ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

Heart Disease Prediction Using Machine Learning

Prof. Farhana Kausar[1], Karan Chaudhary[2]

Associate Professor, Department of Computer Science, Atria Institute of Technology, Bangalore, India [1]

Student, Department of Computer Science, Atria Institute of Technology, Bangalore, India [2]

ABSTRACT: Heart disease has become a major concern in today's lifestyle and in medical field. Due to leading Mortality rate, it has become a major concern. Most of the people of the world suffer from Cardio Vascular Disease. With all these major factors predicting of cardiovascular disease is very much important which helps in reducing the mortality rate. Diagnosing the heart disease is difficult and time-consuming task therefore there is a need of automated system that can predict the likelihood of heart disease of an individual. Such types of prediction help to improves the quality of decision making in medical field. A system that predicts the level of heart disease of a patient is developed based on the attributes such as age, sex, chest pain type, blood pressure, fasting blood sugar etc. The Dataset of the heart disease are introduced to python programming using Machine Learning Algorithm which predicts the disease in term of level of accuracy.

KEYWORDS: Decision Tree algorithm; Naïve Bayes Classifier; Machine Learning; Prediction; Diagnosis

I. INTRODUCTION

Machine Learning is one of the important aspects regarding prediction and analysis. With the help of Machine Learning Algorithm, we will able to predict and analysed any sort of data in an efficient manner. Machine learning has been successfully applied in a wide range area with excellent performance. In heart disease prediction, the data is introduced to various algorithm and the result is predicted based on the observation.

In this system, a heart disease data set is used. The main aim of this system is to predict the possibilities of occurring heart disease of the patients in terms of percentage.This is performed through data mining classification techniques. The classification technique is used for classifying the entire dataset into two categories namely yes and No. Classification technique is applied to the dataset through the machine learning classification algorithm namely Decision tree classification and Naïve Bayes Classification models. These models are used to enhance the accuracy level of the classification technique. This model performs both the classification and prediction methods. These models are performed using python Programming Language [1].

II. LITERATURE SURVEY

1. Efficient Heart Disease Prediction System Using Optimization Technique

The focus is developing a predicting algorithm with the help of datamining and optimization technique [1]. The proposed system makes use of Particle Swarm Optimization (PSO) technique. PSO is used with the constriction factor known as constricted PSO. A brief description about datamining and importance of datamining in health care field is been given [1]. The dataset for the proposed system is taken from the repository of University of California, Irvine (UCI). It is also known as Cleveland dataset. In the dataset, only 14 attributes are made used, namely age, sex, chest pain type, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression, slope of peak exercise ST segment, number of major vesselsand diagnosis of heart disease. Information about Particle Swarm Optimization and Constricted Swarm

Optimization is been given. The main objective is to develop an automated intelligent system that is used to predict the heart disease [1]. ButPSO is only applied to solve the high dimensional and complex problem.PSO is still too slow compared to classical approaches.

2. Heart disease prediction system based on hidden naïve bayes classifier

To apply hidden naïve bayes classifier, they used WEKA 6.4 tool. They ran their experiments on heart data downloaded from UCI [11]. Heart stalog data set consists of 14 attributes and 270 instances. HNB evaluation is performed using 10 fold cross validation. However, it will predict the data, but WEKA is unable to handle large amount of data sets. If we have a large dataset we will most likely run in to out of memory expectation.

3. Heart Attack Prediction System

The main aim is to develop an automated system to predict the heart disease with help of binary classifier [3]. There is a web based graphical user interface built and naïve Bayes algorithm is used to develop the classifier. The proposed system uses the dataset obtained from UCI's machine learning repository. Rapid miner is used for cleaning the dataset and to find the best-fitting algorithm for the given dataset and the 4 algorithms including naïve bayes, decision tree, K-nearest neighbour and random forest were compared by building their processes in Rapid Miner [3]. The outcome of this activity showed that naïve bayes gives the highest accuracy on the given dataset [3]. There is literature survey done by author of this on medical facts and datamining and analysis. The proposed system consists of binary classification model, which predicts the risk level of the patient based on his/her medical data. The system comprises of graphical user interface that is easy to use and understand. The dataset is referred to as Cleveland dataset. The dataset comprises of 14 attributes in total out of which 13 are predictor variables and one feature is a binary response variable. Naïve Bayes is used for classification process [3]. Web based user interface is developed at the end. Implementation involves three steps: Data pre-processing, classification, user interface. The prototype of the system that classifies an individual on his risk factor is implemented [3]. Matters involving web technology can be very complicated, and it would be difficult for someone without relevant experience.

4. Heart Disease Prediction Using Data Mining Techniques

The records were divided equally into 2 datasets: coaching dataset (455 records) and testing datasets (454records) to avoid bias, the records forevery set were hand-picked at random. In this paper, both training and testing data comprises of 50-50 which is not feasible in comparison to the data as training data should always be greater than testing data.

5. An empirical study on applying data mining techniques for the analysis and prediction of heart disease

The heart disease dataset is taken from UCI machine learning repository which contains 303 samples where 164 of them refer to healthy and the remaining 139 refer to heart disease and each of the sample included 14 features. This dataset is initially divided into two equal parts randomly. One set is used for training and the other set is used for testing. After setting up the training models, the results can be determined from the testing data. Here C is used to implement heart disease classification and prediction that is trained through artificial neural network and C# is used for interface. Learning Vector Quantization (LVQ), a prototype-based supervised classification algorithm [5] is also used. The network consists of three layers: input layer with 13 neurons, hidden layer with 6 neurons and output layer with 2 neurons. However, the technology employed uses three layers of neural network, also the use of C language will affect the performance of the system and it is program-oriented language.

III. PROPOSED SYSTEM

1. Design Considerations:

- Dataset should be in CSV file.
- Outliers should be detected and deleted.
- Data having low skewness factor should be skewed.
- Considered all possible way of prediction.
- Model Evaluation is necessary.
- Fine-tuning the model is considered.

2. Description of the Proposed Algorithm:

Aim of the proposed algorithm is to maximize the accuracy level of prediction by using machine learning i.e. Decision Tree Algorithm and Naïve Bayes Classifier. The dataset should be introduced to the decision tree algorithm and Naïve Bayes Classifier.In proposed system, dataset is trained and tested in the ratio of 70:30 with the help of decision tree. Naive Bayes Classifier is used here to classify each data based on the analysis of decision tree algorithm.

Naïve Bayes is probabilistic model that uses Bayes Theorem for classification. The key insight of Bayes' theorem is that the probability of an event can be adjusted as new data is introduced. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Bayes' Theorem Can be given as:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Where,
$P(A)$= Prior Probability: The probability of occurring of A.
$P(B)$= Priori Probability that the evidence itself is true. The probability of occurring of B.
$P(A|B)$=is a posterior probability: the likelihood of event A occurring given that B is true.
$P(B|A)$ = is a likelihood probability: The probability of event B given that A is true.

IV. SYSTEM ARCHITECTURE

1. Data set explanation:

The dataset used in this article is the Cleveland Heart Disease dataset taken from the UCI repository. The dataset consists of 303individual data. There are 14 columns in the dataset, which are age, sex, chest-pain type, resting blood-pressure, serum- cholesterol, fasting blood sugar,electrocardiographic,max_heart_rate,induced_angina,ST_depression,slope,vessels,diagnosis.The diagnosis field refers to presence of heart disease in the patient.

2. **Open the Dataset to perform EDA (Exploratory Data Analysis):**

In the pre-processing, we have to check if any value is missing, as missing values affects the accuracy of the model as training of the model. Replace all the missing values, as a complete dataset can produce better accuracy of the model during training. We have to check the data should be balanced.

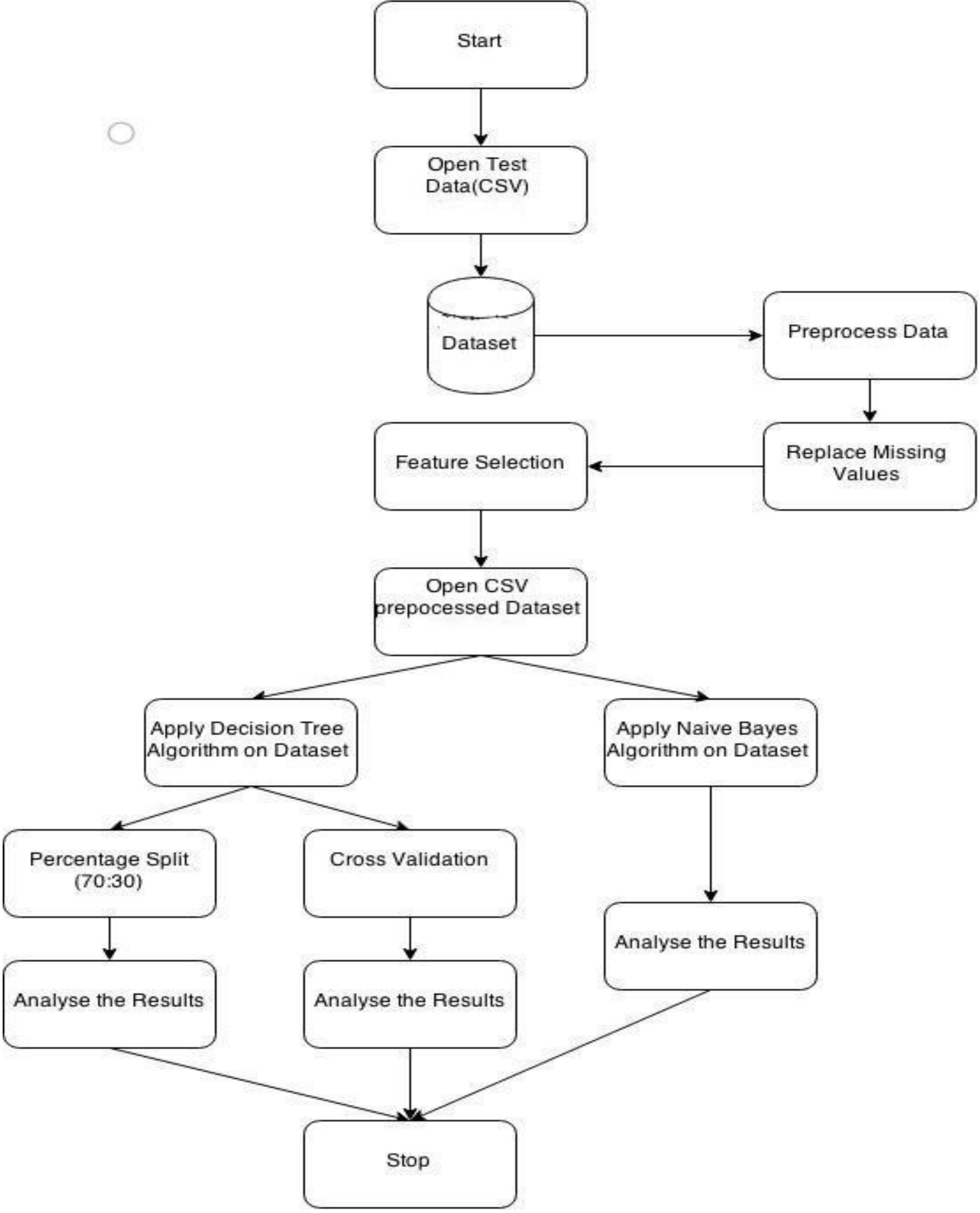

Fig. System Architecture Design

DOI: 10.15680/IJIRCCE.2020. 0803025

3. **Feature Selection***:*

After pre-processing and cleaning of the Dataset we have to perform feature selection. In this process we do analysis on each and every attribute (or feature) about their relevance in the dataset. We are going to check the co-relation between the attributes and we remove the features which have low co-relation value as they have very less effect in decision making and removing those will increase my model accuracy. Only the most affecting attributes can give us more accurate model. The presence of less informatic attributes affect the model accuracy and it also take more space and time.
After selecting the important features from the dataset, we have completed our data pre-processing.

4. **Model Building:**

i. Decision Tree Algorithm:

a. **Percentage split**, In this the dataset is divided into training set and testing set in the ration most preferably 7:3. After applying the decision tree algorithm on dataset analyze the result and accuracy of the model.

b. **Cross-validation** is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.
The general procedure is as follows:

- Shuffle the dataset randomly.
- Split the dataset into k groups
- For each unique group:
  - Take the group as a hold out or test data set
  - Take the remaining groups as a training data set
  - Fit a model on the training set and evaluate it on the test set
  - Retain the evaluation score and discard the model
- Summarize the skill of the model using the sample of model evaluation scores.

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.
After applying cross-validation and decision tree algorithm on dataset analyse the result and accuracy of the model.

ii. Naïve Bayes Classifier:

Naïve Bayes Classifier is used here to classify the data based on the object.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object and analyze the result based on Bayes Theorem.

V. CONCLUSION

The proposed system takes into consideration the data related to age, sex, chest pain type, blood pressure, fasting blood sugar etc and suggests the types of heart disease based on the various algorithm. Decision tree classify the dataset based on training and testing data, and then cross validate the respective data so that the validated data will pass for analysis. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. These two algorithms are applied to the same dataset in order to analyze the best algorithm in terms of accuracy. The combination of machine learning and the medical system help most of the doctor and patient in near future. In comparison with other, Machine Learning is best solution for medical problems (such as prediction of disease). The application of machine learning plays and important role in prediction of the disease based on the certain important factors and it also plays a major role in future evaluation.

VI. FUTURE ENHANCEMENT

The processing speed and efficiency can be improved by providing provision to connect to the server. Various other technique such as Artificial Intelligence, OpenCV, etc. can be implemented and processed to get further beneficial predicted output. Image data can be introduced to the system and real-time prediction can be done using various AI technique.

REFERENCES

1. "Chaitanya Suvarna, Abhishek Sali, Sakina Salmani", "Efficient Heart Disease Prediction System Using Optimization Technique",IEEE,978-1-5090-4890-8/17/$31.00 ©2017 IEEE,pp, 375-379,2017
2. "M.A Jabbar, Sirini Samreen", " Heart disease prediction system based on hidden naïve bayes classifier",,IEEE,2016
3. "Sushmita Manikandan", "Heart Attack Prediction System",IEEE,978-1-5386-1887-5/17/$31.00 ©2017 IEEE,pp. 817-820, 2017
4. "Abhishek Rairikar, Vedant Kulkarni, Vikas Sabale, Harshavardhan Kale","Heart Disease Prediction Using Data Mining Techniques",2017 International Conference on Intelligent Computing and Control (I2C2),pp. 1-8,2017
5. "S. Sivagowry, M. Durairaj, A. Persia", "An empirical study on applying data mining techniques for the analysis and prediction of heart disease", *Information Communication and Embedded Systems (ICICES) 2013 International Conference on*, pp. 265-270, 2013.
6. Ankur Gupta, Rahul Kumar, Harkirat Singh Arora, Balasubramanian Raman, "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis", *Access IEEE*, vol. 8, pp. 14659-14674, 2020
7. Gandhi, Monika, and Shailendra Narayan Singh."Predictions in heart disease using techniques of datamining." In 2015 International Conference on Futuristic Trends on Computational Analysis andKnowledge Management (ABLAZE), pp. 520-525. IEEE, 2015.
8. "A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, M. Abdar", "Using PSO algorithm for producing best rules in diagnosis of heart disease", *Proc. Int. Conf. Comput. Appl. (ICCA)*, pp. 306-311, Sep. 2017.
9. "C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, R. Deepika", "Analysis of neural networks based heart disease prediction system", *Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI)*, pp. 233-239, Jul. 2018.
10. "Ankur Gupta, Rahul Kumar, Harkirat Singh Arora, Balasubramanian Raman", "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis", *Access IEEE*, vol. 8, pp. 14659-14674, 2020
11. "C. Sowmiya and P. Sumitra", "Analytical study of heart disease diagnosis using classification techniques," *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Srivilliputhur, 2, pp. 1-5,2017.
12. "Purushottam, K. Saxena and R. Sharma", "Efficient heart disease prediction system using decision tree," International Conference on Computing, Communication & Automation, Noida, pp. 72-77,2015.