International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Heart Disease Prediction using Machine Learning Techniques

Raparthi Yaswanth, Y. Md. Riyazuddin

Abstract*: Recent advancement of technology allows the automation of things to be done using machine learning techniques. These machine learning techniques can also be used for detecting or predicting the heart disease in the early phase. The health care industry produces a huge amount of data which is in unstructured manner that cannot be understood by a machine. Due to development of modern technology, health care industries also managing the data in a structured manner which can be understood by machine learning technology. In this environment if we use machine learning algorithms for prediction of heart disease, then there is a chance to detect the heart disease status in the early phase and to alert patient to get a better treatment to cure that disease. This paper implements seven supervised learning algorithms which are KNN, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine and Neural Networks for heart disease prediction. This paper generates algorithm performance metrics like Accuracy, Precision, Recall, F-score and ROC values for how the system was predicting accurately. In this paper among those seven algorithms, Neural Networks gave best accuracy as 92.30% and this system provides experimental results for how the model is accurate for heart disease prediction.

Keywords* : Heart disease prediction, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbor, Neural Networks.

I. INTRODUCTION

The rapidly increasing medical data generated from hospital information system signifies the era of machine learning in the healthcare domain. Previously, hospitals used to store medical reports and patient data in paper-based record. By performing clinical test the physician can provide diagnosis details which may take chances to wrong decision. It leads to patient death due to lack of medical components. But the physician who is highly experienced can tell about decisions of diagnosis which is cost expensive and time consuming. In healthcare domain, every patient report will be stored in storage server, so that hospital admin can manage the data in structured way which can be understood by machine. The advantage of machine learning techniques is that the physicians can do analysis on their existing data which leads to lower cost and less diagnosis time. In the modern society, the younger and older generation also getting heart attack diseases. Because of their food habits and life style, risks of heart attack cases are being increased. The physicians are unable to predict the patient disease status in the earlier stage before they were getting in to the final stage. . So in these cases if any machine learning techniques can be used by physicians then there is a chance to predict the heart disease status in the earlier phase which can save many lives with perfect treatment. The hospital administrator can store patient's details which is having huge data in database server. Then the physicians can extract those data and make it useful for heart disease prediction with machine learning algorithms. The supervised machine learning algorithms will work on labeled dataset which contains independent and dependent attributes. The machine learning algorithms can work on huge data or complex data easily by classifying them, so these kind of techniques can reduce unknown diseases and corresponding tasks. These machine learning techniques help physicians to predict heart disease so that they can diagnosis them and check with proper treatment. The heart diseases can be predicted with the help of huge existing datasets and then by applying classifiers like KNN, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine and Neural Networks. These models can give more accuracy with their medical datasets which are stored in repository which shows the information in a meaningful way with machine learning techniques.

In this paper further sections include, section 2: Literature survey, where we discuss the related work of this concept. Section 3: Proposed methodology, we discuss about the system model, database collection, pre-processing, feature extractions and classification of our algorithm steps. Section 4: Results, we compare seven machine learning classification algorithms and show the result in graphs. Section 5: Conclusion, Section 6: References.

II. LITERATURE SURVEY

Sellappan Palaniappan et al. described in [1], how the Intelligent Heart Disease Prediction system was worked based on machine learning techniques such as Naive Bayes and Decision Tree. But it is web-based application and they have used .Net technology which is a dependent platform.

Peter C. Austin et al. [2] described that the doctors can divide the patients into two categories like with disease and without disease. Here to classify they have used tree based technique based on patient disease type, but this type of classification may have chances to get less accuracy. So that for improvement of model accuracy they included machine learning techniques like support vector machine and random forests.

Revised Manuscript Received on March 30, 2020.
***** Correspondence Author
Raparthi Yaswanth*, M Tech (Data Science) Department of CSE, GITAM University, Hdyerabad, India. E: mail: yaswanthraparthi@gmail.com
Dr. Y. Md. Riyazuddin, Assistant Professor, Department of CSE, GITAM University, Hdyerabad, India. E: mail: riyazymd@gmail.com

Shan Xu et al. [3] describes about Cardiovascular Disease prediction which is early detection and it is important to patient for treatment. This system focuses on providing more accuracy with help of support vector machine and naive bayes algorithms. This system consists of four parts like data interface which has hospitals raw data and data preprocessing for data integration and feature selection which is to retrieve only useful attributes for getting more accuracy and performing the classification of those attributes.

Mansoor et al. [4] said that this system is to develop and validate prediction model by implementing multivariate logistic regression, full and reduced random forest models. In this method eleven variables are included in final model based on backward elimination as well as the full random forest model will have 32 variables and they were mitigated up to 17 variables by feature selection for getting more accuracy which can be useful and accurate tool in clinical practice.

III. PROPOSED METHODOLOGY

A. System Model

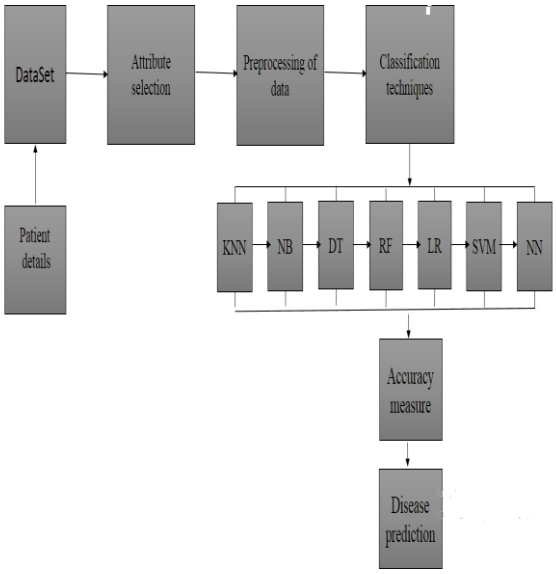
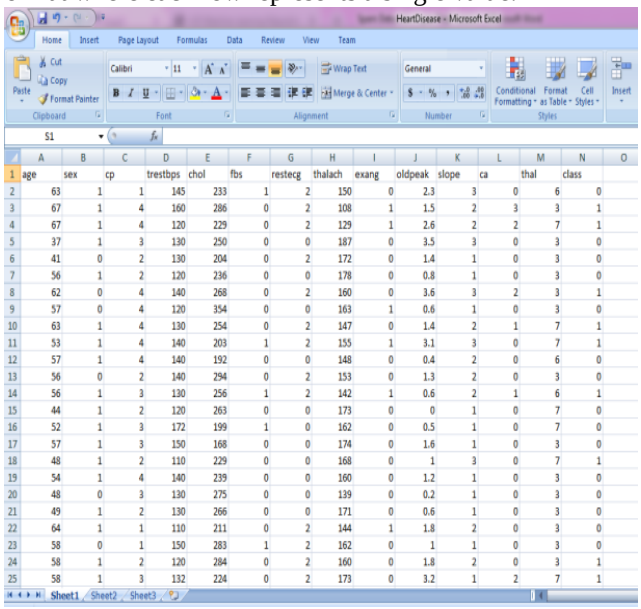


Fig. 1.System Architecture

In Figure.1, the system needs to load heart disease training dataset which is accessed from UCI repository and select best attributes for standardization and then we will do preprocessing if any row has null values or empty cells. Once preprocessing and feature extraction is completed, then it will split the dataset as training and testing and we choose machine learning classifiers like SVM, KNN, NB, DT, RF, LR and NN for heart disease prediction, while the prediction process take test data as input and it returns output as positive or negative with the help of training dataset. We also calculate the accuracy measures by splitting of dataset as 70% training and 30% for testing.

B. Data Collection

In this system the heart disease dataset is shown in Figure 2 which contains 0's and 1's where 0's indicates to negative and 1's indicates to positive status. Dataset has imported from UCI repository [5]. This dataset contains 303rows and among them 164 were negatives, 139 were positives. The dataset is stored in CSV (Comma Separated Value) file format where each row represents a single value.



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	class
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	1
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	1
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	1
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	1
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1
54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
48	0	3	130	275	0	0	139	0	0.2	1	0	3	0
49	1	2	130	266	0	0	171	0	0.6	1	0	3	0
64	1	1	110	211	0	2	144	1	1.8	2	0	3	0
58	0	1	150	283	1	2	162	0	1	1	0	3	0
58	1	2	120	284	0	2	160	0	1.8	2	0	3	1
58	1	3	132	224	0	2	173	0	3.2	1	2	7	1

Fig. 2.Heart Disease Dataset

C. Preprocessing

In this phase first we need to gather the training dataset which is in CSV file format. For that we need to read the file data with the help of pandas library for converting it to list array. The input message which is given by the user should get append to list array, because the machine cannot understand the file format data. On completion of converting process, it will remove null values if they are in training and testing data.

D. Standardization

In this, we need to apply standardization on given training dataset for scaling of each column data for getting best accuracy model. To perform scaling we use StandardScaler python library class which contains fit transform function and it takes input as dataset columns data which have higher values.

E. Classification Techniques

Naive Bayes

The Naive bayes machine learning algorithm is useful for categorizing documents and email spam filtering and this algorithm is working based on Baye's rule [6].

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (1)$$

Wherein:
A is a class
B is a message
P (A) is a class probability

P (B) is the probability of a message

P (B|A) is conditional probability of the class for the given message B

P (A|B) is conditional probability where message B belongs to class A.

In our system, for implementing Naive bayes algorithm we use python library which is named as sklearn. naive_ bayes. Gaussian NB class. This algorithm have function like fit () which will build the training model whose inputs are independent and dependent values of dataset and other function like predict() function which takes input as testing values and then it can predict the heart disease status as positive or negative.

K-Nearest Neighbor:

The K-nearest neighbor classifier [7] is a simplest algorithm for prediction of any dataset with the help of Euclidean distance. Here K means number of neighbor's, so if we take k=1 then it gives very nearest neighbor as output. It works on voting based system that means while prediction it takes nearest neighbors votes. This output gets more votes that will be the predicted output value.

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

Whereas:

D is a distance between x and y

K is a nearest neighbor

x and y are independent attribute values

In our system, for implementing K-nearest neighbor algorithm we use python library from the sklearn API for classification which is named as sklearn.neighbors. KNeighborsClassifier class. This algorithm has function like fit() which will build the training model whose inputs are independent and dependent values of dataset and another function predict() which takes input as testing values then it can predict the heart disease status as positive or negative.

Support Vector Machine:

The support vector machine [8] is supervised machine learning algorithm which is used for classification of instances. It can separate the data linearly and for non-linear data it can use kernel functions. The SVM classifies two classes with the help of hyper plane which has the largest margin to separate the dataset in to classes. The margin between the two classes represents the longest distance between closets data points of those classes which are called support vectors.

$$\vec{w}.\vec{u}+b \geq 1 \quad (3)$$

$$\vec{w}.\vec{u}+b \leq -1 \quad (4)$$

$$y\,(\vec{w}.\vec{u}+b)-1 \geq 0 \quad (5)$$

Wherein

b is a constant distance.

$\vec{w}$, $\vec{u}$ are vectors.

y is the output, where 1 is for positive samples and -1 for negative samples.

In our system, for implementing support vector machine algorithm we use python library which name was sklearn. Svm. SVR Class. This algorithm has function like fit () which is used to build the training model whose inputs are independent and dependent values of dataset and other function predict () which takes input as test message.

Decision Tree:

A tree has a nodes and branches. Decision tree is a classifier in the form of tree structure. The tree has two types of nodes which are decision nodes and leaf nodes. Here decision nodes specify a choice or test, based on this it can decide which direction it can go and leaf nodes indicates the classification of example or the value of example. The decision tree goes well with both classification and regression problems. The classification means, having a group of data and we are supposed to classify the data into predefined set of classes. The representation for the classification and regression tree (CART) is a binary tree. Each root node represents a single input variable (x) and a split on that variable. The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. In decision tree the greedy approach [9] is used to divide the input space called recursive binary splitting. In this procedure, different split points are trained and tested using a cost function. The split with the best cost is selected and all input variables and all possible split points are evaluated and chosen in a greedy manner [10] based on the cost function. For classification Gini cost function indicates how pure the nodes will become split points. The Gini index follows the below equation:

$$G=1-\sum_{i=1}^{m}(p_i)^2 \quad (6)$$

If the decision tree applies the Information Gain then it can follow the below equation.

$$E = -\sum_{i}^{m} p_i log2(p_i) \quad (7)$$

Where p_i denotes probability of class, E denotes entropy and G denotes Gini Index.In our system for implementing Decision Tree algorithm we use python library which is sklearn.tree.DecisionTreeClassifier class.

Logistic Regression:

The Logistic Regression [12] is a supervised machine learning algorithm where it can solve the classification problem. It works same as linear regression. It can be used for describing the data and explains about relationship between one dependent variable and one variable. It is a regression model where it can build to predict the probability with given input data which belongs to category "1". It can classify when a decision threshold came to picture which is very important to set the threshold value. The Threshold value can affect by the majority of precision and recall values.

$$P = \frac{1}{1+e^{-value}} \quad (8)$$

In our system for implementing Logistic Regression algorithm we use python library as sklearn. linear_ model. Logistic Regression class.

Random Forest:

Random Forest [12] is a supervised machine learning technique that constructs multiple decision trees. The final decision is made based on the outcome of the majority of the decision trees. The decision tree depends on a single tree which leads to suffer with low bias and high variances. To overcome this problem the random forest relies on many trees which introduces flexibility and converts high variance to low variance. The random forest process can follow the below steps:

Step-1: Construct bootstrapped dataset: Observe the randomness involved in constructing the bootstrapped data set (random sampling with replacement).

Step-2: Construct Decision tree using the bootstrapped dataset. While constructing the decision tree, the candidates for root node and for the rest of the nodes can be randomly selected.

Step-3: Repeat step-1 & step-2 to get more number of decision trees. This method takes many base learners i.e. decision trees and aggregating the decision of the majority population is called bagging.

The randomness involved in the training dataset makes the random forest classifier more accurate than unseen test dataset. Thereby it is yielding in low variance compared to decision trees. In our system for implementing Random Forest algorithm we use python library which is named as sklearn. ensemble. RandomForestClassifier class.

Neural Networks:

The neural network [11]technique is a collection of nodes where every node is associated with their respective weights. Nodes calculate sum of weights and it forwards to the activation function. This activation function generate or defines a particular output for a given node based on input.

$$Y=f\ (\sum_i^n x_i W_i) \qquad (9)$$

The activation function is of three types: linear function, heaviside step function and sigmoid function. The linear function is simple where it can calculate the sum of weights and then it forwards linearly. The heaviside step function is a condition based function. It provides two states only either 1 or 0. It can return 1 if the weighted sum (v) is greater than or equal to threshold (a) value otherwise it returns 0. The sigmoid function is a complexity function. It can depend on input.

$$f(v) = \frac{1}{1+e^{-v}} \qquad (10)$$

In our system for implementing Neural Networks algorithm we use python library named as sklearn. neural_ network.MLPClassifier class.

F. Evaluation Metrics

The metrics shows the performance of machine learning classifiers and it calculates how many misclassifications it can generate. The below metrics will show how it effects on our proposed system:

True Positive (TP): In our system, given input is expected as positive and also it predicts output as positive.

False Positive (FP): The input is expected as negative but it predicts as positive.

False Negative (FN): It expects input as negative but it predicts as positive.

True Negative (TP): In our system, given input is expected as negative and also it predicts output as negative.

So, by using above metrics we can calculate Accuracy, Precision, Recall, F-score of our proposed classifiers.

Accuracy: It correctly classifies the classes as True Positive and True Negative over total number of classification.

$$Accuracy = \frac{(TN+TP)}{TN+TP+FN+FP} \times 100 \qquad (11)$$

Precision: It retrieves fractions of messages which are relevant for users.

$$Precision = \frac{TP}{TP+FP} \qquad (12)$$

Recall: It retrieves fractions of messages that are relevant to query.

$$Recall = \frac{TP}{TP+FN} \qquad (13)$$

F-Score: It represents the mean of precision and recall.

$$\text{F-score} = 2 \times \frac{P \times R}{P+R} \qquad (14)$$

Where P is a precision, R is a recall.

IV. RESULTS

In our experimental results of heart disease prediction model, we compared seven different algorithms and generated the performance metrics like accuracy, precision, recall, f-score and roc. The results were shown in table 1. We designed and developed our implementation in the Windows operating system having 8 GB Ram and 1 TB HDD. Our Results concluded that Neural Networks algorithm gives best results compared to KNN, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine.

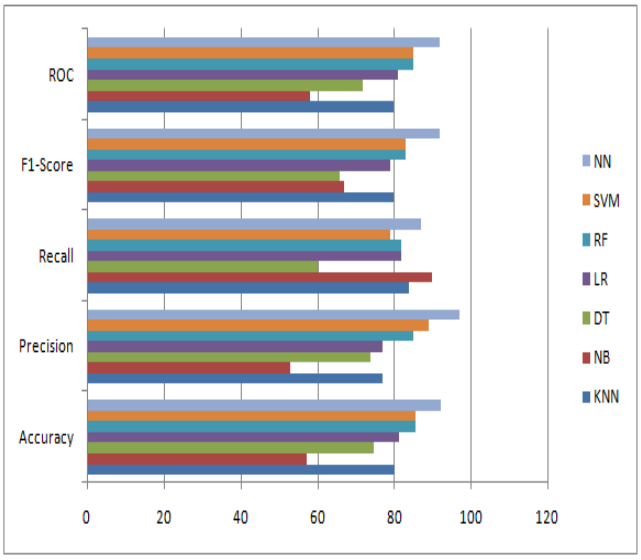


Fig.3. Classification Performance measurements

Table 1: Results Table

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC
KNN	80.21	0.77	0.84	0.80	0.80
NB	57.14	0.53	0.90	0.67	0.58
DT	74.72	0.74	0.60	0.66	0.72
LR	81.31	0.77	0.82	0.79	0.81
RF	85.71	0.85	0.82	0.83	0.85
SVM	85.71	0.89	0.79	0.83	0.85
NN	92.30	0.97	0.87	0.92	0.92

V. CONCLUSION

This paper deals with various machine learning techniques and comparing them to know which technique among them gives best results for accurate prediction of heart diseases. The algorithms and techniques involve ensemble methods and the combination of artificial Neural Networks provide better accurate results. In future, it is expected to use the above techniques for eliminating the existing drawbacks and improve the survival rate for the wellbeing of mankind.

REFERENCES

1. S. Palaniappan, R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," IJCSNS International Journal of Computer Science and Network Security, vol. 8, no. 8, August 2008.
2. P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, D. S. Lee, "Using Methods from Data Mining and Machine Learning Literature for Disease Classification and Prediction: a Case Study Examining Classification of Heart Failure Subtypes," Journal of Clinical Epidemiology 66 (2013) pp. 398-407, 2013.
3. S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan and T. Zhu, "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework," 2017 IEEE 2nd International Conference on Big Data Analysis, 2017.
4. H. M. Islam, Y. Elgendy, R. Segal, A. A. Bavry and J. Bian, "Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach," Journal of Heart & Lung, pp. 1-7, 2017.
5. Heart disease Data Set from UCI Machine Learning Repository,https://www.kaggle.com/ronitf/heart-disease-uci.
6. F. Peng, "Augmenting Naive Bayes Classifiers with Statistical Language Models", Computer Science Department Faculty Publication Series", Paper 91, 2003.
7. M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.
8. J. S. Raikwal,Kanak Saxena "Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set",pp. 35-39 July 2012.
9. A. A. Pathan, M. Hasan, M. F. Ahmed, and D. M. Farid, "Educational Data Mining: A Mining Model for Developing Students Programming Skills," 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2014.
10. M. Kamber and P. J. Han, Data Mining Concepts, and Techniques, 3rd ed., 2012.
11. S.P. Rajamhoana, C. Akalya Devi "Analysis of Neural Networks Based Heart Disease Prediction System",4-6 July 2018.
12. S. M. M. Hasan1, M. A. Mamun2, M. P. Uddin3 "Comparative Analysis of Classification Approaches for Heart Disease Prediction", 8-9 Feb. 2018.

AUTHORS PROFILE

Raparthi Yaswanth, is currently persuing his M.Tech (Data Science) in the department of Computer Science Engineering at GITAM University, Hyderabad. He has a degree of B.Tech in Information Technology from Gitam University, Visakhapatnam. His main area of interest include Machine Learning and Deep Learning. E:mail: yaswanthraparthi@gmail.com.

Dr. Y. Md. Riyazuddin, completed his PhD in CSE, and working as Assistant Professor in CSE Dept at GITAM University,Hdyerabad.His Research Areas include Networks and Security,Machine Learning, Neural Networks.He has published good number of Publications in Reputed Journals. He is Reviewer Member for various Reputed Journals. E: mail: riyazymd@gmail.com.