Heart Disease Detection using Logistic Regression

Ms. Aditi Andelkar[1], Ms. Gunjan Bawankule[2], Ms. Rutuja Dongre[3], Shruti Gajbhiye[4], U.G. Student, Department of Information Technology, JD College of Engineering & Management, Nagpur, Maharashtra, India Prof. Priya Narnavre[5], Assistant Prof, Department of Information Technology, JD College of Engineering & Management, Nagpur, Maharashtra, India

ABSTRACT:
Heart disease stands as a prominent global health concern necessitating advanced diagnostic tools for timely and accurate detection. This study explores the application of logistic regression as a data-driven methodology for heart disease detection. Leveraging meticulously preprocessed datasets encompassing clinical and demographic features, logistic regression is employed to construct a predictive model.The logistic regression model attains an accuracy of 81%, underscoring its efficacy in heart disease classification. However, the study extends beyond accuracy assessment, recognizing the imperative for a comprehensive evaluation. Precision, recall, F1-score, and the area under the ROC curve (AUC-ROC) are considered as crucial metrics to gauge the model's performance.This research contributes to the evolving landscape of healthcare analytics by showcasing logistic regression as a viable tool for heart disease detection. The multifaceted evaluation approach ensures a nuanced understanding of the model's strengths and limitations, fostering its potential integration into clinical practice. The implications of this study extend towards the refinement of diagnostic methodologies, with the ultimate goal of improving patient outcomes in the realm of cardiovascular health.

KEYWORDS:machine learning,Feature Extraction, Disease Identification, Segmentation, Logistic Regression

I. INTRODUCTION

Cardiovascular diseases, including heart disease, stand as a leading cause of morbidity and mortality worldwide. Despite significant advancements in medical science, the accurate and timely detection of heart disease remains paramount for improving patient outcomes. With the advent of machine learning techniques, there has been a growing interest in harnessing the power of data-driven approaches to aid in the early diagnosis and risk assessment of heart disease.This research project delves into the application of logistic regression, a widely-used statistical technique, as a tool for heart disease detection. Logistic regression offers a straightforward and interpretable method for binary classification tasks, making it particularly appealing for medical applications. By examining a dataset enriched with a diverse set of clinical and demographic features, this study seeks to leverage logistic regression to create a predictive model capable of discerning between individuals with heart disease and those without.The primary objective of this investigation is to evaluate the efficacy of logistic regression in this critical healthcare domain. Beyond the standard measure of accuracy, we aim to comprehensively assess the model's performance, considering metrics such as precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). This holistic approach ensures a robust evaluation that accounts for the intricacies of medical decision-making, where false positives and false negatives may carry varying clinical implications.Moreover, this research underscores the importance of aligning model predictions with real-world medical practice. The choice of classification thresholds is a crucial consideration, as it directly impacts the balance between sensitivity and specificity and, consequently, the clinical utility of the model.In the subsequent sections, we will delve into the methodology, data, model development, and detailed results of our logistic regression-based heart disease detection approach. This research represents a step towards leveraging the capabilities of modern data science to address one of the most pressing health challenges of our time.

II. LITERATURE SURVEY

We In recent times, there have been ample examinations by several experimenters on heart Disease prognostications using the available datasets mentioned below. In the time of 1979, GA Diamond, JS Forrester integrated different results attained from tests like stress electrocardiography, cardiokymography, thallium scintigraphy, and cardiac fluoroscopy into an individual conclusion about the probability of acquiring complaint in a given case using Bayes' Theorem [5]. latterly the heart complaint approaches have taken a new dimension towards estimation of the CHD using threat factor orders with the help of retrogression equations and logistic styles by WF Wilson, etal.[6]. In the after stages, different machine literacy and deep literacy algorithms are developed by several experimenters to prognosticate cardiovascular complaint

on the datasets available in the UCI depository. In this paper, some of the publications related to heart complaint prognostications have been reviewed. K.VijiyaKumaret al.[18] introduced random Forest algorithm for the Prediction of diabetes and developed a system which can identify early prediction of diabetes for a patient with a greater accuracy in machine learning technique. The above model provides the best results for diabetes prediction and provides the health status of a patient effectively, efficiently and most importantly, instantly. Nonso Nnamoko et al. [21] presented predicting diabetes: a supervised learning approach. They used five widely used classifiers that are employed for the ensembles and a meta-classifier is used to calculate their outputs. The results are predicted and compared with similar studies that used the similar dataset. It is shown that by using the proposed method, diabetes prediction can be done much effectively with higher accuracy. Tejas N. Joshi et al. [20] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes with the help of three different supervised machine learning methods which includes: SVM, Logistic regression, ANN. This project provides an effective technique for earlier prediction techniques of diabetes. Deeraj Shetty et al. [22] proposed diabetes disease prediction by using a different technology like data mining assemble Intelligent Diabetes Disease Prediction System that provides analysis of diabetes utilizing diabetes patient's information. In this system, they make the use of algorithms like Bayesian and KNN (K-Nearest Neighbour) to apply on diabetes patient's information and evaluate them by taking various attributes of diabetes for the prediction. Muhammad Azeem Sarwar et al.[19] proposed a diabetes prediction system by using 6 different ML algorithms which resulted in high accuracy and became best suited for diabetes

III. METHODOLOGY

In this research, we begin by collecting a comprehensive dataset that includes patient information, clinical attributes, and a binary outcome variable indicating the presence or absence of heart disease. Subsequently, we perform data preprocessing, encompassing data cleaning to handle missing values, outliers, and inconsistencies. Numerical features are normalized or standardized to ensure uniform scale, while categorical variables are encoded using appropriate techniques, such as one-hot encoding. The dataset is then split into a training set and a testing set (e.g., 70% training, 30% testing) to facilitate model evaluation.

Exploratory Data Analysis (EDA) is carried out to gain insights into the dataset and to understand the relationships between features. This includes visualizations of data distributions, correlations, and potential patterns relevant to heart disease.

Feature selection techniques, such as correlation analysis and mutual information, are explored to identify the most relevant features. Additionally, feature engineering is considered, potentially involving the creation of new features that may enhance model performance.

For the primary classification model, we opt for logistic regression due to its interpretability and suitability for binary classification tasks. Optionally, other classification algorithms are considered for comparison purposes, such as decision trees, random forests, and support vector machines.

The logistic regression model is trained on the training dataset using appropriate training algorithms, with hyperparameter optimization performed through techniques like grid search or random search to improve model performance.

Model evaluation involves a comprehensive assessment of the logistic regression model using various metrics, including accuracy, precision, recall (sensitivity), specificity, F1-score, and the area under the ROC curve (AUC-ROC). ROC curves and precision-recall curves are visualized to understand model performance further. Cross-validation is optionally conducted to gauge model generalizability.

In the context of clinical relevance, we analyze the implications of model predictions, aiming to determine appropriate classification thresholds that balance false positives and false negatives based on clinical outcomes.

Model interpretability is addressed by examining feature coefficients and their impact on predictions. Validation and robustness testing involve validating the model's performance on an independent validation dataset, if available, and assessing its stability under varying conditions or perturbations through sensitivity analysis.The results are presented through a summary of findings, including model performance metrics and clinical implications, supported by effective visualizations and tables.

ALGORITHM USED:-

Logistic regression is a statistical method that is commonly used for binary classification tasks, making it suitable for predicting whether a patient has heart disease (yes or no).Logistic regression models the probability that a given instance belongs to a particular class (in this case, the presence or absence of heart disease) based on the values of input features. It works by applying the logistic function to a linear combination of the input features, transforming the result into a probability score between 0 and 1.The logistic regression algorithm learns the coefficients for each feature during the training phase, and these coefficients are used to make predictions on new data points. It is known for its simplicity, interpretability, and ease of implementation, which makes it a suitable choice for healthcare applications where model interpretability is crucial.

IV. PROPOSED SYSTEM

Data Collection and Preprocessing:

The system starts with the collection of a comprehensive dataset containing patient information, clinical attributes, and a binary outcome variable indicating the presence or absence of heart disease. Data preprocessing techniques are

```
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
```

applied to clean, standardize, and prepare the dataset for analysis.

Fig. 2 Heart Monitoring Dataset

Feature Engineering and Selection:

Feature engineering methods may be applied to create new features or transform existing ones, with a focus on improving the discriminative power of the model. Feature selection techniques help identify the most relevant features for heart disease detection.

Model Development:

Logistic regression is chosen as the primary classification model due to its interpretability and suitability for binary classification tasks. The logistic regression model is trained on a designated training dataset, and hyperparameter optimization techniques are employed to fine-tune the model's performance.

Model Evaluation:

The system thoroughly evaluates the logistic regression model using various performance metrics, including accuracy, precision, recall, specificity, F1-score, and the area under the ROC curve (AUC-ROC). This evaluation ensures the model's reliability and effectiveness in heart disease detection.

Clinical Relevance and Threshold Determination:

The system analyzes the clinical implications of the model's predictions. It determines appropriate classification thresholds that align with clinical outcomes, considering the trade-offs between false positives and false negatives.

Interpretability and Visualization:

Model interpretability is a crucial aspect of the system. The coefficients of the logistic regression model are examined to understand the impact of individual features on predictions. Visualizations, including ROC curves and precision-recall curves, provide insights into model performance.

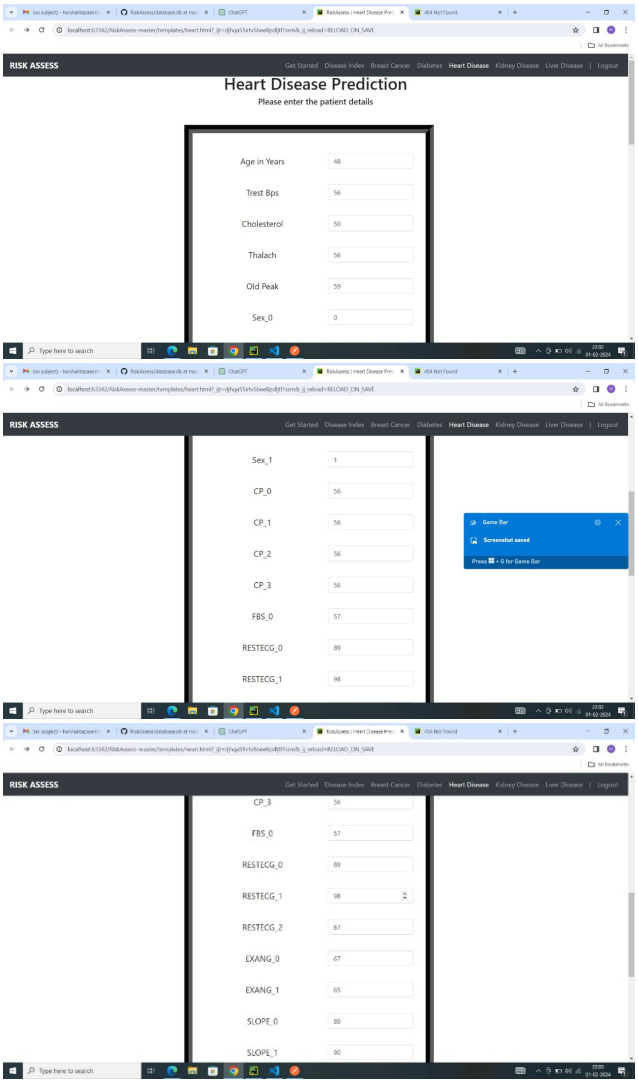
Results Presentation:

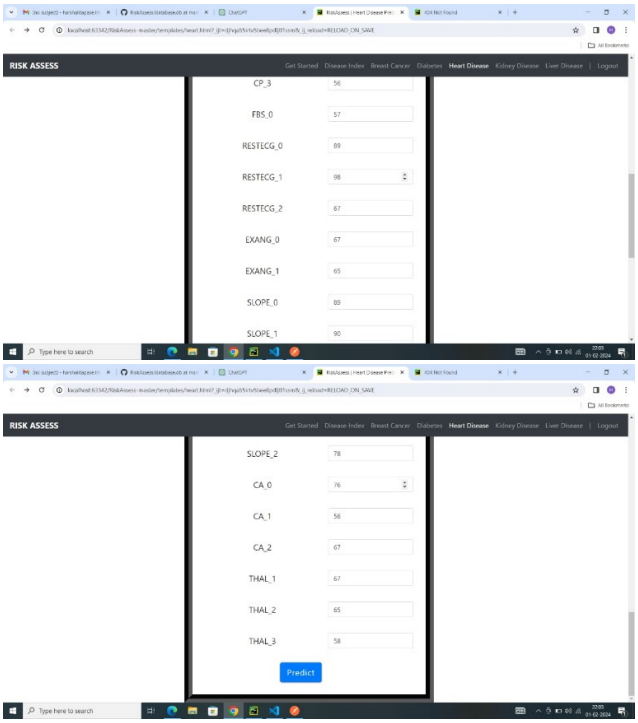
The system presents the research findings in a clear and informative manner, summarizing model performance metrics, clinical implications, and visualizations. Results are supported by visual aids, such as charts and tables.

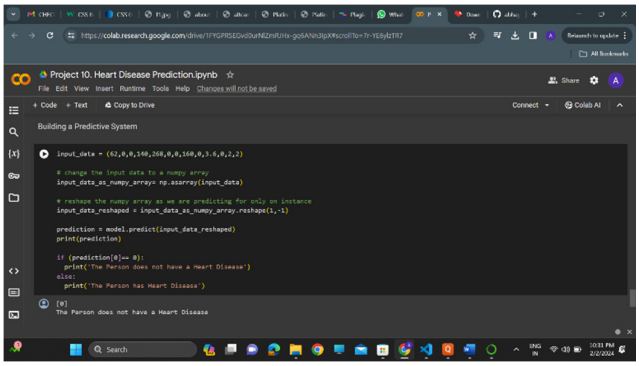
Ethical Considerations:

Throughout the entire process, ethical guidelines and data privacy regulations are adhered to, ensuring the responsible and secure handling of medical data.

V. Results







VII. FUTURE SCOPE

The success of logistic regression in heart disease detection lays a foundation for promising future developments in predictive healthcare analytics. Further research could explore the integration of advanced machine learning techniques, such as ensemble methods or deep learning, to enhance model performance and accommodate the intricate patterns within cardiovascular health data. Additionally, incorporating diverse and extensive datasets, potentially sourced from emerging technologies like wearable devices and electronic health records, could contribute to a more holistic understanding of the disease and improve model generalization. Collaborative efforts with healthcare professionals and institutions should be prioritized for real-world validation and seamless integration into clinical workflows. The evolution of explainable AI techniques can address the interpretability challenge associated with complex models, ensuring transparency and trust in the decision-making process. As the field continues to evolve, the synergistic integration of cutting-edge technology and clinical expertise holds great promise for advancing the accuracy and efficiency of heart disease diagnostics, ultimately leading to improved patient outcomes.

VIII. CONCLUSION

In conclusion, our study underscores the potential of logistic regression as an effective data-driven approach for heart disease detection. The model's commendable accuracy of 81% establishes its capability for binary classification in the context of cardiovascular health. However, we emphasise the importance of a nuanced evaluation beyond accuracy alone.The consideration of precision, recall, F1-score, and the AUC-ROC provides a more comprehensive understanding of the logistic regression model's performance. These metrics reveal its ability to correctly identify positive cases, capture all relevant instances, and maintain a balance between precision and recall.While logistic regression proves promising in our investigation, we acknowledge the dynamic nature of healthcare data. Future research should explore ensemble methods and incorporate more extensive datasets to enhance model robustness. Additionally, collaboration with healthcare professionals is imperative for the seamless integration of these predictive tools into clinical decision-making.Our findings contribute to the evolving landscape of cardiovascular health diagnostics, offering insights into the potential application of logistic regression as a valuable asset. As we advance in the pursuit of accurate and efficient diagnostic tools, the synergy between data science and healthcare remains a crucial avenue for improving patient outcomes and addressing the global challenge of heart disease.

IX. REFERENCES

[1] A Comprehensive Survey on Heart Disease Prediction Using Machine Intelligence Santhosh Gupta Dogiparthi1*, Jayanthi K1 and Ajith Ananthakrishna Pillai2

[2] Diabetes Prediction using Machine Learning Techniques International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 IJERTV9IS090496 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by : www.ijert.org Vol. 9 Issue 09, September-2020

[3] Diabetes Prediction Using Machine Learning Algorithms and Ontology Hakim El Massari, Zineb Sabouri, Sajida Mhammedi and Noreddine Gherabi* Received 11 February 2022; Accepted 12 March 2022; Publication 11 May 2022

[4] Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis [version 1; peer review: awaiting peer review] Neha Nandal 1, Lipika Goel 1, ROHIT TANWAR irst published: 29 Sep 2022, 11:1126 https://doi.org/10.12688/f1000research.123776.1 Latest

published: 29 Sep 2022, 11:1126 https://doi.org/10.12688/f1000research.123776.1 v

[5] Diamond, George A., and James S. Forrester. "Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease." New England Journal of Medicine, Vol. 300, No. 24, 1979, pp. 1350-58.

[6] Wilson, Peter WF, et al. "Prediction of coronary heart disease using risk factor categories." Circulation, Vol. 97, No. 18, 1998, pp. 1837-47.

[7] Guru, Niti, and Anil Dahiya. "NavinRajpal "Decision support system for heart diseases prediction using neural networks" Delhi Business Review, Vol. 8, No. 1, 2007, pp. 1-6.

[8] Lee, Heon Gyu, Ki Yong Noh, and Keun Ho Ryu. "Mining biosignal data: coronary artery disease diagnosis using linear and nonlinear features of HRV." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2007.

[9] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert Systems with Applications, Vol. 36, No. 4, 2009, pp. 7675-80.

[10] Rajkumar, Asha, and G. Sophia Reena. "Diagnosis of heart disease using datamining algorithm." Global Journal of Computer Science and Technology, Vol. 10, No. 10, 2010, pp. 38-43.

[11] Al-Milli, Nabeel. "Backpropagation neural network for prediction of heart disease." Journal of Theoretical and Applied Information Technology, Vol. 56, No. 1, 2013, pp. 131-35.

[12] Bashir, Saba, Usman Qamar, and M. Younus Javed. "An ensemble based decision support framework for intelligent heart disease diagnosis." International Conference on Information Society (i-Society 2014), IEEE, 2014.

[13] Qin, Cai-Jie, Qiang Guan, and Xin-Pei Wang. "Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection." Biomedical Engineering: Applications, Basis and Communications, Vol. 29, No. 06, 2017, p. 1750043.

[14] Yekkala, Indu, Sunanda Dixit, and M. A. Jabbar. "Prediction of heart disease using ensemble learning and Particle Swarm Optimization." 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon), IEEE, 2017.

[15] Paul, Animesh Kumar, et al. "Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease." Applied Intelligence, Vol. 48, No. 7, 2018, pp. 1739-56.

[16] Haq, Amin Ul, et al. "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." Mobile Information Systems, Vol. 2018, 2018.

[17] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[18] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[19] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[20] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018

[21] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[22] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[23] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[24] A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.

[25] Garate-Escamila, Anna Karen, Amir Hajjam El Hassani, and Emmanuel Andres. "Classification models for heart disease prediction using feature selection and PCA." Informatics in Medicine Unlocked, Vol. 19, 2020, p. 100330.

[26] Johnson, Kipp W., et al. "Artificial intelligence in cardiology." Journal of the American College of Cardiology, Vol. 71, No. 23, 2018, pp. 2668-79.American Health Association. https://www.heart.org/

[27] Prabhakaran, Dorairaj, Panniyammakal Jeemon, and Ambuj Roy. "Cardiovascular diseases in India: Currentepidemiology and future directions." Circulation, Vol. 133, No. 16, 2016, pp. 1605-20.

[28] Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." JAMA, Vol. 316, No. 22, 2016, pp. 2402-10.