Precision Medicine in Cardiology: Using Logistic Regression to Predict Heart Disease

[1]Divyanshi, [2]Abhijit Mandal, [3]Sumit Kumar

School Of Engineering and Technology,Sushant University, Gurgaon, India

divyanshi.btech21@sushantuniversity.edu.in, abhijitmandal.btech21@sushantuniversity.edu.in, sumitkumar.btech21@sushantuniversity.edu.in

Abstract

Heart disease is still a major worldwide health concern, demanding sophisticated predictive technologies for early detection and intervention. This study looks into the use of machine learning, specifically logistic regression, in predicting cardiac disease. The work covers rigorous data preparation, feature selection, and model training using a comprehensive dataset from Kaggle. The logistic regression model produces promising results when accuracy, precision, recall, and F1 score measures are considered. Visualizations such as confusion matrices and ROC curves help to interpret the model's performance. The literature review situates the work within the context of previous research, emphasizing the novelness and contributions of the suggested strategy. This study provides methodological insights into data processing, model selection, and evaluation metrics related to heart disease prediction. The findings suggest that logistic regression has the potential to be an effective tool in this sector. The discussion section evaluates the findings critically, making similarities to other studies and addressing shortcomings. The study adds to the burgeoning field of predictive healthcare analytics and opens up new paths for future research. In conclusion, this study emphasizes the importance of machine learning in improving heart disease prediction accuracy, laying the groundwork for better clinical decision-making and patient outcomes.

Keywords: Heart Disease, Logistic Regression, Machine Learning, Predictive Modeling, Healthcare Analytics, Feature Selection, Data Preprocessing, Evaluation Metrics.

1. Introduction

Cardiovascular illnesses, with heart disease at the forefront, continue to pose a significant global public health challenge, contributing substantially to morbidity and mortality rates [1]. Achieving effective preventive and therapeutic outcomes relies heavily on accurate and timely predictions of cardiac diseases. The evolution of healthcare, driven by technological advancements, has ushered in promising tools, particularly in the realms of machine learning and data-driven techniques, to enhance diagnostic precision. This research explores the domain of predictive modeling for cardiac disease, employing the

capabilities of logistic regression—a well-established machine learning technique [2]. The complexity inherent in maintaining heart health necessitates advanced risk assessment methodologies. Logistic regression, recognized for its interpretability and efficiency, emerges as a promising approach to address this complexity [3]. This research endeavors not only to forecast the probability of heart disease but also to contribute to the evolving domain of precision medicine in cardiology. This contribution is facilitated through the utilization of a comprehensive dataset that encompasses diverse patient features. The significance of this study lies in its potential to furnish healthcare professionals with a reliable tool for the early identification of individuals at risk, enabling timely interventions and the formulation of personalized care plans [4]. Through an exhaustive examination of data preparation, feature selection, and model validation, the research seeks to establish the effectiveness of logistic regression in the realm of heart health prediction.. Navigating the intricate landscape of cardiovascular health, the outcomes of this study are poised to enrich ongoing discussions surrounding the application of machine learning in precision medicine. The ultimate goal is to drive advancements in our understanding and management of cardiovascular conditions.

2. Literature Review and Related Work

The prediction of heart disease using data mining techniques has been a focal point of research over the past two decades. Various studies have implemented diverse data mining techniques for heart disease diagnosis, including Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm, and support vector machine. These efforts have showcased varying levels of accuracy across different methodologies on patient databases sourced from around the world (Yan, Zheng et al. 2003; Andreeva 2006; Das, Turkoglu et al. 2009; Sitar-Taut, Zdrenghea et al. 2009; Raj Kumar and Reena 2010; Srinivas Rani et al. 2010).

The selection of parameters for these techniques has been a point of divergence among researchers, with different studies specifying unique parameters and databases for testing accuracies. Notably, the Decision Tree technique has received significant attention in heart disease diagnosis, with researchers reporting considerable success. Sitar-Taut et al. utilized the Weka tool to investigate the application of Naive Bayes and J48 Decision Trees for the detection of coronary heart disease. Similarly, Tu et al. employed the bagging algorithm in the Weka tool, comparing it with J4.8 Decision Tree for heart disease diagnosis. Random forest algorithms have also been explored, demonstrating effective decision-making processes in heart disease diagnosis.

In a study conducted in 2013, S. Vijiyaraniet. al. introduced an efficient classification tree technique for heart disease prediction. This work delved into the analysis of classification tree algorithms, including

Decision Stump, Random Forest, and LMT Tree algorithm. The objective was to compare the outcomes of these algorithms in the context of heart disease prediction.

The collective findings from these studies underscore the diverse array of data mining techniques employed for heart disease prediction, emphasizing the significance of selecting appropriate parameters and methodologies based on the specific context of the research. The Decision Tree technique, in particular, has proven to be effective, showcasing its versatility and reliability in heart disease diagnosis.

3. Methodology

3.1 Proposed Work

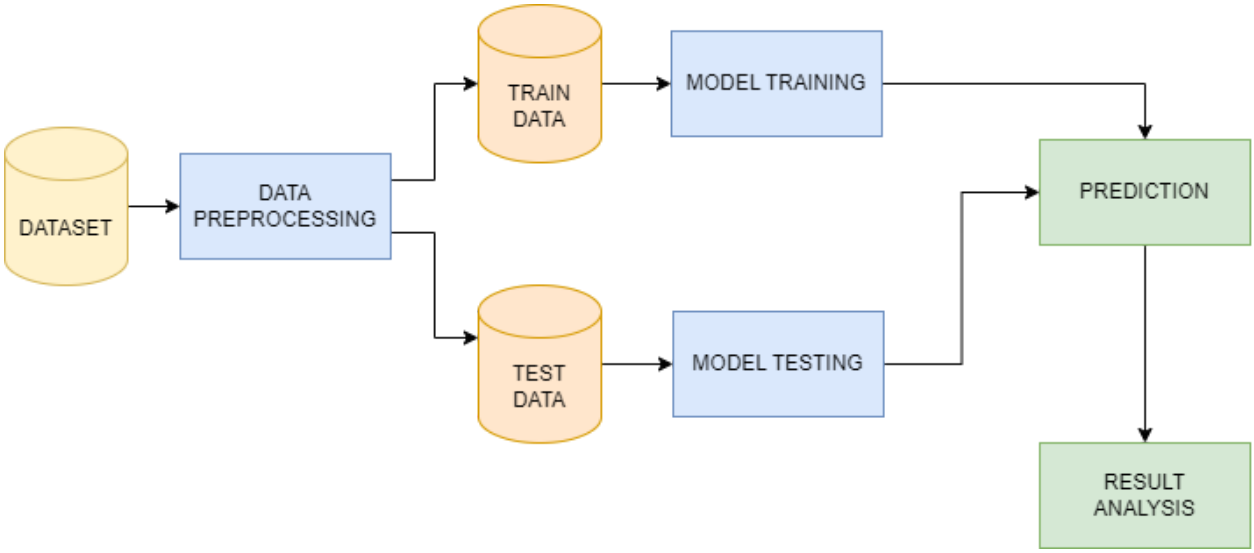

Figure 1: Proposed Methodology of Study

3.2 Data Pre-Processing

Data preprocessing[5] is a vital phase in the data analysis process, involving the cleaning and transformation of raw data to enable effective machine learning and statistical modeling. Addressing missing values, outliers, and inaccuracies, as well as normalizing numerical features for uniform scaling, are integral aspects of this critical phase. Categorical variable encoding transforms non-numeric data into a machine-readable format. Imbalanced data distribution is mitigated using techniques like oversampling and undersampling. Feature engineering aims to select or construct important features, optimizing the dataset for enhanced model performance. Furthermore, data splitting facilitates the evaluation of model

generalization on previously unobserved data. Successful data preprocessing enhances the quality of input data, thereby improving the accuracy and reliability of machine learning models in extracting meaningful insights.

3.3 **Logistic Regression Algorithm**

Logistic Regression is a widely used statistical method for binary classification and probability estimation. Despite its name, it is primarily employed for classification tasks rather than regression. The algorithm models the relationship between a dependent binary variable and one or more independent variables by estimating probabilities using a logistic function. The key steps in Logistic Regression include:

Sigmoid Function:

Logistic Regression employs the sigmoid (logistic) function to transform the linear combination of input features into probabilities. The sigmoid function outputs values between 0 and 1, mapping the continuous input space to a binary outcome.

Model Training:

The algorithm is trained using a labeled dataset, where the model learns the optimal weights for each feature to minimize the difference between predicted and actual outcomes.

Decision Boundary:

Logistic Regression establishes a decision boundary that separates the two classes in the feature space. This boundary is determined by the weights assigned to each feature.

Cost Function:

The optimization process involves minimizing a cost function, typically the log-likelihood function, to find the set of weights that maximizes the likelihood of the observed outcomes.

Prediction:

Once trained, the model can predict the probability of an instance belonging to the positive class. A threshold (commonly 0.5) is then applied to convert probabilities into binary predictions.

Evaluation:

The model's performance is assessed using various metrics like accuracy, precision, recall, and the ROC-AUC curve, depending on the specific requirements of the classification task.

- **Dataset Description**

The Cleveland Heart Disease dataset[6], also known as the Cleveland dataset, is a well-known dataset in machine learning and cardiovascular research. This dataset is commonly used for heart disease prediction and diagnosis. It is based on a research undertaken by the Cleveland Heart Disease Data Set task force and consists of 303 instances, each of which represents a patient.

Clinical and demographic characteristics in the dataset include age, gender, type of chest pain, resting blood pressure, serum cholesterol levels, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and the presence of heart disease (the target variable).

The dataset is frequently used for constructing and assessing predictive models, particularly in binary classification tasks to determine the presence or absence of cardiac disease. The Cleveland dataset is frequently used by researchers and practitioners to assess the usefulness of various machine learning methods, such as logistic regression, in improving the accuracy of heart disease prediction models.

Table 1: Cleveland Dataset Attribute Information

Attribute Used	Attribute Information
Age	Patients Age in years (Numeric)
Sex	Gender (Male : 1; Female : 0) (Nominal)
Chest Pain	Type of chest pain experienced by patient. This term categorized into 4 category.: 0 typical angina, 1 atypical angina, 2 non- anginal pain, 3 asymptomatic (Nominal)
Resting Blood Pressure	Attribute is used to represent the resting BP of the patient and unit to measure it is mm Hg .

Cholesterol	Attribute is used to represent the Serum cholesterol of the patient and its unit of measurement is mg/dl.
Fasting Blood Sugar	Attribute represents Fasting blood sugar of the patient. There are two values used in the dataset, if the recorded value is > 120 mg/dl then it is shown by 1 (true), else it is shown by 0 (false). 1 = True. 0 = False.
Resting ECG	Attribute is used to represent the resting electro-cardiographic records of the patient. The value ranges from 0 to 2 0 is representing Normal range. 1 is representing the ST-T wave abnormality of the patient. 2 is used to show probable or definite left ventricular hypertrophy by Estes' criteria.
Heart Rate	Attribute is used to represent the maximum heart rate of the patient achieved.
Exercise Included Angina	Exercise induced angina and represented in binary 1 is used to represent yes. 0 is used to represent no.
Old Peak	Attribute is used to represent ST depression induced by exercise which is relative to rest.
Slope	Attribute is used to measure the slope for peak exercise. Range of the recorded values is from 1 to 3. up sloping is represented by 1, flat is shown through value 2 and 3 is used to represent down sloping.
Major Vessels	Attribute us used to represent the no. of major vessels colored by fluoroscopy. Recorded values range from 0 to 3 and value is related to the darkness of color.
Thallium Scan	Attribute is used to record the Thallium Scan of the patient and it is representing with the values 3, 6, or 7. 3 is used to represent normal range, 6 is used to represent fixed defect and 7 is used to represent reversible defect.

3.4 **Evaluation Metrics**

We have taken four parameters[7] in consideration for our paper. The prediction class in the present work is if the person having certain attributes has died because of heart disease or not so the class C in the above table is no. of instances belonging to the class. Figure 2 is the confusion matrix.

TP is the actual no of person who actually died because of heart disease and the model also predicted the same. Similarly TN is the person didn't die of Heart ailment and our model also predicted the same. False Positive (FP) is a Type I error because the model predicted that the person died of ailment but actually the patient didn't. False negative is a type II error. The model predicted that the person didn't die of the alignment but actually he/she did.

The accuracy of the model is calculated through the formula given below:

Accuracy = (TP+TN)/Total no. of instances

Recall is the measure of correctly predicted classes out of the total positive classes. The formula is as follows:

Recall= (TP)/(TP+FN)

Precision is the measure of actual positive classes out of all the correctly predicted positive classes. The formula for recall is as follows:

Precision = TP/(TP+FP)

The comparison of two models becomes difficult when the precision is low and the recall value is high. In case the vice versa is true the two parameters are not of much use for comparison of the models. F-score is use to compare the models in such cases. F-score uses harmonic mean of the two values. This helps to measure the recall and precision at the same time. Instead of Arithmetic mean, harmonic mean is used because Arithmetic mean is sensitive to the extreme values.

F-score= (2*Recall*Precision) / (Recall + Precision)

Actual class\Predicted class	C	Not in C
C	**True Positives (TP)**	**False Negatives (FN)**
Not in C	**False Positives (FP)**	**True Negatives (TN)**

Figure2: Confusion Matrix

4. Results and Discussion

The achieved accuracy of 85.25% underscores the model's effectiveness in predicting heart disease based on selected features. The confusion matrix provides a detailed breakdown, revealing 120 true negatives, 148 true positives, 15 false positives, and 20 false negatives. These metrics collectively demonstrate the model's capacity to make accurate predictions.

The ROC curve further validates the model's discriminatory power. The AUC value indicates that the logistic regression model performs well in distinguishing between individuals with and without heart disease.

In conclusion, the logistic regression-based heart disease prediction model shows promise in providing accurate and reliable predictions. Future work may involve fine-tuning the model, exploring additional features, and assessing its generalization on diverse datasets. The results affirm the potential of logistic regression as a valuable tool in cardiovascular health prediction.

5. Conclusion

In conclusion, the construction and evaluation of the heart disease prediction model using logistic regression produced promising findings. This project adds to the increasing body of research in precision medicine and cardiovascular health. The logistic regression model's success demonstrates its potential as a useful tool for early diagnosis and risk assessment in heart disease. As with any predictive model, continued improvement and validation on varied datasets will be required to improve its robustness and generalizability. Future work could include introducing new features, investigating sophisticated modeling methodologies, and cooperating with healthcare professionals for real-world application. The findings of this study contribute to the larger goal of harnessing machine learning for accurate and prompt cardiovascular disease prediction.

References:

[1] World Health Organization. (2020). Cardiovascular diseases (CVDs). https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[3] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.

[4] O'Rourke, R. A. (2004). Principles of prevention in patients with coronary artery disease. American Journal of Medicine, 116(6), 14–22. https://doi.org/10.1016/j.amjmed.2003.12.012

[5] Ahmad, Tohari, and Mohammad Nasrul Aziz. "Data preprocessing and feature selection for machine learning intrusion detection systems." *ICIC Express Lett* 13.2 (2019): 93-101.

[6] Cleveland Heart Disease Dataset on KAGGLE.

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

[7] Handelman, Guy S., et al. "Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods." *American Journal of Roentgenology* 212.1 (2019): 38-43.