International Journal of Latest Trends in Engineering and Technology
Vol.(16)Issue(1), pp.073-077
DOI: http://dx.doi.org/10.21172/1.162.12
e-ISSN:2278-621X

A PROGNOSIS ON CARDIAC INFARCTION USING IMPLIED DATA MINING CLASSIFICATION ALGORITHMS

S.Padma[1], K.Yasudha[2]

Abstract: - The health care industry produces a huge amount of data. This data is always made use to the full extent. Using this data, a disease can be detected, predicted or even cured. A big threat to human kind is caused by diseases like heart disease, Cancer and Tumor. The aim is to develop a system for heart disease prediction using machine learning techniques. Machine learning algorithms produce quality results using health care data that help us to predict the heart disease in less amount of time in an efficient manner. Heart disease is the Leading cause of death worldwide. The medical data parameters such as Blood pressure, hypertension, diabetes, and cigarette smoked per day and so on is taken as input and then these features are modeled for prediction. This model helps us to predict future medical requirements in data. The aim is to predict the outcome feature of the data set. The outcome can contain only two values that is 0 and 1. 0 means disease not exist and 1 means disease. The system is built by using the classification model that can predict the Outcome feature of the test dataset with good accuracy among all. The accuracy of the model along with the accuracy of the algorithm is calculated. Then the one with a good accuracy is taken as the model for predicting the heart disease.

Keywords: KNN, Logistic Regression Algorithm, Decision Trees, Random Forest, Naive Bayes, SVM, SGD classifier, XG Booster, GBM.

I. INTRODUCTION

Healthcare is the improvement of health via the prevention, treatment, recovery, or cure of disease, illness, injury, and other physical and mental impairments in people. Daily, healthcare industry generates a large amount of data about patients, disease, and treatment, etc. The data- powered revolution in health care is well under way and looks forward to seeing how innovations continue to shape and improve patient care.

In today's era heart disease is the most common disease and primary reason for deaths. WHO- World Health Organization has estimated 17.9 million deaths globally in each year because of heart disease. Machine learning is the process of teaching machines to recognize patterns by providing data and an algorithm to work with the data. One of the most important domains using machine learning is the healthcare industry.ML algorithms can process massive amount of medical data at rapid speed. Data mining helps physicians in making the appropriate decision for treatment and prediction of heart disease in early stages which help in preventing disease or reduce its effects.

II. CLASSIFICATION ALGORITHMS

A. K-Nearest Neighbors Classifier –

KNN is one of the supervised machine learning algorithms that can be used for data mining as well as machine learning. Based on the similar data, this classifier learns the patterns present in it. It is a non-parametric and a lazy learning algorithm. By non-parametric, it means that the assumption for underlying data distribution does not hold

[1] *PG Student, Department of Computer Science, GIS, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India.*
[2] *Assistant Professor, Department of Computer Science, GIS, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India.*

valid. In lazy loading, there is no requirement for training data points for generating models. The training data is utilized in testing phase causing the testing phase slower and costlier as compared with the training phase.

B. Logistic Regression –

This is the most popular ML algorithm for binary classification of the data-points. With the help of logistic regression, we obtain a categorical classification that result in the output belonging to one of the two classes. For example, predicting whether the price of oil would increase or not based on several predictor variables is an example of Logistic Regression.

Logistic Regression has two components – Hypothesis and Sigmoid Curve. Based on this hypothesis, one can derive the resultant likelihood of the event. Data obtained from the hypothesis is then fit into the log function that forms the S- shaped curve called „sigmoid". Through this log function, one can determine the category to which the output data belong.

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

b0 and b1 are the two coefficients of the input x. We estimate these coefficients using the maximum likelihood function.

C. Decision Trees Classification –

Decision trees can be used to classify data, and they cut off possibilities of what a given instance of data might be by examining a data points features. A decision tree is a series of nodes, a directional graph that starts at the base with a single node and extends to the many leaf nodes that represent the categories that the tree can classify. It uses a tree-like graph to show the predictions that result from a series of feature- based splits.

The output of decision trees is interpretable. It can be understood by people without analytical or mathematical backgrounds. It does not require any statistical knowledge to interpret them. It can enable analysts to identify significant variables and important relations between two or more variables.

Decision trees are resilient to outliers and missing values, they require less data cleaning than some other algorithms. It can make classifications based on both numerical and categorical variables. It is a non-parametric algorithm, as opposed to neural networks, which process input data transformed into a tensor, via tensor multiplication using large number of coefficients, known as parameters.

D. Random Forest Classifier –

Random forests are made of many decision trees. They are ensembles of decision trees, each decision tree created by using a subset of the attributes used to classify a given population (they are sub-trees, see above). Those decision trees vote on how to classify a given instance of input data, and the random forest bootstraps those votes to choose the best prediction. This is done to prevent over fitting, a common flaw of decision trees. A random forest is a supervised classification algorithm. It creates a forest (many decision trees) and orders their nodes and splits randomly. The more trees in the forest, the better the results it can produce.

E. Naïve Bayes Classifier–

Naive Bayes are a class of conditional probability classifiers that are based on the Bayes Theorem. They assume independence of assumptions between the features. Bayes Theorem has many advantages. They can be easily implemented. Furthermore, Naive Bayes requires a small amount of training data and the results are generally accurate. Bayes Theorem lays down a standard methodology for the calculation of posterior probability P(c|x), from P(c), P(x), and P(x|c). In a Naive Bayes classifier, there is an assumption that the effect of the values of the predictor on a given class(c) is independent of other predictor values.

F. Support Vector machine (SVM) –

Support Vector Machines are a type of supervised machine learning algorithms that facilitate modeling for data analysis through regression and classification. SVMs are used mostly for classification. In SVM, we plot our data in an n-dimensional space. The value of each feature in SVM is same as that of specific coordinate. Then, we proceed to find the ideal hyper plane differentiating between the two classes. Support Vectors represent the

coordinate representation of individual observation. Therefore, it is a frontier method that we utilize for segregating the two classes.

G. Stochastic Gradient Descent (SGD) –

The word 'stochastic' means a system or a process that is linked with a random probability. Hence, in Stochastic Gradient Descent, a few samples are selected randomly instead of the whole data set for each iteration. SGD is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as Support Vector Machines and Logistic Regression. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention in the context of large-scale learning. SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing.

H. XGBoost Classifier –

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also referred to as GBDT, GBM) that solve many data science problems during a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and may solve problems beyond billions of examples.

I. Gradient Boosting Machine (GBM) –

GBM may be a boosting algorithm used once we affect many data to form a prediction with high prediction power. Boosting is really an ensemble of learning algorithms which mixes the prediction of several base estimators so as to enhance robustness over one estimator. It combines multiple weak or average predictors to a build strong predictor. Light GBM is nearly 7 times faster than XGBOOST and may be a far better approach when handling large datasets. This seems to be an enormous advantage once you are performing on large datasets in limited time competitions.

III. About Data-set And its Attributes

Heart Disease informational index has been downloaded from the kaggle website which is recently updated few months ago. This dataset contain 14 attributes namely age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target. the dataset contains information about heart patients in Cleveland, Hungary, Switzerland, and Long Beach V. To identify the best classifier, the data set is divided into trained data and test data. Preparing data includes cleansing the data, altering the data and splitting the data. Cleans and alter data is already done in the dataset. Splitting the data into two parts dependent feature X and an output feature Y. The X and Y are categorised into X_train, X_test, and Y_train, Y_test.

	age	bp	sg	al	su	hemo	pcv
count	396.000000	396.000000	396.000000	396.000000	396.000000	396.000000	396.000000
mean	51.654040	75.813131	1.026904	1.255051	0.734848	12.295707	38.257576
std	16.713339	14.483353	0.066565	1.513852	1.377173	3.166023	9.321233
min	3.000000	32.000000	1.005000	0.000000	0.000000	2.600000	9.000000
25%	42.000000	70.000000	1.015000	0.000000	0.000000	10.300000	32.000000
50%	54.000000	80.000000	1.020000	1.000000	0.000000	12.500000	39.000000
75%	64.000000	80.000000	1.025000	2.000000	1.000000	14.700000	45.000000
max	90.000000	180.000000	2.025000	7.000000	8.000000	26.000000	85.000000

Fig.1.Dataset

IV. RESULTS AND ANALYSIS

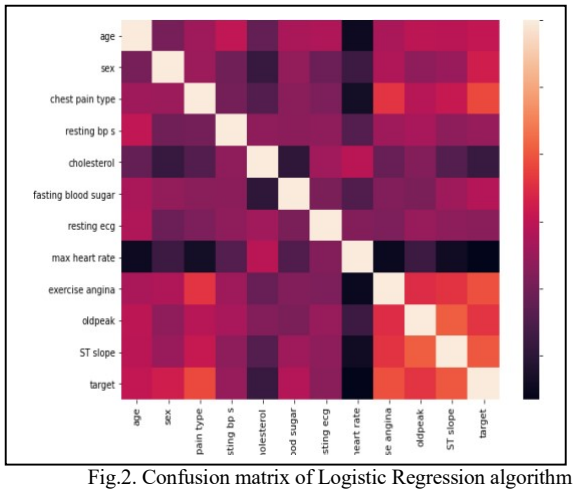

Fig.2. Confusion matrix of Logistic Regression algorithm

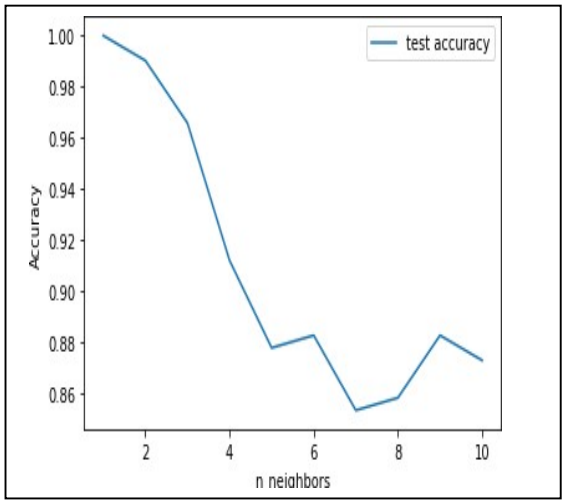

Fig. 3. Value of K for KNN Algorithm

An evaluation of dissimilar algorithms is performed on heart disease data set. To select the best out of all the models created can be done by comparing the accuracy of all the models and selecting the one which gives the maximum accuracy in both training data and test data. The results of the algorithms are shown below:

Table -1 : Comparisonvalues

Algorithm	Train data	Test data
KNN	0.885	0.854
Logistic regression	0.856	0.849
Decision tree	1.000	1.000
Random Forest	1.000	1.000
Naive Bayes	0.827	0.844
SVM	0.841	0.854
SGD	0.838	0.844
XGBoost	0.988	0.980

GBM	1.000	1.000

V.CONCLUSION

The observation of various algorithms and their results when considering the classification model, Decision Tree Algorithm, Random Forest Algorithm and Gradient Boosting machine gives a value 1.000 in both training data and testing data accurately which proves that the data is split into ideal training dataset and test dataset. XGBoos Algorithm also got a value 0.98 in trained dataset and test dataset but the accurate result in classification models is always considered as 1.000 which is acquired by Decision Tree Algorithm, Random Forest Algorithm and Gradient Boosting machine.

REFERENCES

[1] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal ofRecent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue- 1S4, June 2019.

[2] R. Kavitha and E. Kannan et al. "An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature.

[3] J. Vijayashree and N.Ch. Sriman Narayana Iyengar Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review (2016).

[4] Himanshu Sharma,M A Rizvi Prediction of Heart Disease using Machine Learning Algorithms: A Survey (August 2017)

[5] Vikas Chaurasia, Saurabh Pal, "Early Prediction of Heart disease using Data mining Techniques", Caribbean Journal of Science and Technology, 2013.

[6] C. S. Dangare and S. S. Apte, "Improved study of heart disease predictionsystem using data mining classification techniques," International Journal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012.

[7] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniquesin heart disease diagnosis and treatment," pp 173-177, 2012.