict future results and patterns. Anticipation analysis does not provide future projections.Following that, with a suitable level of dependability, evaluate the chance that scenarios and risk evaluation gauge is provided by the expectation examination. The forecast examination completely predicts future prospects. To more conveniently get clients, products, and partners, as well as to identify potential risks and opportunities for company, predictive models are used to analyse current data and recorded realities applied to business.It employs a variety of approaches to create future business gauges, including data mining, factual demonstrating, and AI to assist investigators. Predictive analysis is a branch of insights that extracts data and uses it to forecast trends and norms of behaviour.

The advancements of foresight web evaluation include the estimation of the possibility of future events online. Data displaying, Intelligent systems, Machine learning, deep learning calculations, and data mining are remembered for the Predictive examination factual procedures Predictive examination factual procedures can be applied to an obscure whether it is previously, present, or future frequently the obscure occasion of revenue is later on. Predicting the likely behaviour of people, hardware, or other things Factors that can be calculated and analysed are used in predictive investigation programming applications. Age, gender, location, kind of vehicle, and driving record, as well as analysing and providing collision protection plans, are all regarded possible driving security considerations by the insurance company. With accurate tactics and the ability to create predictive data models Predictive inquiry necessitates a high level of skill.

The final findings of forecast evaluation are frequently the domain of data researchers, analysts, and other talented data investigators.Data scientists aid with social event essential information and put it up for research. As a result, using data perception, dashboards and reports are supported by programming designers and business specialists. The following are categories of clustering techniques:

Partitioning Techniques: The basic operation of this method is the collection of samples in such a way that bunches of similar objects with high likenesses are created. Unique samples are grouped together in different bunches from related ones in this section. These techniques are entirely dependent on the distance between the instances [3].

Hierarchical Methods- A given dataset of articles is progressively degraded inside this method. This approach is completed with the inclusion disintegration. There are two types in this grouping. It is difficult and agglomerative methods are reliant on [4]. The agglomerative technique is a bottom-up strategy in which the first stage is the formation of various gatherings. When the gatherings are near to one another, they are combined.

Methods Based on Thickness- In several methods, the distance between the items is used to divide the articles into groups as a foundation for bunching. Regardless, these approaches must be effective in identifying round-formed bunches. It is difficult to obtain subjective formed using the thickness-based grouping approach.

Methods Based on Networks- The quantization of the item's space to a finite number of cells is known as the age of lattice structure. This method is free since it is not dependent on the availability of a number of data items and it is also speedy.

II.CLASSIFICATION IN DATA MINING

The anticipation of obtaining participation for example data in data mining should be achievable with the aid of the arrangement approach [5]. Forecast research is the cycle in which the outcome is predicted based on existing data. For example, based on current weather data, the day under investigation may be "bright," "stormy," or "shady." Within this cycle, there are two phases..

1. Model Development- Model development clarifies the collecting of predetermined classes. A large number of tuples are utilised in the construction of the prepared set model. In this method, the order of the principles, decision trees, or numerical formulae/relapse appear.

2. Model usage: The following path in the grouping is model use. The preparation set of the obscure from the obscure data for the precision examination is designed to group the test data [6]. The result of the model's grouping is used to contrast in an example test with a known name. The test set is not dependent on the preparation set. Every classifier has some feature that distinguishes it from other classifier structures. The traits are referred to as classifier qualities. These are the characteristics.

3. Correctness - How a classifier order falls precisely depends on these properties. There are various mathematical attributes that are reliant on the number of topple groups correctly and the number of topple orders incorrectly to check for exactness.

4. Time - How long will it take to create the model? This also includes the possibility for the model to utilise to group at that moment number of topples

5. Strength -Ability to properly group a fall, even if the topple is agitated. Commotion cannot be correct value or lacking worth.

6. Data Size -Classifiers should be a free structure the size of the data collection. The model should be flexible. The size of the data base has no bearing on the model's display.

7. Extendibility- A new element can be introduced at any time. This is a difficult element to implement.

1. Svm Classifier

In this investigation, the inventor suggested an SVM classifier for relapse, order, and overall example recognition. This classifier is regarded as satisfactory in comparison to other classifiers because to its high speculative execution without requiring any prior knowledge to include it. The presentation is unexpectedly superior, for example, the information space measurement is really high. For dividing prepared information between the two classes, the SVM requires the best grouping capacity recognised proof.The arrangement work metric may also be expressed mathematically [7]. The direct grouping capability for the straightly distinguishable dataset isolates the hyper plane f(x). This hyperplane goes across the middle of two classes, effectively separating them. xn is sorted by checking the sign capacity of the new data example work f(xn); xn, which denotes the positive class if f(xn) > 0. This is done once another capacity has been assured.

SVM has a key objective of ensuring the best capacity by extending the edge between the two classes. As a result of this fact, there are several direct hyperplanes. A hyperplane is one of two types of room or distance measurements. The edge is closest between the nearest data focuses to a point on the hyperplane with the shortest distance. This can also help us characterise the optimum way to expand the edge, which can help us choose a few of hyper planes for the solution for SVM in any case, when so many hyper planes are available [8]. The goal of

the SVM is to generate direct capacity in order to provide a recognised evidence of objective capacity. The execution of the relapse inquiry can help to widen the SVM. The error models come in handy here for the SVRs. The blunder is defined as zero of the differences between true and expected values inside an epsilon sum. There is a distant chance of a direct development in the epsilon obtuse error. The assistance vectors can be examined by lowering the Lagrangian. Inhumanity for exceptions might be beneficial to the aid vector relapse. The disadvantage of SVM is that the computations are insufficiently effective.

There are a variety of solutions offered for this. One method to addressing this issue is to divide one large issue into a number of smaller concerns. There are only a few elements picked for any issue's productive development. This cycle will continue in an iterative fashion until all concerns have been resolved. The problem of learning SVM will be solved similarly by seeing the estimated least encasing a large number of instances in the programme. This audit report is based on the forecast inquiry, which is nearly complete with the characterization techniques. This article is so well-coordinated that the presentation of the expectation inquiry is presented in part 1 using distinct characterisation techniques. Part 2 consists of a writing study based on the predicted inquiry. Part 3 depicts the outcome evaluation in a large number of papers distributed.

III. LITERATURE REVIEW

Akhiljabbar et al. presented a "Coronary Illness Prediction System Using Associative Classification and Genetic Algorithm" in 2012. They developed a productive affiliated characterisation computation based on hereditary approach for forecasting coronary disease. The primary favoured position of hereditary calculation is the revelation of substantial level forecast chooses is that the discovered guidelines are extremely understandable, with high predictive precision and interesting quality features. The suggested approach aids in the best prediction of cardiac disease, which even aids professionals in their decision-making[1].

Parthiban et al. Using programmed learning techniques, [2] investigated cardiac disease in diabetic individuals. WEKA is used to do Naive Bayes and SVM computations. The data set of 500 patients from the Chennai Research Institute is used. There are 142 people who have the sickness and 358 persons who do not. The Naive Bayes method provides 74 percent accuracy. SVM has the highest exactness of 94.60 percent.

Akhil Jabbar et al. (2013) proposed "Grouping of Heart Disease Using Artificial Neural Network and Feature Subset Selection" in 2013. They presented another component selection approach based on ANN for coronary disease characterisation. They used several element determination approaches to rank the characteristics that provide more towards the order of coronary sickness, therefore reducing the number of determination tests to be taken by a patient. The proposed method eliminates useless and distorted data[3].

N. S. Nithya et al. introduced the "Gain proportion based fluffy weighted affiliation rule digging classifier for clinical demonstrative interface" in 2014. They said that the old methodology, which was based on data collection and fluffy affiliation rule digging computation for distinguishing both affiliation rules and enrollment capabilities, was not feasible. They made use of a large number of distinct characteristics.They enhance classification accuracy by adjusting pick up proportion based fluffy weighted affiliation rule mining. The Decision Tree, Multi-Layer Perceptron (MLP) Neural Network, and Naive Bayes were all subjected to varying data amounts. For the different datasizes, the time consumed by the computations for trainings and the accuracy of their ordering were investigated[5].

Vembandasamy et al. [6] used the Naive Bayes method to identify heart disease in 2015. In Naive Bayes, the Bayes theorem is used. As a result, Naive Bayes has an exceptional autonomy guideline. The data used came from one of the most important diabetes research foundations in Chennai. There are 500 patients in the informative index. WEKA is used as an instrument that performs characterisation using 70% of the Percentage Split. Innocent Bayes has an accuracy of 86.4%.

Otoom et al. [7] developed a strategy for assessment and progress in 2015. The suggested structure perceives and sees coronary course abnormality. The Cleveland Heart data is sourced from the University of California, Irvine. There are 303 instances and 76 credits/features in this informative index. Out of a total of 76 characteristics, 13 are utilised. For revelation purposes, two tests with three computations are performed: Bayes Naive, Support Vector Machine, and Functional Trees FT. The WEKA The gadget is used to achieve enlightenment. As a result of testing the Holdout test, the SVM method produces 88.3 percent accuracy. SVM with Bayes net provide 83.8 percent precision in the cross-endorsement test. After using FT, a precision of 81.5 percent is achieved. The Greatest First assurance count is used to choose the seven best features. Cross-endorsement tests are used to determine endorsement. Bayes Naive obtained 84.5 percent accuracy by applying the test to the 7 best features chosen, SVM achieves 85.1 percent precision, and FT orders 84.5 percent effectively.

Anoopkumar M, et al. (2016) recommended providing a detailed assessment of the examination papers that would have discussed different Data Mining Methods, notably the often used and popular computations applied to EDM settings. The primary focus of this study is on developing methods for boosting scholarly exhibits and institutional

viability. Educational Data Mining is an interdisciplinary candid investigation domain that deals with the improvement of techniques for examining data arising in academic fields (EDM). Thus, it provides inborn data on educating and learning measure for persuasive instruction planning.These assessments provide data into methods in this proposed approach by improving the educational cycle, forecasting understudy execution, correlating the correctness of data mining computations, and displaying the growth of open source actualizes. [8]

Nikhil N. Salvithal and R.B. Kulkarni suggested a"Appraisal Management System Using Data Mining Classification Technique" in 2016. The proposed organised classifier computations were used to the Talent dataset in order to detect the ability set in order to pass judgement on the person's display. Finally, one of the most relevant classifiers is chosen based on precision. This method has been used to construct characterisation guidelines to predict whether or not a potential ability is suitable for advancement[8].

Tanvi Sharma and Anand Sharma proposed the paper "Performance Analysis of Data Mining Classification Techniques on Public Health Care" in 2016. The suggested study focused on the use of several Data mining request techniques based on AI, such as WEKA and Rapid tractor, to the public clinical consideration dataset in order to separate the clinical administrations structure. As a standard for execution, the degree of precision of each used data mining course of action approach is employed. The optimal method for explicit instructional assortment is chosen based on the highest increased accuracy [9].

Lee, E., Jang, et al. [11] suggested using a business game log information competition system for game data mining. The purpose behind the game data mining rivalry is to advance the investigation of game data mining by providing business gaming logs to the overall public. With the help of outside experts, gaming organisations would go to any length to avoid sharing their game data. This strategy enlisted investigators to create and use top-tier data mining algorithms to game log data.The primary goal of this suggested technique was to predict if a player

DOI: http://dx.doi.org/10.1145/12345.67890 would beat and when the player would mix over two periods in which the strategy was changed to a permitted to-play model from a month to month participation. Significantly positioned rivals utilised substantial learning; tree boosting and straight backslide were the results of the resistance discovered in this proposed method by the researchers and manufacturers.

N. Sridevi, S. Anitha[12] (2019) In India, heart infection is one of the most widely perceived diseases. Early detection of heart illness increases survival rate; so, this study effort is designed to predict if the patient has

coronary ailment using clinical data, which will aid in the end cycle. Using the heart diseases dataset, three controlled AI figures explicitly KNN, Naive Bayes, and SVM are examined for precision. The exploratory results clearly show that the Nave Bayes figure predicts coronary disease with an accuracy of 86.6 %.

Youness Khourdifi [13] Mohamed Bahaj (2019) FCBF, PSO, and ACO's suggested advanced model achieves an accuracy score of 99.65 percent with KNN and 99.6 percent with RF. This article may be the underlying step in learning in the detection of cardiovascular disease using tailored learning, and it may be related to future research. There are a couple of prerequisites for this assessment, the most important of which is the maker's database. There are additionally the contraptions employed in this inquiry, for example, the planning force of the PC, and thirdly fairly far accessible for the evaluation. This type of research necessitates cutting-edge resources and fitness in a variety of disciplines.

IV.CLASSIFICATION MODEL

The primary goals of a Classification calculation are to increase the predictive precision of the classification model. Characterization assignment can be seen as a managed method in which each occurrence is assigned to a class. For order, a few model techniques are used, some of which are [8,9,10].

- Decision Tree
- K-Nearest Neighbour
- Support Vector Machines
- Naïve Bayesian Classifier
- Neural Networks
- Random Forest

1. Decision Trees

A decision tree is a classifier that is employed in the event space in a recursive segment. This model has two centre points and a root. Other than the root, the centre points have one going toward the edge.

Midway Center is a test centre that produces dynamic edge after performing a test. Leaves are centre points that lack dynamism (in any case called terminal or decision centre points). In a decision tree, each inner centre divides the event space into at least two sub-spaces, each with its own discrete limit of the information credits esteems.

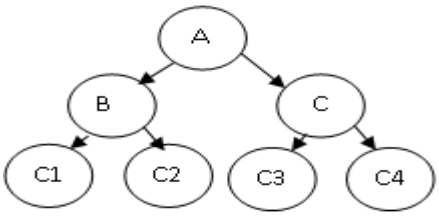


Figure 1.Decision Tree Classifier.

A denotes the root of the tree. B, C are internal nodes denote a test on a particular attribute and C1, C2, C3 and C4.

2. K-Nearest neighbour

This classifier is based on learning from planning tests. Every model corresponds to a point in n-dimensional space. In an n-dimensional model space, all readiness checks are carried out. When given a dark model, a k-Nearest Neighbor classifier searches the model space for the k getting ready tests that are most similar to the dark model. "Closeness" is depicted to the extent that the Euclidean distance, where the Euclidean distance between two centres, X=(x1,x2,.....,xn) and Y=(y1,y2,....,yn) is suggested by d (X, Y).

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

Each attribute is given equal weight in nearest neighbour classifiers. Nearest neighbour classifiers may also be used to forecast, that is, to produce a real-valued prediction for an unknown sample.

3. Support Vector Machine (SVM)

SVM is a very useful approach for backslide, gathering, and general model confirmation. It is regarded as a fair classifier due to its high theoretical execution without the requirement to add from prior information, regardless of how large the component of the information space is. It is regarded as a respectable classifier because to its high theory execution without the requirement for previous information addition, regardless of whether the portion of the information space is particularly large. An instant request work considers to a separating hyper plane f(x) that experiences the focal point of the two classes, segregating the two for a straitly specific dataset. SVMs were originally designed for two-dimensional representation, however they may also be used to solve multiclass problems.

4.Bayesian classifiers

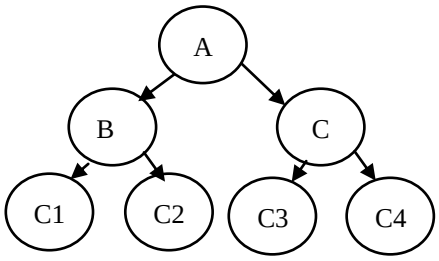
Statistical classifiers are Bayesian classifiers. They can forecast class participation based on probability. When the dimensionality of the information sources is high, the Naive Bayes Classifier approach is ideal. Innocent Bayes can consistently outperform more sophisticated grouping methods. Leave D alone when it comes to preparation set related class names. A n-dimensional ascribe, A1, A2,..., An, speaks to each tuple. Assume you have m classes, C1, C2,..., Cm. Given a tuple, X, the classifier will predict that X belongs to the class with the highest back probability, based on X. In other words, the uninformed Bayesian classifier predicts that tuple x belongs to the class Ci if and only if P (Ci/X) > P (Cj/X) for 1= j = m, j # I. As a result, we extend P(Ci/X). The greatest posteriori hypothesis is the class Ci for which P(Ci/X) is boosted. According to Bayes' theorem

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

P(X) is constant across all classes; only P (X/Ci) P (Ci) must be amplified. If the class earlier probabilities are unknown, it is generally assumed that the classes are similarly likely, that is, P(C1) = P(C2) =..... =P(Cm), and we would therefore enhance P(X/Ci). Otherwise, we increase P(X/Ci)P. (Ci).

5. Neural Networks

Neural networks may communicate about complicated relationships between data sources and yields [19]. The order system based on NNs is divided into three stages: information pre-handling, preparation, and testing. The element selection is alluded to in the information pre-planning. The highlights from the information pre-handling stage are fed into the NN for information preparation, and the NN produces a classifier. Finally, the testing data is used to confirm the classifier's efficiency.



A denotes the root of the tree. B, C are internal nodes denote a test on a particular attribute and C1, C2, C3 and C4.

6.K-Nearest neighbour

This classifier is based on learning from planning tests. Every model corresponds to a point in n-dimensional space. In an n-dimensional model space, all readiness checks are carried out. When given a dark model, a k-

Nearest Neighbor classifier searches the model space for the k getting ready tests that are most similar to the dark model. "Closeness" is depicted to the extent that the Euclidean distance, where the Euclidean distance between two centres, X=(x1,x2,.....,xn) and Y=(y1,y2,....,yn) is suggested by d (X, Y).

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

Each attribute is given equal weight in nearest neighbour classifiers. Nearest neighbour classifiers may also be used to forecast, that is, to produce a real-valued prediction for an unknown sample.

7.Support Vector Machine (SVM)

SVM is a very useful approach for backslide, gathering, and general model confirmation. It is regarded as a fair classifier due to its high theoretical execution without the requirement to add from prior information, regardless of how large the component of the information space is. It is regarded as a respectable classifier because to its high theory execution without the requirement for previous information addition, regardless of whether the portion of the information space is particularly large. An instant request work considers to a separating hyper plane f(x) that experiences the focal point of the two classes, segregating the two for a straitly specific dataset. SVMs were originally designed for two-dimensional representation, however they may also be used to solve multiclass problems.

8.Bayesian classifiers

Statistical classifiers are Bayesian classifiers. They can forecast class participation based on probability. When the dimensionality of the information sources is high, the Naive Bayes Classifier approach is ideal. Innocent Bayes can consistently outperform more sophisticated grouping methods. Leave D alone when it comes to preparation set related class names. A n-dimensional ascribe, A1, A2,..., An, speaks to each tuple. Assume you have m classes, C1, C2,..., Cm. Given a tuple, X, the classifier will predict that X belongs to the class with the highest back probability, based on X. In other words, the uninformed Bayesian classifier predicts that tuple x belongs to the class Ci if and only if P (Ci/X) > P (Cj/X) for 1= j = m, j # I. As a result, we extend P(Ci/X). The greatest posteriori hypothesis is the class Ci for which P(Ci/X) is boosted. According to Bayes' theorem

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

P(X) is constant across all classes; only P (X/Ci) P (Ci) must be amplified. If the class earlier probabilities are unknown, it is generally assumed that the classes are similarly likely, that is, P(C1) = P(C2) =..... =P(Cm), and we would therefore enhance P(X/Ci). Otherwise, we increase P(X/Ci)P. (Ci).

9.Neural Networks

Neural networks may communicate about complicated relationships between data sources and yields [19]. The order system based on NNs is divided into three stages: information pre-handling, preparation, and testing. The element selection is alluded to in the information pre-planning. The highlights from the information pre-handling stage are fed into the NN for information preparation, and the NN produces a classifier. Finally, the testing data is used to confirm the classifier's efficiency.

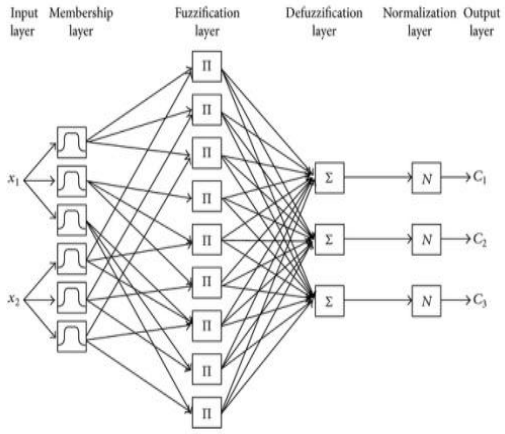


Figure 2.Neural Networks**.**

10.Random Forest (RF)

The random forest, as the name indicates, is made up of a large number of individual decision trees that operate together as a group. Every tree in the erratic backwoods emits a class expectation, and the class with the highest votes becomes our model's forecast.

IV.DATASET DESCRIPTION

The data set for this investigation was acquired from the statLog dataset in the UCI repository. It has 13 ascribes. The coronary sickness dataset remembered for the examination task includes all 270 instances with no characteristics. The dataset is commonly used for many types of cardiac disease, for example, common angina, atypical angina, non-anginal pain, and asymptomatic. This research is aimed at predicting the coronary sickness, which is the least serious of the infection kinds. The characteristic is a number information type that addresses the patient's age, which ranges from 290 to 65 years. The Cp is a characteristic used to determine the type of pain, and it ranges from 1 to 4.

The trestbpd is a resting pulse that is between 92 and 100; the fbs is a fasting glucose level that is either a 1 or a 0, addressing whether Boolean characteristics are genuine or fraudulent. The resting electrocardiogram (resecg) is a realistic resting electrocardiogram with three instances ranging from 0 to 2. The thalach is the most severe pulse,

ranging from 82 to 185. The exang is the activity-induced angina that has a Boolean value. The disease is the dataset's objective class, with a yes or no reflecting the presence of cardiac ailment. Essentially, each credit and its attributes are handled in the table. 1.

Table.1 Attribute and Description of the dataset used for research

Attribute	Descriptin	Details
Age	Age	Patients age, in years
Sex	Sex	0=female; 1=male
Chest pain	Cp	4 types of chest pain (1—typical angina; 2—atypical angina; 3—non-anginal pain; 4—asymptomatic)
Rest blood pressure	Trestbps	Resting systolic blood pressure (in mm Hg on admission to the hospital)
Serum	cholesterol	Chol Serum cholesterol in mg/dl
Fasting blood sugar	Fbs	Fasting blood sugar>120 mg/dl (0—false; 1—true)
Rest electrocardiograph	Restecg	0—normal; 1—having ST-T wave abnormality; 2—left ventricular hypertrophy
MaxHeart rate	Thalch	Maximum heart rate achieved
Exercise-induced angina	Exang	Exercise-induced angina (0—no; 1—yes)
ST depression	Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope	slope of the peak exercise ST segment (1—upsloping; 2—fat; 3—down sloping)
No. of vessels	Ca	No. of major vessels (0–3) colored by fluoroscopy
Thalassemia	Thal	Defect types; 3—normal; 6—fxed defect; 7—reversible defect
Num(class attribute)	Class	diagnosis of heart disease status (0—nil risk; 1—low risk; 2—potential risk; 3— high risk; 4—very high risk)

V. ADVANTAGE AND DISADVANTAGE

Each and every model has some advantage and disadvantage. We give some advantage and disadvantage of these methods

Table 2 Advantages and Disadvantages.

Model	Advantages	Disadvantages
Decision Trees	Easy to understand and explain	Do not work best for uncorrelated variables.
K-Nearest Neighbour	Real if training data is large	Need to control values of parameter
Support Vector Machines	Valuable for non-linearly divisible data	Not suitable for High dimensional data.
Navie Bayesian Classifiers	Handles real and discrete data	Assumption is independence of features
Neural Networks	It is a non-parametric method.	Extracting the knowledge (weights in ANN) is very difficult

Table 3 Comparison of Various Techniques.

Author	Techniques	Disease	Tool	Accuracy
Otoom et al.	Bayes Net	Heart Disease	WEKA	84.5%
	SVM			84.5%
	Functional Trees			84.5%
Vembandasamy et al.	Naive Bayes	Heart Disease	WEKA	86.419%
Parthiban et al.	Naive Bayes	Heart Disease	WEKA	74%
S. Anitha, DN. Sridevi	KNN	Heart Disease	R Programming	76.67%
	SVM	Heart Disease	R Programming	86.6%
	Naïve Bayes	Heart Disease	R Programming	77.7%
YounessKhourdifi, Mohamed Bahaj	K-NN	Heart Disease	WEKA	99.65 %
	SVM	Heart Disease	WEKA	83.55%
	RF	Heart Disease	WEKA	99.6%
	NB	Heart Disease	WEKA	86.15%
	MLP	Heart Disease	WEKA	91.65%

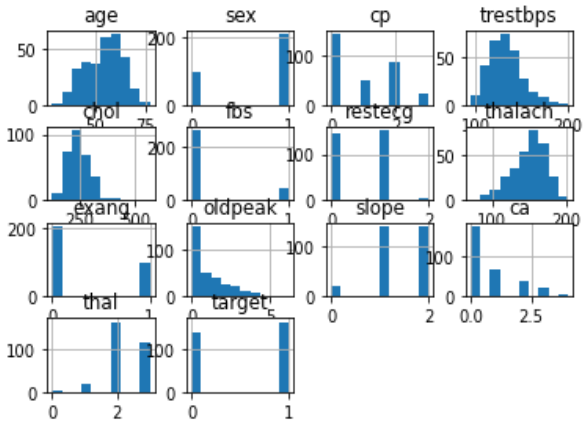


Figure 3. Distribution of data.

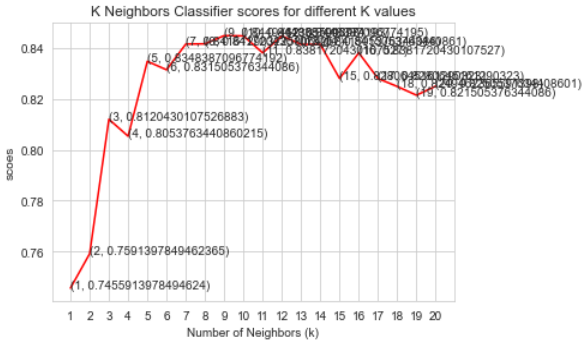


Figure 4. K Neighbors Classifier Scores

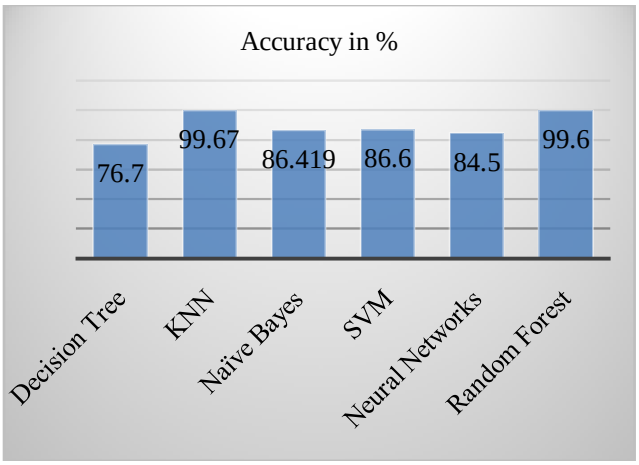


Fig: 5. Comparison of Data Mining Techniques Accuracy.

VI.CONCLUSION

The prediction analysis, which is the data mining approach mentioned, is used to improve future predictions based on existing data. The classifier algorithm assigns a class to the data. This study reviews and analyses different prediction analysis methods in terms of several factors. A literature review is conducted on several classification approaches such as Decision Tree, K-Nearest Neighbor, Support Vector Machines, Naive Bayesian Classifiers, Neural Networks, and Random Forest of prediction analysis from which the challenge is formed. In contrast to other classifiers, Random Forest and K-Nearest Neighbour perform better. In the future, the defined problem can be solved to improve the accuracy of prediction analysis.

REFERENCES

[1] M. Akhiljabbar&Dr.PritiChandrab "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012.

[2] G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients", International Journal of Applied Information Systems, Vol.3, No.7, pp.2249-0868, 2012

[3] M. AkhilJabbar, B.L Deekshatulu&Priti Chandra "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection" Global Journal of Computer Science and Technology Neural & Artificial Intelligence Volume 13 Issue 3 Version 1.0 Year 2013

[4] International Research Journal Publisher: Global Journals Inc. (USA)

[5] N S Nithyaand K Duraiswamy "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface" Sadhana¯ Vol. 39, Part 1, February 2014, pp. 39–52. Indian Academy of Sciences

[6] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", IJISET-International Journal ofInnovative Science, Engineering & Technology,Vol.2, pp.441-444, 2015.

[7] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease", InternationalJournal of Software Engineering and Its Applications, Vol.9, No.1, pp. 143-156, 2015 .

[8] Anoopkumar M, &Rahman, A. M. J. M. Z, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration, International Conference on Data Mining and Advanced Computing (SAPIENCE), (2016)

[9] Nikhil N. Salvithal " Appraisal Management System using Data mining "International Journal of Computer Applications (0975 – 8887) Volume 135 – No.12, February 2016

[10] Tanvi Sharma, Anand Sharma &VibhakarMansotra "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016.

[11] Lee, E., Jang, Y., Yoon, D.-M., Jeon, J., Yang, S., Lee, S, "Kim, K.-JGame Data Mining Competition on Churn Prediction and Survival Analysis" using Commercial Game Log Data Transactions on Games, IEEE, 2018

[12] Dr. S. Anitha, Dr. N. Sridevi"Heart Disease Prediction using Data Mining Techniques" Journal of Analysis and Computation (JAC) (An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861 Volume XIII, Issue II, February 2019.

[13] YounessKhourdifi Mohamed Bahaj "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization" International Journal of Intelligent Engineering and Systems, Vol.12, No.1, 2019.

[14] K. Rajalakshmi, Dr. S. S. Dhenakaran and N. Roobin (2015), "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, 2015, pp. 1023-1028.

[15] BalaSundar V, T Devi and N Saravan, (2012) "Development of a Data Clustering Algorithm for Predicting Heart",International Journal of Computer Applications, vol. 48, 2012, pp. 423-428.

[16] DaljitKaur and KiranJyot (2013), "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, vol. 2 2013, pp. 724-729

[17] Ming, J., Zhang, L., Sun, J.& Zhang, Y, "Analysis models of technical and economic data of mining enterprises based on big data analysis", International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018, IEEE,