International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

Comparative Study of Heart Disease Prediction using Classification

Praneeta Desai[1], Aagam Shah[2], Rohan Shah[3], Mohit Vadsak[4]

[1-4]*Student, Dept. of Information Technology, K.J. Somaiya College of Engineering, Maharashtra, India*

Abstract - *Amongst varied life-threatening diseases, cardiovascular disease is a significant cause of dismal and mortality in modern lifestyle owing to which it has amassed recognition amongst research fields in medical domain. Predicting cardiovascular illness is complex and intricate task, thus, automation in prediction of heart disease in a patient is of utmost benefit so that it aids in further line of treatment promptly and effectively. The diagnosis is based on signs and symptoms of illness along with medical examination performed physically. This paper explores and evaluates different classification algorithms, as it is proven that classification algorithms are significant indicators as compared to clustering algorithms. This paper discusses Classification of Prediction of cardiovascular disease on patients using medical parameters such as age, gender, cholesterol levels, blood pressure monitoring, obesity ratio and so on. This paper surveys and provides a comparative analysis between classifier algorithms, namely Naïve Bayesian and K-Nearest Neighbours. The main rationale of this paper is to demonstrate the correlations between parameters leading to prediction of the disease. Heart Disease UCI Dataset acquired experimental data that is used in the paper to produce the findings which were derived using the Rapid Miner Machine Learning Application.*

Key Words: Data Mining, Heart Disease Prediction, Naïve Bayes Algorithm, K Nearest Neighbours, Medical Diagnosis, Predictive Analysis, Classification Methods, Rapid Miner.

I. Introduction

Cardiovascular disease is a category of diseases that engage the cardiac muscles and vessels. Cardiovascular disease comprises diseases that clog arteries such as angina and cardiopulmonary arrest which is otherwise called as a heart attack. Cardiovascular disease is the common cause of death in many regions, smoking being the prime reason behind cardiovascular diseases. Smoking increases the formation of plaque in blood vessels due to which flow of blood in the blood vessels is thwarted. More than 6 trillion cigarettes are sold round the world annually, that is around 18000 million cigarettes bought/consumed per day. It is a profitable trade and produces its wealth mostly off the rear of the impoverish population all around the world, in regards of offer as well as need. Tobacco use has reached pandemic proportions worldwide, and, although endeavours to back-pedal smoking tendency, the matter alone seems to be dilating per year. Even if cardiovascular disorders have been reported as a major cause of death, they have been identified as the most treatable and controlled disorders. The full and successful management of the condition is dependent on a well-timed measurement of the condition. Every person's body will have different symptoms of cardiac failure, which can differ from time to time. They might have back pain, jaw pain, neck pain, abdominal pain, and tinnitus, chest pain, and discomfort in the shoulders and arms. There are a host of common heart disorders, including cardiac attack and stroke and coronary artery disease.

Data Mining can assist us to evaluate patient data and then help us determine whether or not a person may suffer from Heart Disease. There are very few existing applications that help users predict if they are suffering from cardiovascular disease and are very expensive, for instance, "Apple watch" which suggests if the user should consult a physician if abnormal heart activity is detected, such gadgets are not affordable for all. Hence Heart Disease Prediction with the greatest precision which is also economic is the need for an hour. Heart Disease is not only based on smoking but multiple factors that could include stress, any previous history of medical illness and smoking is often known to be one of the main contributing factors for cardiovascular disease. Extensive work is underway to classify risk factors for heart disease in various patients, various researchers are using specific statistical methods and multiple data mining techniques systems. Statistical research has identified risk factors for cardiovascular disease encompassing smoking, sex, blood pressure, diabetes, cholesterol, hypertension and age, obesity, and lack of exercise. Awareness of heart disease is very important for the prevention and health of patients who are about to suffer from cardiovascular disease.

The next segment to be discussed are mentioned beneath. Section II surveys the Literature Review. Section III explains the Naive Bayes Classifier with its implementation in Rapid Miner Tool. Section IV describes the K-Nearest Neighbours Classifier along with its implementation in Rapid Miner Tool. This paper discusses and draws comparisons amongst the classifiers in Section V. The paper summarizes and concludes in Section VI.

II. Literature Review

Through this analysis we see the particular application of the Data Mining Method that is used in Heart Disease Prediction.

Y. Aslandogani alp, based on 3 separate classifiers such as KNN(K-Nearest Neighbour), Naïve Bayesian, Decision Tree, and utilized the rule of Dempsters as a decisive decision on all three points of view.[1]

Carlos Ordonez (2004), Examined the question of understanding and forecasting the partnership law for heart disease. A dataset including the personal background of

patients with heart failure, including details of risk factors, assessment of restricted artery and cardiac perfusion, was obtained. Each of these limitations was stated in order to reduce the amount of services as follows:

- The attributes may occur at one side of the rule.
- The law would group various characteristics into separate groups.
- The amount of options required under the law is arranged mainly by the personal records of individuals with heart disease. The scientist anticipated the incidence or non-apparition of cardiac disease in four main veins using two clusters of rules.[5]

Using Sellappan Palaniappan data mining software, et. Al. (2008) IHDPS-Intelligent Cardiovascular Prediction Framework, i.e. Naïve Bayes, Neural Network and Decision Trees. Each mechanism has its own authority to bring the best outcomes forward. Paradigming this method has been the unknown designs and association among them.[6]

Chaitrali S.D., (2012), used full sum of input characteristics to explore a theoretical framework for heart syndrome. A few words linked to medical conditions such as blood pressure, age, cholesterol and 13 other characteristics such as this is recycled to identify a single person or patient's heart attack. He has always used two separate characteristics such as smoking and obesity. Besides data mining results, Decision trees, neural networks, and naïve bayes were used to test the database for heart disease.[3]

S. Seema et al.[4] focuses on methods for predicting clinical disease using algorithms like Support Vector Machine(SVM), Decision Tree, Naïve Bayes and Artificial Neural Network(ANN) with the help of information stored in historical patient records. Aiming to determine the optimum value over an appropriate scale, a comparative analysis is performed on the classifiers. SVM offers the peak rating of accuracy from this test, whilst Naïve Bayes provides the best accuracy for diabetes.

Ashok Kumar Dwivedi et al.[4] has suggested various algorithms such as Naive Bayes, K-Nearest Neighbour, Logistic Regression, Classification Tree, Support Vector Machine and ANN. Logistic Regression provides greater precision in comparison to various different algorithms.

Chala Beyene et al.[2] suggested calculating and evaluating the incidence of heart disease utilizing knowledge discovery in databases methods. The primary goal is to envision the occurrence of heart failure way before any complication arises and that too in a brief amount of time. The suggested approach is also appropriate for practitioners who have no further experience and expertise in healthcare institutions. It comprises of various medical measures such as blood sugar, pulse rate, sex and age are some of the criteria used to assess whether or not the patient has a heart failure. Dataset processing is performed using WEKA tools.

III. Naïve Bayes

Naïve Bayes Classifiers is an algorithm centred onto the Bayes Theorem. This is not a singular algorithm, but a collection of algorithms, all of which must follow a shared concept, i.e. each pair of features being defined is independent of each other. Bayes' Theorem states an incident is expected to arise, despite the probability of another case that has already happened. Bayes' theorem is mathematically described as the equation below:

$$P(X|Y)=\frac{P(Y|X)*P(X)}{P(Y)}$$

Naive Bayes Algorithm is used on the UCI Heart Disease Dataset in order to determine the likely hood of cardiovascular disease. The model has been trained on 14 various attributes ranging from the Age, Sex, Resting Blood Pressure and various other factors. The algorithm's accuracy is seen from 60-80% and is based on no of attributes taken into account. The Algorithm has been implemented on Rapid Miner Studio Tool.

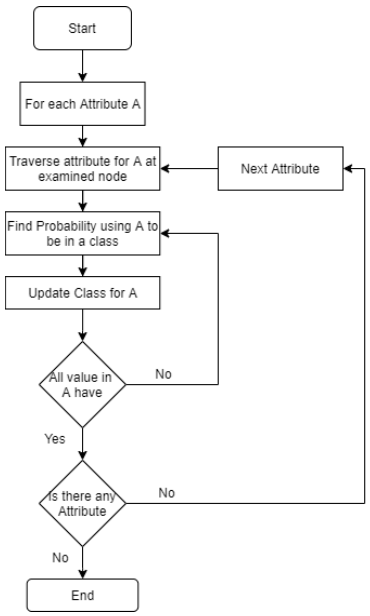

Figure 1 – Naïve Bayes Work Flow

3.1. Components used in Rapid Miner

- Selection
  1. ReadCSV - Dataset is imported and read in csv format.
- Preprocessing
  1. NumericaltoBinomial - Converting particular data fields to Binomial values from Numerical ones

2. SetRole - Setting the 'Target' attribute as the predictive mark
3. SplitData - Splitting the Data into two parts i.e. Test Data and Train Data

- Data Mining
  1. Naïve Bayes Component is used to create the model for prediction.
- Pattern Recognition
  1. ApplyModel – In order to apply the Model and predict the output for the test data
  2. Performance - Performance is used to find out the accuracy of the Model as well as to view the confusion matrix

3.2 Design Setup for Rapid Miner

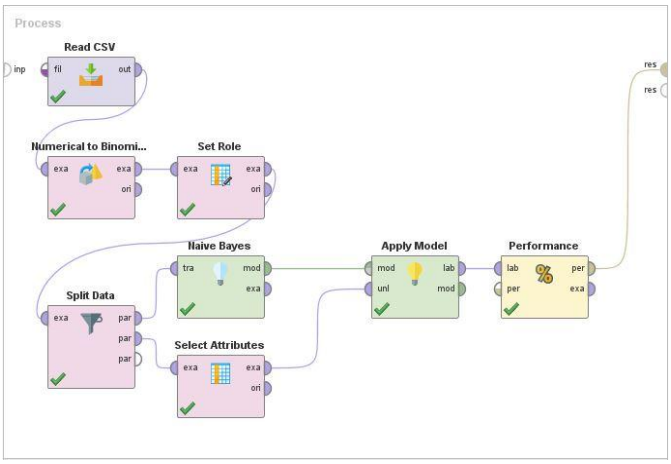

Figure 2 - Design Setup for Rapid Miner

3.3 Accuracy of Naïve Bayes

This Design factor was implemented which yielded the accuracy of about 78.02% when we considered 14 Attributes of the Data set.

•	Label-False	Label-True	Prediction
Pred-False	30	9	76.92
Pred-True	11	41	78.85
recall	73.17	82.00	•

Table 1 – Prediction Matrix for Naïve Bayes

3.4 Advantages of Naïve Bayes

- Naïve Bayes is convenient to execute.
- Requires comparatively less training data and hence decreases the time needed to train the data and correspondingly time required to implement is also lesser.
- If the statement of the above described individual attributes is valid than this algorithm works more effectively than different classification methods.

3.5 Disadvantages of Naïve Bayes

- One of the drawbacks of this algorithm is the fact that the features are independent which in real-life implementations is hardly true.
- If the classification of any categorical factor isn't included in the learning data collection, the algorithm assigns null probability to that class and no prediction can be rendered such that it is considered zero frequency.
- We may use the smoothing technique to solve zero frequency. Another of the simplest smoothing methods is Laplace Estimation.

3.6 Applications of Naïve Bayes

- **Prediction in Real-Time:** Because Naïve Bayes is fairly quick, forecasts can be made in real time.
- **Multi-class foresight:** This approach will estimate the subsequent possibility of achieving a number of groups in the target variable.
- **Spam filtering/ Review of emotions (Sentiment analysis):** Naïve Bayes classifiers are commonly used in textual categorizations with higher rates of performance compared to other algorithms (because of their superior efficiency in multi-class problems and the law of freedom). Spam filtering (identification of spam email) and opinion analysis (identification of positive and negative consumer opinions in social network data) are more commonly utilized.
- **Program of suggestions:** In addition to algorithms such as collective filtering, the Naïve Bayes Classifier provides a Recommendation engine that incorporates neural networks and data mining techniques to process unknown knowledge and decide whether or not a customer would want a service.

IV. K-Nearest Neighbours

One of the most simple and important classification algorithms in Machine Learning is K-Nearest Neighbours. This belongs to the supervised learning and sees heavy use in pattern recognition, data processing and detection of intrusion.

This is commonly applicable in real-life settings because it is non-parametric, meaning it has no inherent assumptions regarding data distribution (as opposed to other algorithms such as GMM, which presume a Gaussian distribution of the data).

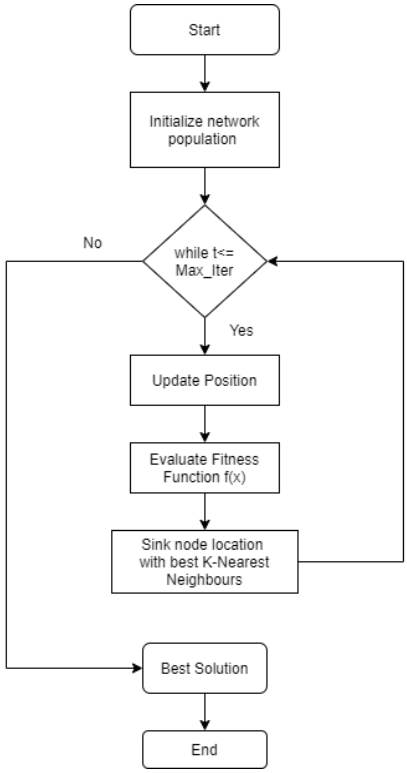

Figure 3 – KNN Work Flow

In K-Means the Distance is calculated using Euclidean Distance. The Algorithm is used as a Prediction Algorithm after it classifies the output parameter Diagnosis. The Formula of the Euclidean Distance is given as:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

4.1. Components used in Rapid Miner

- Selection
  1. ReadCSV - Dataset is imported and read in csv format.
- Preprocessing
  1. NumericaltoBinomial - Converting particular data fields to Binomial values from Numerical ones
  2. SetRole - Setting the 'Target' attribute as the predictive mark
  3. SplitData - Splitting the Data into two parts i.e. Test Data and Train Data
- Data Mining
  1. KNN Component is used to create the model for prediction.
- Pattern Recognition
  1. ApplyModel – In order to apply the Model and predict the output for the test data
  2. Performance - Performance is used to find out the accuracy of the Model as well as to view the confusion matrix

The Value of K was taken as 5 hence 5 neighbour will be considered while training the model.

4.2 Design Setup for Rapid Miner

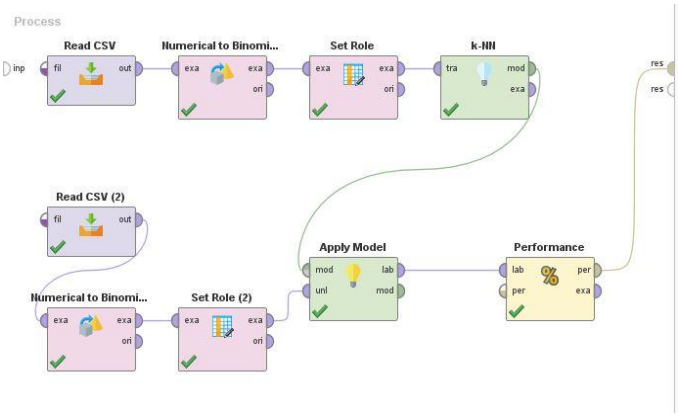

Figure 4 - Design Setup for Rapid Miner

4.3 Accuracy of KNN

This Design factor was implemented which yielded the accuracy of about 71.62% when we considered 14 Attributes of the Data set.

•	Label-False	Label-True	Prediction
Pred-False	89	37	70.63
Pred-True	49	128	72.38
recall	64.49	77.58	•

Table 2 – Prediction Matrix for KNN

The Analysis of Cholesterol value and the prediction where the patent will get Heart Disease or not is shown in the following chart. The X-Axis represents various value of Cholesterol whereas Y - Axis depicts the two outcomes True and False predicting whether the patent will get Heart Disease or not.

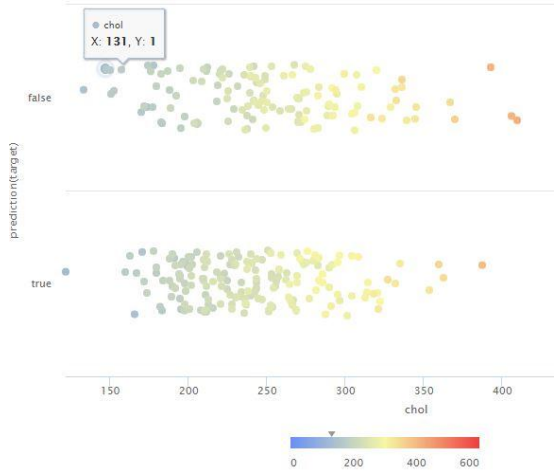

Figure 5 – KNN Prediction Statistics Plotting

4.4 Advantages of KNN

- The method is simple to understand and view.
- As in this algorithm there is no inference about data. it is really good for non-linear data.
- It has fairly high accuracy but other learning models are much better supervised than KNN.
- It is a flexible algorithm, since it can be used both for classification and regression.

4.5 Disadvantages of KNN

- This is a little costly algorithm computationally, since it holds all the training data.
- Large memory capacity needed with other supervised learning algorithms.
- In case of major N, estimation is sluggish.
- This is also prone to the size and insignificant features of the data.

4.6 Applications of KNN

- **Banking System:** In the banking system KNN can be used to determine if an entity is suitable for loan approval.
- **Calculating Credit Ratings:** Through contrasting with individuals with identical characteristics, KNN algorithms may be used to determine an individual's credit ranking.
- **Politics:** Using KNN algorithms, we can identify a prospective elector into various groups such as "Will Vote", "Would Not Vote," "Would Vote X to Party," Will Vote to Party Y " etc.
- **Other areas:** Speech Recognition, Handwriting Identification, Face Recognition and Video Recognition may be used with KNN Algorithm.

V. Discussions

- The Basic difference between the KNN classifier and the Naïve Bayes classifier is that the former is a prejudice classifier, whereas the latter is a phenomenological classifier.
- KNN is a sluggish classifier that is supervised and has local heuristics. As a sluggish classifier, it's hard to implement in real time for analysis. The judgment limits you reach with KNN they're even more complicated than other decision trees, despite a decent rating. In referring to a question that focuses primarily on the discovery of associations between results, KNN is best suited to optimizing locally due to its inherent nature.
- Naïve Bayes is an effective feedback classifier, so its a lot quicker than KNN. Nevertheless, it can be used for tracking in real-time. Usually, the Naïve Bayes classifier utilizes email spam screening.

VI. Conclusion

Cardiovascular Diseases are leading cause of illness and demise in males as well as females of varied ethnicity all over the world. Taking into account all the associated parameters i.e. 14 attributes, the model's accuracy is indicated by the design operator as 78.02\% for Naïve Bayes. The model's accuracy for K-Nearest Neighbours is estimated as 71.62\%. This paper takes 14 attributes into inspection, by which the tool gives a fairly reasonable accuracy. The main approach used for forecasting cardiovascular disease is completely relied on the accuracy of the test. There are several types of treatment for people who have been diagnosed with a rare category of cardiac failure. The Naïve Bayes Classifier is used to evaluate patients as it applies conditional and cumulative probability of two random occurrences. The KNN grouping is used for the study of patients related to each other in terms of characteristics. According to the paper's strategy, Naïve Bayes is more reliable than KNN. In addition, as noted in the literature survey, it is believed that only moderate progress is achieved in building a predictive model for patients with cardiac disease and that, as a result, hybrid and more complicated models are required to enhance the accuracy of predicting early onset of heart diseases.

REFERENCES

[1] Y Alp Aslandogan and Gauri A Mahajani. “Evidence combination in medical data mining”. In: International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.Vol. 2. IEEE.2004, pp. 465–469.

[2] Mr Chala Beyene and Pooja Kamat. “Survey on prediction and analysis the occurrence of heart disease using data mining techniques”. In: International Journal

ofPure and Applied Mathematics118.8 (2018), pp. 165–174

[3] Chaitrali S Dangare and Sulabha S Apte. “Improved study of heart disease prediction system using data min-ing classification techniques”. In: International Journal of Computer Applications47.10 (2012), pp. 44–48.

[4] Kumari Deepika and S Seema. “Predictive analytics toprevent and control chronic diseases”. In:2016 2ndInternational Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).IEEE. 2016, pp. 381–386.

[5] Carlos Ordonez. “Improving heart disease prediction using constrained association rules”. In:Seminar presen-tation at University of Tokyo. 2004.

[6] Sellappan Palaniappan and Rafiah Awang. “Intelligent heart disease prediction system using data mining tech-niques”. In:2008 IEEE/ACS international conference on computer systems and applications. IEEE. 2008,pp. 108–115.

[7] A. Massaro, G. Ricci, S. Selicato, S. Raminelli and A. Galiano, "Decisional Support System with Artificial Intelligence oriented on Health Prediction using a Wearable Device and Big Data," 2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT, Roma, Italy, 2020, pp. 718-723, doi: 10.1109/MetroInd4.0IoT48571.2020.9138258.

[8] V. Laijawala, A. Aachaliya, H. Jatta and V. Pinjarkar, "Classification Algorithms based Mental Health Prediction using Data Mining," 2020 5th International Conference on Communication and Electronics Systems (ICCES), COIMBATORE, India, 2020, pp. 1174-1178, doi: 10.1109/ICCES48766.2020.9137856.

[9] R. B. Jadhav and S. L. Patil, "The cardiovascular health informatics system using LabVIEW," 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, 2017, pp. 70-73, doi: 10.1109/I2CT.2017.8226096.

[10] R. B. Jadhav and S. L. Patil, "The cardiovascular health informatics system using LabVIEW," 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, 2017, pp. 70-73, doi: 10.1109/I2CT.2017.8226096.