# Identification and Implementation Perspective on the Stratification Algorithms in the Prognostication of Heart Disease using Machine Learning Techniques

D. Vijay Lakshmi, K. Yasudha, Vanitha Kakollu

***Abstract:*** ***Analysis of patient's data is always a great idea to get accurate results on using classifiers. A combination of classifiers would give an accurate result than using a single classifier because one single classifier does not give accurate results but always appropriate ones. 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 patient doesn't have heart disease and 1 means patient have heart diseases. So, there is a need to build a classification algorithm that can predict the Outcome feature of the test dataset with good accuracy. For this understanding the data is important, and then various classification algorithm can be tested. Then the best model can be selected which gives highest accuracy among all. The built model can then be given to the software developer for building the end user application using the selected machine learning model that will be able to predict the heart disease in a patient.***

***Keywords: Random Forest Algorithm, SVM, Logistic Regression Algorithm, Machine Learning Classification.***

## I. INTRODUCTION

Machine learning and Data Mining has a main aim of getting more flexible and understandable reports on the basis of various methods. There is an immense growth of data in the real world that needs development of different techniques that could sort the data into different patterns. Healthcare industry contains very large and sensitive data and needs to be handled very carefully. Heart disease is one of the growing extremely a very deadly disease all over the world Medical practitioners wanted to have a prediction model to diagnose and prevent heart attack. Different machine learning techniques are useful for examining the raw data from diverse perspectives and synopsizing it into valuable information.

The accessibility and availability of huge amounts of data will be able to provide us useful knowledge if certain data mining techniques are applied on it.

## II. MACHINE LEARNING ALGORITHMS – CLASSIFICATION

### A. K Nearest Neighbour Algorithm

The K Nearest Neighbour is a simplest machine learning model. Building the model consists of storing the trained dataset. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set-its "nearest neighbors."

### B. Logistic Regression Algorithm

It is a technique to analyze a data-set which has a dependent variable and one or more independent variables to predict the outcome in a binary variable, meaning it will have only two outcomes. The dependent variable is categorical in nature. Dependent variable is also referred as target variable and the independent variables are called the predictors. Logistic regression is a special case of linear regression where we only predict the outcome in a categorical variable. It predicts the probability of the event using the log function.

Logistic regression uses an equation as the representation, which is similar to linear regression. Input values (x) are merged linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to forecast an output value (y). A key difference from linear regression is that the output value being imitated is a binary value (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

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

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's).

**Revised Manuscript Received on January 30, 2020.**
***** Correspondence Author

**D Vijay Lakshmi***, PG Student, Department of CS,GIS,GITAM (Deemedto be University), Visakhapatnam, India, vijjidasari22@gmail.com

**K Yasudha,** Department of CS,GIS,GITAM (Deemed to be University), Visakhapatnam, India, yasudha.p@gmail.com

**Vanitha kakollu**, Department of CS,GIS,GITAM (Deemed to be University), Visakhapatnam, India, vanithagitam@gmail.om

*Retrieval Number: C8663019320/2020©BEIESP*
*DOI: 10.35940/ijitee.C8663.019320*
*Journal Website: www.ijitee.org*

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

### C. Decision Tree Algorithm

A decision tree is a mostly used non-parametric efficacious machine learning modelling technique for regression and classification problems. To find solutions a decision tree makes sequential, hierarchical decision about the resultant's variable based on the forecasted data.

Decision tree constructs regression or classification models in the form of a tree shape. The tree shape formed breaks down a dataset into tiny subsets while at the same time an associated decision tree is gradually developed. The final result is a tree with decision nodes and leaf nodes.

### D. Random Forest Algorithm

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the stronger the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high précised results.

### E. Support Vector Machine

A support vector machine is a supervised learning algorithm that ilk's data into two categories. It is trained with a series of data already arranged into two categories, building the model as it is at first trained. The task of an SVM algorithm is to decide which category a new data point belongs in. This makes SVM a kind of non-binary linear classifier.

## III. DATA SET AND ATTRIBUTES

Here a machine learning classification problem in which a heart disease dataset is given which contain information about various patient and some medical information of each like male, age, current smoker, cigsperday, diabetes, totChol, sysBP, diaBP, heartrate, BMI, Glucose, TenyearCHD. The data set is basically about US Heart patients which is recently updated a month ago downloaded from Kaggle.com for a better accurate result to be generated. The data set is divided trained data and test data. Preparing the data for model generation includes: Cleansing the data, Altering the data and Breaking the data (train dataset, test dataset). The cleansing and altering the data are already performed in the dataset. Hence the main focus is on splitting the data into two part that is relying features i.e. X and output feature i.e. Y. Also, in the data set both X and Y are categorised into training dataset i.e. X_train and Y_train. And test dataset i.e. X_test and Y_test. In the below figure it is clear that we are showing the result in splitting the data set into 75% of training data and 25% of test data.

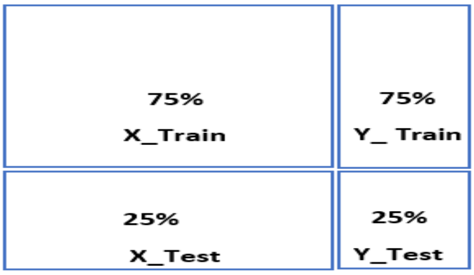

**Fig. 1. Diagrammatic representation of Trained and Test data**

## IV. RESULTS AND ANALYSIS

The results generated with the logistic regression also include the confusion matrix which is as follows:

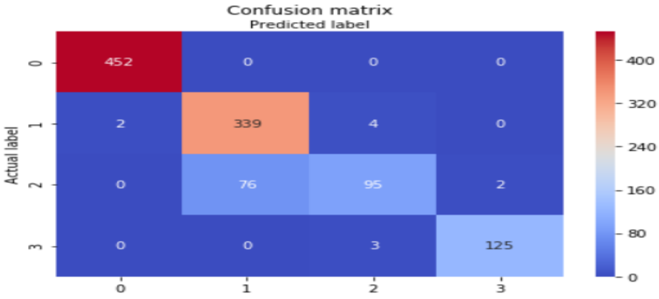

**Fig. 2. Confusion matrix of Logistic Regression Algorithm**

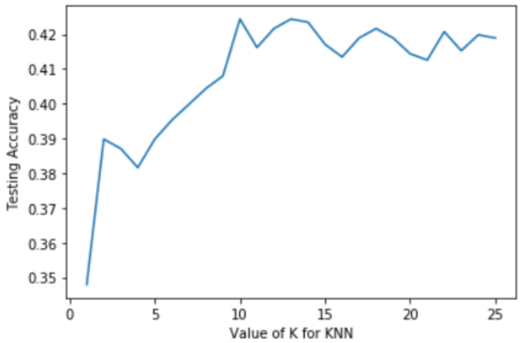

**Fig. 3. Value of K for KNN Algorithm**

The data set results are with the comparison with the other algorithms as well. The training data accuracy and testing data accuracy of each algorithm is generated in order to get exact and effective results with respect to the data set of US patients.

**Table- I: Comparison values**

| Algorithm | Training data Accuracy | Test Data Accuracy |
|---|---|---|
| KNN | 0.96 | 1.00 |
| Decision tree | 0.967 | 0.933 |
| SVM | 0.983 | 0.967 |
| Random Forest | 1.000 | 1.000 |
| Logistic | 0.921 | 0.921 |

## V. CONCLUSION

Taking into observation various algorithms and their end results when considering the classification model as well then, the Random Forest Algorithm gives a précised 1.000 in the training data and testing data accuracy which proves that the data is split into a ideal training dataset and test dataset which is more likely a précised result as expected. Logistic Regression Algorithm also has got an error free with consideration to trained dataset and test dataset i.e. 0.921 but the accurate result in classification models is always considered to as 1.000 which is acquired by Random Forest Algorithm.

## VI. FUTURE SCOPE

There is a possibility of using additional algorithms on this particular heart patient's dataset to assort and provoke a framework for easy utilization in the real time environment by the medical practitioners to prognosticate the number of heart patients in the near future and to take obligatory precautions prior the patients actually suffer from heart diseases. Framework could allow the rudimentary categorization of the dataset and make the best usefulness of it for making accurate upshots in advance.

## REFERENCES

1. Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.
2. K.Sudhakar, Dr. M. Manimekalai "Study of Heart Disease Prediction using Data Mining", IJARCSSE 2016.
3. V. Krishnaiah, G. Narasimha[+], N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review" IJCA 2016.
4. Vikas Chaurasia, Saurabh Pal, "Early Prediction of Heart disease using Data mining Techniques", Caribbean journal of Science and Technology,2013.
5. S. Vijiyrani et. al., "An Efficient Classification Tree Technique for Heart Disease rediction", International Conference on Research Trends in Computer Technologies (ICRTCT - 2013) Proceedings published in International Journal of Computer Applications (IJCA) (0975 – 8887), 2013 (pp 6-9).
6. Chaitrali S. Dangare, Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 888) Volume 47No.10, June 2012.
7. Harsh Vazirani et. al.," Use of Modular Neural Network for Heart Disease", Special Issue of IJCCT[+] Vol.1 Issue 2, 3, 4; 2010 for International Conference [ACCTA-2010], 3-5 August 2010 (pp 88-93).

## AUTHORS PROFILE

**D Vijay Lakshmi,** pursuing Master of Computer Applications, Department of CS, GIS, GITAM (Deemed to be University), Visakhapatnam. Her area of interest in Machine Learning, Data Mining and Parallel Processing .

**K Yasudha,** is currently working as Assistant Professor in the Department of Computer Science, GIS, GITAM (Deemed to be University). Her main areas of research includes Machine Learning and Data Mining.

**Vanitha Kakollu,** is currently working as Assistant Professor in the Department of Computer Science, GIS, GITAM (Deemed to be University). Her main areas of research includes Image Processing, Data Mining and Machine Learning.