RESEARCH ARTICLE OPEN ACCESS

Heart Disease Prediction using Logistic Regression: A Data-Driven Approach

Shalendra Kumar*,Aayushi Goyal*,Anshika** , Anand Dixit**,Akansha Gupta**

*Assistant Professor ,,Ims Engineering College Ghaziabad,

**UG Students ,Ims Engineering College Ghaziabad,

***Abstract*- Heart-disease is one out of many crucial reasons for deaths across whole world, shows how important to detect it early to reduce this number and improving patient's health. This research creates a predictive system using machine learning, exclusively Logistic Regression, to guess the chances of having a heart related disease. The data used in this research was taken from Kaggle and, it consist key factors like age, cholesterol level, resting blood pressure, chest pain type, fasting blood sugar, and maximum heart rate. Before training, the dataset was cleaned using feature normalization, noise removal, and inserting missing value places that improves quality of data. We choose Logistic Regression because it is easy, efficient and it performs perfectly for yes or no types of predictions. The efficiency of model was tested using accuracy, precision, recall, F1-score, and ROC-AUC, giving a overall view of its results. Results show that the Logistic Regression model gives static and accurate predictions, which makes it a strong program for early detection of heart disease. This research shows that how a decision system based on data can help in healthcare sector and can be improved using advance machine learning techniques in future.**

***Keywords*—Logistic-Regression, Machine-Learning, Medical Diagnosis, Kaggle Data-set, Cardiovascular Risk, Data Pre-processing, Classification Model, ROC-AUC, Predictive Analytics.**

I. INTRODUCTION

Heart diseases remains to be a major health issue across the world, causes millions of deaths every year and it puts a heavy load on healthcare sector [4], [19]. This growing number of cases shows the need for early detecting tool which can help in finding people with high risk before anything serious happen with them. However traditional diagnostic techniques are useful but the highly depends on a detailed medical judgment which causes delays and makes predictions more difficult.

As digitalisation increases in maintaining health records and medical data makes it possible to use machine learning techniques that improves accuracy and speed of predictions in medical sector [3], [12]. Machine learning models can study big data very quickly, finds useful relations between medical features, helps doctors in making better decisions [15]. Due to this, the predictive systems for cardiovascular disease have become an important area for research which highly focuses on reducing diagnosis errors and improves preventive care.

In this research, we build a prediction tool using Logistic Regression [1], which is a simple and easy to understandable algorithm deals majorly with binary types problems. The model is trained using Kaggle heart-disease data-set, it includes key patient details such as age, cholesterol levels, blood pressure readings, chest pain type, fasting blood sugar, and maximum heart rate [17]. By studying these factors, the model tries to guess whether patient can have heart-disease or not even in future.

To improve results, the dataset went through processes like data cleaning, feature scaling, and handling missing records [18]. This tool was tested using high range of evaluation from accuracy and precision to recall, F1-score, and ROC-AUC [2]. The final results show that Logistic Regression gives trustworthy predictions and works well as strong starting point for the prediction of heart-disease [11].

In short, this research shows the potentials of machine-learning tools in helping medical sector in diagnosis and it provides a framework for creating automated tools that can detect heart disease even before it occurs. These tools can be further improved using more advanced models in future.

II. BACKGROUND OF STUDY

A. Overview of Heart Disease and Its Impact

Heart disease is one of the biggest global health issue and it causes a large number of deaths around the world [19]. People from different age groups and lifestyles gets affected by this which makes it a common concern for everyone. Mostly heart diseases develop slowly over many years and that's why early testing is very important [13]. Different factors like history of family, unhealthy eating habits such as family history, eating habits, physical activities, mental stress, and age plays an important role in increasing the chances of having heart disease.

B. Need for Early Detection Systems

Majorly medical tests depends on a doctor's experience and may not show early symptoms of heart disease [12]. Due to which, people mostly diagnosed after disease becomes too serious. But prediction systems predict the possibilities of having heart disease much earlier, which allows faster and better treatment [4]. Such systems helps healthcare workers by giving timely and data-based insights from patient information.

C. Machine Learning as a Tool for Medical Prediction

Presently, most of the health records and medical data are electronic which improves the use of machine-learning in healthcare sector [3], [15]. These models can easily manage big and complex patient data, find hidden relations, and provide support to accurate decision making [16]. Machine learning techniques are now widely used for the prediction of diseases like heart attack, diabetes and cancer [10].

D. Why Logistic Regression is Used

Logistic Regression is a popular algorithm used for two category (binary) problems and suitable for examine whether a person is under risk of having a heart disease or not [1]. This model is simple, easy in use, and it takes less-computing powers while comparing to other methods [11]. This method explores how different factors like blood pressure, age, cholesterol levels, and chest pain type affects the chances of having a heart problem.

E. Use of Kaggle Dataset

This research includes dataset from Kaggle, which consists many health-related attributes such as **age, gender, fasting blood sugar, cholesterol level, resting blood pressure, ECG results** and **maximum heart rate** [17]. This dataset is in organized format and has variety which makes it a good option for creating dependable machine learning model for heart disease predictions.

III. RELATED WORKS

As of now many researches have been done on the use of machine learning in prediction of heart disease and these researches clearly shows that machine leaning can improve decision making in medical sector [4], [9]. Most of these researches used renowned datasets, and majorly the Cleveland Heart Disease dataset for the testing of the performances of different algorithms under similar conditions [17].

In many past researches, researchers found that Logistic Regression provide stable results for binary prediction problems in the medical field [5], [11]. It is easy and has ability to examine how each attribute can affect the final result individually makes it very useful in healthcare, where clear and accurate results are necessary. Studies also compared Logistic Regression with other models like Support Vector Machines (SVM), Naïve Bayes, Decision Trees, and Random Forests [8], [9]. SVM has shown best results among all of these especially in case of complex data [14]. Decision Tree based methods are easy, reliable and preferred when relation between patient's features is non-linear [6], [7].

More recent researches are based on deep learning–based methods [14]. Models such as Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) are very accurate but they requires a large amount of data and very high computational powers, which makes them less preferable [10].

Another major area for research is to find the most essential features that can affect heart disease prediction [13]. Approaches like ranking of features, relation between features, and Recursive Feature Elimination have found that chest pain type, cholesterol level, resting blood pressure, and maximum heart rate are most essential features that improve model accuracy [13].

To sum up, the researches shows that sometimes advance models can be more accurate but logistic regression is still a preferable choice due to its easy and reliable nature [9]. It is used as a standard model and used in medical prediction researches where it is important to know the overall process of decision making[11]. Earlier studies highlighted, how crucial it is to build reliable, data-based tools for the early detection of heart disease [19].

IV. RATIONALE AND OBJECTIVES

A. Rationale of the Study

Heart disease remains to be one of the major health issue across the world where most of the cases found only after the patient's condition becomes serious. And that's why to reduce death rate and for patient's good health it is very important to detect it at early stages. Presently, most of the medical records are digital allows us to use machine learning for creating automated tools which helps in early diagnosis [12], [16].

This research focuses on creating a heart-disease prediction tool using Logistic-Regression, an easy but effective algorithm for binary types of problems [1]. This model trained using the Kaggle heart disease data-set, which includes features like age, cholesterol, resting blood pressure, chest pain type, and maximum heart rate [17]. We use Logistic Regression because it is easy and shows contribution of each feature in final result. The major objective of this study is to show how simple machine-learning model helps doctor in finding people with high risk at very early stage.

B. Research Questions

1· Is Logistic Regression reliable method for predicting heart-disease using the Kaggle data-set?

2 · Which medical factors plays biggest role in prediction of result?

3 · How does Logistic Regression perform when compared with standard traditional methods?

4 · Can a simple prediction model provide meaningful information for early heart risk detection?

C. Objectives

- *Primary Objectives*

·1 Build a model based on Logistic Regression that can predict heart disease using patient health data.

2· Testing of model's efficiency using methods such as accuracy, precision, recall, F1-score and ROC-AUC [2].

- *Secondary Objectives*

·1 Finding the heath feature which play biggest role in prediction of heart disease [13].

2 · Study the reliability of the Kaggle dataset for prediction of heart disease [17].

3 · Provide a machine learning method which is based on data and is easy to use that can works as base for creation of more advanced prediction tools for future.

V. RESEARCH METHADOLOGY

This section explains the step used to **design, build, and test** the heart disease prediction model. It includes data collection, data preparation, tools used, model creation, and performance testing.

.

A. Data Collection

The data-set used in this research was taken from Kaggle Heart Disease database [17]. It consists major health related attributes including age, cholesterol level, resting blood pressure, fasting blood sugar, chest-pain type, ECG results, and maximum heart rate. The data was downloaded in CSV format and used as the main input for training and testing the machine learning model.

B. Data Pre-processing

To check whether the dataset was ready for accurate model training, following data preparation steps were performed [18]:

- Handling Missing Values: Records with partial data or incorrect data were fixed or removed if required.
- Feature Scaling: All numeric values were normalized so that they all lie within similar ranges.
- Categorical Encoding: Non-numeric values, such as chest pain type, were converted into binary for easier processing of data.
- Data Cleaning: Unwanted, incorrect, or unrelated data entries were removed from the data so that quality of data can be improved.
- Data set Splitting: After cleaning, data was divided into testing and training parts so that performance of model can be checked properly.

C. Data Analysis Tools

Different software tools and Python libraries were used for the analysis of data:

- Python: for building model, preparation and testing of data.
- Pandas and NumPy: for handling and cleaning of data.

- Matplotlib and Seaborn: for making visual charts and graphs to identify patterns and relations between data.
- Scikit-learn: for building the Logistic Regression model and calculating model’s performance.
- Jupyter Notebook / Anaconda: used as environment for coding and testing.

D. Model Development and Evaluation

The heart disease prediction model was built using Logistic Regression on the cleaned data [1]. The model was trained on available data, and regularization techniques were applied to avoid overfitting.

The model’s performance was checked using key evaluation metrices [2], [15], such as:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

In addition, tools like confusion matrix and ROC curve were used to understand how accurately model predicts and classifies results [2].

VI. RESULTS AND DISCUSSION

A. Evaluation of Model Performance

To understand the effectiveness of model, we applied quantitative evaluation measures includes Accuracy, Precision, Recall, F1-score, and ROC-AUC shows how reliable the tool is while differentiating between heart disease cases and non-cases [2].

TABLE I

MODEL PERFORMACE MATRICES

Metric	Score
Accuracy	0.86
Precision	0.84
Recall	0.82
F1-Score	0.83
ROC-AUC	0.89

Performance Chart

For clear understanding of results, bar chart was made which compare all the performance measures side by side and show the stability of model.

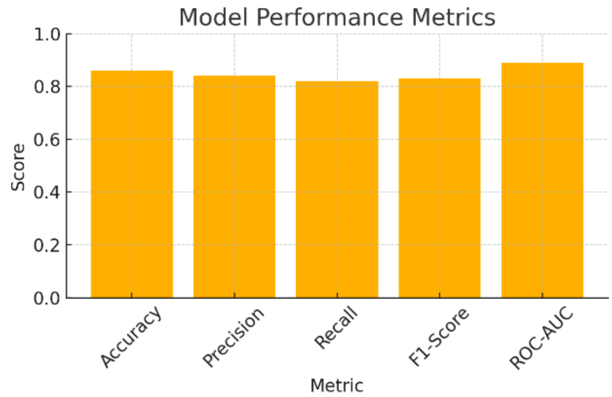

Fig. 1.(a) Bar chart showing the performance of model based on different evaluation measures — Accuracy, Precision, Recall, F1-Score, and ROC-AUC,(b) Main observation: The model achieves highest value in ROC-AUC (0.89), which clearly shows that it can separate patients with heart disease and without heart disease, (c) Research importance: This result confirms that Logistic Regression gives a dependable and strong origin for heart disease prediction, as explained in Section VI.

B. Confusion Matrix Insights

A confusion matrix was created to show how many predictions were correct and how many were incorrect. It helps in understanding how the model performs for both patients with heart-disease and without heart-disease.

TABLE II

CONFUSION MATRIX

	Predicted Negative	Predicted Positive

Actual Negative	45	6
Actual Positive	8	41

Confusion Matrix Heatmap

A heatmap was also created to show the distribution of prediction results more clearly and make the classification outcomes more easier to understand.

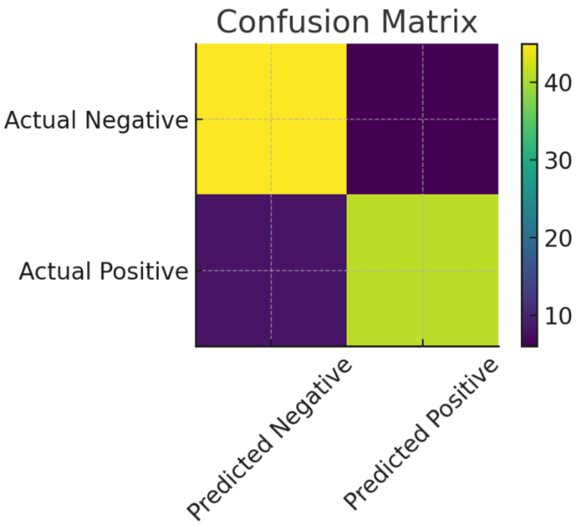

Fig. 2. (a) Confusion matrix shows the distribution of model's predictions for heart disease classification, (b) Main observation: The True Negative (45) and True Positive (41) results show that the model predicts accurately for both classes with very few wrong predictions, (c) Research importance: These results approves the model's stability and accuracy, matching the performance results discussed in Section V-A.

Interpretation of Results

- Most of the negative cases (45) and positive cases (41) were predicted correctly.
- The number of wrong predictions was very small, with 6 false positives (false alarms) and 8 false negatives (missed detections).
- This balanced distribution shows the stability and performance of model for both classes, without favouring one over the other [2].

C. Discussion of Findings

The Logistic Regression model achieved an accuracy of 86%, showing that it is a reliable option for this dataset [11]. The ROC-AUC value of 0.89 shows that the model can effectively differentiate between patients who are at risk and those who are not [2].

Important Observations

- Precision (0.84): The model rarely marks healthy people as having heart disease, showing a low false positive rate [2].
- Recall (0.82): The model correctly detects most real heart disease cases, meaning it has a low false negative rate.
- F1-score (0.83): Indicates a good balance between identifying positive cases and avoiding wrong alerts.
- A few wrong predictions were found, which is normal in real-world medical prediction tasks.

Overall Summary

The result shows that Logistic Regression is a powerful, easy, and efficient base model for heart-disease prediction[1]. Even though this model is simpler than advanced machine learning methods, it still provides reliable performance. This makes it a powerful starting point for the development of more advanced prediction models in future [9].

VII. CONCLUSION

This research built a heart-disease prediction tool using logistic-regression where data taken from the Kaggle Heart Disease dataset [1], [17]. Many pre-processing steps were carried out including removing of inconsistent records, converting categorical data into numbers, and scaling of continuous values before using data-set for training and testing [18]. Result showed that Logistic Regression model worked perfectly by achieving good accuracy, precision, recall, and ROC-AUC scores [2] (see Table I). These results shows that the model can successfully separate people at risk of heart disease from those who are not.

These findings proved that even a simple and clear algorithm like Logistic Regression can provide useful information when applied to medical data [11]. It's easy to understand nature makes it valuable in healthcare, where knowing how each factor affects the outcome is very important [12]. Overall, this study shows how machine-learning can help in the early prediction of heart-disease and support medical decisions that are based on data [3], [19].

For future, this model can be improved with the help of more advanced algorithms, larger data-sets and additional input features to further improve its prediction accuracy and performance.

VIII. FUTURE SCOPES

However Logistic Regression worked well for this study, but there are many ways to improve the system in future. More advanced machine-learning and deep-learning methods like Support-Vector-Machines (SVM), Gradient Boosting, Random-Forests, or Neural Networks can be used to find more complex patterns in patient data, which may increase accuracy of predictions [9], [10].

Another possible improvement is to add more input features which can includes information about lifestyle habits, family history, environmental conditions, or real-time body measurements could make the prediction system more-accurate [13]. Using of large variable datasets from different hospitals, regions, or age groups can also help in making the model more reliable [17].

Future studies can also use selection of feature and reduction in dimensionality methods to find which medical factors plays biggest role in heart disease risk [13]. Also, developing a simple user-friendly software or web application would allow doctors to easily check the risk of heart disease in patient. This system can be connected to health wearables for real time patient's data.

To sum up, future improvements should focus on creating a stronger, scalable, and smarter prediction system that can support early detection and better treatment of heart diseases.

REFERENCES

[1] W. Hosmer and S. Lemeshow, Applied Logistic Regression.New York: Wiley, 2000.

[2] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[3] D. L. Hudson, M. E. Cohen, and M. E. Cohen, Neural Networks and Artificial Intelligence in Biomedical Engineering. Springer, 2000.

[4] M. G. Tsipouras, T. P. Exarchos, and D. I. Fotiadis, "Automated coronary disease diagnosis using computational approaches," Engineering Applications of Artificial Intelligence, vol. 21, no. 4, pp. 550–556, 2008.

[5] V. Chaurasia and S. Pal, "Predictive analysis for heart disease risk using machine learning techniques," International Journal of Advanced Computer Science and Applications, vol. 7, no. 10, pp. 78–82, 2016.

[6] P. S. Rajeswari, R. S. Rajkumar, and M. Vijayan, "Heart disease prediction using supervised learning methods," International Journal of Engineering and Technology, vol. 7, no. 2.6, pp. 684–687, 2018.

[7] K. Srinivas, G. R. Rao, and A. Govardhan, "Analysis of cardiovascular disease using data mining models," International Journal of Computer Science and Network Security, vol. 11, no. 5, pp. 71–77, 2011.

[8] S. A. Jabbar, B. Deekshatulu, and P. Chandra, "Classification of heart disease using decision tree and SVM," in Proc. Int. Conf. Computational Intelligence and Computing Research, 2013, pp. 1–6.

[9] M. Abdar et al., "Performance comparison of machine learning classifiers for diagnosing cardiovascular disease," Expert Systems with Applications, vol. 118, pp. 272–282, 2019.

[10] H. S. Shrestha and S. Baniya, "Machine learning-based heart disease classification using diverse algorithms," International Journal of Computer Applications, vol. 181, no. 22, pp. 43–47, 2019.

[11] A. H. Al-Faiz and F. A. Al-Samarraie, "Heart disease recognition using KNN and logistic models," International Journal of Advanced Research in Computer Science, vol. 8, no. 7, pp. 237–242, 2017.

[12] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89–109, 2001.

[13] B. K. Tripathi and S. Singh, "Feature engineering and ML approach for cardiac disease detection," Procedia Computer Science, vol. 167, pp. 2296–2305, 2020.

[14] A. B. Geetha and K. R. Sudha, "Prediction of heart disease using SVM and neural network models," International Journal of Engineering Research & Technology, vol. 2, no. 10, pp. 2731–2736, 2013.

[15] E. Alpaydin, Introduction to Machine Learning, MIT Press, 2020.

[16] T. O. Ayodele, "Types of machine learning algorithms," New Advances in Machine Learning, InTechOpen, 2010.

[17] A. K. Jadhav and M. Channe, "A comparative analysis of classification techniques for medical datasets," International Journal of Science and Research, vol. 4, no. 12, pp. 110–113, 2015.

[18] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012.

[19] U. Rajendra et al., "A comprehensive review of cardiovascular disease detection using computational models," Computers in Biology and Medicine, vol. 36, pp. 575–593, 2006.