IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LIFESTYLE-DRIVEN HEART DISEASE PREDICTION: A LOGISTIC REGRESSION APPROACH

[1]Syed Adnan, [2]Muteeb Khan, [3]Abdul Baseer

[1]AI Engineer, [2]Student, [3]Student

[1]Computer Science and Engineering,

[1]Nawab Shah Alam Khan College of Engineering & Technology, Hyderabad, India.

Abstract: This report details the work done to build a predictive model using Machine Learning techniques to detect whether a person is prone to heart disease in the next 10 years based on his lifestyle pattern. Heart disease is a major health problem worldwide and is one of the leading causes of death in many countries. Early detection and prevention of heart disease can significantly reduce the risk of mortality and morbidity. In this project, we have developed a predictive machine learning model for heart disease prediction using the various features of patients and detecting whether a patient is prone to heart disease soon using Machine Learning techniques. The early prognosis of cardiovascular disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

Index Terms **- Machine Learning, Logistic regression, Heart Disease Prediction, Cardiovascular Diseases.**

I. INTRODUCTION

Heart disease is a major public health concern, with millions of people around the world suffering from various types of heart conditions. These conditions can range from mild to severe and can lead to serious complications and even death. Identifying individuals at high risk of developing heart disease is important for early prevention and management of the condition. Accurate prediction of heart disease risk can help healthcare providers take preventive measures to reduce the likelihood of heart disease in high-risk individuals.

Overall, the goal of this project is to contribute to the understanding of heart disease risk and to provide a tool that can be used for early prevention and management of the condition.

1.1 Problem Statement

The problem that this project aims to address is the need for a reliable method for predicting an individual's risk of developing heart disease. Heart disease is a major cause of death worldwide, and identifying individuals at high risk of developing the condition can help healthcare providers take preventive measures to reduce the likelihood of heart disease. Current methods for predicting heart disease risk, such as the use of risk calculators or assessment of individual risk factors, can be imprecise and may not accurately predict an individual's risk. This can lead to missed opportunities for prevention and may result in unnecessary treatments or interventions for individuals who are not at high risk.

1.2 Objectives

The objectives of developing this predictive model are following:

1. To help predict the possibility of heart disease in a patient.
2. Understand top features which plays an important role in heart diseases.

I. DATASET

The dataset we have selected for our research is publicly available on the Kaggle Website at [2] which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 14 features and 4000 records.

The features in the dataset includes: age, sex, chest pain type, current smoker, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target containing 0 & 1 , where 0 is absence of heart disease. The dataset is in csv (Comma Separated Values) format.

II. RESEARCH METHODOLOGY

In this model we are predicting whether a person is prone to heart disease in the near future using different features. Since it is a binary classification, we have decided to implement a Logistic Regression algorithm and build a predictive model. We have discussed the algorithm below in more detail.

3.1 Logistic Regression

Logistic regression is a classification algorithm used to predict a binary outcome. It is a widely used method for predicting a binary outcome, such as a customer converting or not converting on a website, or a student being accepted or rejected into a school. Logistic regression works by using an equation as a model to predict the probability that an event will occur.

The resulting probability is then transformed into a binary outcome using a threshold. For example, if the probability is greater than 0.5, the outcome is 1, and if the probability is less than 0.5, the outcome is 0. The coefficients b0 and b1 are learned by the model using an optimization algorithm, such as gradient descent, that minimizes the error between the predicted probabilities and the actual outcomes. The error is usually measured using a loss function, such as the cross-entropy loss or the log loss.

The logistic regression model built in this project provides an understanding of the relationship between the selected features and the target variable, which can be used to make predictions about the likelihood of an individual developing heart disease. The methodology used in this project demonstrates the process of building a predictive model for a binary outcome using logistic regression and the mathematical formulation of the logistic regression equation.

III. EXPERIMENTS

4.1 Exploratory Data Analysis

From exploring data, we can see that patients who are at risk of heart disease in the next ten years are 644, and patients who are not at risk are 3596. The analysis reveals that there isn't a single attribute that has a significant correlation with the target value. Some of the attributes exhibit a positive correlation while others exhibit a negative correlation.

4.2 Splitting Data into Training and Testing

The final step in the process involved dividing the resulting data into two sets - 80% for training and 20% for testing. These two sets were then used to build and evaluate a Logistic Regression model. The model was fitted using the training data, used to make predictions on the test data, and finally evaluated based on its accuracy. This ensured that the model was able to generalize well to new data and provided an accurate

estimate of its performance on unseen data. Additionally, the model's ability to predict accurately was a crucial step in the process, as it helped determine the effectiveness of the algorithm in solving the problem at hand.

IV. EVALUATION METRICS

5.1 Confusion Matrix

A Confusion Matrix is a crucial tool for evaluating the performance of a classification model. It presents the number of correct and incorrect predictions made by the classifier in a visual format, allowing for easy identification of confusion between classes. This matrix provides a breakdown of the number of correct predictions and misclassifications for each class, making it easier to understand the strengths and weaknesses of the model. By looking at the confusion matrix, one can quickly see which classes are commonly mislabeled, which can help inform future improvements to the model. Overall, the confusion matrix provides a valuable, comprehensive overview of the model's performance, making it an essential component of the evaluation process.

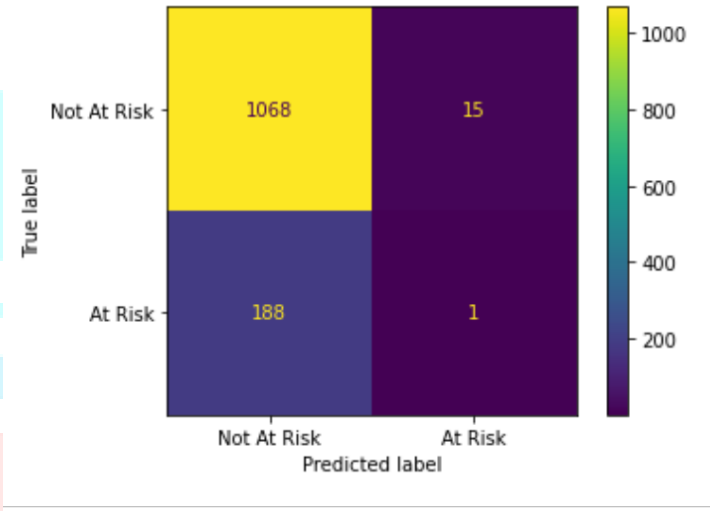

Fig 1 : Confusion Matrix

5.2 Accuracy

Accuracy is a commonly used metric for evaluating the performance of a classification model. It is calculated as the ratio of the number of correct predictions made by the model to the total number of predictions. The accuracy is calculated as :

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Where,

- True Positive (TP) = Observation is positive and it is predicted to be positive.
- False Negative (FN) = Observation is positive, but it is predicted to be negative.
- True Negative (TN) = Observation is negative, and it is predicted to be negative.
- False Positive (FP) = Observation is negative, but it is predicted to be positive

The obtained accuracy after training the model is 85%, meaning that the model correctly predicted 85% of the samples in the test set. However, it's important to note that accuracy is not always the best metric to evaluate the performance of a classifier, especially when the class distribution is imbalanced or when the costs of false positives and false negatives are different. In such cases, other metrics like precision, recall, F1-score, etc., might be more appropriate.

5.3 Precision

Precision is a metric used to evaluate the accuracy of a classifier. It is calculated as the ratio of true positive predictions (correctly classified positive examples) to the total number of positive predictions made by the classifier (true positive plus false positive). High precision means that the classifier produces a small number of false positive predictions, meaning that when the classifier predicts a positive example, it is correct a high percentage of the time. Precision can be calculated using the formula:

Precision = True Positives / (True Positives + False Positives)

The obtained precision after training the model is 0.73.

5.4 Recall

Recall is a measure of a classifier's ability to correctly identify positive examples in a dataset. It is calculated as the number of True Positive (TP) predictions divided by the sum of True Positives and False Negatives (FN). A high recall value indicates a low number of False Negatives, meaning that the classifier is able to identify most positive examples. Mathematically, Recall can be defined as:

Recall = TP / (TP + FN)

The recall after training the model is 0.84.

V. DISCUSSION ON RESULTS

The results of the project where logistic regression was used on the Framingham dataset show an overall accuracy of 85%. This means that out of all the predictions made by the model, 85% of them were correct. This is a good indication that the model is performing well and making accurate predictions.

In terms of precision, the model scored 0.73, which indicates that out of all the positive predictions made by the model, 73% of them were actually positive. Precision is an important metric for evaluating the model's performance, especially in medical applications where false positive predictions can have serious consequences.

The recall score for the model was 0.84, which means that out of all the positive cases in the dataset, the model was able to correctly identify 84% of them. Recall is another important metric in medical applications, as a low recall score can result in missing important cases.

Metrics	Results
Accuracy Score	85%
Precision	0.73
Recall	0.84

VI. CONCLUSIONS

In conclusion, the results of the logistic regression model on the Framingham dataset show good accuracy and precision scores, indicating that the model is performing well and making accurate predictions. However, it is important to consider the trade-off between precision and recall and to use the appropriate metric depending on the requirements of the application.

VII. FUTURE WORKS

Based on the results of the logistic regression model on the Framingham dataset, there are several potential areas for future work:

1. Model improvement: The current model can be further improved by exploring different algorithms or ensembles of algorithms to increase accuracy, precision, and recall.

2. Feature selection: The current model may benefit from feature selection techniques to remove irrelevant or redundant features that could be impacting the performance of the model.

3. Hyperparameter tuning: The model can be optimized by performing hyperparameter tuning to find the best values for the parameters that control the model's behavior.

4. Model interpretability: The model's predictions could be made more interpretable by incorporating feature importance or partial dependence plots. This could help in understanding the model's decisions and how different features contribute to the predictions.

5. Model validation: The model can be validated on independent test data to ensure that its performance on new data is consistent with its performance on the training data.

6. Data augmentation: The model can be trained on augmented data to improve its ability to generalize to new cases and improve its performance on the task at hand.

In summary, these potential future works can help to further improve the performance of the logistic regression model on the Framingham dataset and increase its potential for real-world impact.

VIII. RELATED WORKS

In the field of medical predictive analytics, there have been numerous studies that have used machine learning algorithms on various datasets to make predictions and gain insights into various medical conditions.

[1] M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms"

[2] https://www.kaggle.com/code/ronitf/predicting-heart-disease
[Accessed 5 November 2022].