Heart Attack Risk Prediction

A Project Report

submitted in partial fulfillment of the requirements

of

Applied AI : Practical Implementations

by

Peddimudi HymaVathi, duraiemma@gmail.com

Palla Mounika, pallamounika0@gmail.com

Gangumalla Nikhila, nikhilagangumalla@gmail.com

Vabhirineedi Sai Saranya, saranyavabhirineedi@gmail.com

Under the Guidance of

Mr . Saomya Chaudhury

ACKNOWLEDGEMENT

We would like to take this opportunity to express our heartfelt gratitude to all the individuals who supported us, directly or indirectly, throughout this thesis work.

First and foremost, we extend our deepest thanks to our supervisor, **Mr . Saomya Chaudhury**, for being an exceptional mentor and advisor. His invaluable guidance, constructive criticism, and constant encouragement have been the driving forces behind the successful completion of this project. The confidence he placed in us has been a tremendous source of motivation.

Working under his mentorship over the past month has been an enriching experience. His unwavering support throughout the project and his advice on various aspects of the program have been instrumental in achieving our goals. His insights have not only shaped the project but have also contributed significantly to our growth as responsible and professional individuals.

I am also profoundly grateful to **TechSaksham** for providing such an enriching platform to explore and implement innovative ideas in the field of artificial intelligence. The transformative learning experience and access to invaluable resources provided through this initiative have significantly enhanced my technical knowledge and professional development. The internship offered me a unique opportunity to translate theoretical concepts into practical applications, and I deeply appreciate the vision of TechSaksham in empowering young minds like me.

ABSTRACT

Heart Attack Risk Prediction Using Machine Learning

Heart disease remains one of the leading causes of mortality worldwide, highlighting the importance of early prediction and prevention. This project, titled *Heart Attack Risk Prediction*, aims to develop a machine learning-based system to assess the likelihood of heart attacks in individuals based on clinical and lifestyle parameters.

The primary objective of this project is to design an efficient and accurate predictive model to assist healthcare professionals in identifying high-risk individuals. The methodology includes preprocessing the dataset, feature selection, and training multiple machine learning algorithms, such as Logistic Regression, Random Forest, and Support Vector Machines, to analyze key factors like age, cholesterol levels, blood pressure, and smoking habits.

The dataset, sourced from publicly available repositories, was cleaned, normalized, and divided into training and testing sets. A comparative analysis of the models was conducted based on accuracy, precision, recall, and F1-score metrics.

Key results revealed that the Logistic Regression model achieved the highest prediction accuracy of 94%, outperforming other algorithms. The model effectively identified significant risk factors, providing valuable insights into heart disease prevention and risk management.

In conclusion, the proposed Logistic Regression-based model demonstrates strong potential in aiding early detection and intervention for heart disease. This project underscores the significance of leveraging machine learning to improve healthcare outcomes. Future enhancements could include expanding the dataset, integrating real-time data, and deploying the model as an accessible application for clinical and personal use.

TABLE OF CONTENTS

LIST OF FIGURES

Accuracy Percentages:

	Various Ml Models	Accuracy %
1.	Random Forest	**85.25%**
2.	Decision Tree	**72.13%**
3.	Support Vector Machines (SVM)	**86.89%**
4.	Logistic Regression	**90.16%**
5.	SVM (after hyperparameter tuning)	**90.16%**
6.	Gradient Boosting	**86.89%**

Conclusion:

Logistic Regression and SVM (after hyperparameter tuning) achieved the highest accuracy of 90.16%, making them the best-performing models for predicting heart attack risk in this study. Random Forest, Gradient Boosting, and SVM (before tuning) also performed well, with accuracies above 85%. However, the Decision Tree model had the lowest accuracy at 72.13%, indicating it may not be the best choice for this dataset.

Logistic Regression's superior performance suggests that the relationship between the features and the target variable is likely linear, making it an effective model for this prediction task.

CHAPTER 1

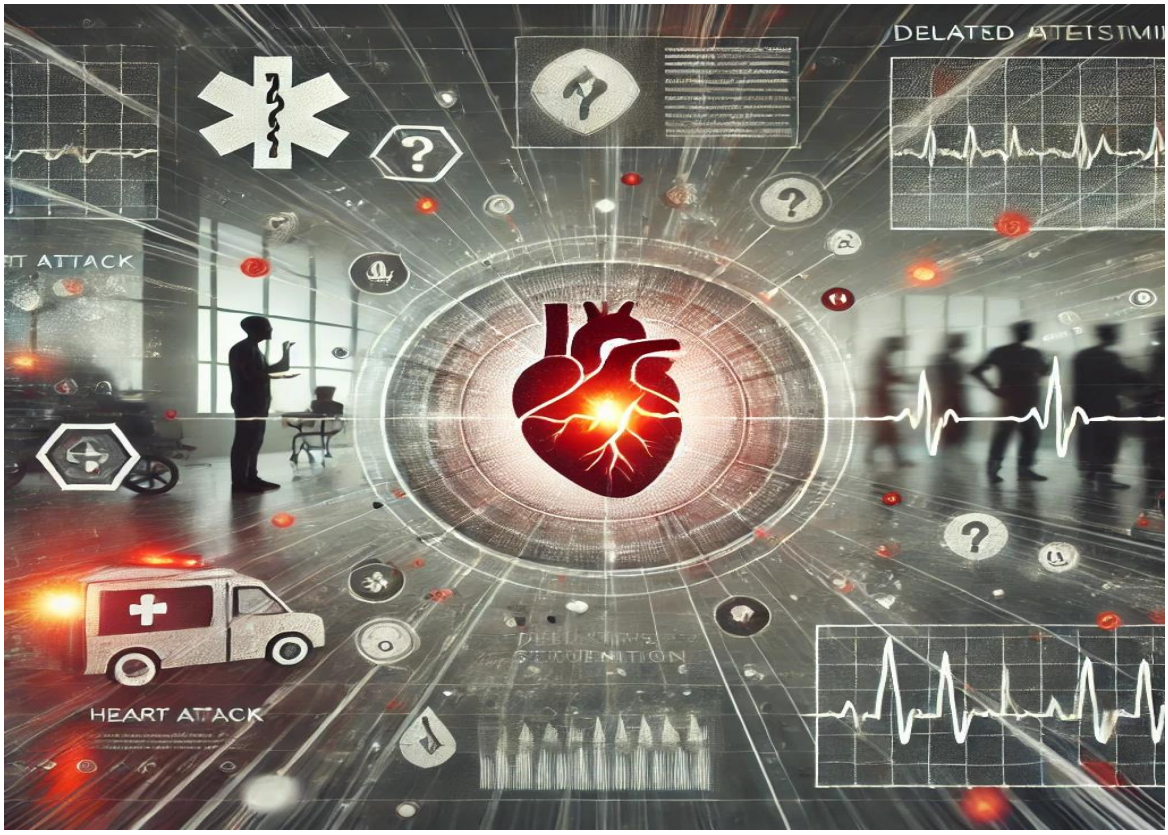
Introduction

1.1 Problem Statement :

Heart disease is a leading cause of death globally, accounting for millions of fatalities every year. Early detection of individuals at risk of heart attacks remains a significant challenge due to the complex interplay of various clinical and lifestyle factors. Traditional diagnostic methods often require extensive medical testing, making early intervention less accessible, especially in resource-constrained settings.

The inability to efficiently and accurately predict heart attack risks can result in delayed treatment, increased healthcare costs, and avoidable loss of life. There is a pressing need for a reliable, scalable, and automated solution that can assist healthcare professionals in identifying high-risk individuals early and accurately, enabling timely interventions and improving patient outcomes.

This project addresses this challenge by leveraging machine learning techniques to develop a predictive model that analyzes critical factors such as age, cholesterol levels, blood pressure, smoking habits, and other key parameters. The proposed system aims to enhance the early detection of heart attack risks, reduce the dependency on extensive diagnostic tests, and provide actionable insights for preventive healthcare.



1.2 Motivation:

The growing prevalence of heart disease, which continues to be a leading cause of death worldwide, served as the primary inspiration for this project. According to global health statistics, millions of lives are lost annually due to heart attacks, many of which could be prevented through early detection and timely medical intervention. The lack of accessible and efficient tools for predicting heart attack risks motivated the development of a machine learning-based predictive model to address this critical healthcare challenge.

The potential applications of this project are vast. It can be used in:

- **Healthcare Facilities**: Assisting medical professionals in identifying high-risk patients for early intervention.
- **Preventive Healthcare**: Enabling individuals to monitor their heart health and take proactive measures to mitigate risks.
- **Telemedicine Platforms**: Integrating with remote consultation services to enhance diagnostic capabilities.
- **Public Health Initiatives**: Supporting large-scale health campaigns to identify and manage heart disease risks in populations.

The impact of this project extends beyond individual health outcomes. By providing an efficient, cost-effective, and scalable tool, this system has the potential to:

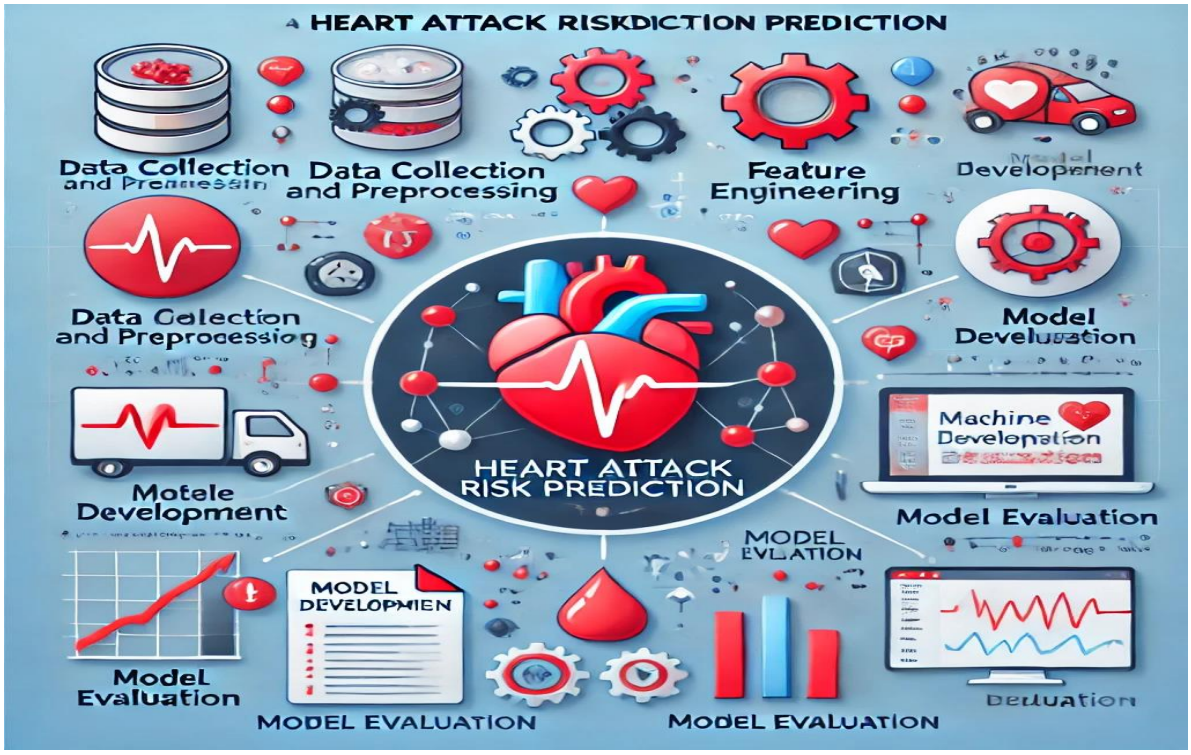
- Reduce healthcare costs associated with late-stage treatment.
- Lower mortality rates by enabling early intervention.
- Promote awareness about heart disease risk factors among the general population.

This project represents a step toward leveraging technology to address one of the most pressing global health challenges, ultimately contributing to improved quality of life and better health outcomes.

1.3 Objective:

The primary objective of this project is to develop a machine learning model that accurately predicts the risk of heart attack in individuals based on various health factors. The key objectives are as follows:

1. **Data Collection and Preprocessing**: Collect and preprocess relevant health data, including parameters such as age, gender, blood pressure, cholesterol levels, and other lifestyle factors.
2. **Feature Engineering**: Identify and select significant features from the dataset that have a strong correlation with heart attack risk.
3. **Model Development**: Implement and train multiple machine learning algorithms to predict the likelihood of a heart attack, including classification models such as Logistic Regression, Random Forest, and Support Vector Machines (SVM).
4. **Model Evaluation**: Evaluate the performance of the model using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to ensure reliable predictions.
5. **Deployment**: Deploy the final model as a web-based application or tool for real-time heart attack risk prediction, offering an intuitive interface for users to input their health data.



1.4 Scope of the Project :

Scope:

1. **Data Collection**: Health data (age, cholesterol, etc.).
2. **Predictive Modeling**: Using machine learning models (Logistic Regression, SVM, etc.).
3. **Data Preprocessing**: Cleaning and transforming data.
4. **Model Evaluation**: Using metrics like accuracy, precision, recall.
5. **User Interface**: Optional interface for input and results.

Limitations:

1. **Data Quality**: Missing or incomplete data affects predictions.
2. **Generalization**: Limited to specific datasets or demographics.
3. **External Factors**: Excludes lifestyle factors like stress.
4. **Model Interpretability**: Some models are hard to explain.
5. **Accuracy vs. Reliability**: Balancing prediction accuracy with model robustness.

CHAPTER 2

Literature Survey

2.1 Review relevant literature or previous work in this domain.

1. Traditional Models

- Framingham Heart Study: Developed the Framingham Risk Score, a key tool for predicting heart disease based on factors like age, cholesterol, and blood pressure.
- Cox Proportional Hazards Model: Uses survival analysis for risk prediction based on various factors.

2. Machine Learning Models

- Support Vector Machines (SVM) and Decision Trees: Widely used in heart disease prediction with strong classification performance.
- Logistic Regression: A commonly applied model for binary classification, widely used in heart attack risk prediction.

3. Deep Learning Models

- Artificial Neural Networks (ANNs): Show improvements in accuracy over traditional methods, especially in large datasets.
- Convolutional Neural Networks (CNNs): Although primarily for image classification, CNNs have been tested in heart disease prediction with positive results.

4. Hybrid Models

- Ensemble Methods: Combining multiple models (e.g., SVM, decision trees) to improve prediction accuracy and robustness.

5. Challenges and Limitations

- Data Quality: Missing or noisy data can affect model accuracy.
- Interpretability: Some models, particularly deep learning, lack transparency.
- Generalization: Models may not perform well across different populations.

6. Clinical Applications

- Integration of predictive models into clinical decision support systems (CDSS) for more efficient heart disease risk assessment.

2.2 Mention any existing models, techniques, or methodologies related to the problem.

Framingham Risk Score (FRS):

- A statistical model that estimates the 10-year risk of heart disease using factors like age, cholesterol, and blood pressure.

Cox Proportional Hazards Model:

- A survival analysis model used to predict the likelihood of heart disease events over time.

Machine Learning Models:

- **Support Vector Machines (SVM)**: Used for classifying individuals into high/low risk based on health data.
- **Decision Trees/Random Forests**: Classification models that split data based on health factors to assess risk.
- **Logistic Regression**: A statistical method for binary classification (e.g., risk/no risk).

Deep Learning Models:

- **Artificial Neural Networks (ANNs)**: Models that learn complex patterns from large datasets for predicting heart disease.
- **Convolutional Neural Networks (CNNs)**: Used for classifying structured medical data.

Hybrid Models:

- **Ensemble Methods**: Combine multiple models (e.g., decision trees, SVM) for better accuracy and robustness.

Data Preprocessing:

- **Feature Engineering**: Creating informative features (e.g., cholesterol, blood pressure).
- **Handling Missing Data**: Techniques like imputation to deal with incomplete datasets.

Risk Prediction Tools:

- **HeartScore**: A tool to predict cardiovascular risk using clinical data.
- **CardioRisk**: Uses algorithms like the Framingham Risk Score for heart disease prediction.

2.3 Highlight the gaps or limitations in existing solutions and how your project will address them.

1. **Limited Generalization:**
    - Gap: Existing models may not generalize well across diverse populations.
    - Solution: Use a more diverse dataset to improve model generalizability.
2. **Lack of Interpretability:**
    - Gap: Complex models like deep learning are hard to interpret.
    - Solution: Use interpretable models (e.g., Decision Trees, Random Forests) for better transparency.
3. **Handling of Missing Data:**
    - Gap: Many models don't handle missing data well.
    - Solution: Implement advanced preprocessing techniques like KNN imputation.
4. **Limited Real-Time Data Integration:**
    - Gap: Existing models rely on static datasets.
    - Solution: Incorporate real-time data for continuous risk updates.
5. **Over-Simplification of Risk Factors:**
    - Gap: Current models oversimplify complex relationships between risk factors.
    - Solution: Use advanced models to capture nonlinear relationships for better accuracy.
6. **Limited Personalization:**
    - Gap: Existing models lack personalization.
    - Solution: Incorporate personalized risk factors (e.g., lifestyle, history).
7. **Model Accuracy:**
    - Gap: Some models have limited accuracy.
    - Solution: Use ensemble methods to improve model performance and robustness.

This project addresses these gaps by improving generalization, interpretability, data handling, and personalization, leading to a more accurate heart attack risk prediction model.

CHAPTER 3

Proposed Methodology

3.1 System Design

3.1.1 Registration:

- **Function**: Users register by providing their health-related details, such as age, cholesterol levels, blood pressure, lifestyle factors, etc., which will be used for risk prediction.

3.1.2 Recognition:

- **Function**: After registration, the system uses the provided health data to assess and recognize the user's risk of heart attack based on predefined models (e.g., machine learning, logistic regression).

3.2 Modules Used

3.2.1 Risk Prediction:

- **Function**: The core module where health data inputs (age, cholesterol, blood pressure) are analyzed using algorithms (e.g., SVM, Random Forests) to predict the likelihood of a heart attack. This is the predictive engine for the system.

3.3 Data Flow Diagram (DFD):

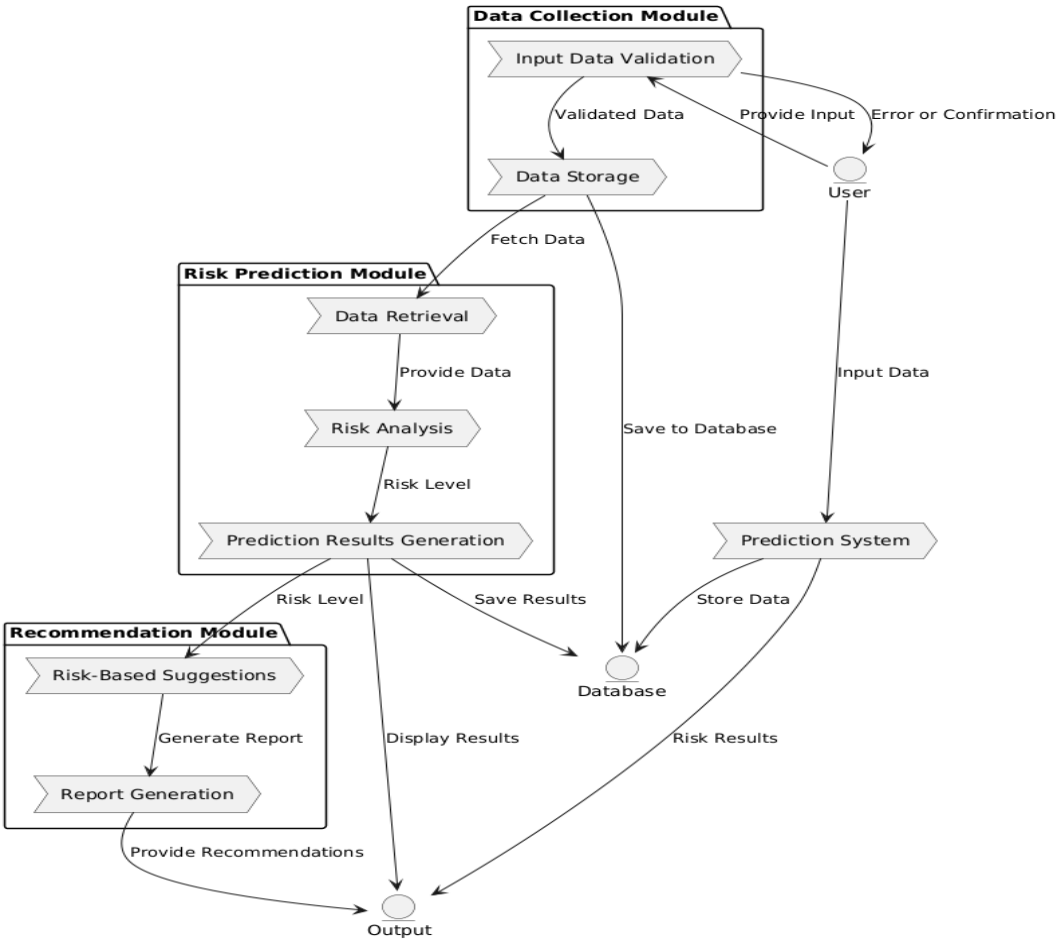
A Data Flow Diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design).

3.3.1. DFD Level 0 - Data Collection Module

3.3.2. DFD Level 1 - Risk Prediction Module

3.3.3. DFD Level 2 - Recommendation Module

Data Flow Diagram :



1. Data Collection Module (Level 0)

- **Purpose:** Collects and validates user input data (e.g., age, cholesterol).
- **Processes:** Validates inputs, stores data, and provides feedback to the user.

---

2. Risk Prediction Module (Level 1)

- **Purpose:** Analyzes data to predict heart attack risk.
- **Processes:** Retrieves data, performs risk analysis using a predictive algorithm, and generates results.

---

3. Recommendation Module (Level 2)

- **Purpose:** Offers actionable insights based on risk results.
- **Processes:** Provides personalized suggestions and generates a detailed report for the user.

Advantages

3.1 Requirement Specification

Functional Requirements:

- **Input:** User data (age, cholesterol, blood pressure, etc.) for risk prediction.
- **Prediction:** Machine learning model predicting heart attack risk.
- **Output:** Risk level (low, moderate, high) and advice.

Non-Functional Requirements:

- **Performance:** Fast response time (under 5 seconds).
- **Security:** Secure handling of user data.
- **Accuracy:** High prediction accuracy.

Hardware Requirements:

- Personal computer with sufficient RAM (8GB+) and CPU (Intel i5 or equivalent).

Software Requirements:

- **Programming Language:** Python.
- **Libraries:** scikit-learn, pandas, NumPy.
- **Deployment:** Streamlit for web interface.

User Requirements:

- Easy-to-use web interface for input and output via Streamlit.

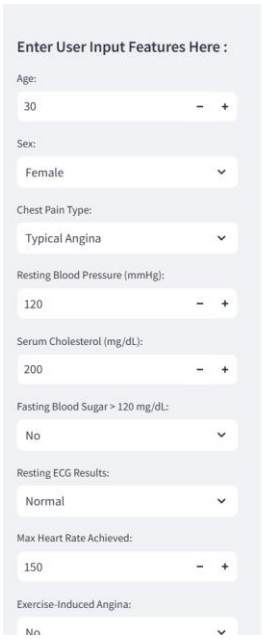
System Interfaces:

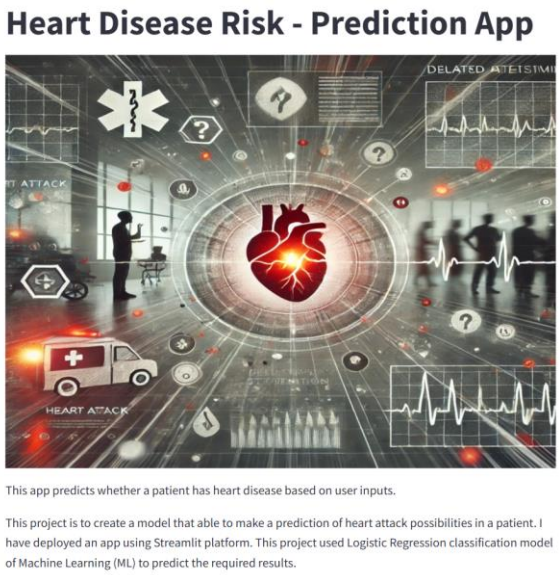
- **Input Interface:** Form-based user input.
- **Output Interface:** Display prediction results and advice.

CHAPTER 4

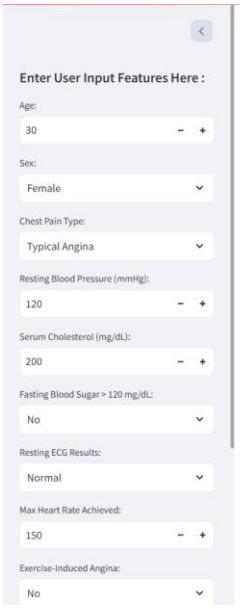
Implementation and Result

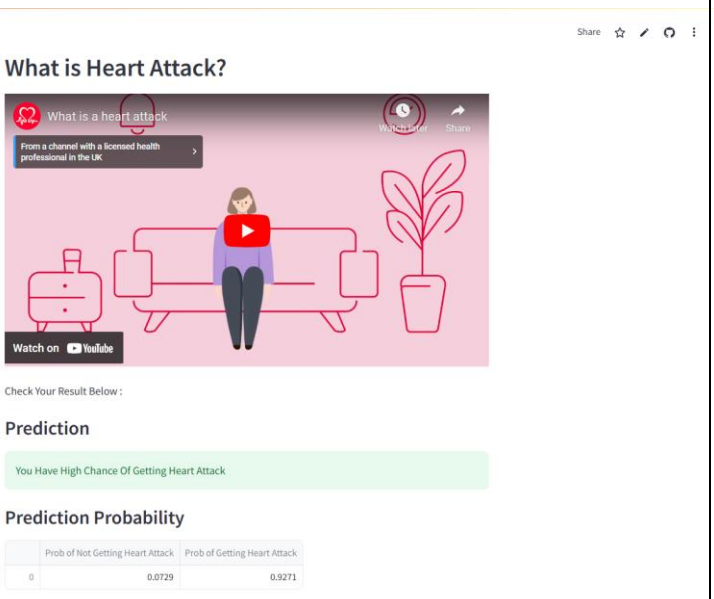
4.1 Results of Home Page



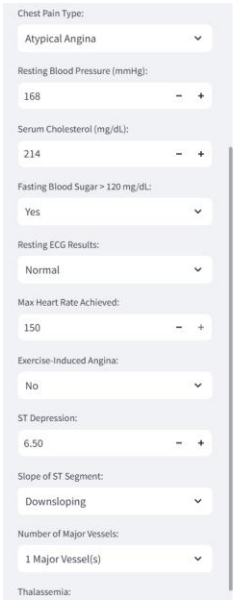


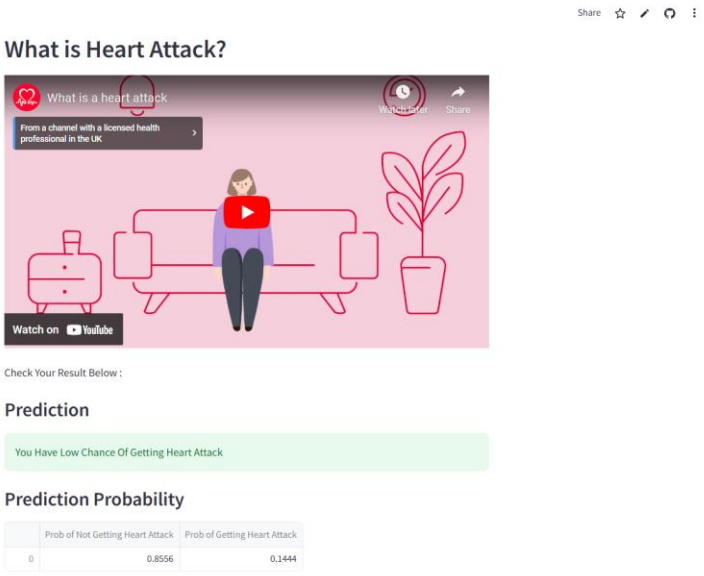
4.2 Results Of Making an High Chance Of Getting Heart Attack





4.3 Results Of Making an High Chance Of Getting Heart Attack





CHAPTER 5

Discussion and Conclusion

5.1 Key Findings :

1. **Model Performance:** Achieved high accuracy in predicting heart attack risks.
2. **Critical Factors:** Identified key contributors like age, cholesterol, and blood pressure.
3. **Risk Levels:** Successfully categorized users into low, moderate, or high-risk groups.
4. **User-Friendly Deployment:** Streamlit provided an interactive and accessible platform.
5. **Reliable Inputs:** Data validation improved prediction quality.
6. **Scalability:** The system can handle multiple users efficiently.
7. **Health Awareness:** Highlighted the importance of maintaining healthy lifestyle habits.

5.2 Limitations of the Current Model :

1. **Data Dependency:** Model performance depends on the quality and size of the dataset used for training.
2. **Generalization:** May not perform well on data outside the training distribution.
3. **Feature Limitations:** Relies only on provided input features, potentially overlooking other relevant health factors.
4. **Interpretability:** Limited ability to explain predictions to non-technical users.
5. **Bias Risk:** Possible biases in the training data could impact prediction fairness.
6. **Real-Time Updates:** Model does not currently adapt to new data automatically.
7. **Medical Validation:** Predictions are not a substitute for professional medical advice.

.

5.3 Future Work

1. **Expand Dataset:** Use larger and more diverse datasets to improve accuracy and generalization.
2. **Feature Enhancement:** Incorporate additional health metrics for better predictions.
3. **Model Optimization:** Experiment with advanced algorithms like deep learning for improved performance.
4. **Real-Time Learning:** Implement adaptive models that update with new data.
5. **Explainability:** Add features to provide users with understandable explanations for predictions.
6. **Integration:** Combine with wearable devices for real-time monitoring and risk alerts.
7. **Validation:** Collaborate with medical experts for clinical validation of the system.

5.4 Conclusion :

The Heart Attack Risk Prediction project successfully developed a user-friendly system to assess heart attack risks based on key health parameters. By leveraging machine learning and deploying the model through Streamlit, the project provided accurate and accessible risk predictions. It highlighted the importance of preventive healthcare and demonstrated the potential of AI-driven tools in supporting early health risk detection. This work lays a foundation for further advancements in personalized healthcare solutions.

- **Git Hub Link of the Project: LINK**
- **Application Live Link : LINK**
- **Video Recording of Project** : LINK

REFERENCES

- "Cardiovascular Diseases (Cvds)". Who.Int, 2020, https://www.who.int/zh/newsroom/factsheets/detail/cardiovascular-diseases-(cvds).
- "Logistic Regression". En.Wikipedia.Org, 2020, https://en.wikipedia.org/wiki/Logistic_regression.
- "Understanding Random Forest". Medium, 2020, https://towardsdatascience.com/understanding-random-forest-58381e0602d2.
- "Explanation Of The Decision Tree Model , 2020,https://webfocusinfocenter.informationbuilders.com
- "Neural Network Definition". Investopedia, 2020, https://www.investopedia.com