iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XII **Month of publication:** December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57336

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XII Dec 2023- Available at www.ijraset.com

Heart Disease Prediction Using ML (CARDIO-CARE)

G. Akshith[1], M. Akshitha[2], K. Akshith[3], E. Akshaya[4], CH. Akshitha[5], N. Ramarao[6], P. Akshitha[7]

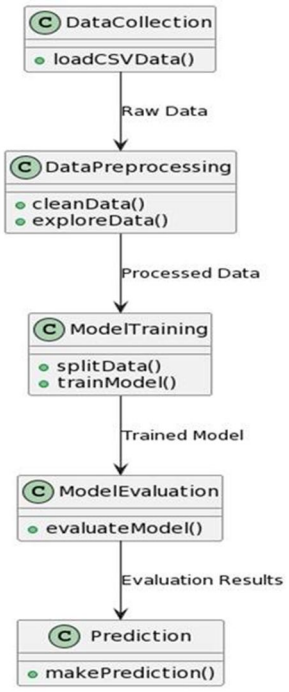
Abstract: ***This project addresses the critical issue of cardiovascular disease by employing machine learning techniques for early heart disease prediction.***
The dataset encompasses diverse patient attributes, including age, gender, blood pressure, and cholesterol levels. Our objectives include data preparation and cleaning, followed by exploratory data analysis to understand the heart disease distribution and feature relationships.
The dataset is then split into training and testing sets, with models trained using Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest algorithms.
The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc.[1]
Model accuracy evaluations reveal that the Random Forest model outperforms the others, offering a compelling choice for potential clinical implementation. Our primary aim is to create a practical predictive system, where individual medical information can be input to determine heart disease risk. This project's ultimate goal is to provide an effective tool for early detection and intervention, potentially reducing the global burden of cardiovascular diseases, thus promoting better patient outcomes.
Keywords: Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest

I. INTRODUCTION

The main goal of the Heart Disease Prediction project is to create a strong machine learning model that can accurately predict the likelihood of individuals having heart disease based on various health-related factors.
The project involves important tasks such as importing necessary dependencies, collecting and pre-processing data using the pandas library.
Exploratory Data Analysis (EDA) is conducted to gain insights into the distribution, correlations, and statistical measures of the dataset. Afterwards, the dataset is divided into training and testing sets for model training and evaluation.
Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest algorithms are utilized to construct predictive models. The project aims to achieve high accuracy on both the training and testing data, ensuring the reliability and generalizability of the models.
Ultimately, a predictive system is established, enabling the input of individual health parameters to determine the presence or absence of heart disease. Overall, the project aims to provide a valuable tool for early detection of heart disease, thereby promoting proactive healthcare interventions.

II. PROBLEM STATEMENT

The prevailing global issue of heart disease, a pervasive and life-threatening condition affecting a significant portion of the population, underscores the critical need for advanced predictive models. Despite medical advancements, the lack of an accurate and interpretable system for early detection and precise prediction of heart disease risk poses a substantial challenge in preventive healthcare.
The complexity of individual health profiles, encompassing diverse demographic and physiological parameters, necessitates a sophisticated solution that can effectively analyze and interpret this multifaceted data. This research project addresses the pressing problem of developing a reliable and adaptable predictive system, utilizing machine learning techniques, specifically logistic regression, to enhance the accuracy of heart disease prediction.
By leveraging comprehensive datasets and advanced analytics, the project aims to contribute to the development of a robust tool that empowers healthcare professionals and individuals alike in proactively managing and mitigating the risks associated with heart disease.



(a) Dataflow diagram

III. LITERATURE REVIEW

The literature surrounding heart disease prediction and machine learning techniques highlights a growing interest in leveraging data-driven approaches for proactive healthcare. Numerous studies have explored the use of diverse datasets containing patient information, with features ranging from demographic details to physiological indicators. Logistic regression, a commonly employed algorithm, has been demonstrated in various research papers as effective for predicting heart disease due to its simplicity and interpretability. This paper contains a brief literature survey. An efficient Cardiovascular disease prediction has been made by using various algorithms some of them include Logistic Regression, KNN, Random Forest Classifier Etc. It can be seen in Results that each algorithm has its strength to register the defined objectives [2].

Recent advancements in machine learning models, such as ensemble methods and deep learning, have also been investigated for their potential to enhance predictive accuracy. Studies emphasize the importance of feature selection and data preprocessing techniques to optimize model performance. Interpretability remains a critical concern, and researchers are actively exploring ways to make machine learning models more transparent and understandable for healthcare practitioners.

Furthermore, the literature underscores the need for collaboration between data scientists and medical professionals to ensure the clinical relevance and applicability of predictive models. Overall, the synthesis of machine learning and cardiovascular research in the literature provides a foundation for the present project, guiding methodologies and inspiring avenues for future exploration.

IV. METHODOLOGY

In this study, a dataset containing important health-related parameters, such as demographic information, cholesterol levels, and medical history, was utilized. Prior to analysis, a rigorous data preprocessing stage was implemented to handle missing values, normalize numerical features, and encode categorical variables. Special attention was given to outlier detection and mitigation. Exploratory Data Analysis (EDA) was then conducted to reveal the distribution of individual features, correlations, and potential patterns. Visualizations, such as heatmaps, were used to enhance understanding. Feature selection techniques, including correlation analysis and domain expertise, were employed to identify the most influential predictors for heart disease, resulting in a refined dataset. The dataset was subsequently divided into training and testing sets, ensuring a balanced distribution of the target variable through stratification. Logistic Regression, chosen for its interpretability, served as the primary predictive model. It underwent thorough training on the designated dataset with ongoing convergence monitoring. Model evaluation included accuracy scores for both training and testing datasets, as well as precision, recall, and F1-score metrics for a comprehensive assessment.

Additionally, the performance of the logistic regression model was compared to alternative models, such as K-Nearest Neighbors, Support Vector Machines, Gaussian Naive Bayes, Decision Trees, and Random Forests. This comparison provided insights into the relative effectiveness of different algorithms. This methodological approach guarantees a systematic and comprehensive investigation into heart disease prediction, incorporating traditional statistical methods and advanced machine learning techniques to achieve accurate and interpretable outcomes.

V. EXPERIMENT RESULTS

```
#accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction,Y_train)

print("Accuracy on training data using logistic regression : ",training_data_accuracy)
```

```
Accuracy on training data using logistic regression :  0.8636363636363636
```

```
from sklearn.metrics import accuracy_score
Y_train_pred = classifier.predict(X_train)
training_accuracy = accuracy_score(Y_train, Y_train_pred)
print("Accuracy on training data (KNN):", training_accuracy)
Y_test_pred = classifier.predict(X_test)
testing_accuracy = accuracy_score(Y_test, Y_test_pred)
print("Accuracy on testing data (KNN):", testing_accuracy)
```

```
Accuracy on training data (KNN): 0.9979338842975206
Accuracy on testing data (KNN): 0.9590163934426229
```

```
from sklearn.metrics import accuracy_score
Y_train_prediction_rf = classifier.predict(X_train)
training_data_accuracy_rf = accuracy_score(Y_train_prediction_rf, Y_train)
print("Accuracy on training data (Random Forest):", training_data_accuracy_rf)
Y_test_prediction_rf = classifier.predict(X_test)
test_data_accuracy_rf = accuracy_score(Y_test_prediction_rf, Y_test)
print("Accuracy on testing data (Random Forest):", test_data_accuracy_rf)
```

```
Accuracy on training data (Random Forest): 0.9979338842975206
Accuracy on testing data (Random Forest): 0.9590163934426229
```

```
input_data = (44,0,2,118,242,0,1,149,0,0.3,1,1,2)
#change input data to a numpy array
input_data_as_numpy_array=np.asarray(input_data)
#reshape the numpy array as we are predicting for only one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)
prediction=model.predict(input_data_reshaped)
print(prediction)
if (prediction[0]==0):
  print("The person does not have Heart Disease ")
else:
  print("The person has Heart Disease")
```

```
[1]
The person has Heart Disease
```

VI. MERITS OF PROPOSED SYSTEM

The heart disease prediction system proposed has numerous advantages that make it effective and practical. To begin with, the use of logistic regression, which is known for its interpretability, enhances the transparency of the model's decision-making process. This provides healthcare practitioners and patients with a clear understanding of the factors that influence predictions. The incorporation of comprehensive data preprocessing techniques, such as handling missing values and addressing outliers, ensures the dataset's integrity, which contributes to the model's robustness. The feature selection process further refines the model by identifying and prioritizing the most influential predictors, streamlining the predictive algorithm.

The comparative analysis with alternative models broadens the perspective on algorithmic performance, enabling informed decisions about the most suitable approach for heart disease prediction. Furthermore, the proposed system's adaptability to real-time health monitoring data and collaboration potential with medical professionals positions it as a dynamic and evolving tool in the realm of predictive healthcare. This lays the groundwork for proactive heart disease management.

VII. CONCLUSION

In conclusion, this research project successfully explored the application of machine learning, particularly logistic regression, for predicting heart disease based on various health parameters. The analysis involved comprehensive data preprocessing, exploratory data analysis, and model training. The logistic regression model demonstrated commendable accuracy on both the training and testing datasets, highlighting its potential as a reliable tool for identifying individuals at risk of heart disease. The findings contribute to the ongoing efforts to leverage data-driven approaches in healthcare, emphasizing the importance of predictive models in early disease detection. Future work could involve further refinement of the model and exploration of additional machine learning techniques to enhance predictive accuracy. Overall, this research sheds light on the intersection of data science and healthcare, showcasing the potential for technology to assist in proactive health management.

VIII. ACKNOWLEDGEMENT

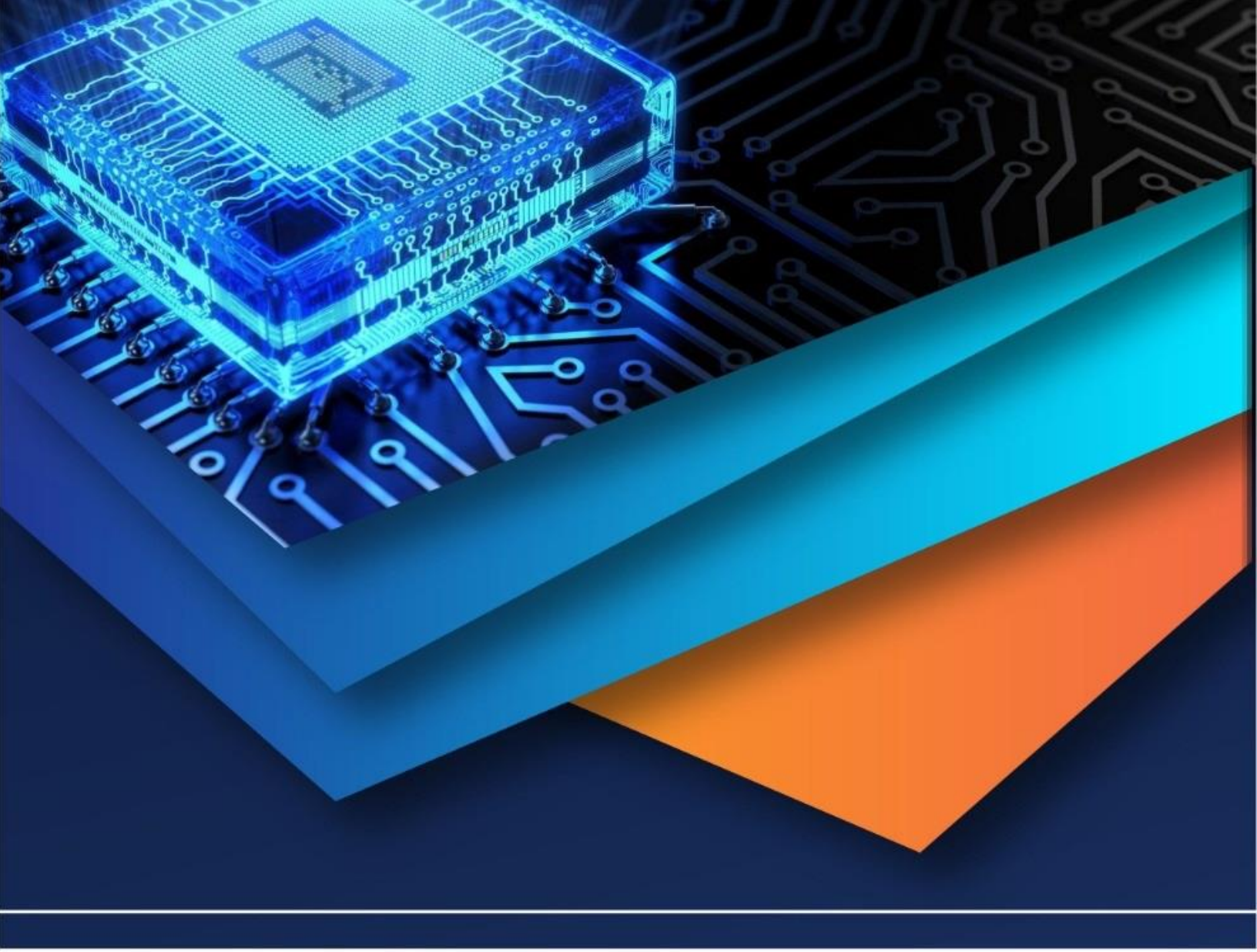
I would like to express my sincere gratitude and appreciation to my project guide and Head of Department for their invaluable support and guidance throughout the development of this project. Their unwavering encouragement and valuable insights have been instrumental in shaping my ideas and ensuring the successful completion of this project.

My guide's expertise and dedication have been pivotal in guiding me through the various stages of this project. They have provided me with the necessary resources and feedback to refine my work and bring out the best in me. Likewise, the Head of Department's constant encouragement and support have been crucial in motivating me to stay focused and dedicated to my work. Without their support, this project would not have been possible, and I am deeply grateful to them for their guidance and mentorship.

REFERENCES

[1] IOP Conference Series: Materials Science and Engineering, Volume 1022, 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India Citation Harshit Jindal et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072 IOP Conference Series

[2] **Citation** Harshit Jindal *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* 1022 012072**DOI** 10.1088/1757-899X/1022/1/012072



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY