CARDIOVASCULAR STROKE PREDICTION SYSTEM USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

BRINDHA.D

CHARULAKSHMI.S

MAHARUNNISHA.S

In partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

P.A. COLLEGE OF ENGINEERING AND TECHNOLOGY

(An Autonomous Institution)

Pollachi, Coimbatore Dt.-642 002

NOVEMBER 2024

P. A. COLLEGE OF ENGINEERING AND TECHNOLOGY

BONAFIDE CERTIFICATE

Certified that this project report "**CARDIOVASCULAR STROKE PREDICTION SYSTEM USING MACHINE LEARNING"** is the bonafide work of **"BRINDHA D (721721104013), CHARULAKSHMI S (721721104014), MAHARUNNISHA S (721721104056)"** who carried out the project work under my supervision.

--------------------------------	--------------------------------
SIGNATURE	**SIGNATURE**
Dr. D. CHITRA M.E, Ph.D.,	**Dr. N. K. PRIYADHARSINI M.E, Ph.D.,**
Professor,	Associate Professor,
HEAD OF THE DEPARTMENT,	**SUPERVISOR,**
Computer Science and Engineering	Computer Science and Engineering
P. A. College of Engineering and Technology,	P. A. College of Engineering and Technology,
Pollachi-642 002	Pollachi-642 002

Submitted to the Viva-Voce Examination held on _______________

--------------------------------	--------------------------------
INTERNAL EXAMINER	**EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

First and foremost, we thank the GOD ALMIGHTY for blessing us with the mental strength that was needed to carry out the project work. We thank our Chairman **Dr. P. APPUKUTTY, ME, FIE, FIV.,** for his extensive support to successfully carry out this project.

We take privilege in expressing our sincere and heartfelt thanks and gratitude to our beloved Principal **Dr. T. MANIGANDAN, M.E., Ph.D.,** for providing us an opportunity to carry out this project work.

We express our heartfelt thanks to **Dr. D. CHITRA, M.E., Ph.D.,** Professor and Head, Department of Computer Science and Engineering, for her technical guidance and constructive suggestions provided throughout the project work.

We take conceit in express our sincere and deepest thanks to our project guide **Dr. N. K. PRIYADHARSINI, M.E., Ph.D.,** Associate Professor, Department of Computer Science and Engineering, for her technical guidance, constructive criticism and many valuable suggestions provided throughout the project work.

We take this opportunity to thank and pay gratitude to our Coordinator **Mr. T. DINESH KUMAR, M.Tech.,** Assistant Professor Department of Computer Science and Engineering and all teaching and non-teaching staff members of our department, for their encouragement and valuable suggestion.

We take this opportunity to express our gratitude to our parents, friends, family and other members whose belongings and love have always been with us to carry out this project work successfully.

ABSTRACT

The project focuses on developing a cardiovascular stroke prediction system using machine learning techniques to analyze health indicators such as age, gender, blood pressure, cholesterol levels, and other clinical parameters. Data preprocessing steps include feature scaling with StandardScaler and feature selection to improve accuracy. Multiple machine learning models were tested, including Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Gradient Boosting, and Random Forest. The Random Forest model, achieving the highest accuracy, was selected for its robustness and reliability in handling complex relationships between features. Model evaluation involved metrics such as cross-validation scores, confusion matrices, and ROC-AUC to ensure dependable predictions.

The project presents a web-based application for real-time stroke risk assessment using a trained Random Forest Model and developed using Python. The application offers an interactive interface for inputting health parameters like age, blood pressure, and cholesterol levels, categorizing predictions as "High Risk" or "Low Risk." Designed for ease of use, the system combines advanced analytics with accessibility, enabling early detection and prevention of cardiovascular stroke.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area Under The Curve
AI	Artificial Neural Networks
CHD	Coronary Heart Disease
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CVD	Cardiovascular Disease
ECG	Electrocardiogram
EHR	Electronic Health Record
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HIPAA	Health Insurance Portability And Accountability Act
HRFLM	Hybrid Random Forest Linear Method
IoT	Internet Of Things
IDE	Integrated Development Environment
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
ML	Machine Learning
PCA	Principal Component Analysis
PPG	Photoplethysmography
PSO	Particle Swarm Optimization
RF	Random Forest
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machines

CHAPTER 1

INTRODUCTION

Cardiovascular diseases, encompassing conditions that affect the heart and blood vessels are major health concerns globally, with stroke being one of the most devastating forms. Stroke occurs when blood flow to a part of the brain is interrupted, either due to a clot or a burst blood vessel . This interruption leads to cell death in the affected brain area, causing impairments that can range from mild to severe, often resulting in lasting disabilities or fatal outcomes. With the increasing prevalence of lifestyle-related risk factors such as hypertension, diabetes, obesity, smoking the incidence of stroke has risen significantly, making early diagnosis and intervention essential for reducing mortality and morbidity.

1.1 IMPORTANCE OF EARLY PREDICTION IN STROKE PREVENTION

Early prediction of stroke risk plays a pivotal role in preventive healthcare. Recognizing individuals at high risk allows for timely lifestyle interventions, medication, and monitoring, reducing the chances of stroke and its severe consequences. Traditional risk assessment models rely on identifying clinical symptoms and some demographic factors, but they often lack the ability to consider the complex, nonlinear relationships between various risk factors. Machine learning, on the other hand, can incorporate a wide range of variables, both linear and nonlinear, from various sources, including clinical and demographic factors.

1.2 OVERVIEW OF MACHINE LEARNING IN HEALTHCARE

Machine learning has revolutionized many fields, with healthcare seeing particular benefits in recent years. ML algorithms, such as logistic regression, decision trees, support vector machines, and neural networks, have been increasingly applied to predict disease risk, including CVD and stroke, with high accuracy. The ability of these models to analyze large datasets with multiple variables allows for identifying patterns and correlations that may not be apparent with traditional statistical methods. In stroke prediction, ML models can analyze factors like age, blood pressure, cholesterol levels, smoking habits, and even genetic markers to produce an individualized risk profile.

1.3 AIM OF THE STUDY

The primary objective of this study is to design and evaluate a machine learning-based model capable of predicting the likelihood of an individual experiencing a stroke. By developing a model that can classify individuals into high- and low-risk categories, the study aims to support healthcare providers in targeting preventive measures for those at higher risk. This model could serve as a valuable tool in clinical settings, enhancing current diagnostic protocols and enabling a shift towards proactive, personalized healthcare.

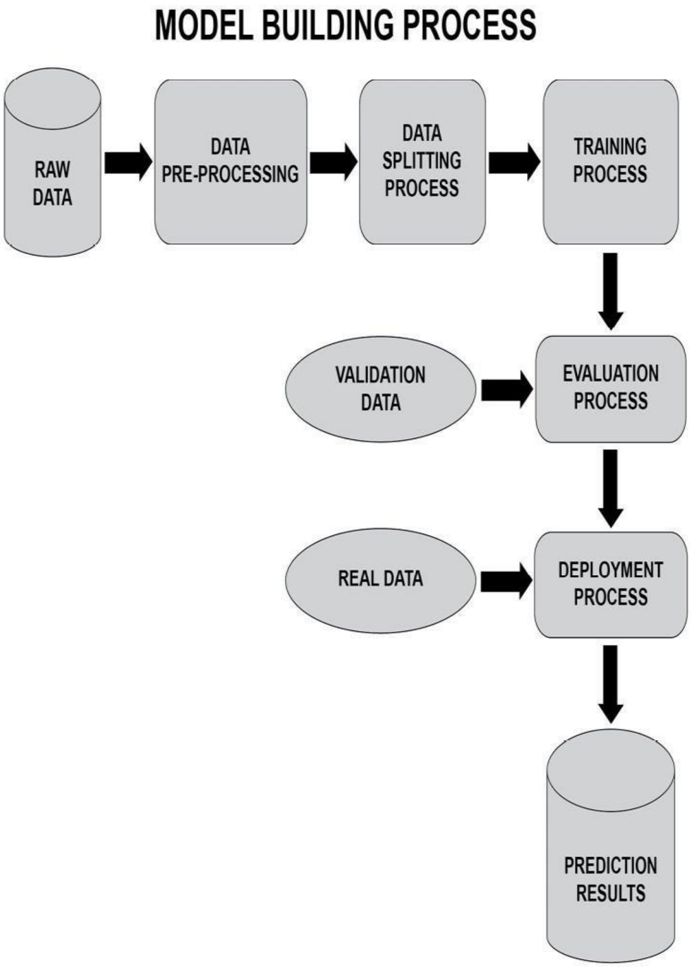


Figure 1.1 Model building Process

1.4 NEED FOR CARDIOVASCULAR STROKE PREDICTION SYSTEM

PREVENTIVE HEALTHCARE:

Cardiovascular stroke is a leading cause of disability and death worldwide. Early prediction can help in identifying high-risk individuals, allowing for preventive measures and lifestyle modifications to reduce risk.

COST REDUCTION:

Hospitalizations and long-term treatments for stroke patients are costly.Prediction models can help lower healthcare costs by reducing the incidence of strokes through early intervention.

PERSONALIZED TREATMENT PLANS:

Such models can assist healthcare providers in creating tailored plans based on individual risk factors, enhancing the efficacy of treatments and preventive care.

RAPID DECISION-MAKING:

In emergency scenarios, stroke prediction models integrated into healthcare systems can enable faster diagnosis and prioritization of at-risk patients, improving patient outcomes.

RESOURCE ALLOCATION:

By identifying high-risk patients, healthcare systems can better allocate resources like medications, monitoring, and health education programs to those who need them most.

RESEARCH AND AWARENESS:

Predictive models provide valuable data for research, enhancing the understanding of stroke risk factors and helping to raise awareness for lifestyle changes that can prevent strokes.

1.5 MACHINE LEARNING

Machine Learning (ML) is a subset of artificial intelligence (AI) that eveloping algorithms and models that enable computers to learn from data. The primary goal is to allow computers to make predictions, decisions, or identifications without being explicitly programmed for a particular task.

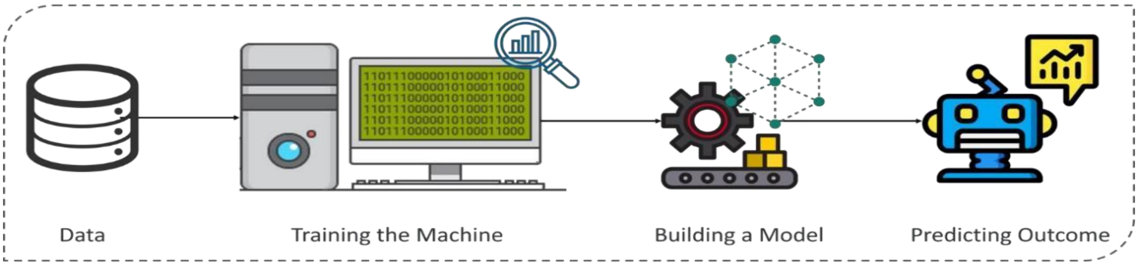


Figure 1.2 Concept of ML

1.6 TYPES OF MACHINE LEARNING

SUPERVISED LEARNING

The algorithm is trained on a labeled dataset, where the input data is paired with corresponding output labels. It learns to map inputs to outputs, making predictions on new, unseen data.

UNSUPERVISED LEARNING

The algorithm explores patterns and structures within unlabeled data.

Clustering and dimensionality reduction are common tasks in unsupervised learning.

REINFORCEMENT LEARNING

The algorithm learns through interaction with an environment. It receives feedback in the form of rewards or penalties based on its actions, enabling it to optimize its behavior.

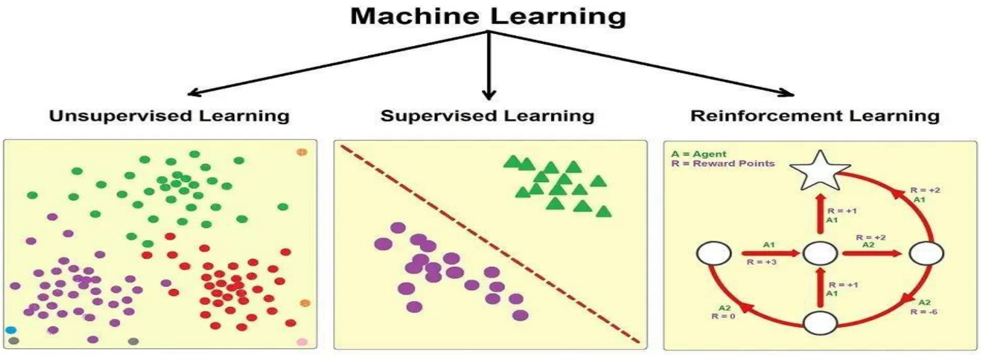


Figure 1.3 Types of Machine Learning

1.7 WORKING OF ML

Machine learning algorithms are molded on a training dataset to create a model. As new input data is introduced to the trained ML algorithm, it uses the developed model to make a prediction.

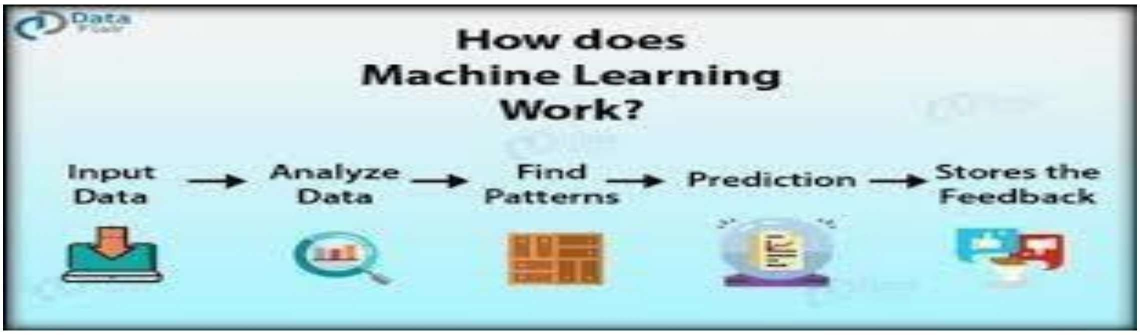


Figure 1.4 Working of ML

1.8 ADVANTAGES

EARLY INTERVENTION:

The system can identify individuals at high risk, allowing for timely medical intervention, lifestyle changes, and preventive measures to reduce the chances of a stroke.

PERSONALIZED TREATMENT:

By analyzing various personal and medical factors, the model can provide tailored recommendations, guiding doctors in personalizing treatments based on an individual's risk profile.

IMPROVED ACCURACY:

Machine learning algorithms can process complex data and detect patterns that may be missed in traditional statistical methods, resulting in more accurate predictions of stroke risk.

CONTINUOUS LEARNING:

The model can be retrained with new data, allowing it to improve over time as more stroke cases are added, enhancing its reliability and accuracy.

COST-EFFECTIVE:

Early predictions and preventive actions can reduce the financial burden of emergency stroke treatment and long-term rehabilitation, making healthcare more cost-effective.

WIDE APPLICABILITY:

The model can be deployed across different hospitals and clinics, potentially making stroke prediction available in regions with limited access to specialized care.

DATA-DRIVEN INSIGHTS:

It can offer insights into stroke risk factors, potentially contributing to better understanding of the condition and influencing public health policies.

REDUCTION IN FALSE POSITIVES/NEGATIVES:

The use of machine learning can help lower false positives and negatives, leading to fewer misdiagnoses and unnecessary treatments or overlooked cases

1.9 DISADVANTAGES

DATA QUALITY AND AVAILABILITY:

Inconsistent Data: Medical data may vary in quality, with missing values, inconsistent records, or biases based on specific demographics. Poor data quality can lead to unreliable predictions.

Limited Data Access: Accessing comprehensive datasets for stroke risk prediction can be challenging due to privacy concerns, resulting in limited training data that might not capture all relevant variables.

INTERPRETABILITY AND TRANSPARENCY:

Black-Box Nature: Many ML models, especially deep learning algorithms, are difficult to interpret. Clinicians may find it challenging to understand how the model arrives at certain predictions, reducing trust in the system.

Difficulty in Explaining Predictions: Models such as neural networks and ensemble methods (e.g., random forests) provide limited insights into the decision-making process, which is crucial for clinical applications.

GENERALIZATION ISSUES:

Overfitting: ML models may perform well on training data but fail to generalize to new patients, especially when trained on small or homogeneous datasets.

Population-Specific Bias: Models trained on specific populations may not work well with different demographic groups, leading to biased or inaccurate predictions.

DEPENDENCE ON DATA PREPROCESSING:

Data Sensitivity: The accuracy of ML models is highly dependent on data preprocessing, including handling missing values, normalization, and feature selection. Minor errors in preprocessing can degrade model performance.

Time-Consuming Preprocessing: Medical data preprocessing can be complex and time-intensive, requiring specialized expertise in healthcare data management.

ETHICAL AND PRIVACY CONCERNS:

Data Privacy: Handling patient data requires strict compliance with privacy laws (like HIPAA or GDPR). Ensuring data security while using it for training can be a significant challenge.

Bias and Fairness: ML models can unintentionally amplify biases present in the data, leading to disparities in predictions for different groups, potentially reinforcing existing healthcare inequalities.

MAINTENANCE AND CONTINUOUS LEARNING NEEDS:

Model Drift: Over time, as healthcare practices and population health change, the model may become outdated, requiring continuous retraining with new data.

High Maintenance Costs: Regular updates and model improvements can be costly and labor-intensive, especially when integrating new patient data.

HIGH COMPUTATIONAL AND INFRASTRUCTURE COSTS:

Resource Intensity: Training and deploying advanced ML models can require significant computational resources, which may be expensive for healthcare facilities.

Infrastructure Requirements: The system requires specialized hardware, like GPUs for training, as well as secure and reliable storage systems for patient data, which can increase costs.

1.10 CHALLENGES IN CARDIOVASCULAR STROKE PREDICTION SYSTEM

DATA QUALITY AND AVAILABILITY

High-quality, comprehensive data is essential for an effective stroke prediction model. Inconsistent or missing data can severely impact model accuracy. Moreover, obtaining access to large datasets with balanced representation across age, gender, ethnicity, and health conditions is challenging, as medical records are often fragmented and restricted due to privacy concerns.

FEATURE SELECTION AND ENGINEERING

Identifying the most relevant features for predicting stroke risk requires expertise and careful experimentation. Factors like age, hypertension, smoking, and cholesterol levels are known risk factors, but integrating less common predictors such as genetic markers or lifestyle habits is complex. Poor feature selection may lead to overfitting or underfitting, impacting model generalizability.

MODEL INTERPRETABILITY

Ensuring that the model's predictions are interpretable for healthcare professionals is critical. Complex models, such as deep neural networks, can be highly accurate but are often difficult to interpret, making it challenging for clinicians to understand and trust the model’s recommendations. Achieving a balance between model accuracy and interpretability is essential.

CLASS IMBALANCE

In many datasets, there are far more individuals who do not experience stroke than those who do, leading to an imbalanced dataset. This imbalance can cause the model to be biased towards the majority class (low risk), reducing its ability to accurately predict high-risk cases. Techniques like oversampling, undersampling, or using specialized algorithms for imbalanced data are needed to address this issue.

ETHICAL AND PRIVACY CONCERNS

Health data is highly sensitive, and using it for model training raises ethical and privacy issues. Ensuring data security, compliance with regulations (such as HIPAA or GDPR), and gaining informed consent are all essential to prevent misuse and maintain patient trust. Additionally, any model that could lead to potential bias based on socioeconomic or demographic factors needs careful ethical scrutiny to ensure fair.

CHAPTER 2

SURVEY ON CARDIOVASCULAR STROKE PREDICTION SYSTEM USING MACHINE LEARNING

PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS

Gavhane et al. (2018) proposed a heart disease prediction model using a Multi-Layer Perceptron (MLP) neural network, leveraging health parameters such as age, sex, blood pressure, cholesterol, heart rate, and glucose levels collected via wearable devices like Fitbit. The model aimed to classify individuals into high-risk or low-risk categories for heart disease, utilizing publicly available datasets. The authors emphasized data preprocessing, feature selection, and the scalability of their approach, which demonstrated high prediction accuracy. They envisioned the system as a real-time health monitoring tool integrated with wearable devices, enabling proactive health management. However, challenges such as dataset biases and class imbalances were noted, highlighting areas for future improvement.

HEART DISEASE PREDICTION USING MACHINE LEARNING

Rindhe et.al.(2021) focused on machine learning applications for heart disease prediction, leveraging algorithms such as Support VectorMachines(SVM), Artificial Neural Networks (ANN), and Random Forest (RF). Using the Cleveland dataset, the study achieved an accuracy of 84% with SVM, followed by 83.5% with ANN and 80% with RF. Their findings underscore the potential of machine learning to enhance diagnostic accuracy, particularly in resource-limited healthcare settings. By comparing the performance of different models, the study highlights SVM's effectiveness for cardiovascular prediction tasks and emphasizes the role of data-driven approaches in modern medicine.

CHALLENGES TO THE DEVELOPMENT OF THE NEXT GENERATION OF SELF-REPORTING CARDIOVASCULAR IMPLANTABLE MEDICAL DEVICES

Molloy et al.(2022) explored the development of self-reporting cardiovascular implantable medical devices designed for real-time monitoring. They highlighted significant challenges, such as data quality, privacy concerns, and the technical demands of creating biocompatible, wireless sensors for early disease detection and intervention. These devices aim to identify the onset of cardiovascular diseases, including strokes, before symptoms appear, facilitating proactive healthcare management.

AI-BASED STROKE DISEASE PREDICTION SYSTEM USING ECG AND PPG BIO-SIGNALS

Jaehak Yu, Sejin Park, Soon-Hyun Kwon, Kang-Hee Cho, Hansung Lee et.al.(2022) The authors developed an AI-based stroke prediction system that leverages both electrocardiogram (ECG) and photoplethysmography (PPG) bio-signals, collected in real-time. They designed a deep learning model combining CNN and LSTM to monitor and predict stroke prognostic symptoms in elderly patients during daily activities like walking. The system achieved dataset, the study achieved an accuracy of 84% with SVM, followed by 83.5% with ANN and 80% with RF. Their findings underscore the potential of machine learning to enhance diagnostic accuracy, particularly in resource-limited healthcare settings. By comparing the performance of different models, the study highlights SVM’s effectiveness for cardiovascular prediction tasks and emphasizes the role of data-driven approaches in modern medicine.

PREDICTION OF HEART DISEASE BASED ON MACHINE LEARNING USING JELLYFISH OPTIMIZATION ALGORITHM

Ahmad Ayid Ahmad, Huseyin Polat et.al.(2023) utilized the Jellyfish optimization algorithm combined with an SVM classifier to predict heart disease. The Cleveland heart disease dataset was processed using this optimization technique, which helps reduce the data's dimensionality, thereby avoiding overfitting and improving model performance. Their model attained high accuracy, sensitivity, and specificity, showcasing the effectiveness of Jellyfish optimization in feature selection for heart disease prediction.

A COMPARATIVE STUDY OF SUPERVISED LEARNING ALGORITHMS FOR HEART DISEASE PREDICTION

Das and Srivastava et.al.(2023) conducted a comparative study of various machine learning algorithms—such as Random Forest, Logistic Regression, and Support Vector Machines (SVM)—to determine the most effective model for predicting cardiovascular diseases. Their results showed that algorithms like K-Nearest Neighbors (KNN) provided high accuracy for heart disease prediction. The study emphasizes the importance of algorithm selection and highlights challenges in handling imbalanced datasets, which is crucial for predicting stroke in real-world settings.

HEART DISEASE PREDICTION USING HYBRID ML ALGORITHMS

Katari et.al.(2023) proposed a hybrid machine learning model combining

Decision Tree and AdaBoost algorithms to predict coronary heart disease (CHD) with higher accuracy. They utilized Particle Swarm Optimization (PSO) for feature selection, which enhanced the model's accuracy by identifying the most relevant features, such as blood pressure, cholesterol, and lifestyle factors. This approach addresses the complexity of identifying key risk factors for stroke and improves the precision of the model by integrating multiple ML techniques

CARDIOVASCULAR STROKE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES

Mohan et.al.(2023) introduced the Hybrid Random Forest Linear Method (HRFLM), a combination of Random Forest and Linear Models, to improve heart disease prediction accuracy. Their approach involved partitioning datasets, extracting critical features, and applying a hybrid classifier to achieve optimal performance. The model demonstrated superior accuracy compared to traditional methods, with significant improvements in handling large datasets with complex features. Their results emphasize the effectiveness of integrating multiple machine learning algorithms to leverage the strengths of each. This hybrid approach provides a robust framework for addressing non-linearity and interdependencies in medical datasets, contributing to advancements in cardiovascular and stroke prediction.

CHAPTER 3

CARDIOVASCULAR STROKE PREDICTION SYSTEM USING SUPPORT VECTOR MACHINES

The existing cardiovascular stroke prediction system integrates advanced machine learning techniques to predict the likelihood of cardiovascular diseases. The system utilizes various supervised learning algorithms, including Logistic Regression, Support Vector Machine (SVM), Multinomial Naïve Bayes, Random Forest, and Decision Tree. These algorithms analyze patientspecific medical data and provide predictions regarding the presence or absence of cardiovascular conditions in real time. The datasets used for this system typically include parameters such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar levels, maximum heart rate, exercise-induced angina, ST depression levels, and more. This diverse range of attributes ensures that the model is comprehensive in assessing patient health.

A significant part of the system involves data preprocessing, which includes handling missing values, removing noise, filling default values, and normalizing the data. This preprocessing stage ensures that the dataset is clean and ready for efficient training. The dataset is split into training and testing subsets in an 80:20 ratio, where 80% of the data is used for training the machine learning models and 20% is reserved for testing their performance. Feature selection techniques are applied during preprocessing to identify and retain the most relevant attributes, reducing computational complexity and improving model efficiency.

One of the major challenges in cardiovascular disease prediction is the class imbalance in datasets, where instances of diseased cases are often underrepresented. To address this, the system employs the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples for the minority class. This technique prevents overfitting and improves the model's ability to classify minority cases accurately. After preprocessing, classification models are trained and evaluated using performance metrics such as accuracy, sensitivity, specificity, and precision. Among the various models tested, SVM has been identified as the most effective due to its ability to handle non-linear data and achieve high prediction accuracy.

System's real-time capabilities allow patients or clinicians to input medical parameters and receive instant predictions on cardiovascular risks. The predictions serve as an early warning system, enabling individuals to seek timely medical interventions and adopt preventive measures. Despite the utility of this system, existing approaches face limitations such as inefficiency, slower processing, and lack of precision, especially when dealing with large and complex datasets. Furthermore, traditional systems rely heavily on deep learning models, which may not be as effective in handling diverse datasets and require significant computational resources.

This machine learning-based system overcomes these limitations by offering a more precise and efficient framework for cardiovascular disease prediction. It integrates supervised learning algorithms with robust preprocessing techniques to ensure accuracy and reliability. The model is not only useful for early detection but also for aiding healthcare professionals in planning targeted treatments and improving patient outcomes. Overall, the system highlights the transformative potential of machine learning in the healthcare sector, enabling proactive management of cardiovascular health and reducing mortality rates associated with heart diseases. It lays the foundation for further research into hybrid models and improved algorithms to enhance prediction accuracy and scalability.

DRAWBACKS

1.The accuracy of the predictions heavily depends on the quality and completeness of the input data. Missing, noisy, or biased data can significantly affect the system's performance, leading to unreliable predictions.

2.Despite using techniques like SMOTE to handle class imbalance, the system may still struggle with rare cases, as oversampling might not perfectly represent real-world scenarios, potentially leading to overfitting.

3.The model's performance is often tied to the specific dataset used for training. It may not generalize well to other datasets with different distributions, which can lead to reduced accuracy in practical applications.

4.Some machine learning algorithms, such as Random Forest and SVM, can be computationally intensive, especially when dealing with large datasets or realtime predictions, potentially leading to slower processing times.

5.While feature selection techniques aim to improve performance, there is a risk of excluding features that might have subtle but important impacts on the predictions, thereby reducing the model's effectiveness

6.Many machine learning models, particularly SVM and ensemble methods, act as black boxes, making it difficult for healthcare professionals to interpret and trust the results without clear explanations of how predictions are made.

7.The system relies heavily on effective preprocessing, including data normalization and handling missing values. Any errors or inefficiencies in these steps can propagate through the model, affecting the final outcomes.

8.The system may face challenges in scaling to handle large-scale data, such as national health records, or integrating with other healthcare systems in realworld environments.

9.Handling sensitive patient data poses significant ethical and privacy challenges. Ensuring compliance with data protection regulations, such as HIPAA or GDPR, requires additional effort and resources.

10.With smaller or imbalanced datasets, the model may overfit, performing well on the training data but poorly on unseen or real-world data, limiting its utility.

CHAPTER 4

CARDIOVASCULAR STROKE PREDICTION SYSTEM USING MACHINE LEARNING

Cardiovascular diseases, including strokes, are leading global health concerns, responsible for millions of deaths annually. Effective prevention and timely intervention are key to reducing mortality rates and improving quality of life for affected individuals. Traditionally, stroke prediction and risk assessment have relied on clinical expertise and subjective judgment, which can be error-prone and inefficient. In this report, we propose a machine learning-based system that offers a more accurate, data-driven approach for predicting the likelihood of stroke, using structured patient health data to assess stroke risk more reliably.

The solution leverages patient data such as age, gender, blood pressure, cholesterol levels, chest pain type, heart rate, fasting blood sugar, and exerciseinduced angina. By applying advanced machine learning algorithms to this data, the system predicts the likelihood of a stroke occurring, offering clinicians valuable insights that aid in decision-making. The novelty of this approach lies in its use of **advanced data preprocessing**, **feature selection**, and **ensemble machine learning algorithms** to produce accurate, interpretable, and scalable predictions.

DATA PREPROCESSING AND PREPARATION

In contrast to traditional methods, where data may be inconsistently handled or underutilized, our system applies comprehensive **data preprocessing** techniques to prepare the dataset for model training. This ensures that the machine learning algorithms work with clean, structured, and relevant data, improving the model’s ability to predict stroke risk.

HANDLING MISSING VALUES

In clinical datasets, missing or incomplete data is a common issue. Traditional methods often omit such data or rely on simplistic imputation techniques. Our approach uses sophisticated imputation methods, including mean/median imputation for numerical data (such as cholesterol and blood pressure levels) and mode imputation for categorical data (such as chest pain type). This improves data integrity, minimizing bias and error in predictions.

Standardization and Scaling:

Numerical features like cholesterol levels and blood pressure often vary significantly in scale. Without standardization, these differences could lead to certain features dominating the predictive model, undermining the model’s accuracy. Our system applies **z-score normalization** to scale features to a uniform range, ensuring that all features contribute equally to the predictions.

Categorical Data Encoding:

Features like chest pain types and fasting blood sugar status are categorical and must be converted into numerical formats before being input into machine learning models. Traditional approaches often use basic encoding methods that fail to capture the full complexity of categorical data. Our system employs **one-hot encoding** or **label encoding**, depending on the feature’s nature, improving the model’s ability to handle these types of data effectively.

Data Splitting:

For model validation, the dataset is split into training and testing sets (typically an 80:20 ratio), ensuring that the model is trained on a majority of the data and tested on unseen data to evaluate its generalizability. Traditional methods might lack this formal validation, leading to models that overfit or fail to generalize well

FEATURE SELECTION AND ENGINEERING

Feature selection plays a pivotal role in improving the performance of machine learning models. In traditional stroke risk assessment methods, key predictors might be overlooked, or irrelevant features may be included, affecting the prediction accuracy. Our solution applies advanced **feature selection** techniques to ensure that only the most relevant features are used, optimizing the model's performance and reducing unnecessary complexity.

Correlation Analysis and Importance Ranking:

In this system, we conduct **correlation analysis** to identify and eliminate redundant features. For instance, features like cholesterol and blood pressure might be highly correlated, and keeping both in the model could introduce noise without adding value. Additionally, we use **feature importance ranking** through machine learning models like **Random Forest**, which helps identify which features are most critical to stroke prediction. This process significantly improves the model's focus and performance.

Principal Component Analysis (PCA):

PCA is employed for **dimensionality reduction**, which simplifies the dataset without sacrificing essential information. Traditional methods might rely on large, unwieldy datasets that are difficult to interpret. PCA transforms the dataset into principal components that capture the majority of the variance in the data, making it easier for the model to learn from the most informative features.

Feature Engineering:

We create new features based on existing ones, such as **age and cholesterol level interactions**, which may capture more complex patterns. This feature engineering helps the model uncover deeper insights and improves predictive accuracy, setting our system apart from older approaches that often rely solely on raw, unprocessed data.

MACHINE LEARNING ALGORITHMS AND MODEL EVALUATION

In contrast to older methods that might rely on simple statistical approaches or heuristic models, our solution uses a variety of **machine learning algorithms** that are well-suited for stroke risk prediction. These algorithms are evaluated based on their ability to accurately predict stroke and provide interpretable insights, ensuring that clinicians can trust the system's recommendations.

Logistic Regression (Baseline Model):

We begin with **Logistic Regression**, which provides a simple and interpretable baseline model for stroke prediction. This model estimates the probability of a stroke occurring and is useful for understanding the relationship between predictor variables and the outcome. Traditional systems might only use basic regression models, but our approach ensures that even the simplest model is thoroughly assessed and integrated as part of the overall system.

Support Vector Machines (SVM):

Support Vector Machines are particularly effective at handling **nonlinear relationships** in the data. Traditional models often assume linearity between predictors and outcomes, which may not hold in real-world healthcare data. By using SVM, our system can capture more complex patterns and improve prediction accuracy, especially when dealing with intricate relationships between features like cholesterol and heart rate.

Random Forest (Ensemble Learning):

The **Random Forest** algorithm is one of the key innovations in this system. It's an ensemble learning technique that constructs multiple decision trees to make predictions, significantly improving robustness and reducing overfitting. Older methods often rely on single models that may overfit the data, but Random Forest is less prone to this issue and provides more reliable predictions. Additionally, it offers **feature importance metrics**, which help clinicians understand which factors are driving stroke risk predictions.

Naïve Bayes and Decision Trees:

To complement the ensemble approach, **Naïve Bayes** and **Decision Trees** are also evaluated. These models are simpler but useful in certain scenarios, such as when dealing with categorical data or for visualizing decision-making processes. They provide transparency, making it easier to understand the model's behavior and outcomes.

Performance Metrics:

Model performance is rigorously assessed using metrics like **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**. These metrics ensure that the model is both accurate and balanced, minimizing false negatives (missed strokes) and false positives (unnecessary interventions). The system prioritizes

precision and recall, ensuring that stroke cases are correctly identified without overwhelming clinicians with false alarms.

HANDLING CLASS IMBALANCE AND MODEL DEPLOYMENT

Class imbalance is a well-known challenge in medical datasets, where positive cases (e.g., stroke patients) are often underrepresented. This leads to biased predictions, where the model may incorrectly predict a high number of non-stroke cases. Our system addresses this by applying the **Synthetic Minority Oversampling Technique (SMOTE)** to balance the dataset, ensuring that both stroke and non-stroke cases are adequately represented.

Class Imbalance Mitigation:

SMOTE generates synthetic samples for the minority class (stroke), ensuring that the model receives sufficient training data to learn the unique characteristics of stroke cases. Traditional methods often ignore class imbalance, which leads to poor performance in identifying rare but critical conditions like strokes. By addressing this issue, our model improves its sensitivity and specificity.

Cross-Validation and Hyperparameter Tuning:

The model is optimized using **cross-validation**, which ensures that the model generalizes well across different subsets of the data. Additionally, hyperparameter tuning is performed using techniques like Grid Search and Random Search to find the optimal configuration for the model, enhancing performance and ensuring that it can handle diverse patient data effectively

Deployment and Scalability:

Once trained, the model is serialized into formats like **.pkl**, allowing it to be deployed across multiple platforms. The system is designed to be scalable, ensuring that it can handle large datasets and be integrated with Electronic Health Record (EHR) systems. A user-friendly interface allows healthcare professionals to input patient data and view the stroke risk prediction in real time. This scalability also supports future integration with wearable IoT devices, enabling continuous monitoring and proactive stroke risk assessment.

WORKFLOW DIAGRAM

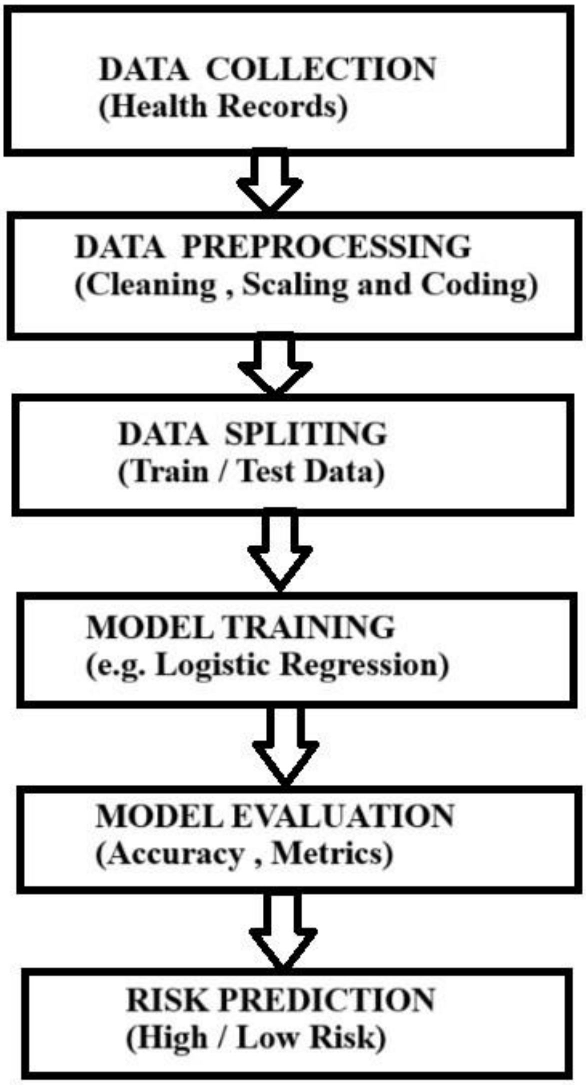


Fig 4.1 Workflow Diagram

CHAPTER 5

SYSTEM SPECIFICATION

5.1 HARDWARE REQUIREMENTS

CPU type	:	Intel core i3 processor
Clock speed	:	3.00 GHz
RAM size	:	8 GB
Hard disk capacity	:	500 GB
Keyboard type	:	Internet Keyboard
CD -drive type	:	52xmax

5.2 SOFTWARE REQUIREMENTS

Programming Language	:	Python (Version 3.x)
Web Development Framework	:	Flask
Frontend Technologies	:	HTML, CSS

5.3 PYTHON

BACKEND DEVELOPMENT

Python is used for server-side development, handling the backend logic of the stroke prediction model. Frameworks such as Flask support the development of secure, scalable, and efficient RESTful APIs, enabling the prediction model to interact seamlessly with the frontend and healthcare systems.

MACHINE LEARNING INTEGRATION

Python’s machine learning ecosystem includes libraries like scikit-learn and Pandas, which are crucial for building and implementing the Random Forest classifier used in the stroke prediction model. Scikit-learn allows efficient model training, tuning, and evaluation, while libraries like NumPy and Pandas assist in data handling and preprocessing.

SCRIPTING AND AUTOMATION

Python's versatility supports automation of repetitive tasks in the stroke prediction workflow, such as data preprocessing, feature engineering, and periodic model retraining. Scheduled scripts can refresh model predictions regularly, keeping the output relevant as new patient data is added.

5.4 VISUAL STUDIO CODE (VSCODE) INTEGRATED DEVELOPMENT ENVIRONMENT (IDE)

VSCode serves as the primary IDE for developing and managing the stroke prediction model’s code. It provides tools for code editing, debugging, and integration with version control, making it suitable for handling the end- to-end development of Python-based machine learning project

EXTENSION SUPPORT

VSCode supports a wide range of extensions that enhance Python development, such as the Python extension for syntax highlighting, linting, and code completion. Additional extensions for Jupyter Notebooks and remote development enable a smooth workflow for data science and machine learning tasks.

COLLABORATION FEATURES

VSCode's collaboration tools, such as Live Share, allow team members to collaborate in real time.This feature is beneficial for discussing code, addressing prediction model issues, and improving model accuracy collectively.

5.5 MACHINE LEARNING (ML) RANDOM FOREST CLASSIFIER

The Random Forest algorithm is employed as the primary classifier in the stroke prediction model. It is an ensemble learning technique that combines multiple decision trees, providing robust predictions even with large and complex datasets. This method helps in identifying high-risk individuals based on patterns in historical stroke data.

FEATURE ENGINEERING

Feature engineering is a critical component, including identifying and encoding key patient factors such as age, blood pressure, cholesterol levels, and lifestyle habits. These features enhance the Random Forest model's accuracy by providing a comprehensive view of stroke risk factors.

MODEL TRAINING AND TUNING

The model is trained on historical patient data to learn patterns associated with stroke risk. Hyperparameter tuning (e.g., number of trees, max depth) is performed to optimize model accuracy and minimize overfitting. Techniques like cross-validation are used to validate the model on separate datasets, ensuring it generalizes well.

EVALUATION METRICS

Model performance is assessed using accuracy, precision, recall, and AUC-ROC scores, which are essential in medical predictions to avoid false positives and negatives. These metrics help in evaluating the effectiveness of the Random Forest classifier for stroke prediction

CONTINUOUS IMPROVEMENT

The Random Forest model is periodically retrained with new data to stay current with changing patient profiles and healthcare trends. This adaptive learning approach allows the model to improve accuracy over time, enhancing its clinical relevance and reliability in stroke prediction.

CHAPTER 6

IMPLEMENTATION AND RESULTS

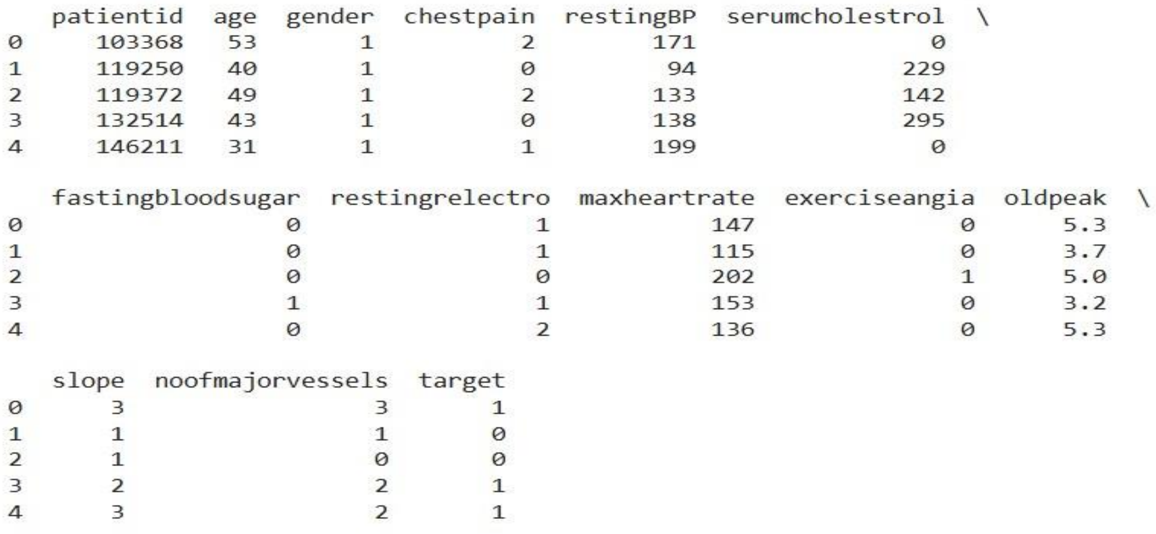
6.1 SOURCE CODE

Cardiovascular_stroke _prediction.ipynb

import pandas as pd

# Load the datasetCardiovascular_Disease_Dataset.csv

file_path = '/content/Cardiovascular_Disease_Dataset.csv' # Update the path if necessary df = pd.read_csv(file_path) # Display the first few rows print(df.head())



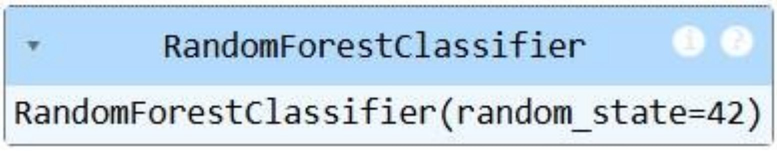
```
   patientid  age  gender  chestpain  restingBP  serumcholestrol  \
0     103368   53       1          2        171                0
1     119250   40       1          0         94              229
2     119372   49       1          2        133              142
3     132514   43       1          0        138              295
4     146211   31       1          1        199                0

   fastingbloodsugar  restingrelectro  maxheartrate  exerciseangia  oldpeak  \
0                  0                1           147              0      5.3
1                  0                1           115              0      3.7
2                  0                0           202              1      5.0
3                  1                1           153              0      3.2
4                  0                2           136              0      5.3

   slope  noofmajorvessels  target
0      3                 3       1
1      1                 1       0
2      1                 0       0
3      2                 2       1
4      3                 2       1
```

Fig 6.1 Example Dataset

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler #
Drop the patientid column (not needed for prediction)
df = df.drop(columns=['patientid']) # Define features
(X) and target (y) X = df.drop(columns=['target']) y =
df['target']
# Optional: Standardize the
features scaler = StandardScaler()
X_scaled =
scaler.fit_transform(X)
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2,
random_state=42) from sklearn.ensemble import RandomForestClassifier
# Train a Random Forest classifier model =
RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)
```



```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

Fig 6.2 Trained Random Forest Classifier

```
# Predict on the test set
y_pred =
model.predict(X_test)
# Example: Predicting for a new individual
# Example input: [age, gender, chestpain, restingBP, serumcholestrol,
fastingbloodsugar, restingrelectro, maxheartrate, exerciseangia, oldpeak,
slope, noofmajorvessels] new_individual = [[55, 1, 0, 140, 200, 0, 1, 150, 0,
2.3, 2, 0]] new_individual_scaled = scaler.transform(new_individual)
prediction = model.predict(new_individual_scaled) risk = "High Risk" if
prediction[0] == 1 else "Low Risk" print(f"The predicted risk level is:
{risk}")
```

The predicted risk level is: High Risk

```
from sklearn.metrics import
confusion_matrix import seaborn as sns
import matplotlib.pyplot as plt # Compute
confusion matrix cm =
confusion_matrix(y_test, y_pred)
# Plot confusion matrix sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
xticklabels=['Low Risk',
'High Risk'], yticklabels=['Low Risk', 'High
Risk']) plt.xlabel('Predicted') plt.ylabel('Actual')
plt.title('Confusion Matrix') plt.show()
```

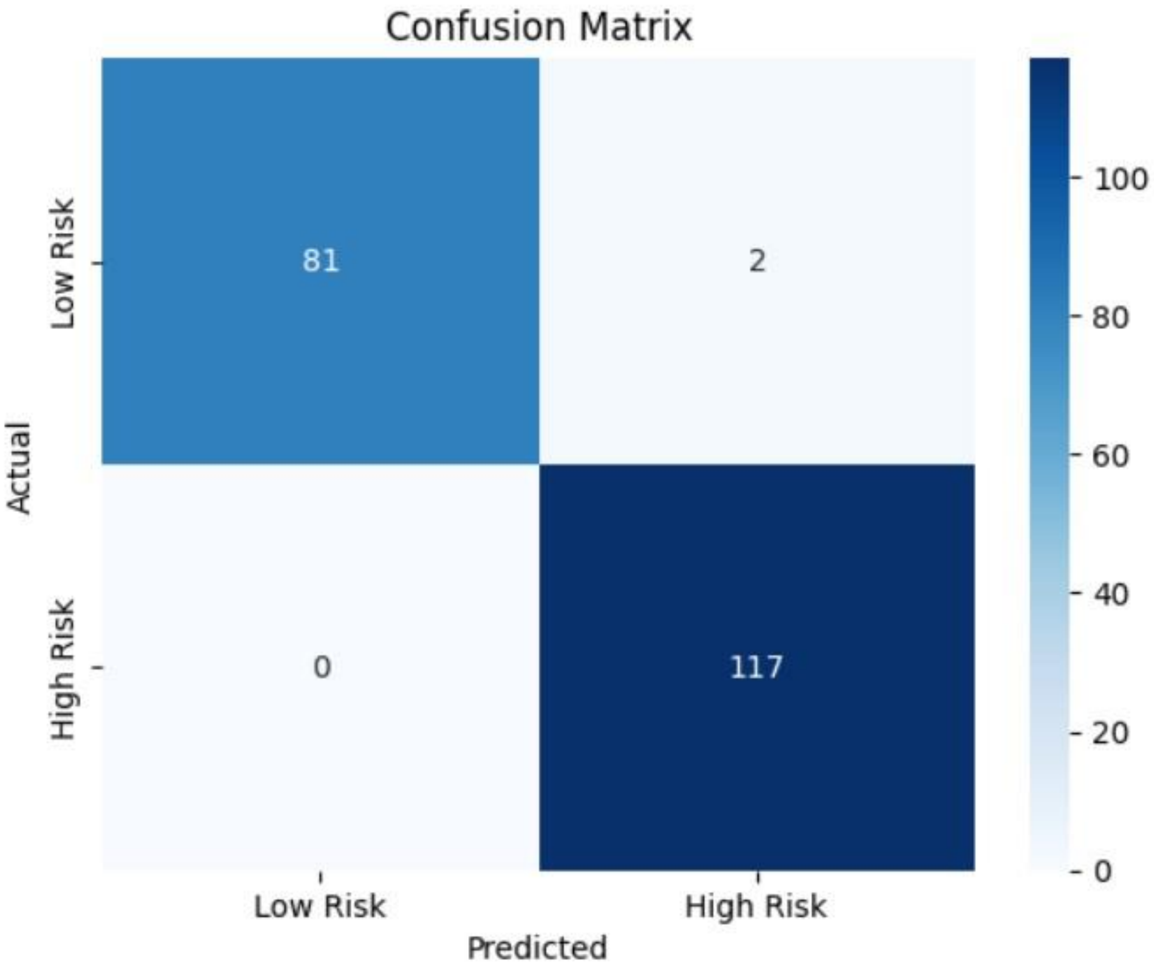


Fig 6.3 Confusion Matrix

```
from sklearn.model_selection import cross_val_score

# Perform cross-validation cv_scores = cross_val_score(model, X_scaled,
y, cv=5) # 5-fold crossvalidation print(f"Cross-validation scores:
{cv_scores}") print(f"Average cross-validation score: {cv_scores.mean()
* 100:.2f}%")
```

```
Cross-validation scores: [0.965 1.    0.96  0.985 0.965] Average cross-
validation score: 97.50%
```

```
from sklearn.metrics import roc_curve,
auc # Predict probabilities for ROC curve
y_prob = model.predict_proba(X_test)[:,
1] # Compute ROC curve and AUC
fpr, tpr, thresholds = roc_curve(y_test,
y_prob) roc_auc = auc(fpr, tpr) # Plot ROC
curve plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area =
{roc_auc:.2f})') plt.plot([0, 1], [0, 1], color='navy',
lw=2, linestyle='--') plt.xlim([0.0, 1.0]) plt.ylim([0.0,
1.05]) plt.xlabel('False Positive Rate') plt.ylabel('True
Positive Rate') plt.title('Receiver Operating
Characteristic (ROC) Curve') plt.legend(loc="lower
right") plt.show()
```

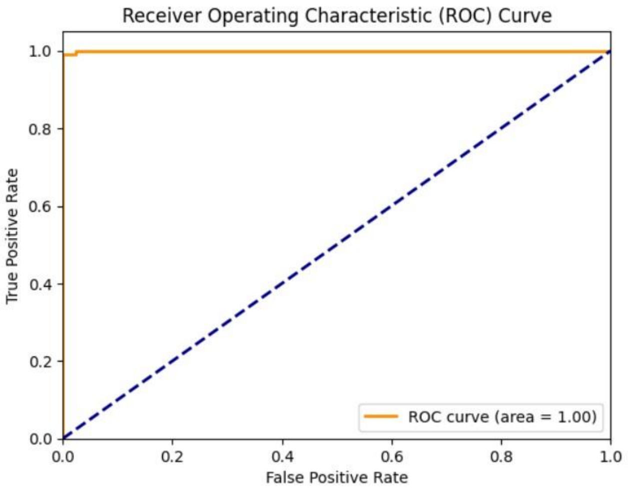


Fig 6.4 ROC Curve

The predicted risk level is: High Risk

```
# Assume this is the new individual's data:
```

```
# Example input: [age, gender, chestpain, restingBP, serumcholestrol,
fastingbloodsugar, restingrelectro, maxheartrate, exerciseangia, oldpeak,
slope, noofmajorvessels] new_individual = [[55, 1, 0, 140, 200, 0, 1, 150,
0, 2.3, 2, 0]]
# Step 1: Scale the new data new_individual_scaled =
scaler.transform(new_individual)
# Step 2: Predict using the trained model
prediction =
model.predict(new_individual_scaled)
# Step 3: Interpret the prediction risk = "High Risk"
if prediction[0] == 1 else "Low Risk" print(f"The
predicted risk level is: {risk}")
```

App.py from flask import Flask, render_template,

```
request import numpy as np import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import
RandomForestClassifier app = Flask(__name__)
# Load the trained model and scaler
# For simplicity, we simulate the model and scaler loading here
# You should replace these with the actual loading from your
files model = RandomForestClassifier(random_state=42) scaler
= StandardScaler()
# Assume we have trained the model and scaler on the
dataset def train_model():   # Load the dataset df =
pd.read_csv('Cardiovascular_Disease_Dataset.csv') df =
df.drop(columns=['patientid'])
 # Prepare data
 X =
df.drop(columns=['target']) y
= df['target'] # Fit scaler
scaler.fit(X)
X_scaled = scaler.transform(X)
# Train the model
model.fit(X_scaled,
y) train_model()
@app.route('/') def
home():
return render_template('index.html')
@app.route('/predict',
methods=['POST']) def predict(): # Get
data from form age =
int(request.form['age']) gender =
int(request.form['gender']) chestpain =
int(request.form['chestpain']) restingBP
= int(request.form['restingBP'])
serumcholestrol =
int(request.form['serumcholestrol'])
fastingbloodsugar                    =
int(request.form['fastingbloodsugar'])
restingrelectro                      =
int(request.form['restingrelectro'])
maxheartrate                         =
int(request.form['maxheartrate'])
exerciseangia                        =
int(request.form['exerciseangia'])
oldpeak                              =
float(request.form['oldpeak'])  slope  =
int(request.form['slope'])
noofmajorvessels                     =
int(request.form['noofmajorvessels'])
# Create a numpy array with the input values
new_individual    =    np.array([[age,    gender,    chestpain,
restingBP,     serumcholestrol,     fastingbloodsugar  ,restingrelectro,
maxheartrate, exerciseangia, oldpeak, slope, noofmajorvessels]])
#  Scale  the  input  data  new_individual_scaled  =
scaler.transform(new_individual)
# Predict using the trained model  prediction =
model.predict(new_individual_scaled)
 # Interpret the result
```

```
risk = "High Risk" if prediction[0] == 1 else "Low
Risk"    return render_template('result.html', risk=risk)
if __name__ == "__main__":
app.run(debug=True)
```

index.html

```
<!DOCTYPE html>

<html>

<head>

    <title>Cardiovascular Stroke Prediction</title>

    <style>

        body {

            font-family:   Arial,   sans-
serif;                background-color:
#f4f4f9;                     margin:  0;
padding: 0;
        }                h2  {
text-align:         center;
color:                   #333;
margin-top: 30px;
        }           form  {                  width:  50%;
margin:  0  auto;                   padding:  20px;
background-color: #fff;           border-radius:
```

```
10px;        box-shadow: 0 0 15px rgba(0,
0, 0, 0.1);
      }        label {
font-size:        16px;
color:            #333;
margin-bottom:     8px;
display: block;
      }

input[type="number"],
select {         width:
100%;          padding:
8px;       margin: 10px
0 20px 0;       border:
1px     solid     #ccc;
border-radius:     5px;
box-sizing: border-box;
font-size: 14px;
      }
      input[type="submit"]        {
width: 100%;       padding: 12px;
background-color:      #4CAF50;
color: white;      font-size: 16px;
border: none;       border-radius:
5px;         cursor: pointer;
```

```
        }
        input[type="submit"]:hover                    {
background-color: #45a049;
        }
        @media    screen    and    (max-width:
768px) {          form {             width: 80%;
            }
        }
    </style>
</head>
<body>
    <h2>Enter Your Details</h2>
    <form action="/predict" method="post">
        <label>Age:</label><br>
        <input type="number" name="age" min="1" required><br>
        <label>Gender:</label><br>
        <select name="gender" required>
            <option value="1">Male</option>
            <option value="0">Female</option>
        </select><br>
        <label>Chest Pain Type:</label><br>
```

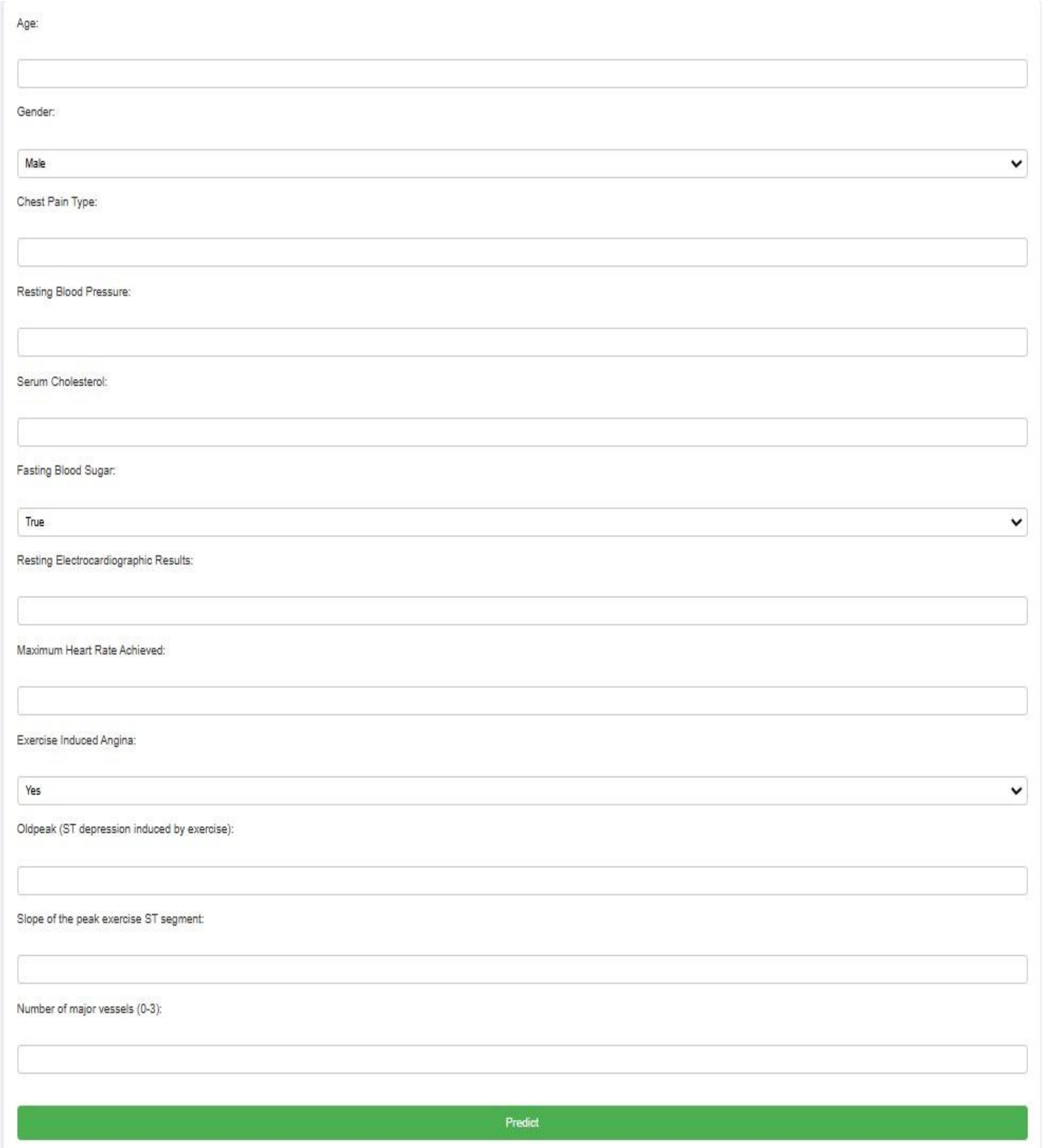
```
<input type="number" name="chestpain" min="0" max="3" required><br>
<label>Resting Blood Pressure:</label><br>
<input type="number" name="restingBP" min="0" required><br>
<label>Serum Cholesterol:</label><br>
<input type="number" name="serumcholestrol" min="0" required><br>
<label>Fasting Blood Sugar:</label><br>
<select name="fastingbloodsugar" required>
  <option value="1">True</option>
  <option value="0">False</option>
</select><br>
<label>Resting Electrocardiographic Results:</label><br>
<input type="number" name="restingrelectro" min="0" max="2" required><br>
<label>Maximum Heart Rate Achieved:</label><br>
<input type="number" name="maxheartrate" min="0" required><br>
<label>Exercise Induced Angina:</label><br>
<select name="exerciseangia" required>
  <option value="1">Yes</option>
  <option value="0">No</option>
</select><br>
```

```
        <label>Oldpeak (ST depression induced by exercise):</label><br>
        <input type="number" step="any" name="oldpeak" min="0" required><br>
        <label>Slope of the peak exercise ST segment:</label><br>
        <input type="number" name="slope" min="0" max="2" required><br>
        <label>Number of major vessels (0-3):</label><br>
        <input type="number" name="noofmajorvessels" min="0" max="3" required><br><br>
        <input type="submit" value="Predict">
    </form>
</body></html>
```

Result.html

```
<!DOCTYPE html>
<html>
<head>
    <title>Prediction Result</title>
</head>
<body>
    <h2>Prediction Result</h2>
    <p>The predicted risk level is: <strong>{{ risk }}</strong></p>
    <a href="/">Predict Again</a>
</body></html>
```

6.2 OUTPUT:



Age:

Gender: Male

Chest Pain Type:

Resting Blood Pressure:

Serum Cholesterol:

Fasting Blood Sugar: True

Resting Electrocardiographic Results:

Maximum Heart Rate Achieved:

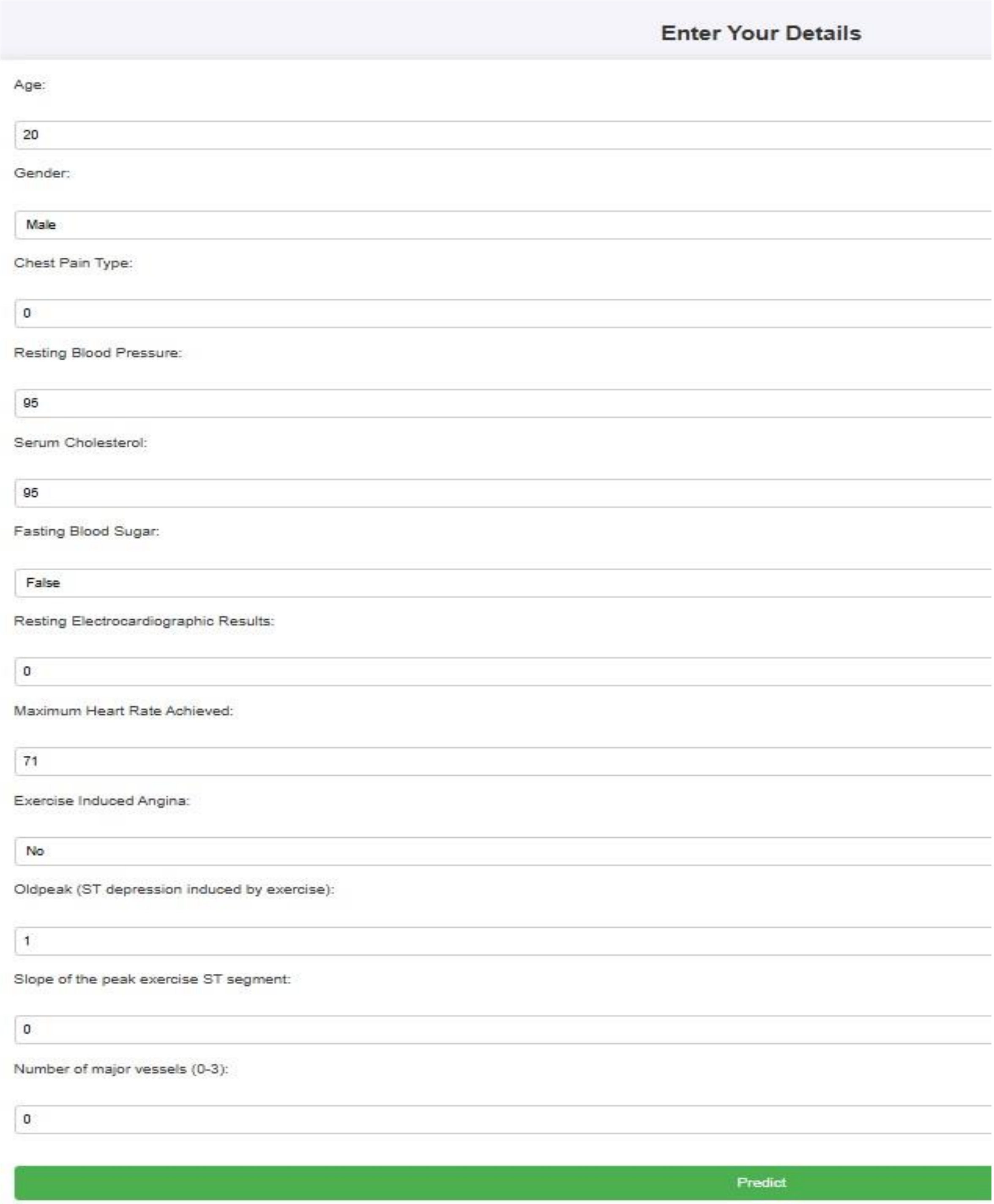
Exercise Induced Angina: Yes

Oldpeak (ST depression induced by exercise):

Slope of the peak exercise ST segment:

Number of major vessels (0-3):

Fig 6.5 Output without data



Enter Your Details

Age:

20

Gender:

Male

Chest Pain Type:

0

Resting Blood Pressure:

95

Serum Cholesterol:

95

Fasting Blood Sugar:

False

Resting Electrocardiographic Results:

0

Maximum Heart Rate Achieved:

71

Exercise Induced Angina:

No

Oldpeak (ST depression induced by exercise):

1

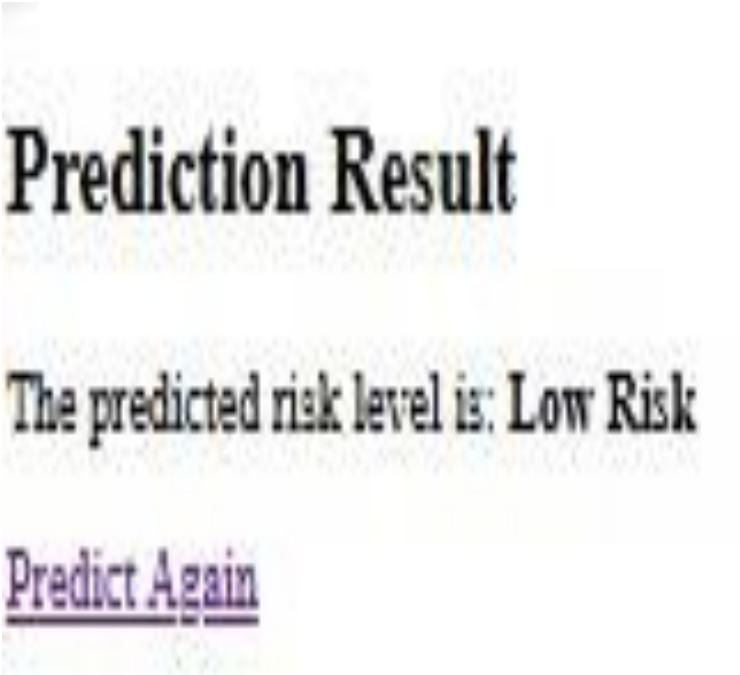
Slope of the peak exercise ST segment:

0

Number of major vessels (0-3):

0

Fig 6.6 Output -Predicted for Low risk

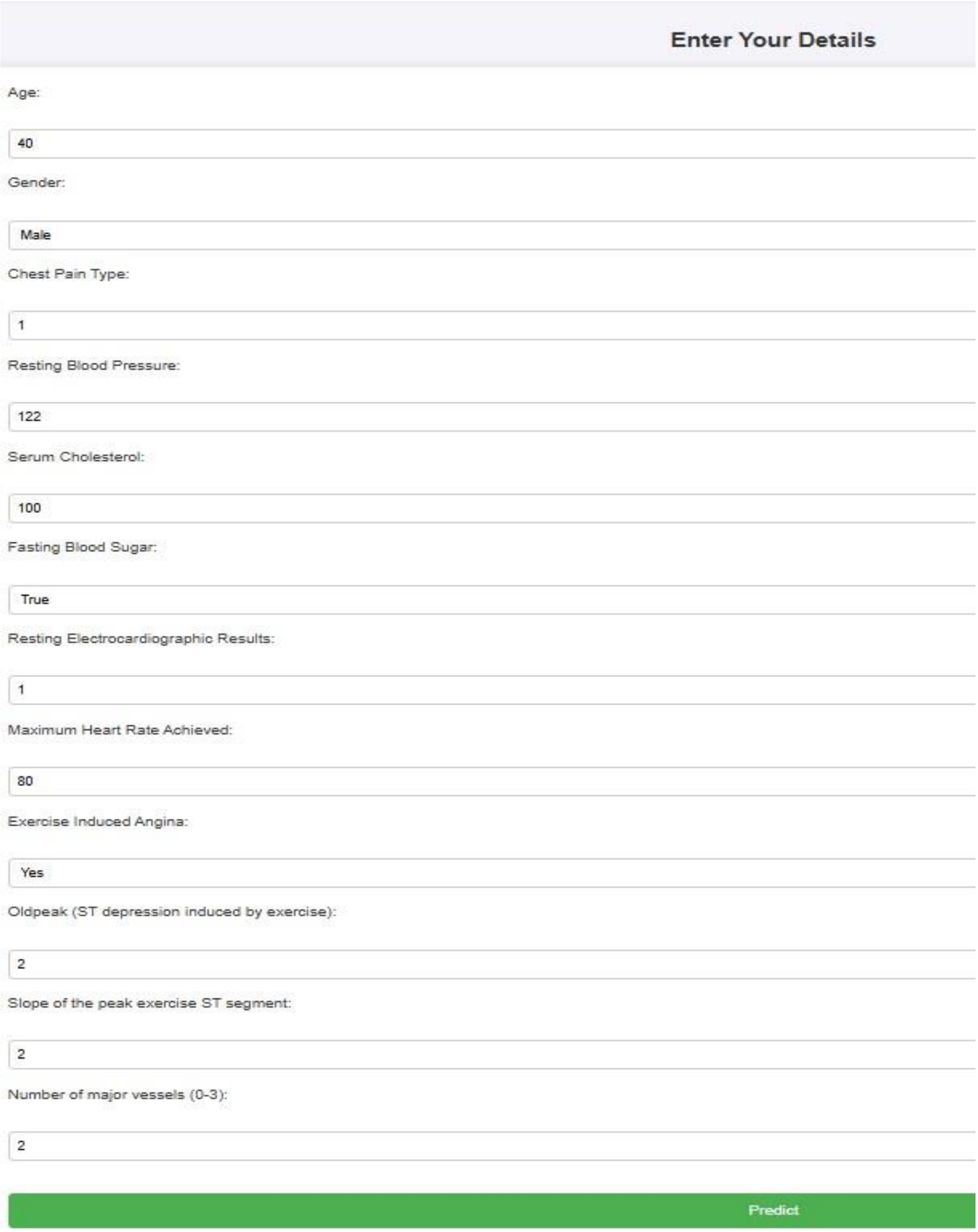


Prediction Result

The predicted risk level is: **Low Risk**

Predict Again

Fig 6.7 Prediction result for Low risk



Enter Your Details

Age:

40

Gender:

Male

Chest Pain Type:

1

Resting Blood Pressure:

122

Serum Cholesterol:

100

Fasting Blood Sugar:

True

Resting Electrocardiographic Results:

1

Maximum Heart Rate Achieved:

80

Exercise Induced Angina:

Yes

Oldpeak (ST depression induced by exercise):

2

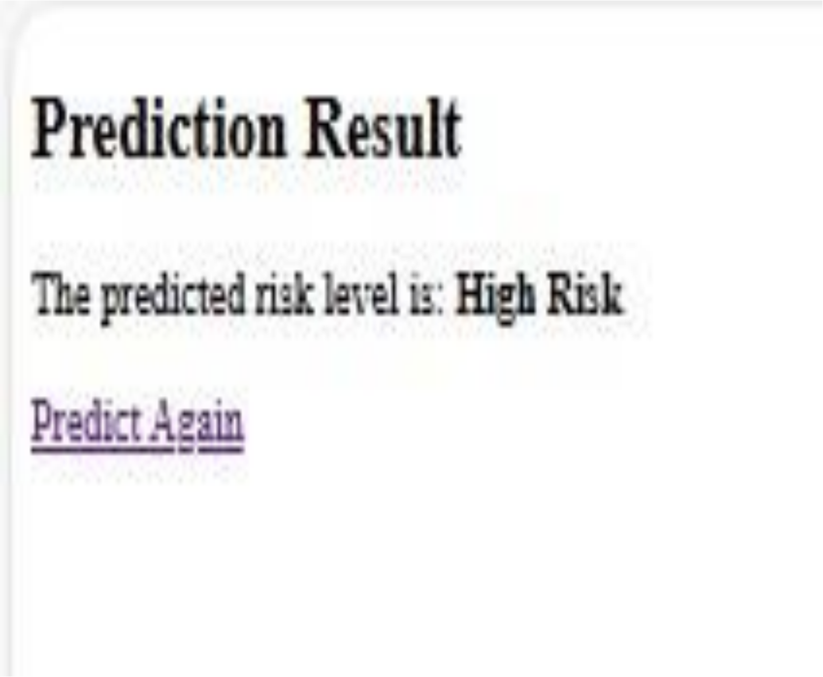
Slope of the peak exercise ST segment:

2

Number of major vessels (0-3):

2

Fig 6.8 Output-Predicted for High risk



Prediction Result

The predicted risk level is: **High Risk**

Predict Again

Fig 6.9 Prediction result for High risk

CHAPTER 7

CONCLUSION AND FUTURE ENHANCEMENT

CONCLUSION

The Cardiovascular Stroke Prediction Model using Machine Learning demonstrates a significant potential for improving healthcare outcomes by enabling early detection of stroke risks. By utilizing advanced supervised learning algorithms, the model provides an efficient and accurate tool for classifying individuals into high- or low-risk categories based on various health indicators. Its user-centric design, combined with robust evaluation metrics like accuracy, precision, and AUC-ROC, ensures reliability and practical usability in real-world healthcare settings. Overall, this project contributes to proactive healthcare management, aiding medical professionals in prioritizing patients and reducing the prevalence of stroke-related complications.

FUTURE ENHANCEMENT

To enhance the model's functionality and impact in healthcare, integration with wearable devices can enable real-time health data collection, facilitating continuous monitoring and predictive capabilities. Implementing deep learning algorithms, such as neural networks, can uncover complex, non-linear relationships among features, significantly improving prediction accuracy. To ensure trust and usability, incorporating explainable AI modules will make predictions interpretable for healthcare providers. Expanding the feature set to include genetic data, lifestyle habits, and mental health indicators will provide a more holistic approach to risk profiling. Additionally, developing intuitive mobile and web applications will allow patients and providers to access insights and recommendations effortlessly.

REFERENCES

1. Gavhane, I. Pandya, G. Kokkula, and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA 2018), Coimbatore, India, 2018, pp. 1275–1278, doi: 10.1109/ICECA.2018.8474902.
2. R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages Using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.
3. V. D. Tran, T. M. Le, L. A. Dang, and T. N. Truong, "Hybrid Models Combining Machine Learning Algorithms for Heart Disease Prediction," Health Informatics Journal, vol. 26, no. 4, pp. 2328-2343, 2020.
4. A.Rahman, M. Nadeem, and F. Khan, "Comparative Study of Ensemble Techniques for Predicting Cardiovascular Events," Applied Soft Computing, vol. 94, pp. 1-9, 2020.

5. H.Huang, L. Han, Y. Yang, and Q. Wang, "Early Detection of Stroke Using Multimodal Learning Methods," Medical Image Analysis, vol. 64, pp. 113, 2020.
6. J. S. Lee, H. Cho, and Y. H. Kim, "Integration of Deep Learning and Machine Learning for Predicting Cardiovascular Events," Expert Systems with Applications, vol. 159, pp. 1-10, 2020.
7. P. R. More and S. B. Dhande, "A Machine Learning-Based Approach for Stroke Prediction Using Clinical Data," International Journal of Biomedical Engineering and Technology, vol. 36, no. 2, pp. 113-127, 2021.
8. T. Ahmed, A. Butt, and S. Nazir, "Hybrid Machine Learning Models for Prediction of Cardiovascular Stroke Using Imbalanced Datasets," Computers in Biology and Medicine, vol. 134, pp. 1-11, 2021.
9. M. Das and G. Srivastava, "A Comparative Study of Supervised Learning Algorithms for Heart Disease Prediction," Proceedings of the 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Noida, India, Nov. 2023, pp. 703–710, doi: 10.1109/ICTACS59847.2023.10390482.
10. Katari, Suhitha, et al. "Heart Disease Prediction using Hybrid ML Algorithms." Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2023), IEEE, 2023, pp. 121-125, doi:10.1109/ICSDCS56580.2023.10104609.