iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5377

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Analysis of Supervised Machine Learning Algorithms

Priyank Bhardwaj[1], Jaydeep Kishore[2]
[1]*B. tech, 4th year student, Galgotias University, U.P., India*
[2]*Assistant Professor, Galgotias University, U.P., India*

Abstract: ***Data is collected by different sectors to enhance the productivity and to decrease the problems that are faced by that sector by analyzing the datasets that were created by different organizations of that sector. Health sector also collects an immense amount of data every day but unfortunately the health sector was not able to use that data to its full potential and that's why that data is underutilized. There are many types of diseases whose causes can be determined and the future generations can use that info for the betterment of the mankind. The cardiovascular diseases are also comes in this category and in this research paper we will use four different types of supervised machine learning algorithms on a single dataset that is collected by different hospitals of its cardiovascular patients that dataset contains the medical data such as age, gender , alcohol, hypertension, diabetes ,cigarette smoked per day and so on is taken as input and then these features are modelled for prediction and after the outcomes on the taken dataset we will be able to recognize that which supervised learning algorithm is best to predict the desirable result. The algorithms like K- nearest neighbour, Naive Bayes, Support vector machine and decision tree are used.***

I. INTRODUCTION

We all are running to find a better place in every aspect weather it is prosperity or wealth. We need to be ascetic but it seems like we have entirely different and opposite road. A road which lead us to a heaven of fortune but do we really know where this road ends? While racing in this fast pacing world on a road of success we often give our health as an alms to anxiety, hypertension and stress. Due to these things increase in use of cigarette and alcohol can be seen which finally leads to various kinds of diseases.

Despite, this much advancement in the medical field heart disease is one of the major cause of death in the world irrespective of age in both men and women. According to WHO 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Most cardiovascular diseases can be prevented by addressing behavioural risk factors like tobacco use, unhealthy diet and obesity, diabetes, BMI. In this research paper we have taken all those factors in account such as age, gender, height, heart rate, diabetes and so on. Since numerous factors are involved in heart disease. In this paper, we have tried prediction and analysis of heart disease by considering the parameters like age, gender, blood pressure, heart rate, diabetes and so on. As we can see that there are so many factors that can lead to cardiovascular disease this makes the prediction of this disease a difficult one.

A. *Major Symptoms Of Heart Attacks Are*

1) Chest pain, chest tightness, chest pressure and chest discomfort (angina)
2) Shortness of breath
3) Pain, numbness, weakness or coldness in your legs or arms if the blood vessels in those parts of your body are narrowed
4) Pain in the neck, jaw, throat, upper abdomen or back

B. *Causes of Cardiovascular Disease*

1) Heart defects you're born with (congenital heart defects)
2) Coronary artery disease
3) High blood pressure
4) Diabetes
5) Smoking
6) Excessive use of alcohol or caffeine
7) Drug abuse
8) Stress
9) Some over-the-counter medications, prescription medications, dietary supplements and herbal remedies
10) Valvular heart disease

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

II. LITERATURE REVIEW

Monika Gandhi used Naïve Bayes, Decision tree and neural network algorithms and analysed the medical dataset. There are a huge number of features involved. So, there is a need to reduce the number of features. This can be done by feature selection. On doing this, they say that time is reduced. They made use of decision tree and neural networks.

J Thomas, R Theresa made use of K nearest neighbour algorithm, neural network, naïve Bayes and decision tree for heart disease prediction. They made use of data mining techniques to detect the heart disease risk rate.

Sana Bharti, Shailendra Narayan Singh made use of Particle Swarm Optimization, Artificial neural network, Genetic algorithm for prediction. Associative classification is a new and efficient technique which integrates association rule mining and classification to a model for prediction and achieved good accuracy.

Purushottam proposed "An automated system in medical diagnosis would enhance medical care and it can also reduce costs. In this study, we have designed a system that can efficiently discover the rules to predict the risk level of patients based on the given parameter about their health. The rules can be prioritized based on the user's requirement. The performance of the system is evaluated in terms of classification accuracy and the results shows that the system has great potential in predicting the heart disease risk level more accurately".

Sellappan Palaniyappan, Rafiah made use of decision tree Naïve Bayes, Decision tree, Artificial Neural Networks to build Intelligent Heart Disease Prediction Systems (IHDPS).To enhance visualization and ease of interpretation, it displays the results both in tabular and graphical forms. By providing effective treatments, it also helps to reduce treatment costs. Discovery of hidden patterns and relationships often has gone unexploited. Advanced data mining techniques helped remedy this situation.

Himanshu Sharma, M A Rizvi made use of Decision tree, support vector machine, deep learning, K nearest neighbour algorithms. Since the datasets contain noise, they tried to reduce the noise by cleaning and pre-processing the dataset and also tried to reduce the dimensionality of the dataset. They found that good accuracy can be achieved with neural networks.

J.Vijayashree and N.Ch.Sriman Narayana Iyengar used data mining. A huge amount of data is produced on a daily basis. As such, it cannot be interpreted manually. Data mining can be effectively used to predict diseases from these datasets. In this paper, different data mining techniques are analysed on heart disease database. In conclusion, this paper analyses and compares how different classification algorithms work on a heart disease database.

Ramandeep Kaur, Er.Prabhsharn Kaur have showed that the heart disease data contains unnecessary, duplicate information. This has to be pre-processed. Also, they say that feature selection has to be done on the dataset for achieving better results.

J.Vijayashree and N.Ch. Sriman Narayana Iyengar used data mining. A huge amount of data is produced on a daily basis. As such, it cannot be interpreted manually. Data mining can be effectively used to predict diseases from these datasets. In this paper, different data mining techniques are analysed on heart disease database. In conclusion, this paper analyses and compares how different classification algorithms work on a heart disease database.

III. PROPOSED METHODOLOGY

Machine Learning is a process to feed machine enough data to train and predict a possible outcome using the algorithms at bay. There are three types of machine learning algorithms supervised learning algorithms, unsupervised learning algorithms and reinforcement algorithms. In this research paper the algorithms like K- nearest neighbour, Naive Bayes, Support vector machine and decision tree are used which are type of supervised learning algorithm.

A. Proposed Work

Dataset of the hearts patients are taken as input. After that the data is loaded in the platform and then the data is explored after exploring the data. The process of splitting the data is done. After that the particular model is generated on which we want to do the process after that the model evaluation is done.

There are four different methods used in this paper. The output is the accuracy metrics of the machine learning models. The model can then be used in prediction.

B. K-Nearest Neighbors (KNN)

K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. KNN doesn't have a discriminative function from the training data but it memorizes the training the training data, there is no learning phase of the model and all the work is done at the time of prediction is requested.

©IJRASET: All Rights are Reserved

A supervised classification algorithm in which we have some data points or data vectors which is separated into different number of several categories and tries to predict the classification of new sample from that particular population set. It classifies any new points on similarity measures that similarity measures for continuous variables, Euclidean distance, Manhattan distance and Minkowski distance measures can be used.

However, the commonly used measure is Euclidean distance. The formula for Euclidean distance is as follows:

$$d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

The data is divided into training and test sets. The train set is used for model building and training. A k- value is decided which is often the square root of the number of observations. Now the test data is predicted on the model built. There are different distance measures. For continuous variables, Euclidean distance, Manhattan distance and Minkowski distance measures can be used.

C. Support Vector Machine (SVM)

Support Vector Machine is a discriminative classifier that is formally designed by a separative hyperplane. It is a representation of examples points in space that are mapped so that the points of different categories are separated by a gap as wide as possible.

The main aim of the SVM is to segregate the given data in the best possible way. When the segregation is done the distance between the nearest points known as margin. The approach is to select a hyperplane with maximum possible margin between the support vector in the given dataset.

In some datasets the hyperplane may not be useful in that the SVM uses a kernel trick to transform the input into a higher dimensional space. Kernel is a set of mathematical functions. In this proposed methodology, linear kernel is used.

$$K(x, x') = \exp((-\|x-x'\|^2)/2\sigma^2)$$

D. Naive Bayes Algorithm (NB)

Naïve Bayes is a simple but surprisingly powerful algorithms for predictive analysis. It is a classification technique based on Bayes theorem with an assumption of independent amount of predictors.

It comprises of two parts Naïve and Bayes. Naïve Bayes classifiers assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if this feature depends on each other or upon to the adjacent of the other feature all these properties independently contribute to the probability. It is easy to built and particularly useful for very large datasets.

Given a hypothesis X and Evidence E, Bayes theorem states that the relationship between the hypothesis before getting the evidence P(X) and the probability of the hypothesis after getting the evidence P(X/Y) is as follows:

P(Y/X) = P(X/Y) P(X)

This calculates the probability of Y given X where X is the prior event and Y is the dependence event.

E. Decision Trees

A decision tree is a graphical representation of all the possible solutions to a decision based on certain conditions. It is called so because it starts with a root and then branches off to a number of solutions just like a tree. Even the tree starts growing its branches once it gets bigger and bigger, similarly in a decision tree it has a root which keeps on growing with increasing number of decisions and the conditions. In this research paper we have chosen a dataset which have several factors such as smoking, age, gender. The factor used in root node must clearly classify the data. We make use of age as the root node. The decision tree is easy to interpret. They are non-parametric and they implicitly do feature selection.

IV. RESULT AND DISCUSSION

A. Data source

The dataset used is taken from Kaggle named as cardio_train.

There were various attributes that were taken into account such as: Gender Age, Gender, Height, Weight, Cholesterol, Glucose, Smoke, Alcohol, Active, Cardio.

B. Results

After downloading the data the data from Kaggle the same dataset is processes using jupyter notebook on the anaconda navigator. The dataset is uploaded and the data is processed using all the four algorithm we chose before and then their accuracy is calculated

©IJRASET: All Rights are Reserved

The accuracy results are tabulated as follows:

Method	Accuracy
KNN	80.60%
NB	78.66%
Decision tree	72.58%
SVM	62.56%

The accuracy of K-nearest neighbor algorithm is good when compared to other algorithms.

V. CONCLUSION AND FUTURE WORK

This paper discusses the various machine learning algorithms such as support vector machine, Naïve Bayes, decision tree and k-nearest neighbour which were applied to the data set. It utilizes the data such as age, cholesterol, alcohol consumptions and then tries to predict the possible coronary heart disease patient in coming years.

Family history of heart disease can also be a reason for developing a heart disease as mentioned earlier. So, this data of the patient can also be included for further increasing the accuracy of the model.

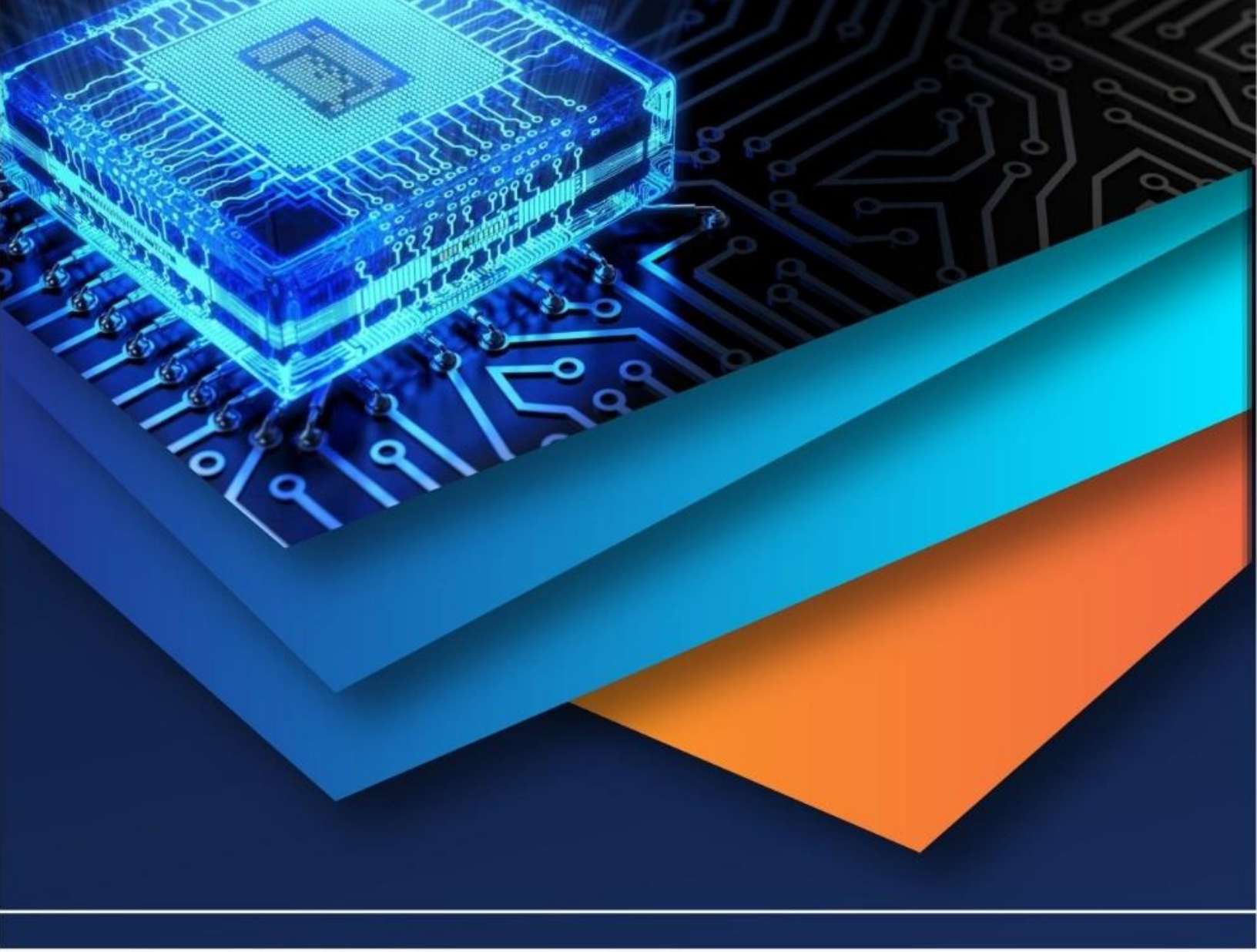
This work will definitely help the current generation and the generations to come in identifying the possible patients who may suffer from heart disease in the coming years. This may help in taking preventive measures and hence try to avoid the possibility of heart disease for the patient. So when a patient is predicted as positive for heart disease, then the medical data for the patient can be closely analysed by the doctors.

In this research paper we had only used four supervised algorithm in future the other types of supervised algorithm and also the different types of machine learning algorithm can be used to predict the data and the accuracy of those algorithm can be better than the algorithm used in this research paper.

REFERENCES

[1] Monika Gandhi, Shailendra Narayanan Singh Predictions in heart disease using techniques of data mining (2015)
[2] J Thomas, R Theresa Princy Human heart disease prediction system using data mining techniques (2016)
[3] Sana Bharti, Shailendra Narayan Singh, Amity university, Noida, India Analytical study of heart disease prediction comparing with different algorithms (May 2015)
[4] Purushottam, Kanak Saxena, Richa Sharma Efficient heart disease prediction system using Decision tree (2015)
[5] Sellappan Palaniyappan, Rafiah Awang Intelligent heart disease prediction using data mining techniques (August 2008)
[6] Himanshu Sharma,M A Rizvi Prediction of Heart Disease using Machine Learning Algorithms: A Survey (August 2017)
[7] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review (2017)
[8] V.Krishnaiah, G.Narsimha, N.Subhash Chandra Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review (February 2016)
[9] Ramandeep Kaur, 2Er. Prabhsharn Kaur A Review - Heart Disease Forecasting Pattern using Various Data Mining Techniques (June 2016)
[10] J.Vijayashree and N.Ch. SrimanNarayanaIyengar Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review (2016)

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)