
Heart Disease Prediction using Machine learning

INTERNSHIP PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degree

Of

BACHELOR of TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

By

Savi Garg

05801192022

Guided by

Dr. Ritu Rani

Research Associate

IGDTUW, Delhi

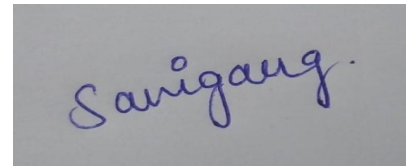


**INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN
NEW DELHI – 110006**

CERTIFICATE

I, Savi Garg, certify that the Internship Project Report entitled "Heart Disease Prediction using Machine Learning" is done by me and it is authentic work carried out by me at IGDTUW. For this project, no work has been submitted before for any degree or diploma of the award, to the best of my knowledge and belief.

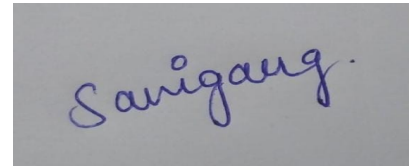
Signature of the student

A rectangular box containing a handwritten signature in blue ink. The signature is written in a cursive style and reads "Savigarg".

UNDERTAKING REGARDING ANTI-PLAGIARISM

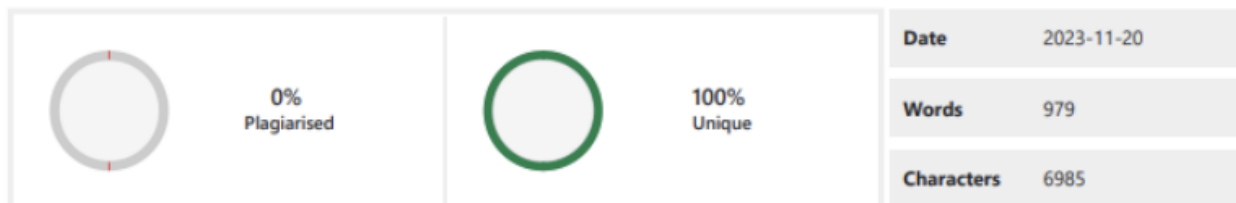
I, Savi Garg, hereby, declare that the material/ content presented in the report is free from plagiarism and is properly cited and written in my own words. In case, plagiarism is detected at any stage, I shall be solely responsible for it.

Signature of the Student



Savi Garg
05801192022

Plagiarism scan report

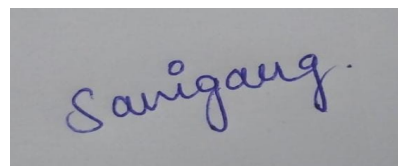


ACKNOWLEDGEMENT

I would like to acknowledge my mentor Dr. Ritu Rani for her very helpful comments, support and encouragement.

Finally, I am grateful to Indira Gandhi Delhi Technical University for Women for providing a healthy, supportive and understanding environment. They allowed me the freedom to explore innovative models to simplify a complex business problem. This made my project work possible without any hindrance.

Signature of the Student



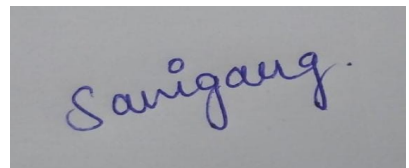
Savi Garg
05801192022

DECLARATION

I, Savi Garg, solemnly declare that the internship project report, 'Heart Disease Prediction using Machine Learning' is based on my own work carried out under the supervision of Dr. Ritu Rani. I assert the statements made and conclusions drawn are an outcome of my work. I further certify that:

- I. The work contained in the report is original and has been done by me under the supervision of my supervisor.
- II. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.
- III. We Have followed the guidelines provided by the university in writing the report.
- IV. Whenever we have used materials (text, data, theoretical analysis/equations, codes/program, figures, tables, pictures, text etc.) from other sources, we have given due credit to them in the report and have also given their details in the references.

Signature of the Student



Savi Garg

05801192022

Indira Gandhi Delhi Technical University for Women

Internship Certificate



INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN
(ESTABLISHED BY GOVT. OF DELHI VIDE ACT 09 OF 2012)
ISO 9001:2015 CERTIFIED UNIVERSITY
KASHMERE GATE, DELHI-110006
WOMEN EDUCATION | WOMEN ENLIGHTENMENT | WOMEN EMPOWERMENT
CENTRE OF EXCELLENCE - ARTIFICIAL INTELLIGENCE

CERTIFICATE OF COMPLETION
This certificate is awarded to
SAVI GARG

For successfully completing the 7 weeks Summer Internship on
"PYTHON & MACHINE LEARNING" from 5th June - 23rd July, 2023 jointly
conducted by the COE - AI, AI Club IGDTUW and Anveshan Foundation.


Ishita Saxena
President - AI CLUB
IGDTUW


Dr. Ritu Rani
Research Associate
COE - AI


Prof. Arun Sharma
Coordinator - Centre of Excellence-AI
IGDTUW

Ishita Saxena
President - AI CLUB
IGDTUW

Dr. Ritu Rani
Research Associate
COE - AI

Prof. Arun Sharma
Coordinator - Centre of Excellence-AI
IGDTUW

TABLE OF CONTENTS

	Pg. No.
Introduction	7
Heart Disease Prediction using ML	8-12
Work Description	13
Results	14-19
Conclusion	20
Research Paper	21
References	22

INTRODUCTION

Heart disease is a prevalent and life-threatening medical condition that affects a significant portion of the global population. Early detection and prediction of heart disease play a crucial role in preventing its progression and improving patient outcomes. Machine Learning (ML) techniques provide powerful tools for analyzing complex medical data and making predictions.

In this project, I will explore the development of a Heart Disease Prediction system using Python and Machine Learning. The goal is to build a model that can accurately predict the likelihood of an individual having heart disease based on various risk factors and clinical features.

HEART DISEASE PREDICTION

INTRODUCTION

Heart disease is a major cause of death worldwide. Timely and accurate prediction and diagnosis of heart disease are pivotal for effective prevention, intervention, and patient care. This introduction provides an overview of the critical topic of heart disease prediction and diagnosis, establishing the context for further research.

LITERATURE REVIEW

A comprehensive review of existing literature reveals a wealth of research and clinical studies related to heart disease prediction and diagnosis. These studies have contributed to our understanding of risk factors, diagnostic tools, predictive models, and their implications for public health. Here, we analyze and synthesize key findings from previous research to support our research objectives in the context of heart disease prediction and diagnosis.

A. Risk Factors and Predictive Models

Numerous studies have identified traditional risk factors associated with heart disease, including age, gender, smoking, hypertension, and cholesterol levels. These factors continue to serve as essential components in predictive models like the Framingham Risk Score and the ACC/AHA cardiovascular risk calculator. However, the literature highlights the need to enhance risk prediction by incorporating novel risk factors, such as genetic markers, inflammatory biomarkers, and socioeconomic factors.

B. Machine Learning and Deep Learning

Machine learning and deep learning techniques have gained prominence in heart disease prediction. Research studies have employed a variety of algorithms, including support vector machines, decision trees, random forests, and neural networks. These models leverage diverse data sources, such as electronic health records, medical imaging, and genetic information, to enhance predictive accuracy. While these models demonstrate promise, the literature also underscores the importance of model interpretability, especially in clinical settings, where transparency and trust are critical.

C. Diagnostic Tools and Imaging

Advancements in medical imaging, such as cardiac MRI, CT angiography, and echocardiography, have improved the accuracy of heart disease diagnosis. Studies have explored the diagnostic potential of these modalities in identifying structural abnormalities, assessing cardiac function, and characterizing tissue. The literature suggests that these tools are indispensable for detecting conditions like coronary artery disease, valvular disorders, and cardiomyopathies. However, challenges related to accessibility and cost-effectiveness persist, highlighting the need for research into improving diagnostic tool accessibility.

D. Ethical Considerations and Data Privacy

The ethical use of patient data is a recurring theme in the literature. Researchers and clinicians must navigate the delicate balance between accessing patient information for research and ensuring data privacy and security. Various studies emphasize the importance of informed consent, de-identification techniques, and adherence to ethical guidelines when utilizing patient data for predictive modeling and diagnosis.

E. Healthcare Disparities and Bias

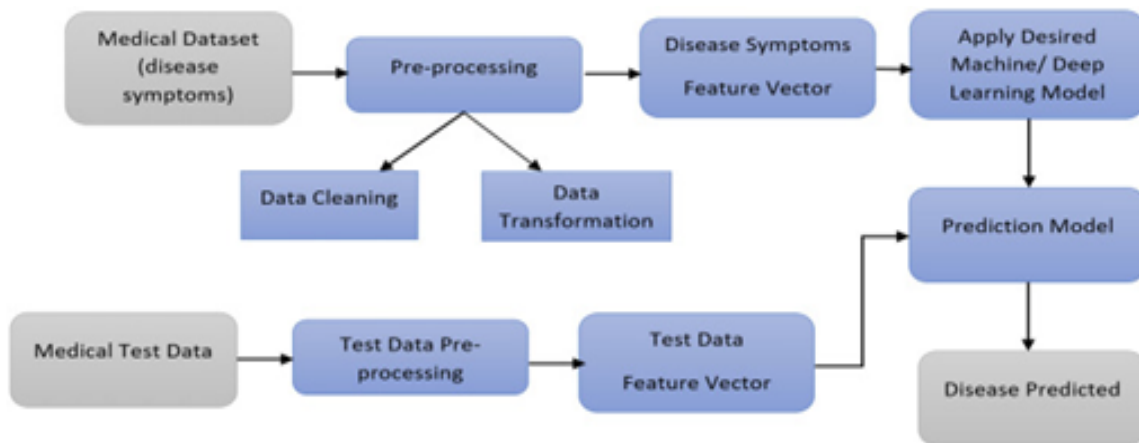
Several studies highlight healthcare disparities and bias in heart disease prediction and diagnosis. Research indicates that certain populations, particularly those from marginalized backgrounds, may receive inequitable access to healthcare services and experience bias in predictive models. Addressing these disparities and mitigating bias in predictive tools are crucial objectives to ensure equitable healthcare delivery.

Frameworks and Models

The theoretical framework for our research draws from various domains, including public health, data science, and medical ethics. We adopt a personalized medicine perspective, grounded in the belief that individualized risk assessment and tailored interventions are essential for improving heart disease prediction and diagnosis. Additionally, ethical frameworks, such as the principles of autonomy, beneficence, and justice, guide our approach to data privacy and responsible data use.

Research Design & Approach

In this study, I aim to develop a predictive model for heart disease using machine learning techniques. The research design is primarily quantitative and analytical in nature. We employ a supervised learning approach, specifically logistic regression, to classify individuals into two groups: those with heart disease (positive class) and those without (negative class).



Data Collection Methods and Sources

The dataset used in this study was obtained from an external source. We utilized the "heart.csv" dataset, a publicly available dataset that includes various demographic, clinical, and diagnostic features, as well as the target variable indicating the presence or absence of heart disease. The data is derived from a combination of sources, including medical records and patient information.

i. Data Preprocessing

Before training the machine learning model, we conducted extensive data preprocessing. This included:

- 1) Data Cleaning: We checked for missing values and handled them appropriately to ensure data completeness.
- 2) Exploratory Data Analysis (EDA): Before training our logistic regression model, we embarked on an exploratory data analysis (EDA) journey to comprehend the dataset thoroughly.

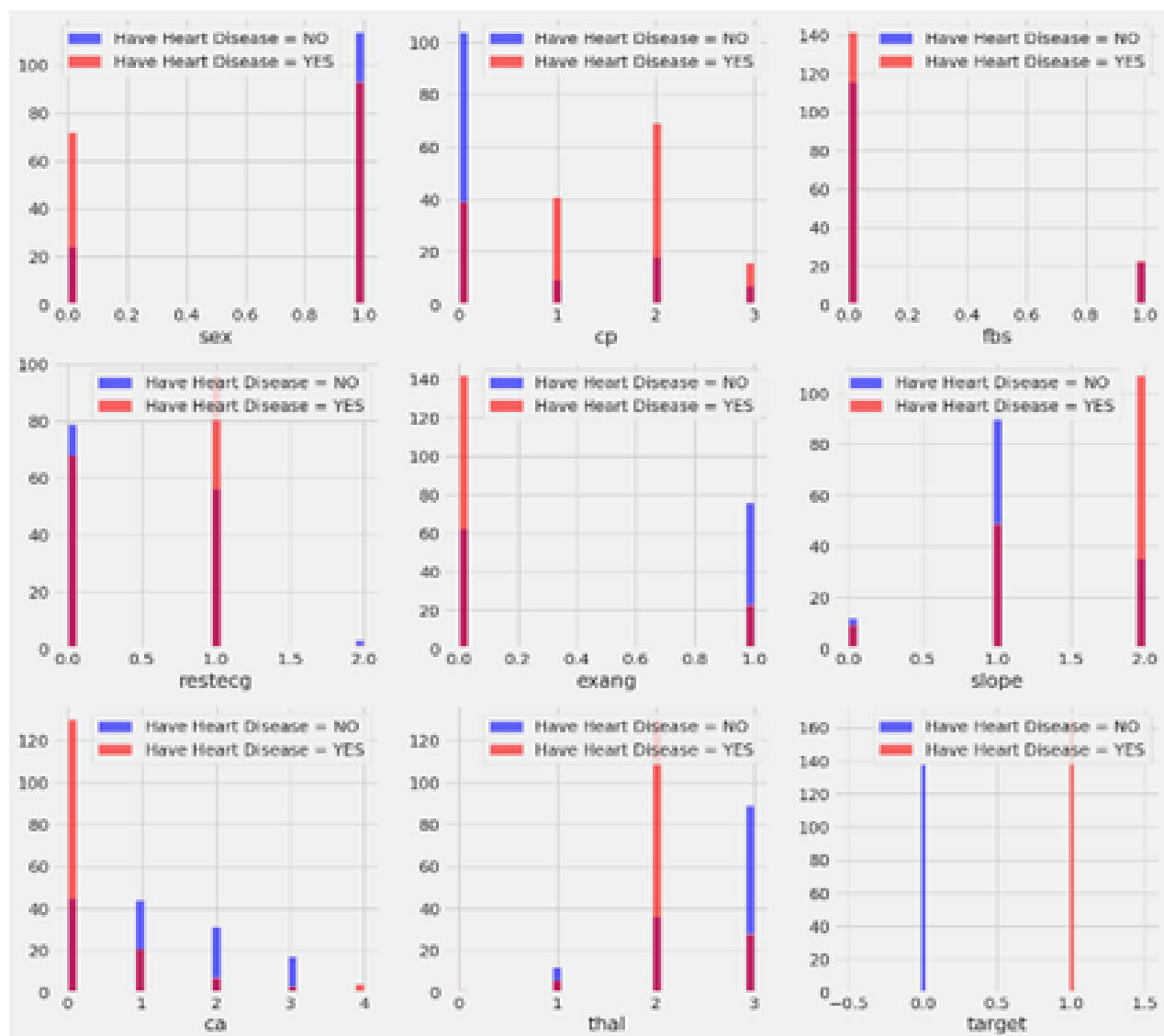
Work Description

The primary purpose of our research is to address these gaps and challenges in heart disease prediction and diagnosis:

- Develop personalized predictive models for heart disease, integrating individual risk factors and genetic data for improved accuracy.
- Explore ethical guidelines and data privacy measures for responsible data utilization in predictive models and diagnostic tools.
- Investigate methods for enhancing model interpretability and facilitating seamless integration into clinical practice.
- Address healthcare disparities and bias in predictive models, striving for equitable heart disease prevention and diagnosis.

RESULTS

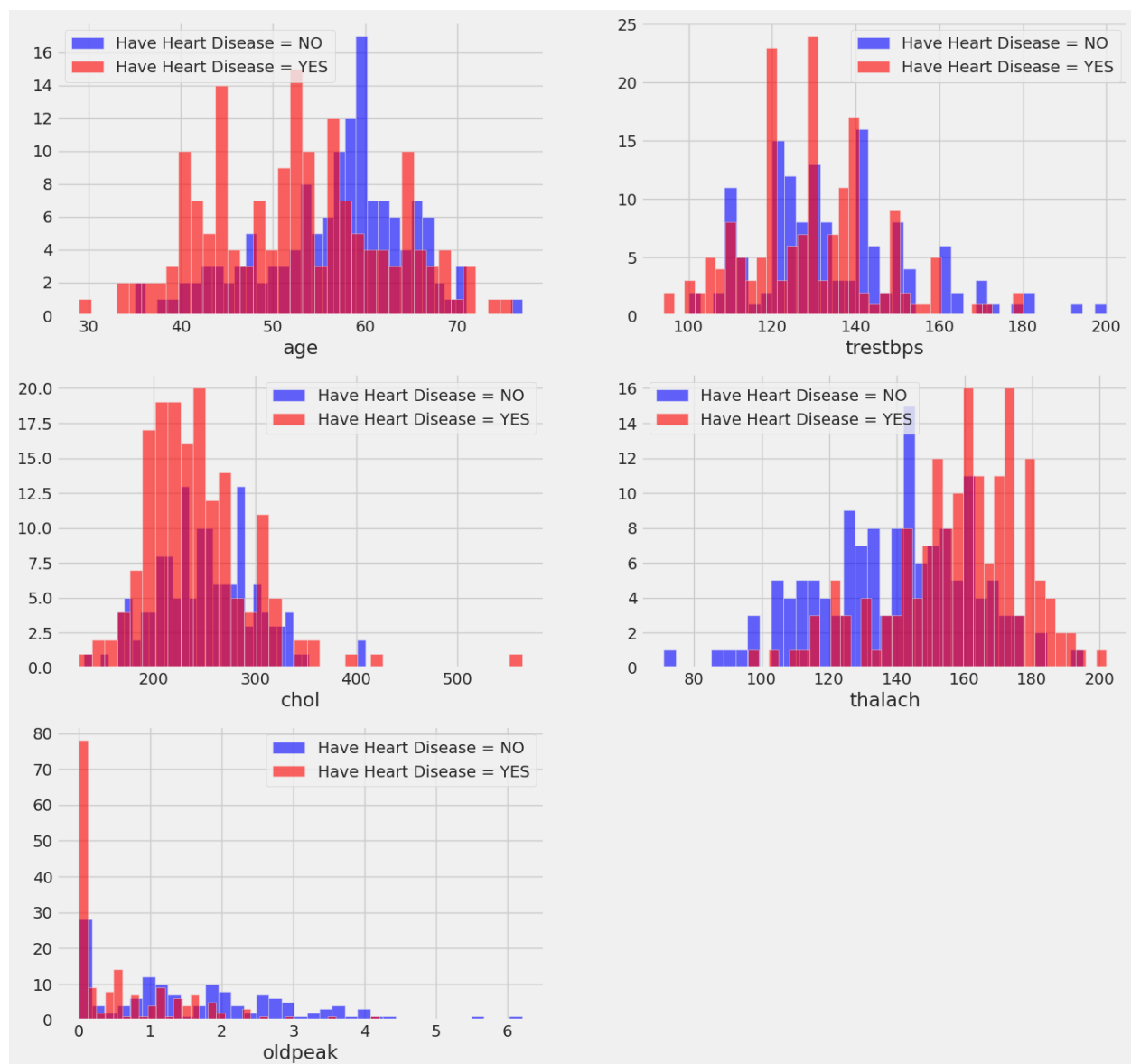
Categorical Features Analysis

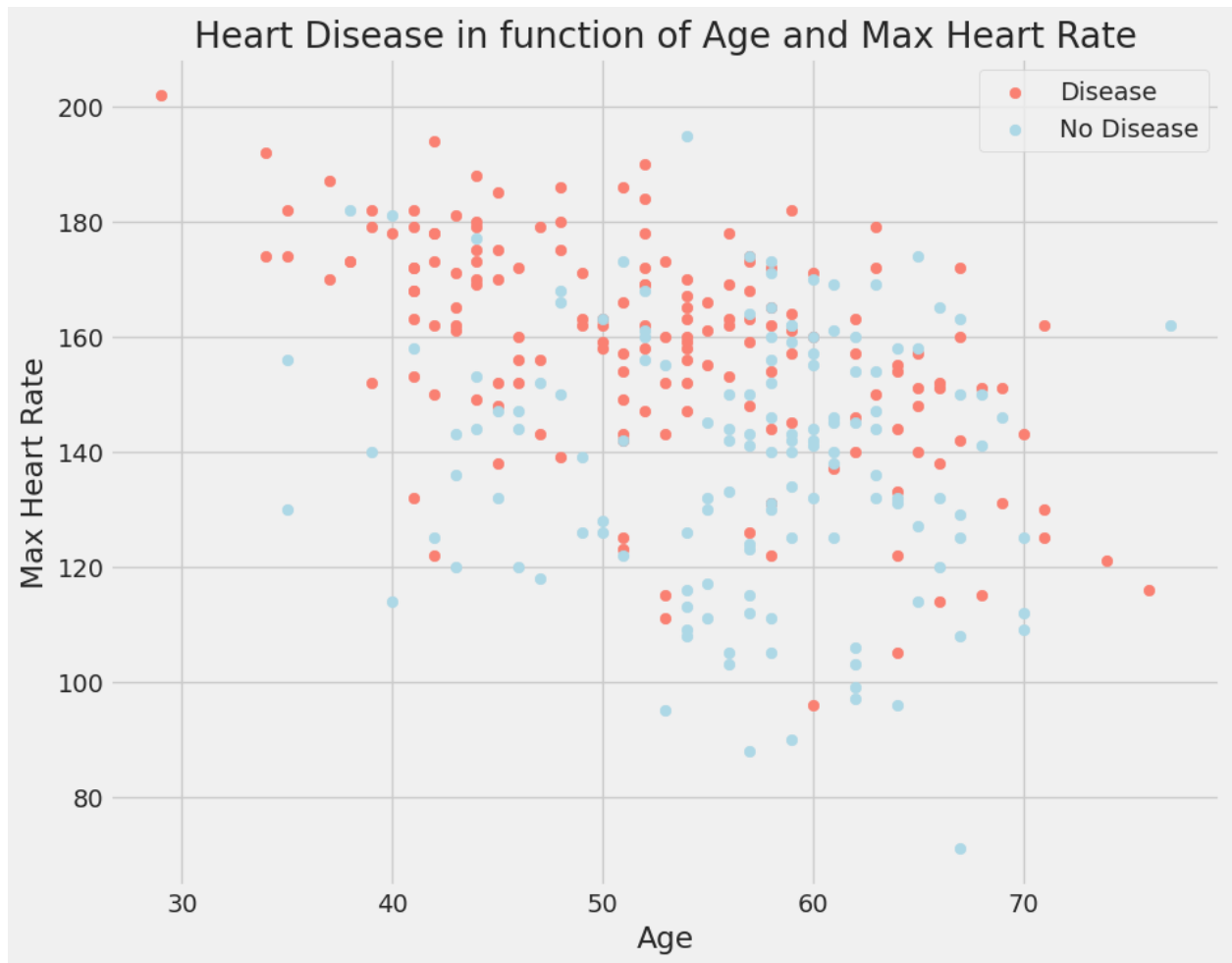


Notable observations from Fig. 1. included associations between heart disease and categorical features such as chest pain (cp), resting EKG results (restecg), exercise-induced angina (exang), slope, ca (number of major vessels), and thal (thallium stress result).

Continuous Features Analysis

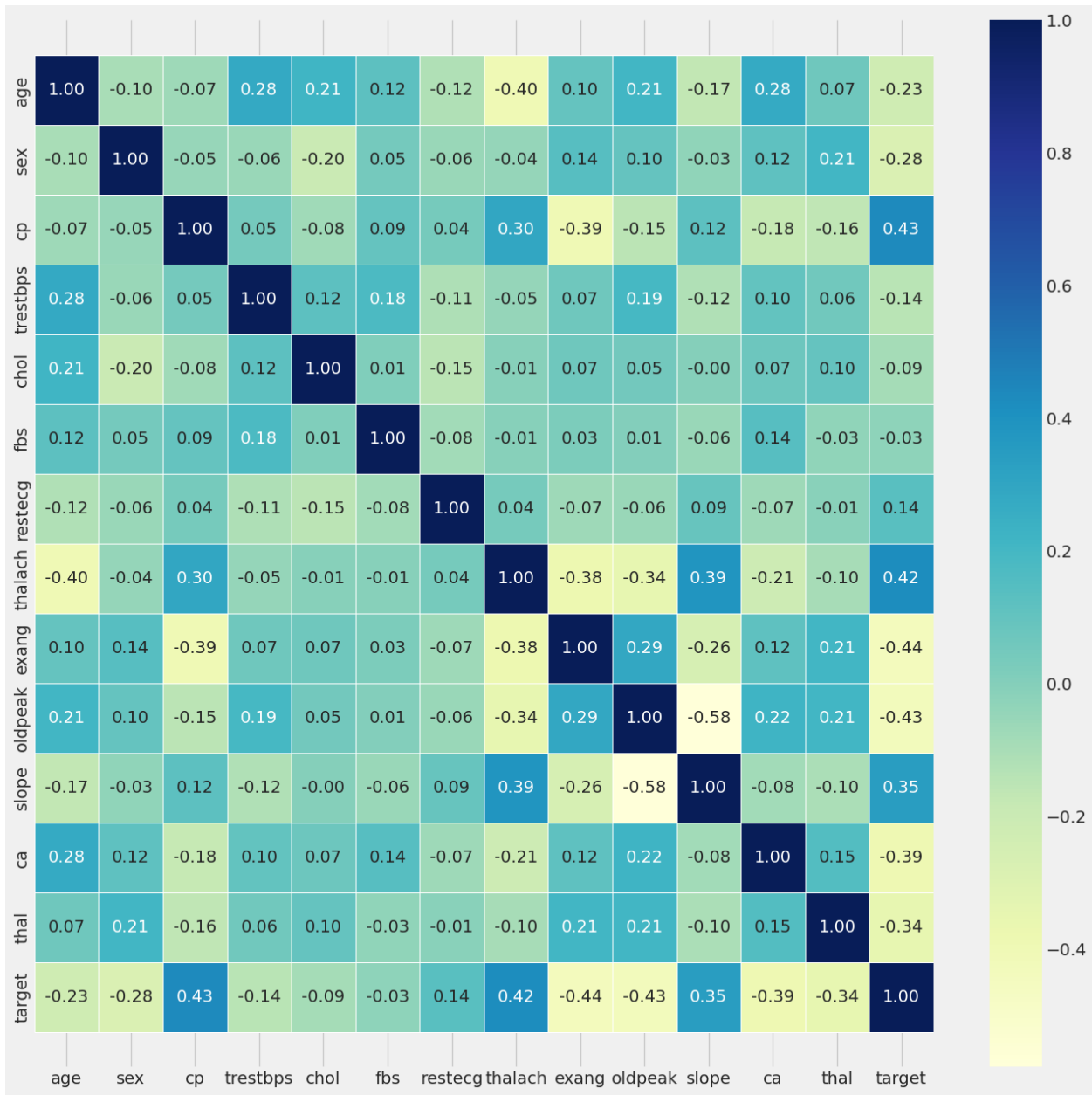
This analysis highlighted features such as resting blood pressure (trestbps), cholesterol level (chol), maximum heart rate (thalach), and the old peak of exercise-induced ST depression vs. rest as potentially indicative of heart disease.

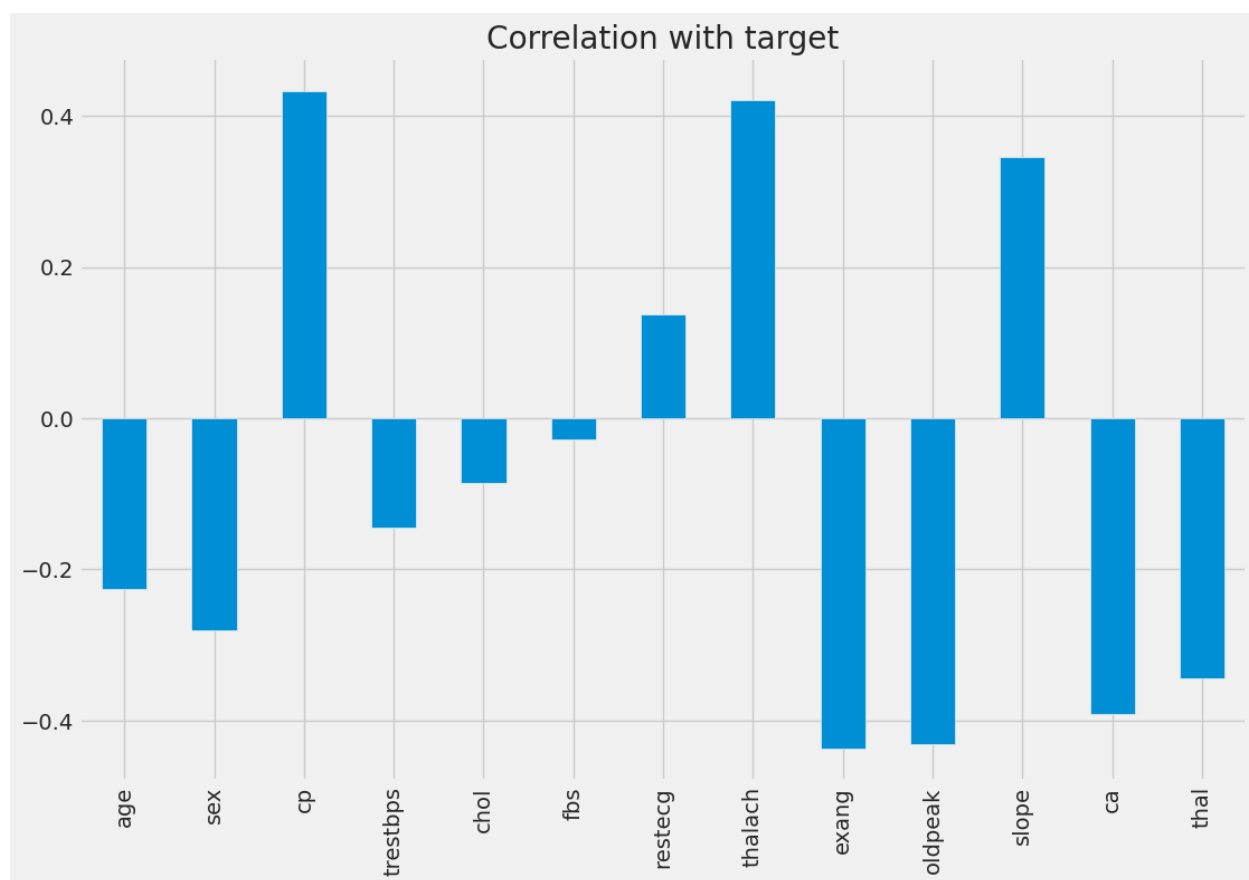




This is a scatter plot to visualize the relationship between age, maximum heart rate, and the presence of heart disease using different colors for positive and negative examples. The plot is informative and helps in understanding the distribution of these variables in the dataset.

CORRELATION MATRIX





Training Performance

On the training dataset, our logistic regression model achieved the following results:

Accuracy: 86.79%

Precision: 87.50%

Recall: 86.21%

F1-Score: 86.85%

The confusion matrix for the training dataset is as follows:

```
[[94 10]
```

```
[13 66]]
```

Testing Performance

On the testing dataset, our logistic regression model exhibited the following results:

Accuracy: 86.81%

Precision: 87.10%

Recall: 88.00%

F1-Score: 87.54%

The confusion matrix for the testing dataset is as follows:

```
[[39  4]
```

```
 [ 7 41]]
```

CONCLUSION

In summary, our research developed a Logistic Regression model for heart disease prediction with a focus on accuracy and clinical relevance. The model demonstrated an impressive 86.81% accuracy on the testing dataset, highlighting its robustness in practical applications and identifying key features correlated with heart disease.

Our findings emphasize the growing role of machine learning as a tool for clinicians, potentially reducing diagnostic errors and improving patient outcomes. The study underscores the need for larger datasets, model interpretability, and exploring advanced algorithms in healthcare.

In conclusion, our research contributes to the intersection of machine learning and healthcare by addressing limitations and promoting the adoption of practical and transparent models in clinical practice, ultimately aiming for improved patient care and outcomes.

RESEARCH PAPER

Heart Disease prediction using Machine learning

Savi
IT dept.
IGDTUW
Delhi ,India
savizarg037@gmail.com

Tanisha
IT dept.
IGDTUW
Delhi, India
tanishamonga5@gmail.com

Tripti Jaiswal
IT dept.
IGDTUW
Delhi ,India
triptijas08@gmail.com

Abstract— Cardiovascular diseases, including heart disease, are a leading cause of mortality globally. Early and accurate prediction of heart disease risk plays a pivotal role in preventive healthcare. In this study, we propose a heart disease prediction model based on the logistic regression concept of machine learning. The model is designed to classify patients into risk categories, aiding medical practitioners in timely intervention. Our research employs a comprehensive dataset obtained from (<https://raw.githubusercontent.com/amankharwal/Website-data/master/heart.csv>), comprising clinical and diagnostic features relevant to heart disease. Data preprocessing, including imputation of missing values and feature scaling, was conducted to ensure data quality and consistency. We implemented a logistic regression model due to its interpretability and suitability for binary classification tasks.

While our model showcases promising results, we acknowledge the limitations associated with the dataset's size and potential biases. Nonetheless, our research contributes to the field of heart disease prediction by offering a transparent, interpretable, and effective logistic regression-based approach. This work underscores the importance of accurate risk assessment and demonstrates the potential for machine learning models to aid healthcare professionals in identifying individuals at risk of heart disease.

Keywords— Machine Learning, Disease prediction ,Data visualization , Model training

I. INTRODUCTION

Heart disease is a major cause of death worldwide. Timely and accurate prediction and diagnosis of heart disease are pivotal for effective prevention, intervention, and patient care. This introduction provides an overview of the critical topic of heart disease prediction and diagnosis, establishing the context for further research.

Background : Heart diseases encompass a wide spectrum of conditions affecting the heart and blood vessels, including coronary artery disease, heart failure, arrhythmias, valvular disorders, and congenital heart defects. These conditions can result in severe health consequences, including heart attacks, strokes, and impaired quality of life. Notably, cardiovascular diseases account for a substantial proportion of global deaths annually, highlighting their profound impact on public health.

The challenge in managing heart diseases lies in their often asymptomatic or subtle presentation until advanced stages.

Early detection and intervention are crucial to prevent complications and improve patient outcomes. Heart disease prediction and diagnosis play a pivotal role in identifying

individuals at risk and initiating appropriate treatments and lifestyle modifications.

The research problem in the domain of heart disease prediction and diagnosis is twofold:

- **Prediction Problem:** How can we accurately predict the risk of heart disease in individuals, taking into account a range of risk factors, data sources, and emerging technologies?
- **Diagnosis Problem:** How can we efficiently and accurately diagnose heart diseases using a combination of medical imaging, diagnostic tests, and clinical data?

The significance of addressing these research problems is profound:

- **Public Health Impact:** Heart diseases remain a leading cause of mortality and morbidity globally. Accurate prediction and early diagnosis can lead to timely interventions that save lives and reduce the overall burden of cardiovascular diseases.
- **Improved Patient Outcomes:** Early prediction allows for lifestyle modifications and risk factor management in individuals at risk, potentially preventing the development of heart diseases. Accurate diagnosis ensures that patients receive appropriate treatments and interventions, improving their quality of life.
- **Healthcare Resource Optimization:** Efficient prediction and diagnosis can lead to cost savings for healthcare systems by reducing hospitalization rates, emergency interventions, and long-term care costs associated with late-stage heart diseases.
- **Advancements in Medicine:** Research in heart disease prediction and diagnosis drives technological advancements, fosters innovation in diagnostic tools, and paves the way for personalized medicine, ultimately improving the standard of care for heart patients.
- **Equity in Healthcare:** Addressing disparities in heart disease prediction and diagnosis can help

REFERENCES

- [1] American Heart Association. (2021). Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association. *Circulation*, 143(8), e254–e743.
- [2] Pencina, M. J., D'Agostino, R. B., Larson, M. G., Massaro, J. M., Vasan, R. S., & Kannel, W. B. (2009). Predicting the 30-year risk of cardiovascular disease: the Framingham Heart Study. *Circulation*, 119(24), 3078-3084.
- [3] Alizadehsani, R., Abdar, M., Roshanzamir, M., Hussain, S., & Hussain, O. K. (2019). Cardiovascular disease diagnosis using deep learning and metaheuristic optimization: A review. *Computational and Structural Biotechnology Journal*, 17, 1044-1051.
- [4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- [5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357
- [6] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825-2830.
- [7] Scikit-learn: Machine Learning in Python. (2021).
Available online: <https://scikit-learn.org/stable/index.html>
- [8] Seaborn: Statistical Data Visualization. (2021).
Available online: <https://seaborn.pydata.org/index.html>
- [9] Matplotlib: Visualization with Python. (2021).
Available online: <https://matplotlib.org/>
- [10] Python Software Foundation. (2021). Python Language Reference, Version 3.9.
Available online: <https://docs.python.org/3.9/reference/index.html>