International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Regression Based Model for Prediction of Heart Disease Recumbent

M.Diviya, G.Malathi, A.Karmel

Abstract*: Supervised Learning, a novel method that figures out how to anticipate the resultant of an input-output pair by inducting data under series of training and testing functions. Regression model is a sub classification of Supervised Machine Learning. In this paper various Regression models such as Logistic Regression, SVM, KNN, Naive Bayes and Random forest have been applied on Heart Disease dataset. The anticipated outcomes draw the deduction on the level of patients inclined to coronary illness dependent on the traits and qualities. In reference to the applied calculations both KNN and Random Forest beats the other relapse calculation with a precision of 88.52%.

Keywords* : Supervised Learning, Logistic Regression, SVM, KNN, Naive Bayes and Random forest.

I. INTRODUCTION

Regression analysis involves analyzing and predicting the data to reveal patterns and relationships. It is the well-known analysis that evaluates the relationship between the variables of the data under study. The existence of various regression based models paves way for the researchers to understand the data and its attributes. It throws focus on deducing the dependent factors in regard with the study. Each algorithm exhibits its own features and is renowned in deriving its accuracy.

II. RELATED WORK

[1] The researchers have drawn limelight on using EDA as a tool to resolve the issues confronted by the researchers in information retrieval. They used visualization tools to analyze the data and put forth the inefficiencies in handling them in a statistical study. A tool titled “Weight of evidence” was deployed, using which a theory of support for Inverse Document Frequency ranking was proposed.[2] The proposed work was carried out in a wind tunnel at United Forest Services Fire Science Laboratory. The work aimed at understanding the spreading of wildfire and the tools utilized by managers to control it.

Manuscript published on November 30, 2019.
*Correspondence Author
M.Diviya*,School of Computing Science and Engineering,VIT Chennai Campus, Chennai, India.
Dr.G.Malathi, School of Computing Science and Engineering,VIT Chennai Campus, Chennai, India.
Dr.A.KArmel, School of Computing Science and Engineering,VIT Chennai Campus, Chennai, India.

The results of this Exploratory Analysis paved the way to generate probability models. The basic study considers the amount of fuel needed and the geometry followed by the fire. The geometry is evaluated based on the fuel tines of the comb placed on a grid format
along the test bed and is ignited by the windward of the bed with a wind speed of 0.67m/s. A thermocouple monitors the fire spread and the time side exploratory data analysis is undertaken to resolve the problem.[3] The main aim of the proposed work is to promote the research as the enhancements in data and various improvements in Exploratory Analysis is at its peak. Several geoscientists and computer scientists collaborated to work with geoscience data. The resultant work engendered a hypothesis in analyzing the data science process by deploying a case driven approach. The data visualization enables facile human visibility to infer the complex relationships present in the data.[4]The proposed work starts with the description of the historical and conceptual background represented by the Exploratory Data Analysis. The main aim of this research is to understand, analyze, detect and empirically study the phenomenon of data. The authors have also discussed the issues related to EDA and their connectivity in evaluating the concerned applications involved in scientific understanding. The initial work focuses on the development of replication and science based requirement of cross-validation and its emphasis on data patterns. The next level of work focuses on the differentiability of the exploratory analysis in practice and identifying the scientifically questionable data. The final result focuses on utmost maximization of data.[5] The researchers handled the work to develop samples of large datasets that are purposive to carry out the research. They started with the study of the data that is intended to be analyzed. Every research starts with collecting datasets from various resources, initially when the analysis is to be carried out in a bigger environment. For example regarding social sites, the researchers need not initially decide the area of interest to be pursued. The exploratory study helps the researcher to identify their data and how far the research infers a valuable conclusion. Vista is the approach used and it results in supporting the qualitative and trustworthy research for the given large dataset.[7]In general data comes in profusion with a variety of formats and dimensionality. Hence the information extraction becomes unwieldy when diffusing the query and the goal to have promptness also plays a crucial role. If the data is to be explored and queried, the query framed shouldn't be complicated .Different data intends to use different study and perusal. Hence the methods are inclined to be unique.[8]

The need for research is to have a glimpse of how Google Classroom enhances learning and resultant following learning through Google Classroom.

A Likert scale of five points with 24 questions were generated and the results prove that [9]Google Classroom enhances the learning skills of the students. The usage of the tool and its ergonomics were the two conditions to explore which contributes to the adequacy of the research. The analysis followed a statistical study, Measurement Model Assessment, and Structural Model Assessment, the competent explication of Spatio-temporal patterns of crime clusters. Researchers inquired the maximality of mapping the three-dimensional cluster events in a space-time cube in supporting the space-time variants of kernel density estimation and scan statistics. The study was carried out using the dataset of snatch and run offenses of Kyoto city from the year 2003-2004. The space-time clusters have been visualized by using the proposed method and it yielded a good cluster analysis. The space statistical approaches reveal that spatial-temporal exploratory analysis of clustered events creates a transgression of crime knowledge.Advancement in areas of medical sciences and biotechnology paved the way for noteworthy research in this area. The data is explored inculcating machine learning algorithms and the available data mining algorithms. For example, in the case of Diabetes Mellitus, the disease under study, the availability of huge datasets facilitated conducting a study consisting of predicting and diagnosing diabetes followed by complications and genetic effects, healthcare and management. The dataset under supervised learning was 85% and 15% under unsupervised learning. SVM was employed as a classifier and results were analyzed

III . EXPERIMENTAL ANALYSIS

A. Dataset Attributes

This work has been carried out using a dataset pertaining to Heart diseases. The sequence of attributes considered are as follows : Age of the person understudy in years , sex - (1 = male; 0 = female) cp - chest pain type restbps - resting blood pressure chol - serum cholesterol in mg/dl fbs - fasting blood sugar > 120 mg/dl which is a boolean value (1 = true; 0 = false), restecg - resting electrocardiographic results ,thalach - maximum heart rate achieved, exang - exercise induced angina having yes or no values (1 = yes; 0 = no), oldpeak - ST depression induced by exercise relative to rest ,slope - the slope of the peak exercise ST segment ,ca - number of major vessels (0-3) colored by fluoroscopy, thal - 3 = normal; 6 = fixed defect; 7 = reversible defect and finally these attributes are analysed to arrive at a result whether the person has heart disease or not. target - have disease or not (1=yes, 0=no).

B. Exploratory Analysis of Dataset

Exploratory analysis is waved out using the given set of attributes. Primarily the data set is cleansed and the normalization of the values were done. Data cleaning has been the most influential part when the analysis of a data set is needed. The generic method of normalization introduced to the data is min-max normalization. Then the data has been split for testing and training, by considering 80% of data to undergo training and 20% of it to be tested in the presence of the regression models. The regression models imputed for study are Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes and Random Forest Classification.

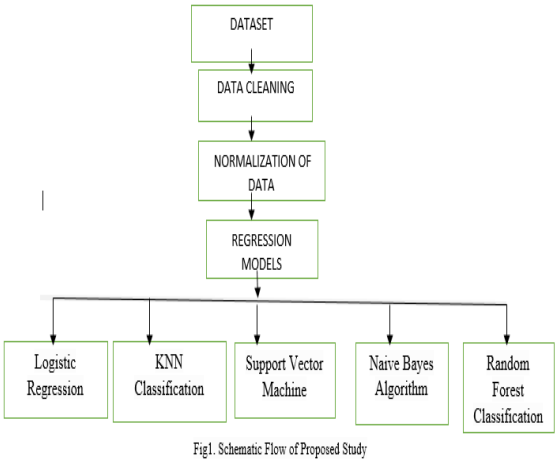

Fig1. Schematic Flow of Proposed Study

C. Attribute relationship

The pool of attributes in regard with the dataset are considered for study. From Fig.2 shows that the situation of a person is linked to the diagnosis of heart disease according to the heart rate frequency of a person and their fasting blood sugar value. But this isn't the major concern since the true value is very low. The inference in containment with the fasting blood sugar of a person doesn't play a major role in heart disease prediction. Graphical representation (Fig 3) of a person having heart disease based on maximum heart rate and the age factor infers that the heart rate value could influence the arena, thus having a higher risk of having a heart disease.

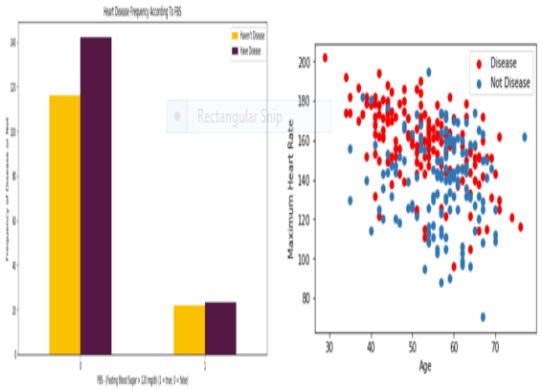

Fig 2.Graphical Plot of Heart Disease Frequency Vs FBS

Fig 3.Person having heart disease(red) and Person Not having heart disease(blue) based on Maximum Heart Rate

Consider Fig 4. Here a contingent value analysis is done for persons having heart disease with regard to the age factor. Believing the dataset values to be true, the true value 1 infers that at the age of 54 the primary risk of encountering as well as not encountering heart disease have an equal proportion. The risk of facing heart diseases based on chest pain type 2 is more persistent and it is around 65%.

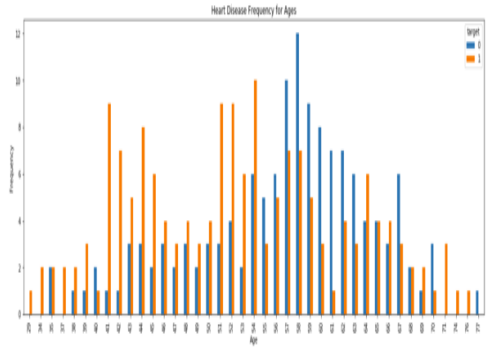

Fig 4.Heart Disease Frequency Based on Age Factor

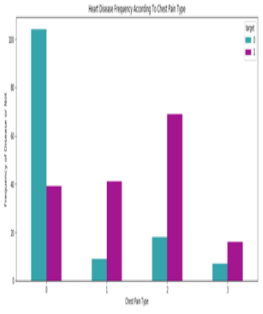

Fig 4.Heart disease Frequency According to Type of Chest Pain

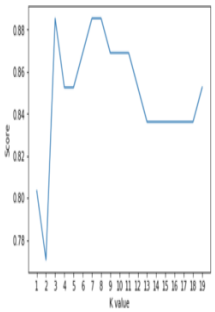

Fig 5.Graphical Depiction of K Value Vs. score

D. Regression Based Models

Logistic regression is a regression analysis that predicts the probabilistic model of the class of variables under scrutiny. In the contained dataset the analysis infers from the resultant value of whether the person has heart disease or not. Embedding Logistic regression, the test accuracy increases to 86.89%. The study is continued by employing K-Nearest Neighbor Classification. It is an elemental level of understanding of the non-parametric algorithm for classification. The value for K can be any integer value. Each preceding value of the neighbor constitutes of a vote for the label. The model initially starts with a value of 2 for k and the accuracy is 77.05. The accuracy is at the highest if K value is stepped up as 3-7-8 and it will attain the maximum score of 88.52.

Support Vector Machine is a linear classification model and it is a classic regression model that can also be implemented in nonlinear problems. In the given dataset the regression model behaves as a linear classifier, because of the existence of only two classes depicting if the person will encounter with a heart disease or not . The algorithm landed with an accuracy of 86.89% which is marginally lower than KNN. In the ladder of Regression model the next is the Naïve Bayes algorithm, which is also a probabilistic model of the data. This arrived to an accuracy 86.89%, hence the model that co-performs the Logistic regression model. The final regression algorithm for study is the Random forest which is an ensemble learning. The strong learner starts with weak decision trees and with the size of the forest. It outperforms and avoids the limitation with over fitting and ends as a good predictive model.

Fig 6Table representing Accuracy Value of Regression Algorithm

Algorithm	Accuracy
Logistic Regression	86.89
KNN	88.52
SVM	86.89
Naive Bayes Algorithm	86.89
Random Forest Classification	88.52

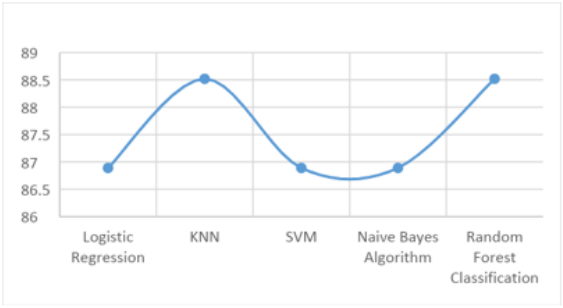

Fig 7 Comparison Chart of accuracy of the Regression Algorithms

E. Error rate Deduction and Accuracy

$$Err = \frac{FP+FN}{TP+TN+FP+FN} = 0.13$$

$$Err = \frac{FP+FN}{TP+TN+FP+FN} = 0.11$$

$$Err = \frac{FP+FN}{TP+TN+FP+FN} = 0.19$$

The confusion matrix(Fig 7) of regression algorithms were obtained. In connection with the regression algorithms, the false positive and false negative values are plotted. In case of Logistic regression, the false positive is 4 and the false negative is 4.The false positive and false negative in the case of KNN is 4 and 3 and the case continues based on regression algorithm. When we consider based on accuracy Random Forest algorithm has false-positive and false negative values as 4 and 3 which is similar to KNN. The Error Rate is calculated based on the following. The Error rate for Logistic Regression, SVM and Naïve will be the same, while the Error rate for Random forest and KNN are optimal. This gives better accuracy.

III. CONCLUSION AND FUTURE WORK

The dataset exhibits execution based on all the given algorithms and the models worked fine. Among all the models, KNN and Random Forest models displayed increased efficiency with an accuracy of 88.52%. KNN Classification is similar toVoronoi Diagram, since it can separate multiple regions in a non-convex plane. KNN exhibited better performance than SVM, indicating that the data cannot be easily separated using decision planes. In the case of Random Forest Classification, it erodes overfittingof data and when handled with precise seed hence making them good regressors. In future, the results and accuracy can be improved by using Neural Networks for the same data values and the results can be predicted with utmost efficiency.

REFERENCES

1. G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
3. H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
4. B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
5. E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," IEEE Trans. Antennas Propagat., to be published.
6. J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," IEEE J. Quantum Electron., submitted for publication.
7. C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
8. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].
9. M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.
10. (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). Title (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)
11. J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com
12. (Journal Online Sources style) K. Author. (year, month). Title. Journal [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)

AUTHORS PROFILE

Ms.M.Diviya is a Research Scholar in School of Computing Science and Engineering in VIT Chennai Campus .Her area of interest and research includes Machine learning, Natural Language Processing .She has published papers in Scopus Indexed journals and conferences.

Dr.G.Malathi is working as an Associate Professor in the School of Computing Science and Engineering, VIT Chennai. She has over 15 years of experience in teaching and research. Her area of specialization is Image Processing and Healthcare Analytics. She has been a resource person in FDPs and Seminars. Currently she is guiding 3 PhD scholars and has number of publications in International Journals. She has authored few bookchapters. She has filed patent in novel Biometrics. She received 'Best Outstanding Faculty Award in Computer Science' by VENUS Foundation in the July 2018.

Dr.Karmel.A is working as an Associate Professor in the School of Computing Science and Engineering, VIT Chennai. She has over 15 years of experience in teaching and research.Her area of interest includes Networking,High Speed Networks and Machine Learning.She has done many funded research projects and owns papers in various National and International Journals.