Exploring Heart Disease Diagnosis using Multivariate Data Analysis: A Comparative Study of Naive Bayes and Logistic Regression

Suriboina Manichandhana[1], G V Ramesh Babu[2]

[1]PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati
[2]Associate Professor, Dept of Computer Science, SV University, Tirupati

***Abstract*—** *This research paper delves into the intricate task of heart disease diagnosis, utilizing a multivariate dataset encompassing 14 distinct attributes. The dataset, commonly referred to as the Cleveland database, has been widely adopted in machine learning research. With 606 instances and attributes spanning age, sex, medical parameters, and electrocardiographic results, the primary objectives of this study are two-fold: Firstly, to develop predictive models that accurately identify the presence of heart disease based on patient attributes, and secondly, to unearth valuable insights from the dataset that contribute to a deeper understanding of this critical health concern. Two classification algorithms, Naive Bayes and Logistic Regression, are employed and their performance in terms of accuracy, precision, and recall are compared.*

I. INTRODUCTION

Step by step the instances of heart infections are expanding at a quick rate and it's vital and worried to foresee any such sicknesses ahead of time. This conclusion is a troublesome errand for example it ought to be performed exactly and productively. Cardiovascular sicknesses are extremely normal nowadays, they depict a scope of conditions that could influence your heart. World wellbeing association assesses that 17.9 million worldwide passings from (Cardiovascular illnesses) CVDs [1][2]. It is the essential explanation of passings in grown-ups. Our undertaking can assist with foreseeing individuals who are probably going to determine to have a coronary illness by help of their clinical history [5][6]. It perceives who all are having any side effects of coronary illness, for example, chest torment or hypertension and can assist in diagnosing sickness with less clinical trials and successful therapies, so they can be restored appropriately [12]. The exploration paper fundamentally centers around which patient is bound to have a coronary illness in light of different clinical qualities. We arranged a coronary illness forecast framework to foresee whether the patient is probably going to be determined to have a coronary illness or not utilizing the clinical history of the patient.

II. METHODOLOGY

2.1 Naïve Bayes

The Naive Bayes is a vivacious methodology for blueprint of quantifiable farsighted models. NB relies on the Bayesian speculation [3][4]. This estimation uses class prohibitive independence and has ability to change quickly. This portrayal technique assessments the connection between every property and the class for every manual for pick a prohibitive probability for the connection between the brand name characteristics and the class [7][11]. In the midst of setting up, the probability of each class is enrolled by checking how regularly it occurs in the arrangement dataset. This is known as the "prior probability" P(C=c). No matter what the previous probability, the estimation besides enlists the probability for the event x given c with the assumption that the properties are free. This probability changes into the possible result of the probabilities of each single quality. The probabilities would then have the decision to be evaluated from the frequencies of the events in the organizing set.

2.2 Bayesian Speculation

Given planning data X, back probability of a theory H, P(H|X), follows the Bayes speculation

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Let X be data tuple and H be a theory so much that the data tuple X has a spot with a foreordained class C. For course of action issues, we want to choose P (H|X), the probability that the theory H holds the given verification or saw data tuple X.

P (H|X) is the back probability of H shaped on X

P (H) is the prior probability of H

P (X|H) is the back probability of X adjusted on H

P(X) is prior probability of X

III. LOGISTIC REGRESSION (LR)

LR is a directed AI calculation that is principally used to gauge the likelihood of an occasion having two potential results in view of the given free factors.

In Direct space, Calculated relapse tracks down a straight choice limit, a line, or a hyperplane, to separate between the classes. Principally, the strategic relapse assesses the likelihood for parallel results as it were [7][9]. For this, it utilizes the sigmoid change capability (sorcery capability) as default. The strategic relapse does this by assessing the ideal qualities for the coefficients that amplify the probability of the noticed information.

Key break faith is an evaluation used to foresee a twofold result: either something occurs, or doesn't. This can be displayed as Yes/No, Significant/Bogus. Free factors are bankrupt down to decide the twofold result with the outcome tending to be requested as one of two groupings [7][10]. The free factors can be immovable or numeric, yet the reliant variable is continually self-evident. Made thusly:

P(Y=1|X) or P(Y=0|X)

It decides the likelihood of ward variable Y, given free component X. This can be utilized to handle the likelihood of a word having a respectable or lamentable essential importance (0, 1, or on a scale between). Of course it will overall be utilized to close the article contained in a photograph (tree, bloom, grass, and so forth), with all that given a likelihood a few spot in the extent of 0 and 1

IV. EXPERIMENTAL RESULTS

We have used the Python Language to implement the experiment our proposed algorithms. The dataset, commonly referred to as the Cleveland database has taken from Kaggle [8], has been widely adopted in machine learning research. This dataset consists of 606 instances and 14 attributes, two class labels (Present (276) and absent (330).

4.1 Results

In our experimental analysis, Naive Bayes achieved an impressive 92.67% accuracy, with precision and recall rates of 92.6% and 93% respectively. On the other hand, Logistic Regression outperformed, boasting a 94.25% accuracy rate, complemented by precision and recall rates both standing at 94.3%. These results indicate that both models exhibit substantial promise in diagnosing heart disease from multivariate patient data shown in the table-1.

Table-1
Experimental Results

Algorithm	Accuracy	Precision	Recall
Logistic Regression	94.25	94.3	94.3
Naive Bayes	92.67	92.6	93

4.2 Discussion

Algorithm Performance: The findings suggest that both Naive Bayes and Logistic Regression can be effective tools in heart disease diagnosis. The relatively high accuracy, precision, and recall values underscore their potential in this critical medical application.

Model Interpretability: Logistic Regression, in particular, offers interpretability advantages, enabling clinicians to identify the most influential attributes in diagnosing heart disease. This knowledge can contribute to more informed medical decisions.

Dataset Insights: Beyond predictive accuracy, our study also uncovers valuable insights from the dataset, aiding in a deeper understanding of heart disease factors. This can pave the way for further research and improved diagnostic methodologies.

Clinical Application: These results hold promise for real-world clinical applications, where early and accurate heart disease diagnosis is crucial for patient health. Healthcare practitioners may leverage these models as decision support tools.

V. CONCLUSION

This study showcases the potential of Naive Bayes and Logistic Regression in heart disease diagnosis using multivariate data analysis. The high performance metrics and the insights gained from the dataset underscore the significance of machine learning in healthcare and motivate further research to improve diagnostic accuracy and patient outcomes.

Future work could explore the incorporation of additional advanced machine learning techniques and feature engineering to enhance diagnostic accuracy further. Moreover, extending this research to diverse datasets and incorporating more attributes may improve the robustness and generalizability of the models.

REFERENCES

[1] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heartdisease/, 2004.

[2] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[3] G. Ravi Kumar, K. Venkata Sheshanna, S. Rahamat Basha, and P. Kiran Kumar Redd, "An Improved Decision Tree Classification Approach for Expectation of Cardiotocogram", Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing, Lecture Notes on Data Engineering and Communications Technologies 62, https://doi.org/10.1007/978-981-33-4968-1_26

[4] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[5] HeonGyu Lee, Ki Yong Noh, KeunHoRyu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[7] J.Han and M.Kamber,"Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2nd ed.San Mateo, CA; Morgan Kaufmann, 2006.

[8] https://www.kaggle.com/code/nimapourmoradi/heartattack-classification

[9] N.Michael, "Artificial Intelligence – A Guide to Intelligent Systems", 2nd Edition, Addison Wesley 2005

[10] P.N.Tan, M.Steinbach and V.Kumar "Introduction to Data Mining", A: Addision-Wesley, 2005.

[11] M. V. Lakshmaiah, G. Ravi Kumar and G. Pakardin, "Frame work for Finding Association Rules in Bid Data by using Hadoop Map/Reduce Tool", International Journal of Advance and Innovative Research, Volume 2, Issue1(1), PP:6-9,2015, ISSN: 2394-7780

[12] Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.