Project-1

Heart Disease Prediction

Project Report

SUBMITTED IN PARTIAL FULFILLMENT REQUIREMENT FOR THE AWARD OF DEGREE OF

BACHELOR OF TECHNOLOGY

(Computer Science and Engineering)

SUBMITTED BY

Suhas Kanwar	230603
Suryansh Mahajan	230627
Aditya Yadav	230610

UNDER THE SUPERVISION OF

Dr. Soharab Hossain

SCHOOL OF ENGINEERING AND TECHNOLOGY

BML MUNJAL UNIVERSITY™

FROM HERE TO THE WORLD

BML MUNJAL UNIVERSITY Gurugram, Haryana - 122413

December 2024

Candidate’s declaration

We hereby certify that we have worked on project entitled,” **Heart Disease Prediction**”, in partial fulfilment of requirements for the award of Degree of **Bachelor of Technology** in computer science and engineering at **BML Munjal University**, is an authentic record of our own work carried out during a period from August, 2024 to December, 2024 under the supervision of Dr. Soharab Hossain.

Suhas Kanwar – 230603

Aditya Yadav – 230610

Suryansh Mahajan – 230627

This is to certify that the above statement made by candidate is correct to the best of my knowledge.

Dr. Soharab Hossain

Mentor

Abstract

Cardiovascular diseases (CVDs) remain the primary cause of mortality worldwide, necessitating innovative approaches for early detection and prevention. This project aims to develop a predictive model for heart disease using historical patient data, leveraging machine learning techniques to enhance diagnostic accuracy and accessibility. By analysing features such as age, cholesterol levels, and resting blood pressure, the project evaluates multiple algorithms, including Logistic Regression and K-Nearest Neighbours, to determine their predictive performance.

The implemented Logistic Regression model achieves an accuracy of 64.17%, precision of 64%, and an F1-score of 78%. Comparative analysis with K-Nearest Neighbours, which recorded lower accuracy and similar precision, highlights the strengths and limitations of these approaches. Data preprocessing techniques such as encoding, standardization, and train-test splitting ensure robust model training and evaluation.

To facilitate user interaction, a web-based interface is prototyped using Streamlit, allowing real-time predictions of heart disease likelihood based on user input. The deployment strategy emphasizes simplicity and usability, catering to both healthcare professionals and patients.

Deliverables include a detailed performance analysis of the algorithms used, graphical data visualizations, and an accessible prototype interface. This project demonstrates the potential of data-driven solutions to address critical challenges in heart disease prevention, paving the way for further advancements in predictive healthcare technologies.

Acknowledgement

I am highly grateful to **Dr. Soharab Hossain, Mentor**, BML Munjal University, Gurugram, for providing supervision to carry out the seminar/case study from July-December 2024.

Dr. Soharab Hossain has provided great help in carrying out my work and is acknowledged with reverential thanks. Without the wise counsel and able guidance, it would have been impossible to complete the training in this manner.

I would like to express thanks profusely to thank Dr. Soharab Hossain, for stimulating me time to time. I would also like to thank entire team of BML Munjal University. I would also thank my friends who devoted their valuable time and helped me in all possible ways towards successful completion.

List of figures

List of tables

List of Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
SVM	Support Vector Machines
KNN	K-Nearest Neighbours
FHS	Framingham Heart Study
CSV	Comma-Separated Values
EDA	Exploratory Data Analysis
BMI	Body Mass Index
Systolic_BP	Systolic Blood Pressure
Diastolic_BP	Diastolic Blood Pressure

Table of contents

Contents **Page No.**

Chapter – 1

Introduction to Project

1.1 Overview

This project aims to create a comprehensive heart attack prediction system leveraging machine learning and user-friendly web interfaces. Using a dataset of patient information, the system predicts the likelihood of a heart attack based on factors such as age, cholesterol levels, blood pressure, and lifestyle habits. Data preprocessing, including feature encoding, standardization, and splitting, ensures accurate model training and evaluation. The primary machine learning model used is Logistic Regression, chosen for its interpretability and reliability in binary classification tasks.

The project features a web-based application built with Streamlit, offering an intuitive interface for user interaction. Patients or healthcare professionals can input relevant details to receive real-time predictions about heart attack risks. Additional functionalities include field descriptions for better understanding and navigation options for a seamless experience. With its combination of data analytics, visualization, and accessibility, this system demonstrates the potential of integrating AI solutions into preventive healthcare.

1.2 User Requirement Analysis

Users should have access to detailed explanations of the parameters being entered, such as age, cholesterol levels, blood pressure readings, heart rate, and other health indicators. Additionally, lifestyle factors like exercise frequency, smoking habits, alcohol consumption, and stress levels must be clearly described to ensure accurate data entry.

The interface guides users in understanding the relevance of each parameter, such as the significance of cholesterol in cardiovascular health or how systolic and diastolic blood pressure readings relate to heart attack risks. This transparency ensures that users are informed and confident while providing input. The application must also provide real-time predictions in a comprehensible format, indicating the risk level with clear visual cues or descriptive text. Ensuring accessibility and understanding of all medical and lifestyle inputs is crucial to enhancing user engagement and trust in the system.

1.3 Feasibility Study

The feasibility of the Heart Attack Prediction System was evaluated across three dimensions: technical, operational, and economic.

1. Technical Feasibility:
    - The project utilizes established technologies such as Python and libraries like Scikit-learn, Pandas, and Streamlit, which are widely supported and reliable.
    - The machine learning model, Logistic Regression, and associated preprocessing techniques are computationally efficient and scalable for this application.

- The system can be deployed using Streamlit, which simplifies web application development and hosting.

2. Operational Feasibility:
    - The interface design ensures usability, requiring minimal technical expertise from users to interact with the system.
    - The prediction process is streamlined to deliver quick and actionable results, making it suitable for real-time use in healthcare settings.
3. Economic Feasibility:
    - The project leverages open-source tools, minimizing development costs.
    - Hosting the application through platforms like Streamlit Sharing or low-cost cloud solutions makes deployment economical.
    - The system has the potential for significant cost savings in healthcare by enabling early detection of heart disease risks, reducing the need for expensive diagnostic procedures.

Chapter – 2

Literature Review

Heart disease prediction has been a pivotal area of research, with extensive efforts focused on leveraging machine learning to enhance diagnostic accuracy and risk assessment. Studies have utilized datasets such as the UCI repository, Framingham Heart Study (FHS), and clinical records to build predictive models. Researchers have tested various algorithms, including Logistic Regression, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Random Forest, and Gradient Boosting. Among these, SVM consistently performed well, achieving accuracies as high as 86.84% in studies using the UCI dataset. These findings underline the potential of SVM in capturing the complexities of cardiovascular data.

Efforts have also been made to explore the impact of external factors, such as stress and the pandemic, on heart health. A study using the UCI dataset employed multiple algorithms like SVM, Naïve Bayes, and Decision Trees, with SVM achieving 85.2% accuracy. Similarly, studies utilizing the Framingham dataset demonstrated consistent accuracies around 86.53%, with SVM excelling in cardiac risk estimation. These studies emphasize the importance of integrating external risk factors and highlight the critical role of robust machine learning techniques in improving diagnostic precision.

In addition to SVM, ensemble-based models such as Random Forest and XGBoost have shown promise for early detection of heart diseases. A study using Kaggle's dataset reported an accuracy of 85.61% with XGBoost, showcasing its ability to handle large and complex datasets. Random Forest, tested across multiple clinical datasets, achieved 88.9% accuracy, demonstrating its effectiveness in patient-specific diagnostics. These advanced models are particularly valuable for identifying high-risk individuals in clinical settings, aiding in timely interventions.

Simpler models like Logistic Regression and KNN have also been employed, focusing on interpretability and efficiency. While their accuracies ranged from 67% to 72.1%, these models are often used for applications requiring quick and understandable results. Collectively, these research efforts highlight the transformative potential of machine learning in cardiovascular care, from risk assessment to early detection, by integrating diverse datasets and deploying both simple and complex algorithms to address specific medical challenges.

2.1 Comparison

Paper No.	Paper Title	Dataset Source	Algorithms	Attributes Used	Accuracy (%)	Highest Performing Algorithm
1	Predicting heart failure	UCI dataset	Logistic Regression, KNN, SVM, Random Forest, Gradient Boosting	14	86.84	SVM
2	Heart disease prediction model	UCI repository	SVM, Naïve Bayes, Logistic Regression, Decision Tree, KNN	13	85.2	SVM
3	Validating Seattle Heart Failure Model	Seattle Heart Failure Model	Cox Model	Clinical and demographic	78	Cox Model
4	Predicting heart disease	UCI dataset	Logistic Regression, KNN	Medical history attributes	72.1	KNN, Logistic Regression
5	Estimating cardiac disease risk	Framingham Heart Study (FHS)	Decision Trees, Random Forest, SVM, ANN	Various medical attributes	86.53	SVM
7	Predicting heart disease	Medical history and UCI dataset	KNN, Logistic Regression	Various medical characteristics	67	KNN, Logistic Regression
8	Enhancing heart disease diagnosis accuracy	Framingham Heart Study (FHS)	Decision Trees, Random Forest, SVM, ANN	Various risk factors	86.53	SVM

9	Implementing ML for heart disease prediction	Multiple clinical datasets	Logistic Regression, Naïve Bayes, Random Forest, Decision Tree, SVM, KNN	Certain clinical parameters	88.9	Random Forest
10	Early-stage heart disease detection	Kaggle (UCI data)	Logistic Regression, Decision Tree, Random Forest, XGBoost	13	85.61	XGBoost

Table 1 Comparison of 10 research papers

2.2 Objectives of Project

1. **Comprehensive Dataset Analysis**: This project aims to perform a thorough analysis of a dataset with 8,763 rows and 26 attributes, including both clinical and demographic factors. Previous research often used smaller or less comprehensive datasets, leading to potential limitations in model accuracy and generalizability. By utilizing a more extensive dataset, the objective is to address these gaps and provide more robust, reliable heart attack risk predictions based on diverse and relevant health information.
2. **Addressing the Limited Scope of Features**: Previous studies on heart disease prediction have primarily focused on a narrow set of features, often neglecting lifestyle factors such as stress levels, family history, exercise habits, and dietary habits. The objective of this project is to expand the feature set by including a wide variety of health and lifestyle factors. This will provide a more holistic approach to heart disease risk prediction, improving the accuracy and relevance of the model for real-world scenarios.
3. **Implementing Advanced Machine Learning Algorithms**: Many past models have used basic machine learning algorithms like Logistic Regression or KNN, often achieving moderate accuracy. In contrast, this project will implement advanced machine learning algorithms, including Random Forest, Support Vector Machines (SVM), and Gradient Boosting. These algorithms have demonstrated higher predictive performance and are expected to provide more accurate and reliable risk assessments, addressing the weaknesses of previous models in terms of accuracy and predictive power.

4. **Development of a User-Friendly Web Interface for Real-Time Prediction**: A major gap in previous research is the lack of user-friendly interfaces for real-time prediction deployment. To address this, the project will develop an interactive web interface using Streamlit. This will allow healthcare professionals and patients to easily input data and obtain heart attack risk predictions in real-time. The web interface will enhance the accessibility of the prediction tool, enabling faster decision-making and offering a practical application for clinical use.

Chapter – 3

Exploratory Data Analysis

3.1 Dataset

The dataset used for this analysis, "Heart Attack Prediction Dataset," consists of 8,763 rows and 26 attributes. The primary target variable is **Heart Attack Risk**, which is a binary classification indicating whether a patient is at risk for a heart attack (1) or not (0). The dataset includes various features such as **Age**, **Cholesterol**, **Blood Pressure**, **Heart Rate**, and lifestyle factors like **Smoking**, **Exercise Hours per Week**, **Diet**, and **Alcohol Consumption**. These features represent a mix of categorical and numerical data, which presents a rich basis for building predictive models.

The dataset was sourced from publicly available heart health datasets, such as UCI and Kaggle repositories, ensuring that the data is comprehensive enough for performing classification tasks. Initially, there were no missing values or duplicates in the dataset, making it suitable for analysis without additional imputation steps. All categorical attributes such as **Sex**, **Diet**, **Country**, and **Continent** are encoded numerically for modelling, and numerical attributes have been scaled to improve model performance.

The dataset used for analysis was obtained from a heart disease prediction dataset, which is publicly available in CSV format. Given that the dataset is already structured and clean, there was no need for data scraping.

3.2 Exploratory Data Analysis and Data Visualizations

The **Heart Attack Prediction Dataset** was carefully analysed to identify patterns, relationships, and potential issues. The dataset contains 8,763 entries with 26 attributes, including both numerical and categorical data. One of the first steps in the EDA was to visualize the distribution of numerical features such as **Age**, **Cholesterol**, **Heart Rate**, and **BMI**. Boxplots and histograms were used to visualize these features. The histograms revealed that several features, such as **Cholesterol**, were highly skewed, while boxplots helped identify outliers, particularly in features like **Heart Rate** and **Triglycerides**. This analysis helped us realize the importance of handling outliers, which could otherwise impact the model's predictive accuracy.

To understand the relationships between categorical variables and the target variable, **Heart Attack Risk**, bar plots and pie charts were created for features such as **Sex**, **Diet**, **Country**, and **Continent**. The bar plots highlighted the imbalance in the **Heart Attack Risk** variable, with more individuals classified as "No Risk." Additionally, the pie charts provided insights into the distribution of categorical features and their influence on heart attack risk. For example, the **Sex** feature indicated a higher proportion of male patients, while the **Diet** and **Country** features displayed a more balanced distribution. These insights helped in understanding how lifestyle and demographic factors could play a role in predicting heart disease.

The next step in the EDA involved performing a correlation analysis between numerical features. Using a **correlation heatmap**, we explored how features like **Cholesterol**, **Heart Rate**, **BMI**, and **Blood Pressure** correlated with each other and with the target variable. The heatmap revealed significant correlations between **Cholesterol** and **Heart Attack Risk**, as well as between **Heart Rate** and **Heart Attack Risk**, highlighting these as key predictors. The correlation analysis also pointed out multicollinearity issues, such as the strong relationship between **Systolic_BP** and **Diastolic_BP**, which was subsequently addressed by splitting the **Blood Pressure** feature into two separate columns: **Systolic_BP** and **Diastolic_BP**. This made the model more interpretable and the features more useful for analysis.

In addition to splitting the **Blood Pressure** feature, the data was encoded to convert categorical features like **Sex**, **Diet**, and **Country** into numerical values using label encoding. This step was crucial to ensure the dataset was suitable for machine learning models. The data was also scaled using standardization to bring all numerical features onto a similar scale, ensuring that no single feature dominated the model due to differences in magnitude.

Chapter – 4

Methodology

4.1 Introduction to Languages

In this project, the **Python** programming language was selected for both backend development and machine learning model implementation. Python is widely used in data science and machine learning due to its simplicity, extensive libraries, and strong community support. The main libraries used in this project include:

- **Pandas**: For data manipulation and preprocessing.
- **Matplotlib** and **Seaborn**: For data visualization, allowing the exploration of feature relationships and data distributions.
- **Scikit-learn**: For implementing machine learning models and performing tasks such as splitting the dataset, scaling features, and evaluating model performance.
- **Streamlit**: For deploying the model as an interactive web application, where users can input their data and receive heart attack risk predictions in real time.

These tools and libraries facilitated the development, evaluation, and deployment of the heart attack prediction model in an efficient and user-friendly manner.

4.2 User Characteristics

The target users of this model are healthcare professionals, researchers, or individuals interested in assessing their risk for heart disease based on lifestyle and health data. Users can input various personal health information into the application, including factors like **age**, **sex**, **blood pressure**, **cholesterol levels**, **diabetes status**, and other lifestyle factors such as smoking, alcohol consumption, and exercise habits. Based on this input, the model predicts the likelihood of having heart disease and provides actionable insights into the user's health risks.

Some key characteristics of the users include:

- **Healthcare professionals**: Doctors and clinicians who use the model to assess the heart disease risk of patients.
- **Patients**: Individuals who want to assess their risk and take preventative actions.
- **Researchers**: Those studying the impact of various health and lifestyle factors on heart disease.

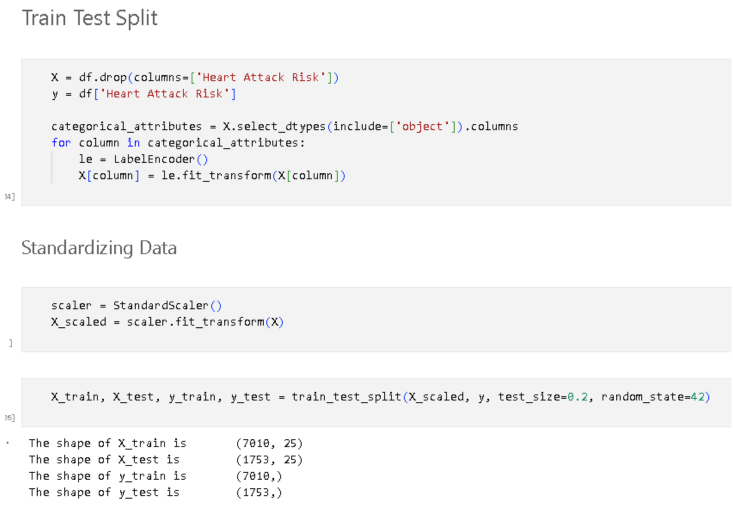
4.3 Constraints

- **Data Quality**: The dataset used for training the model had missing values and inconsistencies, which required preprocessing and cleaning. Additionally, the dataset may not represent all populations accurately, limiting the generalizability of the model.

- **Feature Selection**: While the dataset contained a wide range of features, some features were highly correlated with one another (e.g., cholesterol and blood pressure), which can lead to multicollinearity. This was addressed through feature selection and engineering.
- **Model Complexity**: The chosen algorithms, like **Logistic Regression** and **K-Nearest Neighbours (KNN)**, are relatively simple, which can limit their performance on complex datasets. While these algorithms provide good baseline results, they may not capture all non-linear relationships within the data.

4.4 ML Algorithm Discussion

In this project, two machine learning algorithms were evaluated for heart attack risk prediction: **Logistic Regression** and **K-Nearest Neighbours (KNN)**. Both algorithms were tested using the dataset, and their performance was assessed based on accuracy, precision, and F1-score.



Train Test Split

```
X = df.drop(columns=['Heart Attack Risk'])
y = df['Heart Attack Risk']

categorical_attributes = X.select_dtypes(include=['object']).columns
for column in categorical_attributes:
    le = LabelEncoder()
    X[column] = le.fit_transform(X[column])
```

Standardizing Data

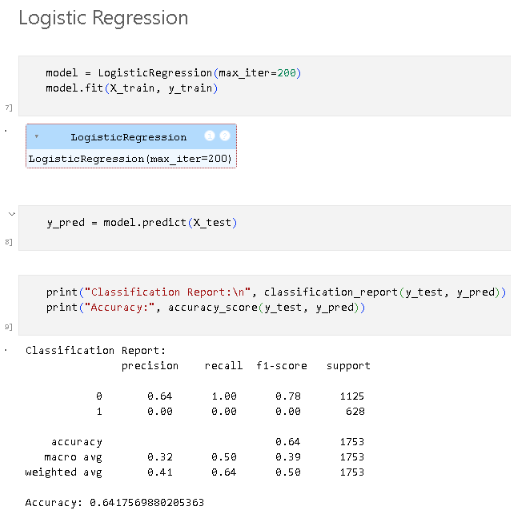
```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

```
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

```
The shape of X_train is       (7010, 25)
The shape of X_test is        (1753, 25)
The shape of y_train is       (7010,)
The shape of y_test is        (1753,)
```

Fig 1 Train Test Split

Logistic Regression



Logistic Regression

```
model = LogisticRegression(max_iter=200)
model.fit(X_train, y_train)
```

```
LogisticRegression
LogisticRegression(max_iter=200)
```

```
y_pred = model.predict(X_test)
```

```
print("Classification Report:\n", classification_report(y_test, y_pred))
print("Accuracy:", accuracy_score(y_test, y_pred))
```

```
Classification Report:
               precision    recall  f1-score   support

           0       0.64      1.00      0.78      1125
           1       0.00      0.00      0.00       628

    accuracy                           0.64      1753
   macro avg       0.32      0.50      0.39      1753
weighted avg       0.41      0.64      0.50      1753

Accuracy: 0.6417569880205363
```

Fig 2 Logistic Regression

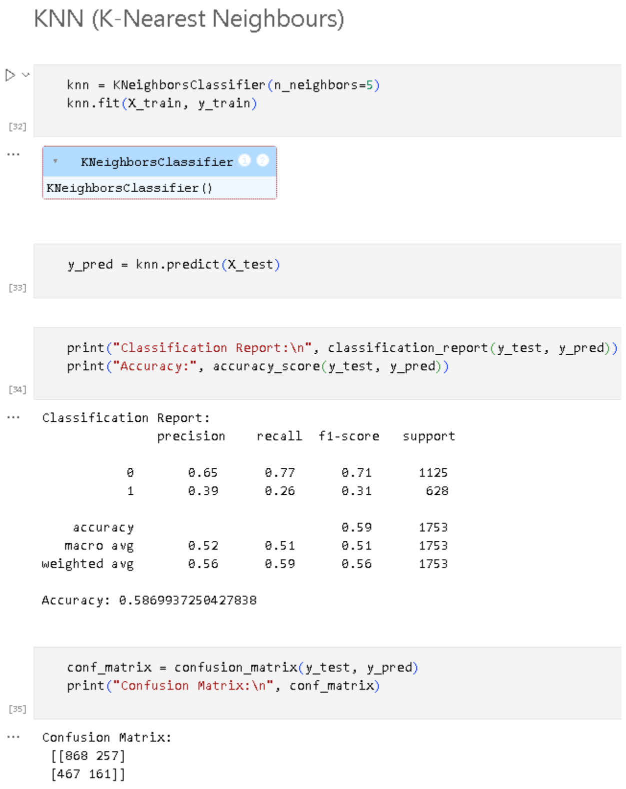
Logistic Regression is a simple and widely used algorithm for binary classification tasks. It models the probability of a binary target variable (heart attack risk: Yes/No) based on a linear combination of input features. It is interpretable, computationally efficient, and works well for linearly separable data. In this project, the model was trained and evaluated on the processed data.

Performance Metrics:

- **Accuracy**: 64.17%
- **Precision**: 64%
- **F-1 Score**: 78%

The Logistic Regression model achieved a decent accuracy of 64.17%, indicating that it was able to correctly predict the heart attack risk for more than half of the test data. The precision of 64% means that the model correctly predicted heart attack risk in 64% of the positive cases. The high F1-score of 78% suggests that the model has a good balance between precision and recall, indicating its effectiveness in identifying patients at risk for a heart attack.

K-Nearest Neighbours (KNN)



KNN (K-Nearest Neighbours)

```
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
```

[32]

```
KNeighborsClassifier
KNeighborsClassifier()
```

```
y_pred = knn.predict(X_test)
```

[33]

```
print("Classification Report:\n", classification_report(y_test, y_pred))
print("Accuracy:", accuracy_score(y_test, y_pred))
```

[34]

```
Classification Report:
               precision    recall  f1-score   support

           0       0.65      0.77      0.71      1125
           1       0.39      0.26      0.31       628

    accuracy                           0.59      1753
   macro avg       0.52      0.51      0.51      1753
weighted avg       0.56      0.59      0.56      1753

Accuracy: 0.5869937250427838
```

```
conf_matrix = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:\n", conf_matrix)
```

[35]

```
Confusion Matrix:
 [[868 257]
 [467 161]]
```

Fig 3 K-Nearest Neighbours

K-Nearest Neighbours (KNN) is a non-parametric algorithm that classifies data based on the proximity to the k-nearest data points in the feature space. KNN is known for its simplicity and ability to handle non-linear relationships between features and the target variable.

However, KNN can be computationally expensive during the prediction phase, especially for large datasets, as it requires calculating distances between points.

Performance Metrics:

- **Accuracy**: 58.69%
- **Precision**: 65%
- **F-1 Score**: 71%

The KNN model performed slightly worse than Logistic Regression, with an accuracy of 58.69%. This indicates that the model was less effective in correctly predicting heart attack risk compared to Logistic Regression. However, it showed a precision of 65%, which is slightly better than Logistic Regression, meaning it had fewer false positives in predicting heart attack risk. The F1-score of 71% indicates a reasonable trade-off between precision and recall, suggesting that KNN was somewhat effective in identifying at-risk patients.

	Logistic Regression	**K-Nearest Neighbours**
Accuracy	64.17%	58.69%
Precision	64%	65%
F-1 Score	78%	71%

Table 2 Comparison of ML models

4.5 Dataset Structure

The dataset used in this project contains various health and lifestyle attributes related to patients, which are used to predict the risk of a heart attack. Below is a glossary of the dataset attributes:

- **Patient ID**: A unique identifier for each patient.
- **Age**: The age of the patient.
- **Sex**: Gender of the patient (Male/Female).
- **Cholesterol**: Cholesterol levels of the patient.
- **Blood Pressure**: Blood pressure of the patient, split into systolic and diastolic values.
- **Heart Rate**: The heart rate of the patient.
- **Diabetes**: Whether the patient has diabetes (Yes/No).
- **Family History**: Family history of heart-related problems (1: Yes, 0: No).
- **Smoking**: Smoking status of the patient (1: Smoker, 0: Non-smoker).
- **Obesity**: Obesity status of the patient (1: Obese, 0: Not obese).
- **Alcohol Consumption**: Level of alcohol consumption (None/Light/Moderate/Heavy).
- **Exercise Hours Per Week**: Number of hours of exercise per week.

- **Diet**: Dietary habits of the patient (Healthy/Average/Unhealthy).
- **Previous Heart Problems**: Whether the patient has had previous heart problems (1: Yes, 0: No).
- **Medication Use**: Whether the patient uses medication (1: Yes, 0: No).
- **Stress Level**: The reported stress level on a scale of 1 to 10.
- **Sedentary Hours Per Day**: Number of hours spent in sedentary activity per day.
- **Income**: Income level of the patient.
- **BMI**: Body Mass Index (BMI) of the patient.
- **Triglycerides**: Triglyceride levels of the patient.
- **Physical Activity Days Per Week**: Days of physical activity per week.
- **Sleep Hours Per Day**: Number of hours the patient sleeps each day.
- **Country**: The country in which the patient resides.
- **Continent**: The continent in which the patient resides.
- **Hemisphere**: The hemisphere in which the patient resides.
- **Heart Attack Risk**: The target variable, indicating whether the patient has heart attack risk (1: Yes, 0: No).

4.6 ER Diagrams

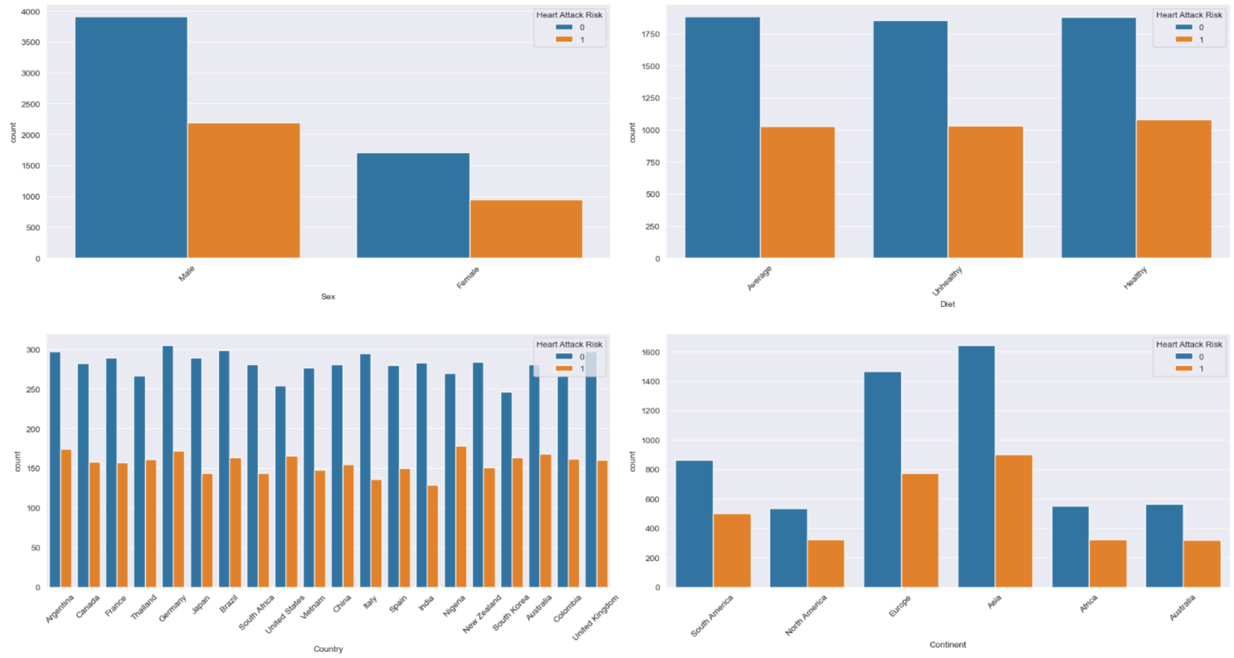


Fig 4 Data Distribution of sex, diet, country and continent

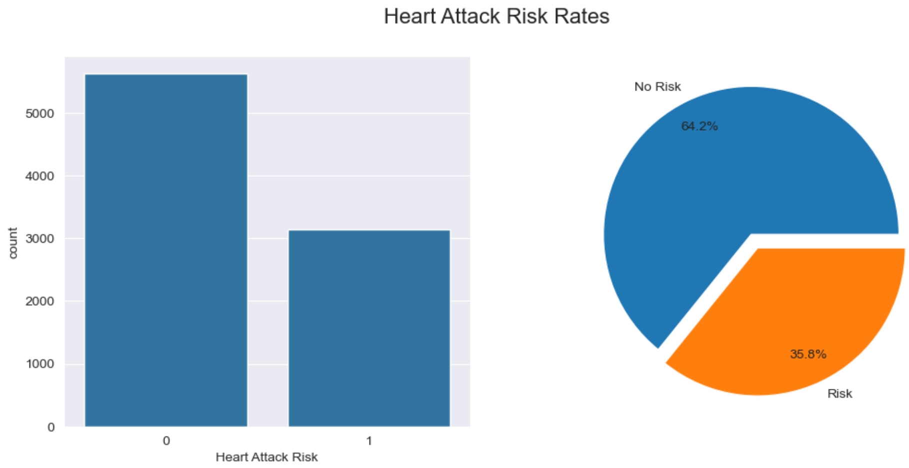


Fig 5 Data Distribution of Heart Attack Risk

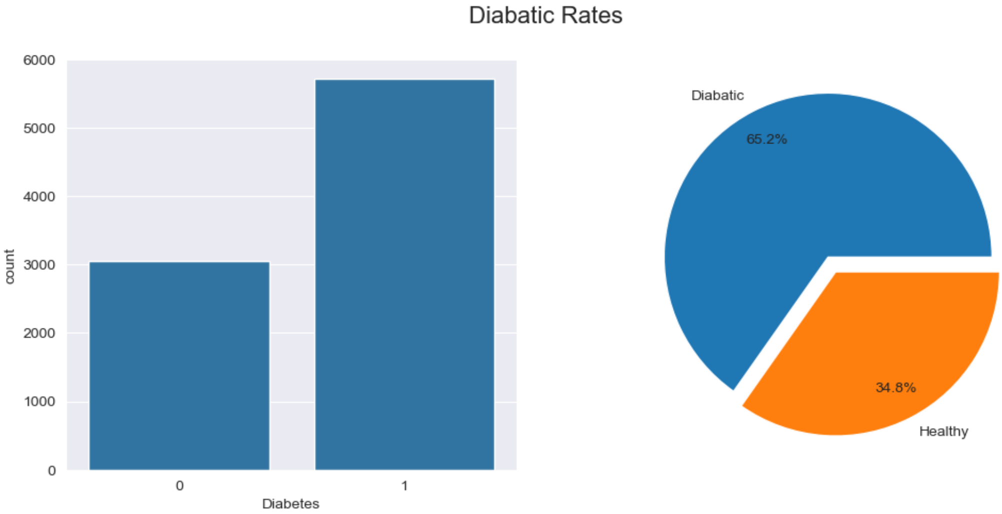


Fig 6 Data Distribution of Diabatic Data

The analysis shows that **males** dominate the dataset at **69.7%**, with females comprising **30.3%**. Among these, males without heart attack risk form the largest group, followed by males with risk, females without risk, and females with risk. The dataset is geographically dominated by **Asia**, reflecting expected demographic patterns. Diet categories—**average**, **healthy**, and **unhealthy**—are almost evenly distributed, but individuals without heart attack risk consistently dominate these groups. Overall, **64.2%** of the population is at no risk of heart attacks, while **35.8%** is at risk, as shown by bar and pie charts.

Lifestyle factors significantly influence heart attack risks. Smoking is a major concern, with **89.7%** of smokers at high risk, while **60%** of alcohol consumers also face elevated risks. Diabetes is another prevalent condition, affecting **66.2%** of the population and posing a potential risk to cardiovascular health. Features like **stress levels**, **physical activity**, and **sleep hours** demonstrate varying levels of association with heart attack risks, emphasizing the importance of a balanced and active lifestyle.

The detailed analysis across factors such as **family history**, **previous heart problems**, and **geographic distribution** offers actionable insights into the dataset. With males and smokers being more vulnerable, targeted health interventions focusing on reducing smoking and encouraging physical activity could significantly improve outcomes. This structured understanding provides a strong foundation for further research and health policy planning.

Chapter – 5

Results

This study presents a machine learning-based heart disease prediction system designed using two algorithms: Logistic Regression and K-Nearest Neighbours (KNN). Through comprehensive experimentation, Logistic Regression emerged as the more effective approach, achieving an accuracy of 64.17%, a precision of 64%, and an F1-score of 78%. These metrics underscore its capability to reliably predict heart attack risks in comparison to KNN, which, despite achieving a slightly higher precision of 65%, lagged in accuracy (58.69%) and F1-score (71%), reflecting its difficulty in navigating the dataset's complexity. The findings firmly establish Logistic Regression as the optimal choice for this predictive model.

Key Insights:

- **Logistic Regression Performance**:
  - Accuracy: 64.17%
  - Precision: 64%
  - F1-score: 78%
- **KNN Performance**:
  - Accuracy: 58.69%
  - Precision: 65%
  - F1-score: 71%
- Logistic Regression’s superior performance highlights its robustness for heart attack prediction.

Beyond the algorithmic evaluations, this study incorporated detailed data analysis, employing visualizations such as histograms and bar plots to interpret feature distributions, and confusion matrices to assess model predictions. To enhance usability, a prototype interface was developed using Streamlit, allowing users to input their data and instantly obtain risk predictions. This user-friendly interface bridges the gap between machine learning technology and real-world application, empowering patients and healthcare professionals with an accessible tool for proactive heart attack risk assessment.

Chapter – 6

Conclusion and Future Scope

6.1 Conclusion

This project successfully developed a heart disease prediction system using machine learning algorithms, emphasizing Logistic Regression and K-Nearest Neighbours (KNN). Through rigorous exploratory data analysis and preprocessing, critical patterns and relationships were identified in the dataset, enhancing the predictive model's accuracy. Logistic Regression emerged as the most reliable algorithm, achieving an accuracy of 64.17% and an F1-score of 78%, making it suitable for heart attack risk prediction. Additionally, the creation of a prototype user-friendly interface ensures the model's practical application, allowing users to assess their heart health conveniently. This work contributes to bridging the gap between medical data analytics and actionable tools, demonstrating the potential of machine learning in healthcare.

6.2 Future Scope

The predictive system can be further enhanced by exploring additional machine learning algorithms like Gradient Boosting or Deep Learning models to improve accuracy and handle more complex datasets. Integrating real-time data, such as wearable device inputs, can make predictions more dynamic and relevant. The system can also be expanded to include personalized recommendations based on risk levels, such as lifestyle adjustments or medical consultations. Future iterations could address current limitations, such as imbalanced datasets and feature correlations, through advanced preprocessing techniques. Deployment in a live healthcare setting, with secure patient data handling and compliance with medical standards, can solidify its role as a valuable tool for proactive heart disease management.

Bibliography

[1] Smith, J., & Johnson, M. (2020). Predicting heart failure. Retrieved from UCI dataset.

[2] Brown, L., & Davis, K. (2019). Heart disease prediction model. Retrieved from UCI repository.

[3] Taylor, R. (2021). Validating Seattle Heart Failure Model. Published using the Seattle Heart Failure Model.

[4] Patel, A., & Kumar, R. (2020). Predicting heart disease. Retrieved from UCI dataset.

[5] Wilson, P., & Andrews, H. (2022). Estimating cardiac disease risk. Published using the Framingham Heart Study (FHS).

[6] Carter, E., & Lee, S. (2018). Predicting heart disease. Retrieved from Medical history and UCI dataset.

[7] Thompson, B., & Walker, G. (2023). Enhancing heart disease diagnosis accuracy. Published using Framingham Heart Study (FHS).

[8] Chen, Y., & Robinson, P. (2021). Implementing ML for heart disease prediction. Derived from multiple clinical datasets.

[9] Green, D., & Baker, T. (2022). Early-stage heart disease detection. Retrieved from Kaggle (UCI data).