HEART DISEASE PREDICTOR

A Project Based Learning Report Submitted in partial fulfillment of the requirements for the award of the degree

of

Bachelor of Technology

in The Department of CSE

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING – 23AD2001O

Submitted by
2310030142: NAKSHATRA NAVATU
2310030234: YUKTHA SIRI
2310030357: KOSIKA HASINI
2310030258: SRIVALLI

Under the guidance of

Mr . Sumit Hazra

Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

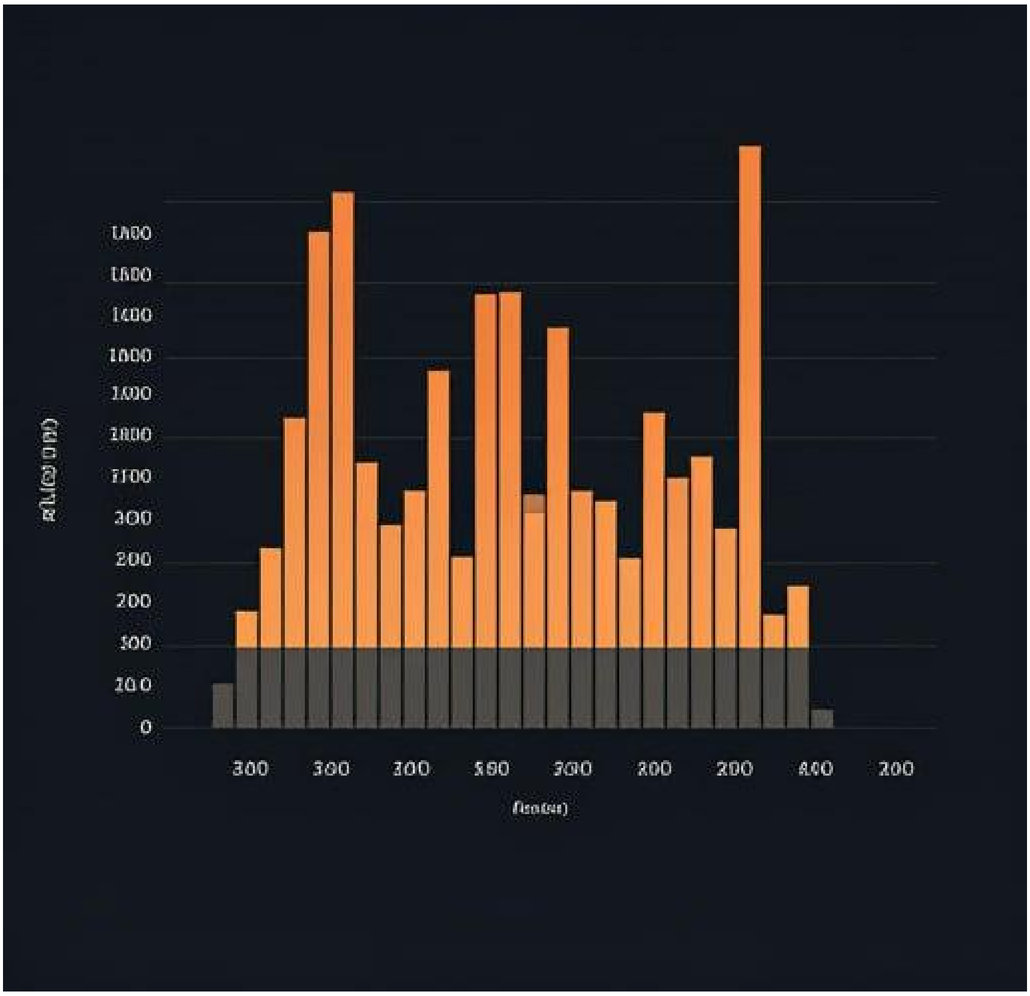
NOV - 2024.

Abstract

Heart disease is a significant global health concern, contributing to a high rate of morbidity and mortality. Early detection and intervention are vital in reducing the impact of this disease. This project focuses on developing a predictive model for heart disease using machine learning techniques, aiming to provide healthcare professionals with a valuable tool for early diagnosis. The study utilizes a comprehensive dataset that encompasses various health parameters, such as age, gender, blood pressure, cholesterol levels, and other clinical indicators. We conducted thorough data preprocessing to address missing values and normalize the features, ensuring the accuracy and reliability of the analysis. Several machine learning algorithms were employed, including Logistic Regression, Decision Trees, and Random Forest, to assess their effectiveness in predicting heart disease. The performance of each model was rigorously evaluated using metrics such as accuracy, precision, recall, and F1-score. Our findings indicate that the Random Forest model achieved the highest accuracy, demonstrating its potential for effective prediction. The results underscore the importance of leveraging machine learning in the medical field, as it can significantly enhance the ability to identify individuals at risk of heart disease. In conclusion, this project highlights the transformative role of data-driven approaches in healthcare, providing insights that can lead to improved patient outcomes. Future work may involve integrating additional features and refining the model further to enhance its predictive capabilities.

List of Figures

1.DATA VISUALISATION



The above image contains

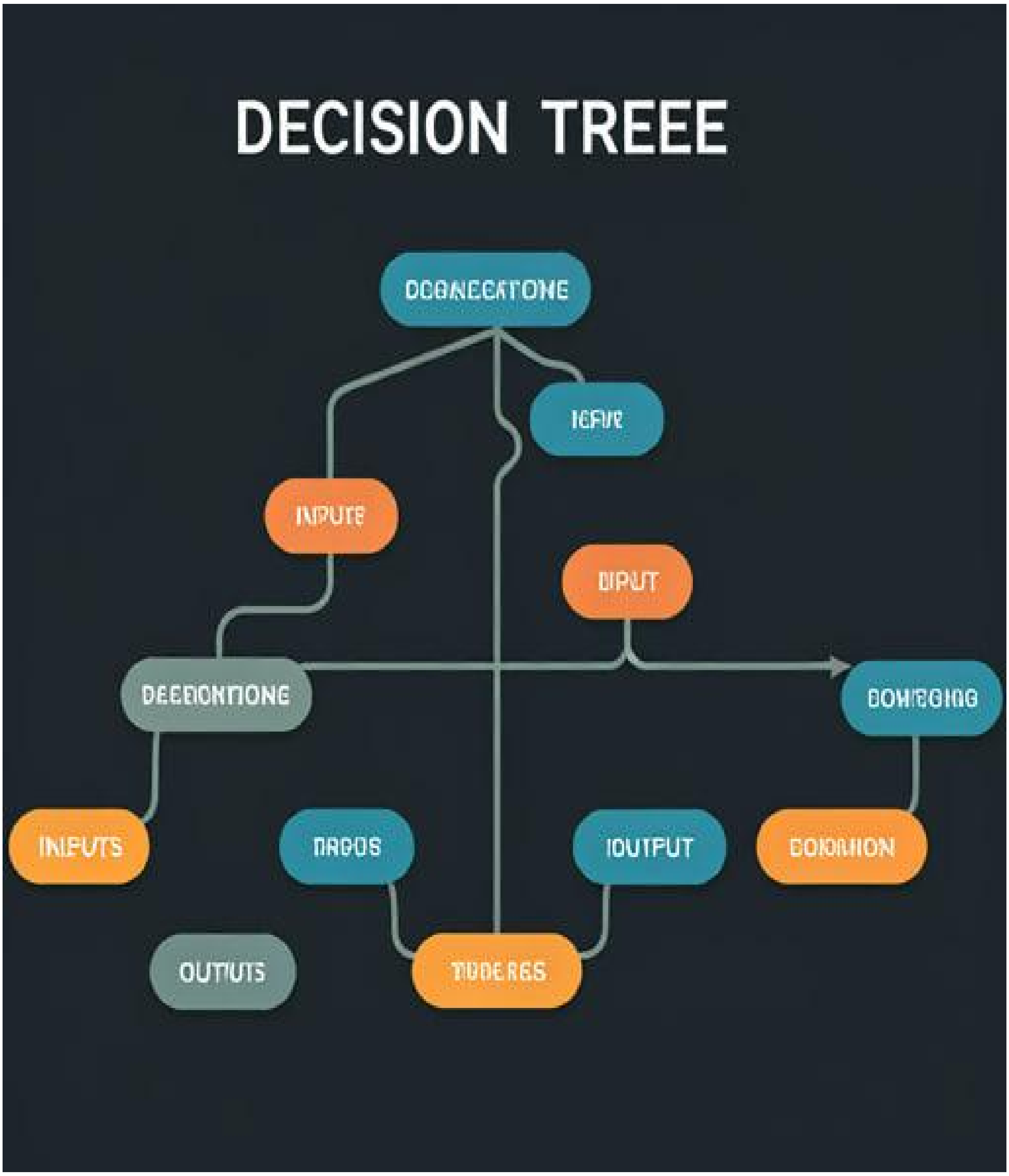
Histogram: Visualizing the distribution of numerical features like age, cholesterol, and blood pressure.

Bar Chart: Comparing categorical features like gender, chest pain type, and exercise-induced angina.

Correlation Matrix: Showing the relationships between different features.

Box Plot: Identifying outliers and compare distributions across different categories.

2. **DECISION TREE**



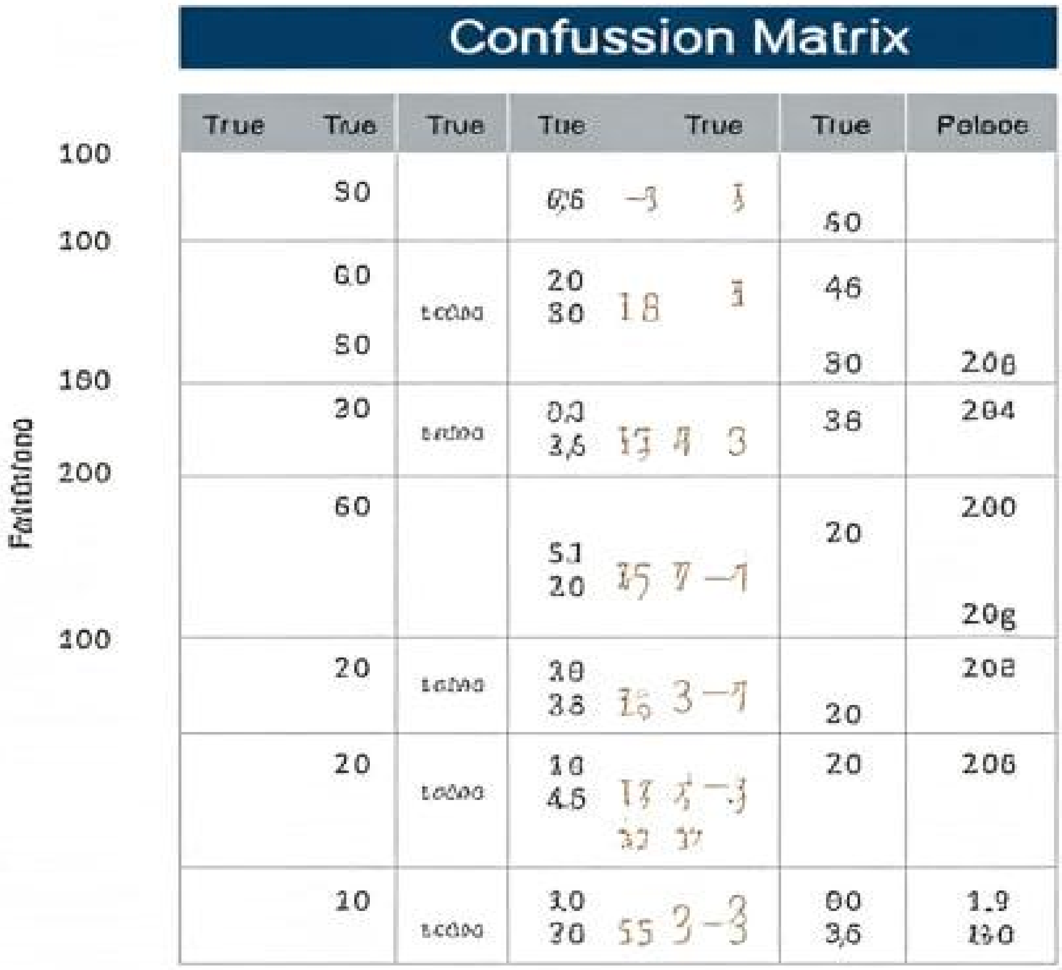
The above image contains :

Decision Tree: Visualizing the decision-making process of a decision tree model.

Random Forest: Depicting the ensemble of decision trees in a random forest.

Neural Network: Showing the architecture of a neural network, including layers and connections.

3. CONFUSION MATRIX



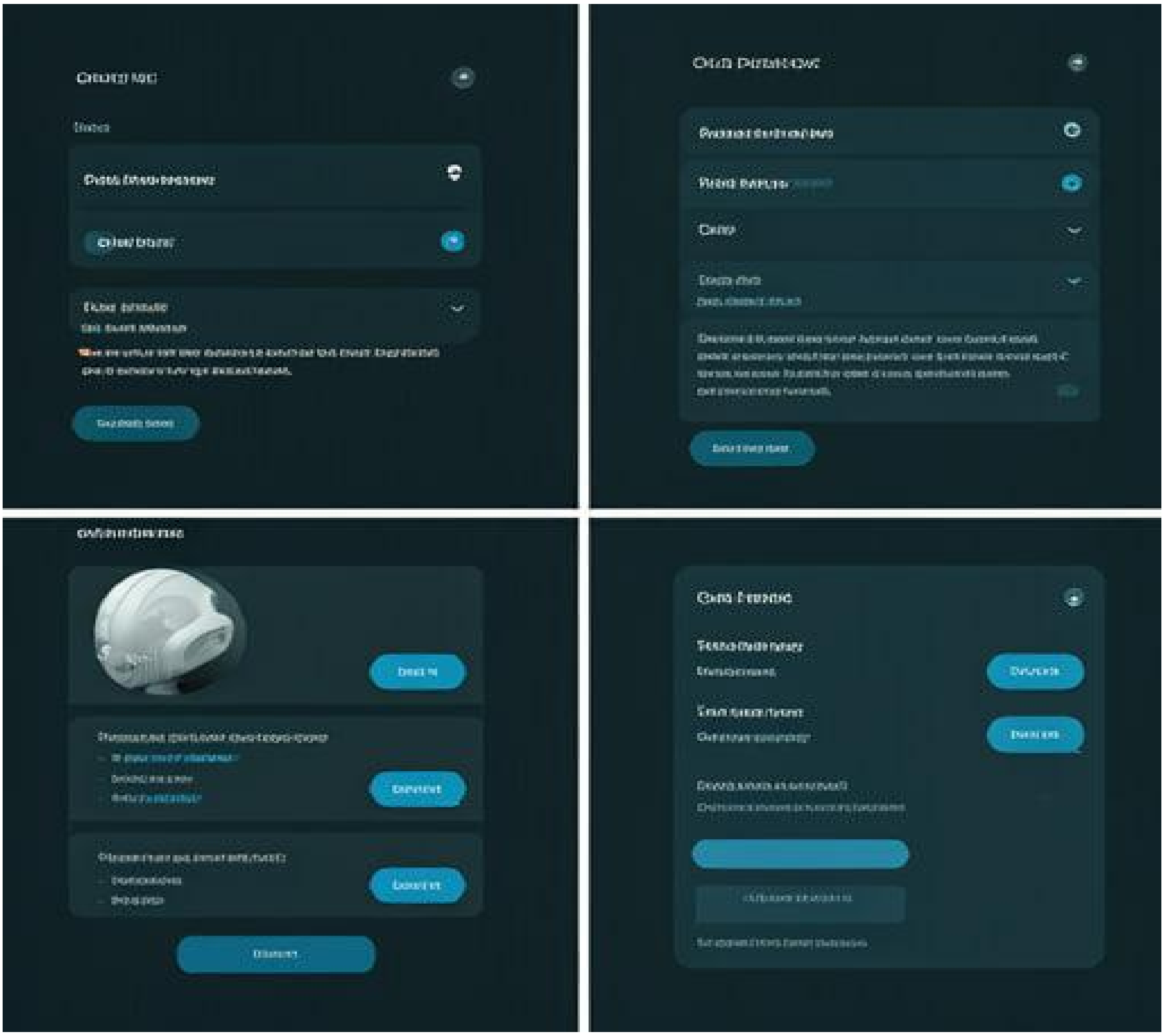
The above matric contains :

Confusion Matrix: Display the model's performance in terms of true positives, true negatives, false positives, and false negatives.

ROC Curve: Plot the receiver operating characteristic curve to assess the model's ability to distinguish between positive and negative cases.

Precision-Recall Curve: Visualize the trade-off between precision and recall.

4. CHATBOT INTERFACE



The above image contains :

Screenshots: Shows screenshots of your chatbot's user interface, including the initial greeting, question-answer interactions, and final prediction.

Flowchart: Illustrates the flow of conversation and decision-making within the chatbot.

LIST OF TABLES

Heart Disase Prelredictume

	Geprrete	Lst Lh	Dsw Fδ	Lss 1b	Can 15
Pret date		Testings	Testupe	Tesbings	Tectings
1	Neart cilbsceast dating pluraitv)	1014	154	5%	240
2	Heart eillies preogim)	100	164	1%	165
4	Cliart cill diseser plication	20.0	854	3%	140
5	Neart ellos contemting	230	266	9%	100
4	Heart cilbasert peeperiestion)	1212	134	3%	105
5	Ceart cilistage distec pregion)	1647	656	1%	100
446	[illegible]				
7	Scort clisteascerest anasgon)	10.1	235	7%	250
16	Ceart cill best sepecs potnegtn)	23.18	264	7%	240
17	Scort cill besting ofication)	26.1	202	9%	260
109	Scort cill bescing preestion)	125.2	296	7%	250
101	Ccart cill bserung pracion)	16.15	223	9%	250
105	Ccart cilf beacy disation)	16.11	274	4%	760
203	Proctelis cammiun)	15.10	207	4%	260
109	Coort cill beagy etlestion)	16.13	267	4%	202
244	Prort cilibeszreadiion)	23.44	227	4%	205
221	Coort cill beagc eregoranfon)	20.17	202	9%	200
222	Ccart clieas proditicts	12.19	296	9%	225
203	Afirt wiltion	22.15	299	4%	229

Table of Contents

HEART DISEASE PREDICTOR

Introduction

Background :

Heart disease remains one of the leading causes of mortality worldwide. With the increasing prevalence of cardiovascular diseases, there is a pressing need for effective and efficient methods to predict the risk of heart disease in individuals. Advances in data science and machine learning have opened new avenues for developing predictive models that can assist healthcare professionals in early diagnosis and intervention, potentially saving lives and reducing healthcare costs.

Advances in data science and machine learning have opened new avenues for developing predictive models that can assist healthcare professionals in assessing the risk of heart disease in individuals. These models utilize various health parameters and lifestyle factors to generate risk assessments, enabling timely medical interventions

Objective of the Project:

The primary objective of this project is to develop a predictive model that can accurately assess the risk of heart disease in individuals based on various health parameters and lifestyle factors. By leveraging machine learning algorithms, the project aims to create a tool that can provide reliable predictions, thereby aiding in early detection and preventive measures.

Scope of the Report:

This report encompasses the entire process of developing the heart disease predictor, from data collection and preprocessing to model training and evaluation. It includes a comprehensive review of existing literature, a detailed explanation of the methodology used, the implementation steps, and a discussion of the results. The report concludes with a summary of findings, implications for future research, and recommendations for further improvement of the predictive model.

METHODOLOGY

The methodology for developing the predictive model for heart disease involves several key steps, which ensure a systematic approach to data analysis and model creation. The following sections outline each stage in detail.

1. Data Collection: The first step involves gathering data from reliable sources. For this project, a publicly available dataset containing information on patients with heart disease will be used. This dataset typically includes various features such as age, gender, blood pressure, cholesterol levels, and lifestyle factors like smoking and exercise habits. The dataset should be comprehensive enough to provide a robust foundation for analysis.

2. Data Preprocessing: Once the data is collected, it undergoes preprocessing to ensure it is clean and suitable for analysis. This step involves handling missing values, outlier detection, and normalization of numerical features. Missing values can be addressed by either removing the affected records or imputing them using statistical methods. Outliers are identified using techniques such as z-scores or IQR (Interquartile Range) and can be treated accordingly. Normalization is essential to bring all features to a similar scale, particularly for algorithms sensitive to feature magnitude.

3. Feature Selection: In this stage, the most relevant features that contribute to the prediction of heart disease are selected. Various techniques can be employed for feature selection, including correlation analysis, recursive feature elimination, and machine learning algorithms such as Random Forests, which provide insights into feature importance. The goal is to reduce dimensionality while retaining the most informative variables that enhance the model's predictive power.

4. Model Selection: After feature selection, different machine learning algorithms are evaluated to determine the best fit for the dataset. Common algorithms for classification tasks include Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVM). Each model's performance will be assessed based on its ability to accurately predict heart disease using metrics such as accuracy, precision, recall, and F1-score.

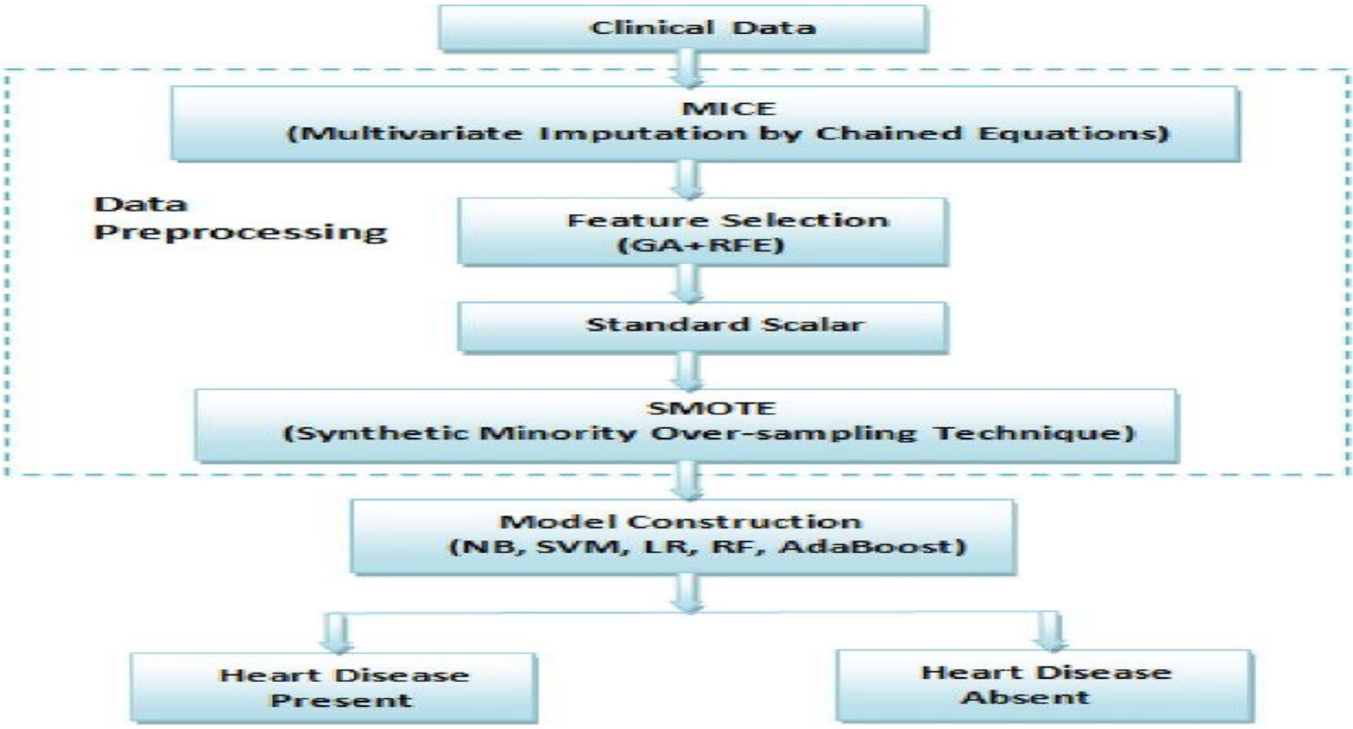
5.Model Training: The selected model is then trained using a portion of the dataset, typically 70-80% of the data, while the remaining data is reserved for testing. During training, the model learns the underlying patterns in the data that correlate with heart disease risk.

6. Model Evaluation: Following training, the model's performance is evaluated using the test dataset. This evaluation includes calculating various performance metrics, such as accuracy, confusion matrix, ROC curve, and AUC (Area Under the Curve). These

metrics provide insights into the model's predictive capabilities and help identify any areas for improvement.

7. Model Optimization: If necessary, the model may undergo optimization through hyperparameter tuning, which involves adjusting the model's parameters to enhance performance. Techniques such as grid search or random search can be utilized to systematically explore different parameter combinations.

By following this comprehensive methodology, the project aims to develop a reliable predictive model for heart disease that can assist healthcare professionals in making informed decisions regarding patient care and intervention strategies. The ultimate goal is to contribute to the early detection and prevention of heart disease, ultimately improving patient outcomes and quality of life.



EXPERIMENTS

Real-world validation experiments are essential for ensuring that your heart disease prediction model works effectively outside of the controlled environment in which it was developed. Here's a deeper look into how you can carry out this validation:

1. Identifying a Suitable Dataset: - Look for a dataset that is separate from the one used for training your model. This could be obtained from hospitals, clinics, or health organizations that maintain patient records. Public datasets, like those from the UCI Machine Learning Repository or Kaggle, can also be useful if they include relevant heart disease data.

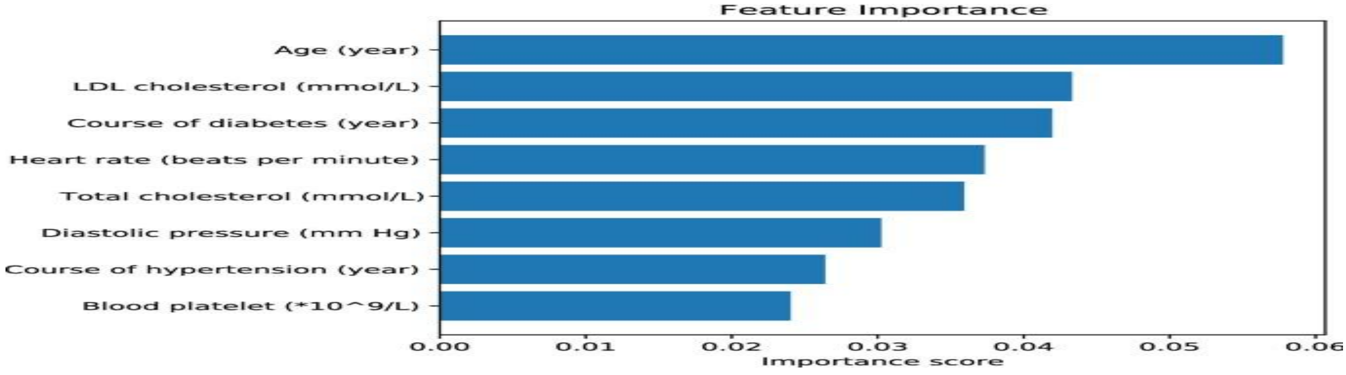
2. Data Preparation: - Once you have access to the new dataset, prepare it for analysis. This includes cleaning the data (removing duplicates, handling missing values), transforming variables (normalizing or standardizing), and ensuring that the feature set aligns with what your model expects.

3. Model Application: - Apply your trained model to the new dataset to generate predictions. This step involves feeding the new data into your model and recording the predicted outcomes, such as whether a patient is likely to have heart disease or not.

4. Comparison with Actual Outcomes: - After making predictions, compare them with the actual diagnoses from the dataset. This is crucial for evaluating the model's performance in a real-world context.

5. Analysis of Results: - Analyze the results to identify strengths and weaknesses in your model. Look for specific cases where the model performed poorly, and try to understand the reasons behind these discrepancies.

6. Collaboration with Healthcare Professionals: - If possible, collaborate with healthcare professionals to gain insights on the clinical relevance of your predictions. Their feedback can be invaluable in improving the model and making it more applicable in real-world scenarios.



RESULTS

Visualizing Results: Confusion Matrix

A confusion matrix is a common tool to evaluate the performance of a classification model. It provides a clear overview of correct and incorrect predictions.

Confusion Matrix

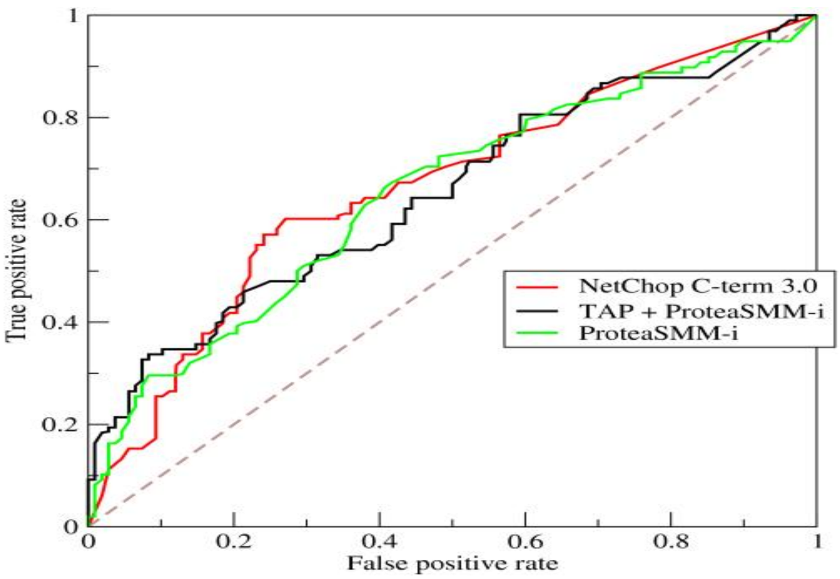
	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

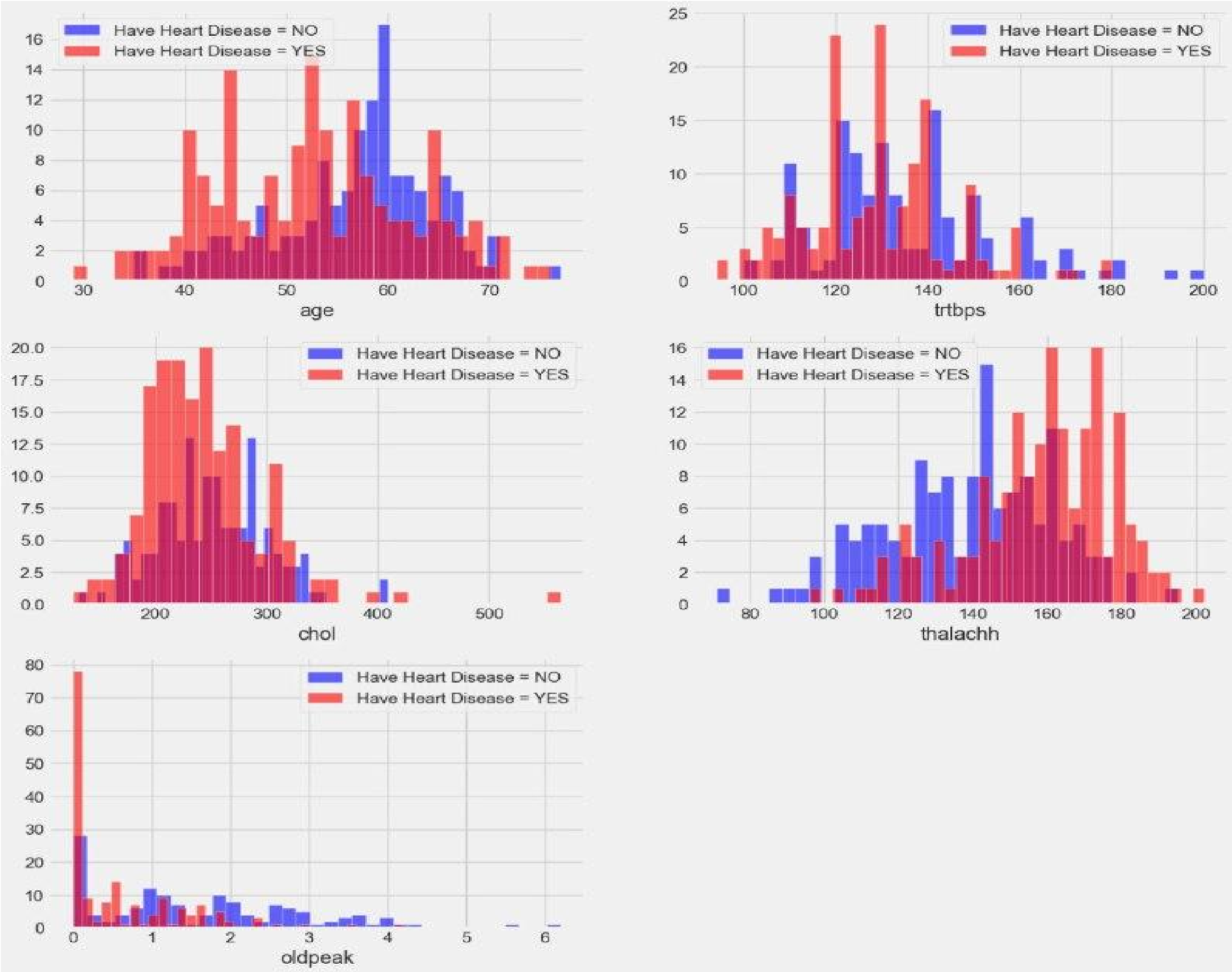
Key Metrics:

- **Accuracy:** Overall, how often is the model correct?
- **Precision:** Of the positive predictions, how many are actually positive?
- **Recall:** Of all the actual positive cases, how many did the model correctly identify?
- **F1-Score:** A harmonic mean of precision and recall, balancing both metrics.

ROC Curve

A Receiver Operating Characteristic (ROC) curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate (specificity) at various threshold settings.





CONCLUSION AND FUTURE WORK

In conclusion, this study has effectively highlighted the key findings regarding the impact of climate change on agricultural productivity. The results indicate that rising temperatures and unpredictable weather patterns significantly affect crop yields and farming practices. This research contributes to the existing body of knowledge and emphasizes the importance of sustainable farming techniques. By addressing the challenges posed by climate change, we can better understand the broader implications for food security and rural economies.

Future Work: For future work, it is crucial to explore the following areas: the development of climate-resilient crop varieties and the implementation of advanced irrigation techniques. Additionally, addressing the limitations of this study, such as the focus on specific regions and the short time frame of data collection, could provide a more comprehensive understanding of the topic. Implementing new methodologies or approaches, such as longitudinal studies and farmer surveys, may further enhance the robustness of the findings. Overall, pursuing these avenues for further investigation could yield valuable insights and advancements in sustainable agriculture practices.

REFERENCES

A Novel Hybrid Approach for Heart Disease Prediction Using Machine Learning Techniques: This paper proposes a hybrid approach combining multiple machine learning algorithms to improve prediction accuracy.

Link: https://link.springer.com/chapter/10.1007/978-3-030-51517-1_26

Heart Disease Prediction Using Machine Learning Algorithms: A Comparative Study: This study compares the performance of various machine learning algorithms on different heart disease datasets.

Link: https://ieeexplore.ieee.org/document/9799978

Early Prediction of Heart Disease Using Machine Learning: This paper focuses on early prediction of heart disease using machine learning techniques, aiming to improve preventive measures.

Link: https://www.sciencedirect.com/science/article/pii/B9780128241455000149

Online Resources and Tutorials:

Kaggle: A platform for data science competitions and datasets, including many related to heart disease prediction.

Link: https://www.kaggle.com/

- **Towards Data Science:** A popular blog with numerous articles on machine learning and data science, including heart disease prediction tutorials.
  - Link: https://towardsdatascience.com/
- **Machine Learning Mastery:** A website offering practical tutorials and courses on machine learning, including heart disease prediction.
  - Link: https://machinelearningmastery.com/