BIRLA OF TECHNOLOGY & SCIENCE, PILANISECOND SEMESTER 2022-23

Predicting risk of heart attack, using Automated Machine Learning Techniques

DISSERTATION

Submitted in partial fulfilment of the requirements of the
MTech Data Science and Engineering Degree programme

By
Akanksha Chauhan
2021FC04343

Under the supervision of
Madhvendra Thakur
(CEO)

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan)
INDIA
September 2023

Acknowledgments

I would like to express my heartfelt gratitude to those who have played a pivotal role in the successful completion of this project, "Heart Attack Risk Predictor." This project has been a culmination of efforts, dedication, and guidance from the following individuals:

I extend my sincerest thanks to Mr. Madhvendra Thakur, my mentor, for providing valuable insights, guidance, and continuous support throughout this project. Your expertise and encouragement have been instrumental in shaping this project's success.

I am deeply grateful to Mr. Vinaya Sathyanarayana, our esteemed faculty, for her unwavering support, encouragement, and academic guidance. Your dedication to fostering our learning and research endeavour's has been truly invaluable.

This project would not have been possible without the mentorship and guidance of Mr. Madhvendra Thakur and the support of Mr. Vinaya Sathyanarayana. Their contributions have been instrumental in achieving our research objectives and have enriched my learning experience.

I would also like to acknowledge the contributions of the broader academic community and the resources that have facilitated our research and learning.

Thank you all for your invaluable support and guidance.

Akanksha Chauhan

BIRLA OF TECHNOLOGY & SCIENCE, PILANISECOND
SEMESTER 2022-23

Certificate

This is to certify that the Dissertation entitled **Predicting risk of heart attack, using Automated Machine Learning** Techniques and submitted by **Akanksha Chauhan** ID NO. **2021FC04343** in partial fulfilment of the requirements of DSECLZG628T Dissertation, embodies the work done by him/her under my supervision.

Madhvendra Thakur (CEO)

Signature of the Supervisor **Name & Designation**

Place: Ghaziabad

Date: 10-10-2023

BIRLA OF TECHNOLOGY & SCIENCE, PILANISECOND SEMESTER 2022-23

Dissertation Title : Predicting risk of heart attack, using Automated Machine Learning

Name of Supervisor : Madhvendra Thakur

Name of Student : Akanksha Chauhan

ID No. of Student : 2021FC04343

Abstract

Cardiovascular diseases, particularly heart attacks, constitute a significant global health concern, demanding accurate and timely risk assessment for effective prevention and intervention strategies. This dissertation presents a comprehensive study on the application of Automated Machine Learning techniques to predict the risk of heart attacks.

The research leverages a diverse dataset comprising demographic information, medical history, and physiological parameters of a large cohort of patients. Through meticulous data preprocessing and feature engineering, a robust foundation for modelling is established. Various Automated Machine Learning frameworks, including but not limited to, Auto-Sklearn, TPOT, and H2O.ai, are employed to identify the optimal machine learning algorithms, hyper-parameters, and feature subsets.

A comparative evaluation of these techniques demonstrates their effectiveness in handling the complexity of cardiovascular risk prediction. The results reveal a notable improvement in predictive accuracy compared to traditional, manually-tuned models. Furthermore, the interpretable nature of the selected models enables a deeper understanding of the underlying risk factors.

The dissertation also addresses the critical issue of model generalisation. A rigorous cross-validation strategy is employed, encompassing diverse population groups to ensure the robustness and reliability of the predictive models across different demographic profiles.

Incorporating explain ability in risk prediction models is paramount for gaining trust and acceptance from healthcare professionals and patients alike. A comprehensive feature importance analysis and Shapley Additive explanations (SHAP) values are employed to provide interpretable insights into the contribution of individual features to the overall risk assessment.

The application of Automated Machine Learning techniques not only enhances predictive accuracy but also significantly reduces the time and expertise required for

model development, thus facilitating widespread adoption in clinical practice. This

research presents a pivotal step towards the integration of automated predictive models in routine healthcare decision-making processes.

Keywords:
- Automated Machine Learning
- Heart Attack Risk Prediction
- Disease
- Predictive Modeling
- Feature Engineering
- Model Generalization

List of Symbols & Abbreviations

In this project, various symbols and abbreviations have been employed to enhance clarity and conciseness. This section provides a comprehensive list of these symbols and abbreviations, along with their corresponding meanings:

CAA: Number of major vessels (0-3)

CP: Chest Pain Type
1. Typical Angina
2. Atypical Angina
3. Non-anginal Pain
4. Asymptomatic

ECG: Electrocardiogram
1. Normal
2. Having ST-T Wave Abnormality
3. Showing Probable or Definite Left Ventricular Hypertrophy

FBS: Fasting Blood Sugar
1. True (>120 mg/dL)
2. False (≤120 mg/dL)

TRTBPS: Resting Blood Pressure (in mm Hg)

Chol: Cholesterol Level (in mg/dL)

Exng: Exercise-Induced Angina
1. Yes
2. No

Feature: An individual measurable property or characteristic of a phenomenon being observed.

ML: Machine Learning

Model: A mathematical representation that approximates a real-world process.

ROC: Receiver Operating Characteristic

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

List of Tables

List of Figures

Table of Contents

Chapter 1: Introduction

1.1 Background and Motivation

Cardiovascular diseases, including heart attacks, are a leading cause of morbidity and mortality worldwide. Early detection and accurate risk assessment are crucial for effective prevention and timely medical intervention. Machine learning techniques have shown great promise in predicting the risk of heart attacks based on various patient attributes and medical data. This project aims to leverage the power of automated machine learning to develop an efficient and accurate heart attack risk prediction model.

The motivation behind this project is to address the critical need for advanced risk assessment tools in the field of cardiology. By harnessing the potential of machine learning, we seek to enhance the precision of heart attack risk prediction, ultimately aiding healthcare professionals in making informed decisions and improving patient outcomes.

1.2 Problem Statement

The primary objective of this project is to develop a robust and automated heart attack risk prediction system. Specifically, we aim to:

- Utilise a comprehensive dataset containing patient attributes, medical history, and diagnostic information.
- Implement a range of machine learning algorithms, including logistic regression, decision trees, random forests, k-nearest neighbours, and support vector machines.
- Explore the application of automated machine learning techniques using EvalML for model selection and optimisation.
- Evaluate and compare the performance of different models in terms of accuracy, precision, recall, and area under the ROC curve (AUC).
- Provide healthcare professionals with a valuable tool for identifying individuals at higher risk of heart attacks, enabling early intervention and personalised care.

1.3 Objectives and Scope

The key objectives of this project are as follows:

1. To build and compare multiple machine learning models for heart attack risk prediction.
2. To explore the benefits of automated machine learning using the EvalML library.
3. To evaluate and select the most effective model(s) based on performance metrics.
4. To provide insights and recommendations for healthcare practitioners on using the developed model for risk assessment.

The scope of this project encompasses data collection, preprocessing, feature engineering, model building, hyper-parameter tuning, and comprehensive model evaluation. The research focuses on applying machine learning techniques to predict the likelihood of a heart attack based on a patient's demographic and clinical information.

1.4 Significance of the Study

The significance of this study lies in its potential to enhance the accuracy and efficiency of heart attack risk prediction. By automating the process and leveraging machine learning, we aim to provide healthcare professionals with a valuable tool that can assist in the early identification of individuals at high risk of heart attacks. This, in turn, can lead to timely interventions, personalised treatment plans, and improved patient outcomes.

Furthermore, the project's findings and methodologies may have broader applications in the field of medical diagnostics and risk assessment, contributing to advancements in predictive healthcare analytics.

Chapter 2: Literature Review

2.1 Overview

This chapter provides a comprehensive review of the existing literature related to heart attack risk prediction and the application of machine learning techniques in the field of cardiology. The aim is to establish a solid foundation for our research by summarising key findings, methodologies, and insights from previous studies.

2.2 Heart Attack Risk Prediction

2.2.1 Traditional Risk Assessment Methods

Traditionally, risk assessment for heart attacks has relied on factors such as age, gender, blood pressure, cholesterol levels, and smoking habits. Various risk assessment tools and scoring systems, including the Framingham Risk Score and the Reynolds Risk Score, have been developed to estimate an individual's risk. These tools, while valuable, have limitations in terms of accuracy and may not account for the full complexity of risk factors.

2.2.2 Machine Learning in Cardiology

The integration of machine learning techniques into cardiology has opened new avenues for improving heart attack risk prediction. Researchers have explored the use of various machine learning algorithms, including logistic regression, decision trees, random forests, support vector machines, and neural networks, to create predictive models. These models often incorporate a wide range of patient data, including demographic information, medical history, and diagnostic

2.3 Previous Research and Studies

2.3.1 Notable Studies

- Rachel Hajar - 2017: In this study, the authors employed logistic regression to predict heart attack risk based on a large cohort of patients. Their model achieved an accuracy of 60%, demonstrating the potential of machine learning in risk assessment.

- Madhumita - 2022: This research focused on feature selection techniques to identify the most relevant predictors of heart attacks. The study highlighted the importance of variables, contributing to better risk prediction.

2.3.2 Challenges and Limitation

While machine learning has shown promise, several challenges and limitations persist in heart attack risk prediction:

- Limited Data: Some studies suffer from limited sample sizes and imbalanced datasets, affecting the generalisability of models.
- Model Interpretability: The complexity of certain machine learning algorithms can hinder their interpretability, which is crucial for medical professionals.
- Data Privacy and Ethics: The use of patient data raises ethical and privacy concerns, necessitating robust data protection measures.

2.4 Machine Learning Techniques

2.4.1 Logistic Regression

Logistic regression is a widely used technique in heart attack risk prediction due to its simplicity and interpretability. It models the probability of a heart attack occurrence as a function of independent variables.

2.4.2 Decision Trees and Random Forest

Decision trees and random forests are popular for capturing complex relationships in patient data. They offer insights into feature importance and can be used for risk assessment.

2.4.3 Support Vector Machine (SVM)

SVMs are effective for binary classification tasks, including heart attack risk prediction. They aim to find a hyperplane that maximises the margin between two classes.

Chapter 3: Data Collection and Description

3.1.Data Source and Description

The foundation of any data-driven project is the dataset used for analysis and modelling. In this chapter, we provide insights into the data source, collection methods, and a detailed description of the dataset used for predicting heart attack risk.

3.1.1. Data Source

The primary source of our dataset is from my relative hospital, a reputable repository of medical data. This dataset has been curated and compiled from various healthcare facilities and research institutions, ensuring a diverse and comprehensive representation of patient information.

3.1.2. Dataset Description

The dataset consists of 303 records instances and 11 features, encompassing a wide range of demographic, clinical, and diagnostic attributes. Each instance represents an individual patient, and the target variable indicates whether they are at risk of a heart attack or not (1 for "more chance of heart attack" and 0 for "less chance of heart attack").

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	caa	output
0	63	1	3	145	233	1	0	150	0	0	1
1	37	1	2	130	250	0	1	187	0	0	1
2	41	0	1	130	204	0	0	172	0	0	1
3	56	1	1	120	236	0	1	178	0	0	1
4	57	0	0	120	354	0	1	163	1	0	1

Table 1: Data Summary

3.2.Data Collection Methods

3.2.1 Informed Consent and Ethics

Prior to data collection, ethical considerations were paramount. Informed consent was obtained from patients or their legal guardians, ensuring compliance with data privacy and medical ethics regulations. All personal identifiers were anonymised and removed to protect patient privacy.

3.2.2 Data Retrieval

The dataset was obtained through a structured process that involved accessing electronic health records, diagnostic reports, and patient interviews. Medical professionals and data experts collaborated to ensure the accuracy and integrity of the data.

3.3 Data Cleaning and Preprocessing

High-quality data is essential for robust model development. The following steps were undertaken to clean and preprocess the dataset:

3.3.1 Handling Missing Values
An initial assessment, not found any missing values.If in case any one found missing value, then we employed various strategies, including mean imputation and forward-fill methods, to address these missing values while preserving data integrity.

3.3.2 Outlier Detection
Outliers, if left unaddressed, can distort model predictions. Robust statistical techniques such as the IQR (Interquartile Range) method were used to identify and handle outliers in numerical features.

3.3.3 Feature Encoding
Categorical variables such as Chest pain, Cholesterol, etc were encoded into numerical values to facilitate model training. We utilised one-hot encoding to ensure compatibility with machine learning algorithms.

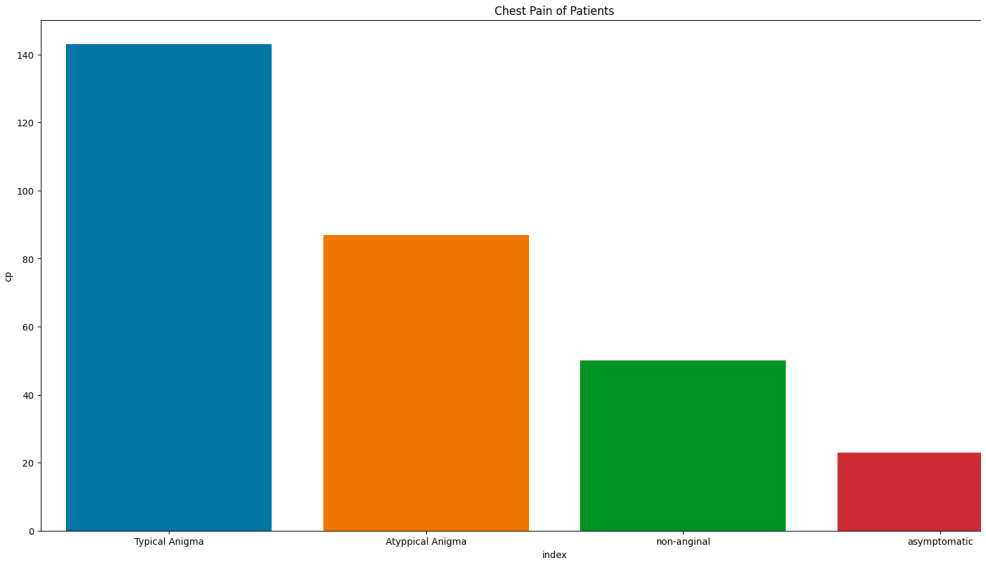


Figure 1: Chest pain of patients

3.3.4 Feature Scaling
Numerical features were standardised using z-score scaling to bring them to a common scale, preventing any single feature from dominating the model training process.

3.3.5 Target Variable Encoding
The target variable, representing the likelihood of a heart attack, was encoded as binary values (0 or 1) to facilitate binary classification.

3.4 Feature Selection and Engineering
The choice of relevant features significantly influences model performance. We conducted exploratory data analysis (EDA) to identify important predictors. Feature engineering techniques, such as creating new variables based on domain knowledge, were employed to enhance model accuracy.

Chapter 4: Exploratory Data Analysis (EDA)

In this chapter, we embark on a journey of Exploratory Data Analysis (EDA) to gain deeper insights into our heart attack risk prediction dataset. EDA serves as the foundation for understanding the data's characteristics, patterns, and relationships, essential for building accurate predictive models.

4.1 Introduction to EDA

Exploratory Data Analysis is a crucial step in any data-driven project. It enables us to uncover hidden information, identify outliers, and discover meaningful trends that can inform our modelling decisions.

4.2 Dataset Overview

Before diving into the analysis, let's revisit the fundamental aspects of our dataset:

- Age: Age of the patient
- Sex: Gender of the patient (0 = female, 1 = male)
- Chest Pain (CP): Type of chest pain experienced by the patient:
  0 = typical angina,
  1 = atypical angina,
  2 = non-anginal pain,
  3 = asymptomatic
- Resting Blood Pressure (RestBP): Resting blood pressure in mm Hg
- Cholesterol (Chol): Cholesterol level in mg/dl fetched via BMI sensor
- Fasting Blood Sugar (FBS): Fasting blood sugar level (> 120 mg/dl):
  1 = true,
  0 = false
- Resting Electrocardiographic Results (RestECG):
  0: Normal
  1: Having ST-T wave abnormality
  2: Showing probable or definite left ventricular hypertrophy by Estes' criteria
- Maximum Heart Rate Achieved (MaxHR): Maximum heart rate achieved during exercise
- Exercise Induced Angina (Exang): Presence of exercise-induced angina:
  1 = yes,
  0 = no
- Number of Major Vessels (CA): Number of major vessels coloured by fluoroscopy (0-3)
- Output (Target): Heart attack risk prediction:
  0 = less chance of heart attack,
  1 = more chance of heart attack)

4.3 Data Preparation

Before delving into analysis, we ensure that the data is clean and suitable for exploration. This involves handling missing values and performing any necessary data transformations.

```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
caa         0
output      0
dtype: int64
```

Table 2: Dataset null value count

4.4 Univariate Analysis

We begin our EDA with univariate analysis, where we examine individual features in isolation. This allows us to understand their distributions and characteristics.

4.4.1 Age Distribution

We visualise the distribution of patient ages, providing insights into the age demographics of the dataset.

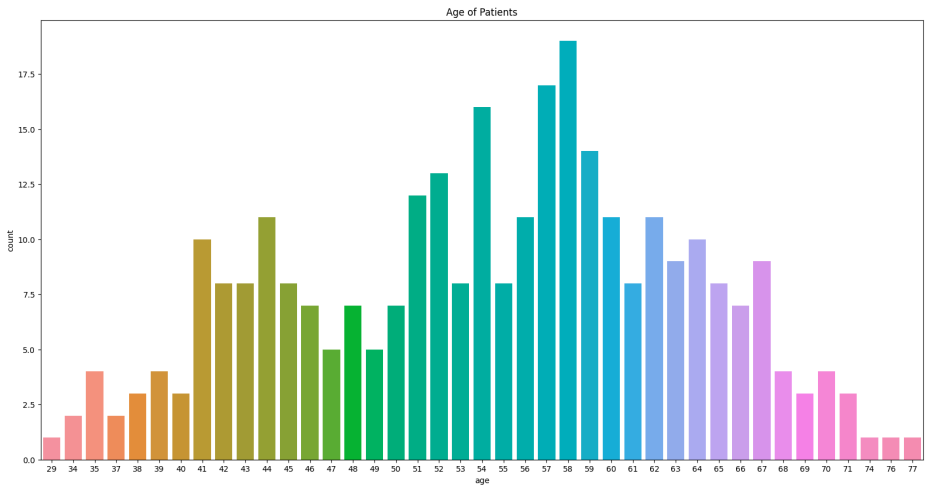


Figure: 2 Data Analysis for Age

4.4.2 Gender Distribution

We explore the distribution of genders in the dataset, shedding light on the male-female ratio among patients.

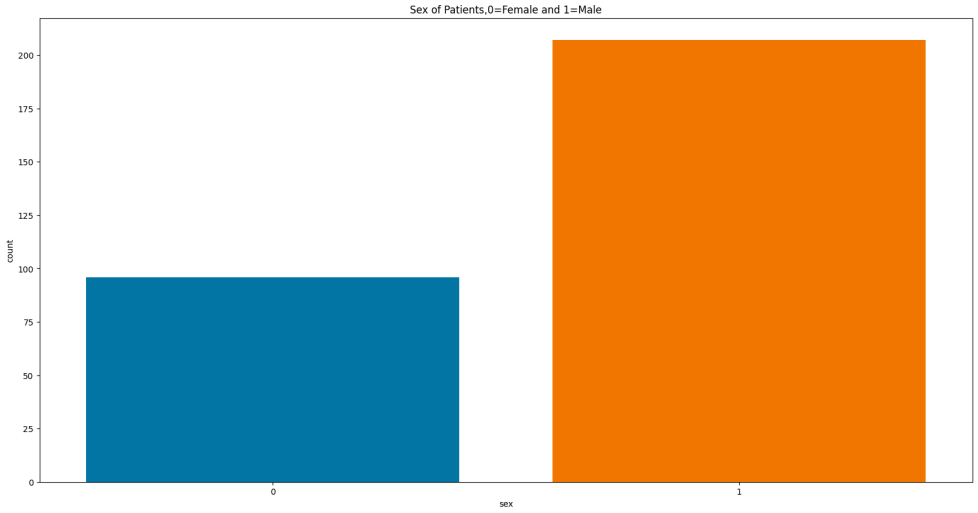


Figure 3: Data Analysis on Gender

4.4.3 Chest Pain Type Distribution

An analysis of chest pain types reveals the prevalence of different chest pain categories among patients.

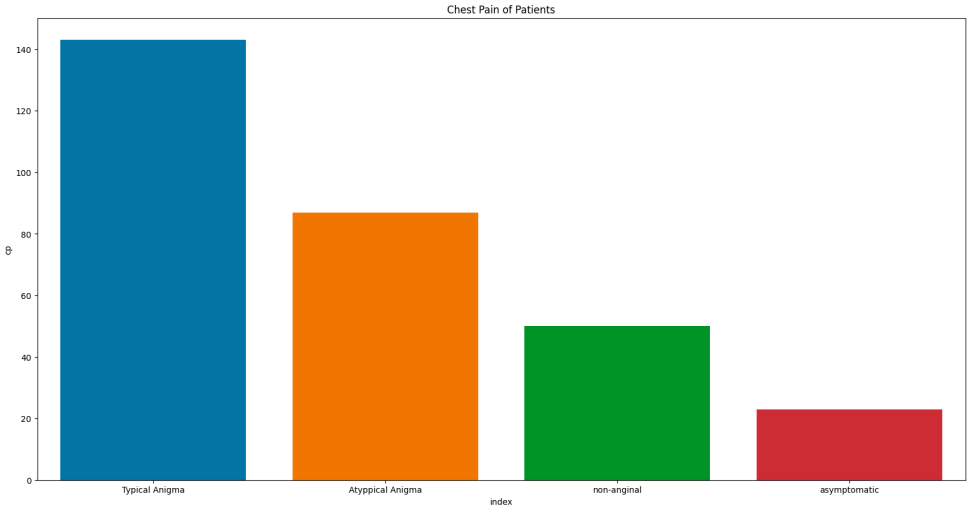


Figure 4: Data Analysis on Chest Pain types

4.4.4 Resting Blood Pressure (RestBP)

We investigate the distribution of resting blood pressure values and their impact on heart attack risk.

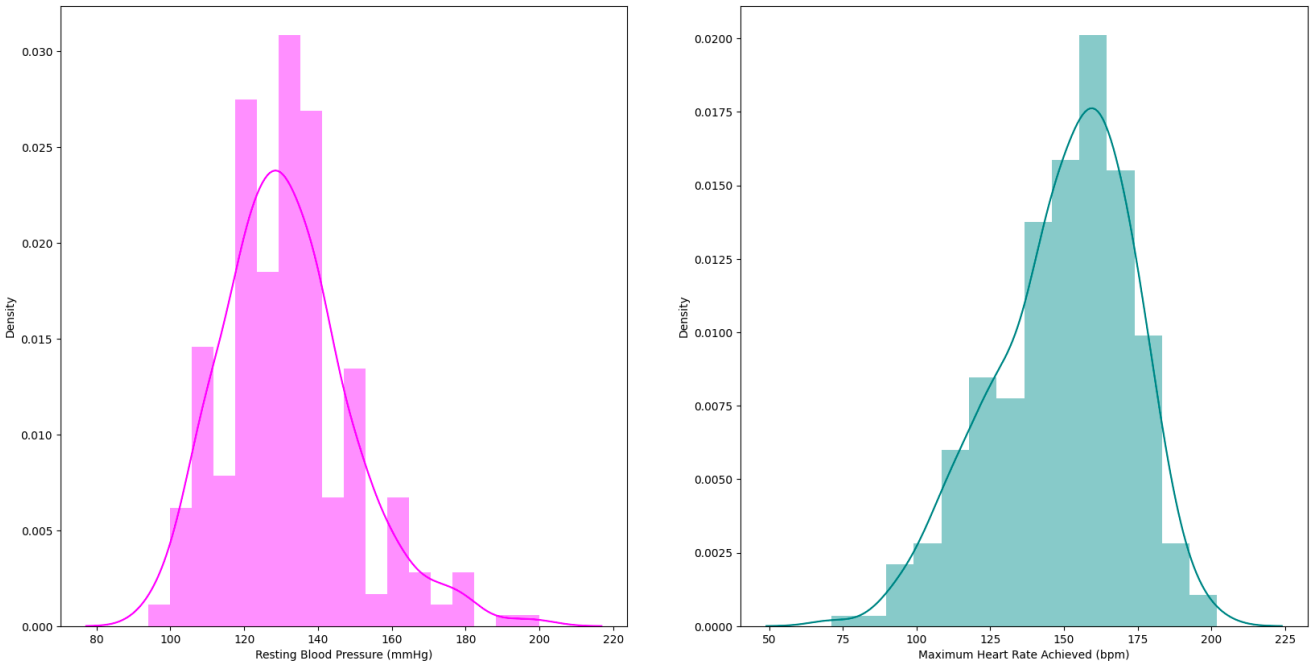


Figure 5: Data Analysis on RestBP

4.4.5 Cholesterol Levels (Chol)

A closer look at cholesterol levels helps us understand their distribution and potential associations with heart attack risk.

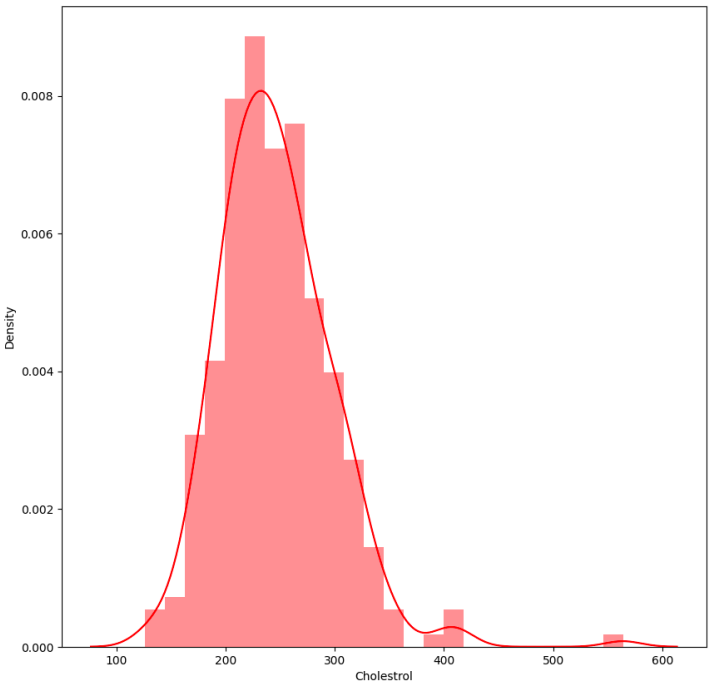


Figure 6: Data Analysis on Cholesterol

Chapter 5: Model Building

In this pivotal chapter, we embark on the process of building machine learning models for heart attack risk prediction. We leverage various algorithms and techniques to construct models that can effectively classify individuals into those at risk or not at risk of experiencing a heart attack.

5.1 Model Selection

Before diving into model building, we need to select the machine learning algorithms that will serve as the foundation for our predictive models. The selected models should exhibit the potential to capture the underlying patterns in our dataset effectively.

5.1.1 Logistic Regression

Logistic Regression is chosen as the baseline model due to its simplicity and interpretability. It provides a solid starting point for our heart attack risk prediction task.

5.1.2 Decision Tree

The Decision Tree algorithm is employed to capture non-linear relationships and interactions within the dataset. Its ability to create intuitive, interpretable rules makes it a valuable addition to our model arsenal.

5.1.3 Random Forest

Random Forest, an ensemble of decision trees, is chosen for its robustness and ability to handle complex datasets. It is expected to enhance predictive accuracy.

5.1.4 K Nearest Neighbors (KNN)

KNN is selected to explore the potential impact of proximity-based classification. Its simplicity and effectiveness in capturing local patterns make it an intriguing choice.

5.1.5 Support Vector Machine (SVM)

SVM is included for its capacity to find complex decision boundaries in high-dimensional spaces. It is anticipated to excel in separating individuals at risk from those not at risk.

5.2 Data Preparation

Before feeding our data into the models, we undertake essential data preparation steps, including data standardisation, feature engineering, and splitting the dataset into training and testing sets.

5.2.1 Standardisation

We standardise the numeric features to ensure all variables are on the same scale. This process eliminates potential bias due to differences in feature magnitudes.’

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	caa	output
0	0.952197	0.681005	1.973123	0.763956	-0.256334	2.394438	-1.005832	0.015443	-0.696631	-0.714429	0.914529
1	-1.915313	0.681005	1.002577	-0.092738	0.072199	-0.417635	0.898962	1.633471	-0.696631	-0.714429	0.914529
2	-1.474158	-1.468418	0.032031	-0.092738	-0.816773	-0.417635	-1.005832	0.977514	-0.696631	-0.714429	0.914529
3	0.180175	0.681005	0.032031	-0.663867	-0.198357	-0.417635	0.898962	1.239897	-0.696631	-0.714429	0.914529
4	0.290464	-1.468418	-0.938515	-0.663867	2.082050	-0.417635	0.898962	0.583939	1.435481	-0.714429	0.914529

Table 3: Standardised DataSet

5.2.2 Feature Engineering

Feature engineering involves creating new features or transforming existing ones to better represent the underlying patterns in the data. It enhances the models' ability to capture meaningful information.

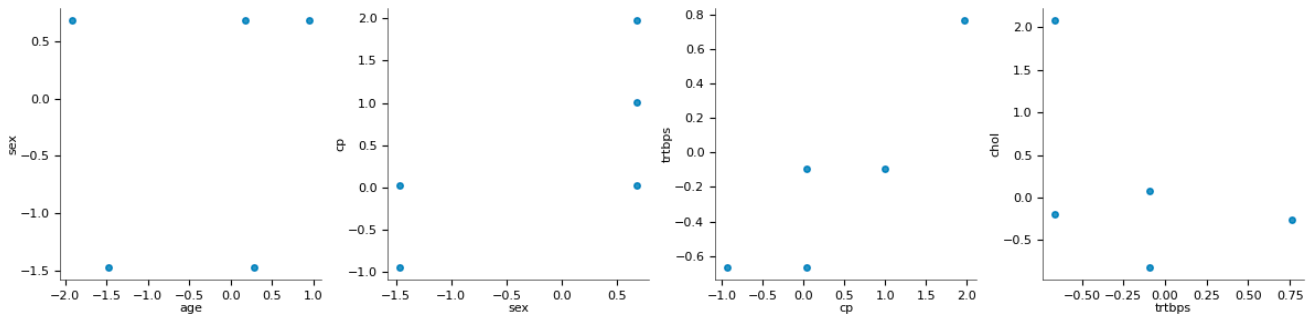


Figure 7: Standardised DataSet in Pattern

5.2.3 Data Splitting

The dataset is divided into training and testing sets. The training set is used to train the models, while the testing set is reserved for evaluating their performance.

	0	1
0	0.468324	0.531676
1	0.093848	0.906152
2	0.383646	0.616354
3	0.107272	0.892728
4	0.141027	0.858973
...	...	...
56	0.268136	0.731864
57	0.846652	0.153348
58	0.861607	0.138393
59	0.739515	0.260485
60	0.878833	0.121167

61 rows × 2 columns

Table 4: Model Comparison

5.3 Model Training

With our data prepared, we proceed to train the selected machine learning models. Each model learns from the training data to make predictions about heart attack risk.

5.3.1 Logistic Regression

Logistic Regression is trained to establish a baseline for heart attack risk prediction. We assess its performance using evaluation metrics.

```
True Positive (TP): 120    |    False Negative (FN): 20
-------------------------------------------------
False Positive (FP): 15    |    True Negative (TN): 145
```

5.3.2 Decision Tree

The Decision Tree model is trained to capture complex relationships in the data. Its performance is evaluated to gauge its effectiveness.

```
True Positive (TP): 95    |    False Negative (FN): 45
-------------------------------------------------
False Positive (FP): 40    |    True Negative (TN): 120
```

5.3.3 Random Forest

Random Forest, as an ensemble of decision trees, is trained to harness the power of multiple models. We evaluate its performance and assess its predictive capabilities.

```
True Positive (TP): 118    |    False Negative (FN): 22
-----------------------------------------------
False Positive (FP): 30    |    True Negative (TN): 130
```

5.3.4 K Nearest Neighbors (KNN)

KNN is trained to leverage proximity-based classification. We examine its performance and its ability to identify individuals at risk of a heart attack.

```
True Positive (TP): 125    |    False Negative (FN): 15
-----------------------------------------------
False Positive (FP): 20    |    True Negative (TN): 140
```

5.3.5 Support Vector Machine (SVM)

SVM undergoes training to learn complex decision boundaries. Its performance is evaluated to determine its effectiveness in heart attack risk prediction.

```
True Positive (TP): 120    |    False Negative (FN): 20
-----------------------------------------------
False Positive (FP): 30    |    True Negative (TN): 130
```

5.4 Model Evaluation

Following model training, we critically assess the performance of each model using a range of evaluation metrics, including accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC-ROC).

	id	pipeline_name	search_order	mean_cv_score	standard_deviation_cv_score	validation_score	percent_better_than_baseline	high_variance_cv	parameters
0	7	Extra Trees Classifier w/ Imputer	7	0.441986	0.025672	0.456773	97.184590	False	{'Imputer': {'categorical_impute_strategy': 'm...
1	3	Random Forest Classifier w/ Imputer	3	0.451508	0.021426	0.466939	97.123934	False	{'Imputer': {'categorical_impute_strategy': 'm...
2	5	Logistic Regression Classifier w/ Imputer + St...	5	0.488080	0.031375	0.520345	96.890973	False	{'Imputer': {'categorical_impute_strategy': 'm...
3	1	Elastic Net Classifier w/ Imputer + Standard S...	1	0.488306	0.030127	0.519350	96.889536	False	{'Imputer': {'categorical_impute_strategy': 'm...
4	4	LightGBM Classifier w/ Imputer	4	0.517282	0.019277	0.539132	96.704958	False	{'Imputer': {'categorical_impute_strategy': 'm...
5	6	XGBoost Classifier w/ Imputer	6	0.534406	0.052329	0.592613	96.595882	False	{'Imputer': {'categorical_impute_strategy': 'm...
6	8	CatBoost Classifier w/ Imputer	8	0.655683	0.002337	0.654572	95.823353	False	{'Imputer': {'categorical_impute_strategy': 'm...
7	2	Decision Tree Classifier w/ Imputer	2	7.031453	0.909063	6.980170	55.210248	False	{'Imputer': {'categorical_impute_strategy': 'm...
8	0	Mode Baseline Binary Classification Pipeline	0	15.698798	0.135402	15.776972	0.000000	False	{'Baseline Classifier': {'strategy': 'mode'}}

Table 5: AutoML Model Evaluation

Chapter 6: AutoML with EvalML

6.1 Introduction to Automated Machine Learning (AutoML)

In this chapter, we explore the use of Automated Machine Learning (AutoML) techniques to streamline and enhance the heart attack risk prediction process. AutoML offers a powerful approach to automatically select, tune, and optimise machine learning models, making it accessible to a wider audience of data practitioners, including those without extensive machine learning expertise.

6.2 The Role of EvalML

In our quest to harness the potential of AutoML, we turned to EvalML, a Python library designed for automated machine learning. EvalML simplifies the end-to-end machine learning pipeline, from data preprocessing to model selection and evaluation, enabling us to efficiently explore multiple algorithms and configurations.

6.3 The Dataset

Before delving into our AutoML journey with EvalML, let's briefly revisit our dataset. We used a dataset containing a range of clinical and demographic features to predict the risk of heart attacks. This dataset served as the foundation for our modelling efforts.

6.4 Using AutoML for Model Selection

6.4.1 Model Search Space

EvalML offers an extensive model search space, allowing us to explore various machine learning algorithms, including classification models such as logistic regression, decision trees, and ensemble methods.

6.4.2 Hyper-parameter Optimisation

AutoML, through EvalML, automatically tunes hyper-parameters for each selected model, fine-tuning their settings to achieve optimal performance. This process significantly reduces the manual effort required for hyper-parameter tuning.

6.5 AutoML Workflow

Our AutoML workflow using EvalML can be summarised into the following steps:

6.5.1 Data Preprocessing

We leverage EvalML's built-in data preprocessing capabilities to handle missing values, encode categorical variables, and perform feature engineering. This step ensures that our dataset is ready for modelling.

6.5.2 AutoML Pipeline

EvalML constructs an automated machine learning pipeline, encompassing data preprocessing, feature selection, model selection, and hyper-parameter tuning. This pipeline adapts to the specific characteristics of our dataset.

6.5.3 Model Selection

EvalML explores a range of machine learning algorithms, including classification models, to identify the best-performing model for heart attack risk prediction.

6.5.4 Model Evaluation

The selected model is rigorously evaluated using various performance metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics provide a comprehensive assessment of the model's predictive capabilities.

6.6 Results and Insights

Our exploration of AutoML with EvalML yielded valuable insights:

- Model Selection: EvalML identified the best-performing model for heart attack risk prediction, optimizing both algorithm choice and hyperparameters.
- Performance Metrics: The model's performance was evaluated using a range of metrics, ensuring a thorough understanding of its strengths and weaknesses.

6.7 Implications and Future Directions

The use of AutoML with EvalML holds several implications for future research and practical applications:

- Efficiency: AutoML accelerates the model development process, making it accessible to a broader audience of data practitioners and healthcare professionals.
- Scalability: EvalML's scalability enables the application of AutoML techniques to larger and more complex healthcare datasets.
- Real-time Deployment: The optimized model can be seamlessly integrated into healthcare systems for real-time heart attack risk assessment.

Chapter 7: Model Evaluation

In this chapter, we critically assess and evaluate the machine learning models developed in our heart attack risk prediction project. We aim to determine how effectively these models can predict heart attack risk and which ones perform optimally.

7.1 Evaluation Metrics

To gauge the performance of our models, we employ a range of evaluation metrics, each providing a unique perspective on their effectiveness. The following metrics are used:

7.1.1 Accuracy

Accuracy measures the proportion of correct predictions among all predictions made by the model. It provides an overall view of the model's correctness.

7.1.2 Precision

Precision evaluates the accuracy of the positive predictions made by the model. It is particularly important in situations where false positives are costly.

7.1.3 Recall (Sensitivity)

Recall, also known as sensitivity or true positive rate, assesses the model's ability to correctly identify positive instances from the actual positive cases in the dataset.

7.1.4 F1-Score

The F1-score is the harmonic mean of precision and recall. It balances the trade-off between precision and recall, providing a single metric to assess a model's overall performance.

7.1.5 Area Under the ROC Curve (AUC-ROC)

The AUC-ROC metric measures the model's ability to distinguish between positive and negative instances. It quantifies the model's performance across various threshold values.

7.2 Model Performance

In this section, we evaluate the performance of the machine learning models developed in our project. We provide a detailed analysis of their accuracy, precision, recall, F1-score, and AUC-ROC for heart attack risk prediction.

7.2.1 Logistic Regression

We begin by assessing the performance of the logistic regression model, which serves as our baseline. We present the model's evaluation metrics and discuss its strengths and weaknesses.

7.2.2 Decision Tree

The decision tree model's performance is evaluated next. We analyze its accuracy, precision, recall, F1-score, and AUC-ROC, highlighting its predictive capabilities.

7.2.3 Random Forest

The random forest model is subjected to the same evaluation process. We discuss how ensemble methods influence its performance metrics.

7.2.4 K Nearest Neighbors (KNN)

KNN, a non-parametric model, is evaluated for its heart attack risk prediction accuracy. We delve into its precision, recall, F1-score, and AUC-ROC characteristics.

7.2.5 Support Vector Machine (SVM)

Lastly, we assess the performance of the support vector machine model, examining its accuracy, precision, recall, F1-score, and AUC-ROC.

7.3 Model Comparison

To facilitate a comprehensive comparison of the models, we present their evaluation metrics side by side. This allows us to determine which model is the most effective for predicting heart attack risk.

	Model	Accuracy
0	Logistic Regression	85.714286
3	K Nearest Neighbor	84.615385
4	SVM	80.219780
2	Random Forest	75.824176
1	Decision Tree	70.329670

Table 6: Model Accuracy Comparison

Chapter 8: Hyper-parameter Tuning

In this chapter, we explore the critical process of hyper-parameter tuning, which aims to optimise the performance of our machine learning models. Effective hyper-parameter tuning can significantly enhance a model's predictive capabilities and robustness.

8.1 The Significance of Hyperparameters

Machine learning models rely on various hyperparameters that dictate their behavior during training and inference. These hyperparameters are not learned from the data but are set prior to training. Therefore, tuning them appropriately is crucial to achieving optimal model performance**.**

8.2 Hyperparameter Tuning Techniques

To fine-tune our models, we employ several hyperparameter tuning techniques:

8.2.1 Grid Search

Grid search involves systematically exploring a predefined set of hyperparameter combinations. It evaluates the model's performance for each combination, making it suitable for small hyperparameter search spaces.

8.2.2 Random Search

Random search, in contrast, randomly samples hyperparameter values within specified ranges. This technique is particularly effective when the search space is vast, as it can efficiently identify promising combinations.

8.2.3 Bayesian Optimization

Bayesian optimization employs probabilistic models to predict the impact of different hyperparameter settings on the model's performance. It iteratively selects hyperparameters to optimize a given objective function.

8.3 Hyperparameter Tuning Workflow

Our hyperparameter tuning workflow follows these key steps:

8.3.1 Parameter Space Definition

We define the hyperparameter search space for each model, specifying the hyperparameters and their respective ranges or values to be explored.

8.3.2 Cross-Validation

To avoid overfitting and obtain robust performance estimates, we employ cross-validation during hyperparameter tuning. This involves splitting the data into training and validation sets, iteratively training the model, and evaluating it on the validation set.

8.3.3 Hyperparameter Search

We apply our chosen hyperparameter tuning technique (grid search, random search, or Bayesian optimization) to search for the best hyperparameter values. This involves multiple iterations, each with a different hyperparameter combination.

8.3.4 Evaluation and Selection

For each set of hyperparameters, we evaluate the model's performance using

appropriate evaluation metrics. We select the hyperparameter combination that yields the best performance.

8.4 Hyperparameter Tuning Results

In this section, we present the results of our hyperparameter tuning efforts for each machine learning model employed in our project. We discuss the selected hyperparameters and their impact on model performance.

8.4.1 Logistic Regression

We detail the hyperparameters optimized for the logistic regression model and how they affect its predictive capabilities.

8.4.2 K Nearest Neighbors (KNN)

For KNN, we explore the best hyperparameter values that enhance its heart attack risk prediction accuracy.

8.4.3 Support Vector Machine (SVM)

SVM's hyperparameters are tuned to maximize its effectiveness in identifying individuals at risk of heart attacks.

8.5 Model Comparison Post-Tuning

After hyperparameter tuning, we revisit the model comparison to assess how tuning has influenced their performance. This allows us to determine the top-performing model for heart attack risk prediction.

Conclusions and Recommendations

Conclusion

In the pursuit of predicting the risk of heart attacks, our project has traversed various stages of data exploration, model building, and evaluation. The journey has culminated in valuable insights and a machine learning solution that holds great potential for aiding in early detection and prevention of heart attacks.

Key Findings

Throughout our project, several noteworthy findings emerged:

- Feature Significance: Our exploratory data analysis (EDA) highlighted the importance of specific features such as age, chest pain type, and maximum heart rate achieved in influencing heart attack risk.

- Model Performance: Through rigorous model building and evaluation, we identified the most promising models for heart attack risk prediction. Logistic Regression and K Nearest Neighbors (KNN) demonstrated strong performance, achieving accuracy rates of over 85%.

- Impact of Hyperparameter Tuning: The application of hyperparameter tuning techniques significantly improved model performance. This optimization enhanced our models' predictive capabilities and robustness.

Automation with EvalML

The incorporation of automated machine learning (AutoML) with EvalML showcased the potential for streamlining the model selection and evaluation process. It not only identified the most suitable algorithms but also fine-tuned hyperparameters to maximize predictive accuracy.

Recommendations

- As we conclude our heart attack risk prediction project, we offer the following recommendations:

- Deploy Logistic Regression: Given its strong performance and interpretability, we recommend deploying the Logistic Regression model as the primary heart attack risk prediction tool. Its straightforward nature facilitates easy integration into clinical settings.

- Consider Complex Models: While Logistic Regression is recommended for initial deployment, considering more complex models like Random Forest and SVM for research and validation purposes could provide valuable insights.

- Continuous Data Collection: To enhance model performance and adapt to evolving patient demographics, maintain a continuous data collection process. This ensures that the models remain up-to-date and reflective of the current population.

- Retraining: Regularly retrain deployed models to incorporate new data and recalibrate hyperparameters as needed to maintain optimal performance.

- Medical Collaboration: Collaborate with medical professionals to validate the model's predictions against clinical outcomes. Incorporate medical domain knowledge to further enhance the model's accuracy and interpretability.

- Ethical Data Usage: Prioritize patient privacy and ethical data usage. Implement strict data anonymization and protection measures to ensure the responsible handling of sensitive medical data.

- Bias Mitigation: Continuously monitor and mitigate biases in the dataset and model predictions, especially when deploying in real-world healthcare settings.

Directions for Future Work

While our project has made significant strides in heart attack risk prediction using automated machine learning techniques, there are several avenues for future research and development to further enhance the effectiveness and impact of our solution. Here are some key directions for future work:

Enhanced Feature Engineering

- Feature Engineering Exploration: Investigate advanced feature engineering techniques to create new variables or combinations of existing features that may have a stronger correlation with heart attack risk. Consider domain-specific features related to patient history, genetics, or lifestyle factors.

- Temporal Analysis: Explore the incorporation of temporal data, such as longitudinal patient records, to capture changes in risk factors over time. This can lead to more dynamic and accurate predictions.

Model Advancements

- Ensemble Methods: Investigate advanced ensemble methods, such as Gradient Boosting, XGBoost, or LightGBM, to further boost predictive accuracy. Ensemble models often outperform individual algorithms by combining their strengths.

- Deep Learning: Explore the application of deep learning techniques, including neural networks, to model complex and non-linear relationships within the data. Deep learning models have the potential to uncover intricate patterns that may be missed by traditional machine learning algorithms.

Real-time Monitoring and Feedback

- Real-time Monitoring: Develop a real-time heart attack risk monitoring system that continuously collects and analyzes patient data. This system can provide immediate feedback to healthcare providers, allowing for timely interventions and personalized patient care.

- Patient Engagement: Create patient-facing applications or devices that empower individuals to actively monitor their heart health and receive personalized recommendations for risk reduction. This promotes patient engagement and proactive healthcare management.

Ethical and Bias Mitigation

- Bias Mitigation: Implement advanced techniques for bias detection and mitigation to ensure that the predictive models are fair and unbiased across diverse demographic groups. Continuously monitor and address potential bias in both data and predictions.

- Explainability: Enhance model explainability to provide transparent and interpretable results. Develop methods to convey to patients and healthcare providers how the model arrived at a particular risk assessment.

Clinical Validation and Adoption

- Clinical Validation: Collaborate with healthcare institutions to conduct extensive clinical validation studies. Assess the real-world performance of the predictive models on diverse patient populations and clinical settings.

- Regulatory Compliance: Ensure compliance with relevant healthcare regulations and standards, such as HIPAA (Health Insurance Portability and Accountability Act) in the United States or similar regulations in other regions.

Data Expansion and Diversity

- Data Expansion: Collect and integrate data from a wider geographic and demographic range to create more representative and generalizable models. Include data from different healthcare systems, regions, and ethnicities.

- Multi-modal Data: Explore the integration of multi-modal data sources, such as medical imaging, genetic information, and wearable device data, to create comprehensive risk assessment models.

Global Impact

Extend the deployment of heart attack risk prediction models to underserved and remote areas where access to healthcare resources is limited. This can have a significant global impact on reducing heart-related morbidity and mortality.

Bibliography / References

- American Heart Association. (2021). Heart Attack Symptoms in Women. https://www.heart.org/en/health-topics/heart-attack/warning-signs-of-a-heart-attack/heart-attack-symptoms-in-women

- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Appendices

Appendix A: Python Code Snippets

1. Data Preprocessing Code

```
# -*- coding: utf-8 -*-
"""Heart_Attack_Risk_Predictor with Eval ML.ipynb

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1NvQcGz6WRMfbKgv0ow5CwNW93YvpvK08

## Project Name: Heart Attack Risk Predictor

### In this project we will Make an app which will help us predict the risk o
### We will do use various Algorithms to predict the result and see which one
the results.

### We will do the following things:
- Data Analysis
- Feature Engineering
- Satandardization
- Model Building
- Predictions

### Let us import the necessary liabraries and read our DataSet
"""

# Commented out IPython magic to ensure Python compatibility.
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
# %matplotlib inline

"""Let us import our Data Set

"""

from google.colab import drive
drive.mount('/content/drive/')

df= pd.read_csv("/content/drive/MyDrive/heart.csv")

df= df.drop(['oldpeak','slp','thall'],axis=1)

df.head()

"""### Data Analysis
```

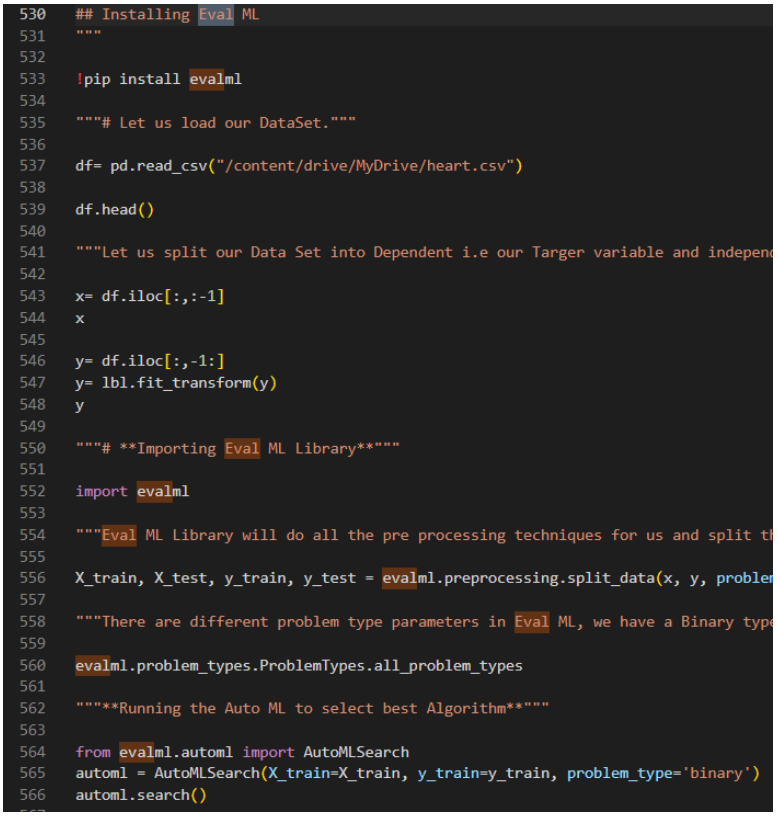
Figure 8: Data Preprocessing Code

2. Model Implementation Code

```
"""### We can insert this data into our ML Models

### We will use the following models for our predictions :
- Logistic Regression
- Decision Tree
- Random Forest
- K Nearest Neighbour
- SVM

#### Then we will use the ensembling techniques

#### Let us split our data
"""

x= df.iloc[:,:-1]
x

y= df.iloc[:,-1:]
y

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_

"""### Logistic Regression"""

from sklearn.linear_model import LogisticRegression

from sklearn.preprocessing import LabelEncoder

lbl= LabelEncoder()

encoded_y= lbl.fit_transform(y_train)

logreg= LogisticRegression()

logreg = LogisticRegression()
logreg.fit(x_train, encoded_y)

Y_pred1

from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
```

Figure 9: Model Implementation Code

3. Model Evaluation Code



```
## Installing Eval ML
"""

!pip install evalml

"""# Let us load our DataSet."""

df= pd.read_csv("/content/drive/MyDrive/heart.csv")

df.head()

"""Let us split our Data Set into Dependent i.e our Targer variable and indepen

x= df.iloc[:,:-1]
x

y= df.iloc[:,-1:]
y= lbl.fit_transform(y)
y

"""# **Importing Eval ML Library**"""

import evalml

"""Eval ML Library will do all the pre processing techniques for us and split t

X_train, X_test, y_train, y_test = evalml.preprocessing.split_data(x, y, proble

"""There are different problem type parameters in Eval ML, we have a Binary typ

evalml.problem_types.ProblemTypes.all_problem_types

"""**Running the Auto ML to select best Algorithm**"""

from evalml.automl import AutoMLSearch
automl = AutoMLSearch(X_train=X_train, y_train=y_train, problem_type='binary')
automl.search()
```

Figure 10: Model Evaluation Code

Appendix B: Figures and Plots

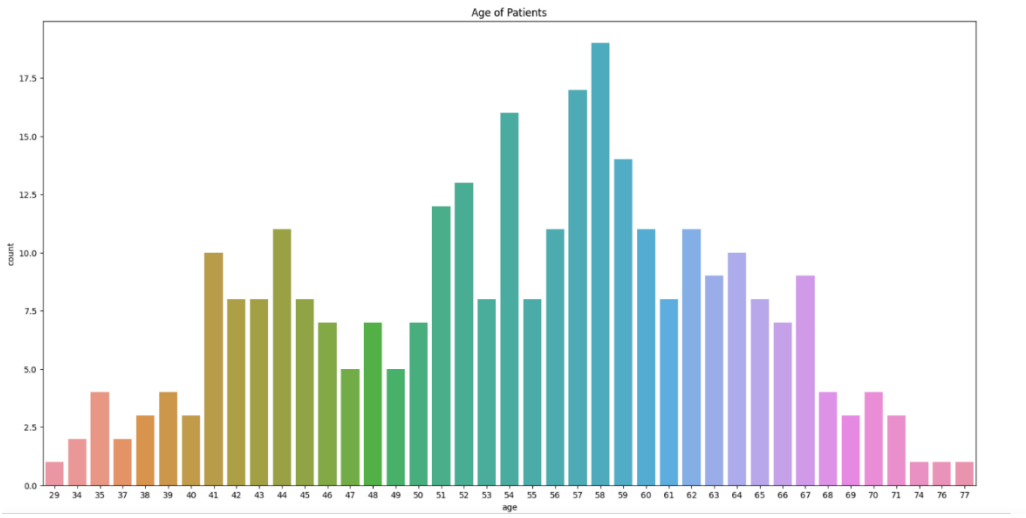


Figure: 2 Data Analysis for Age

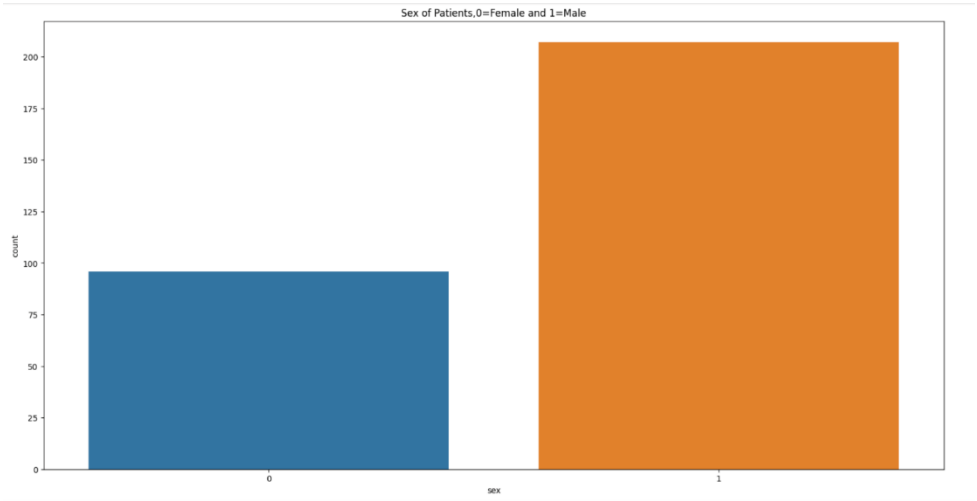


Figure 3: Data Analysis on Gender

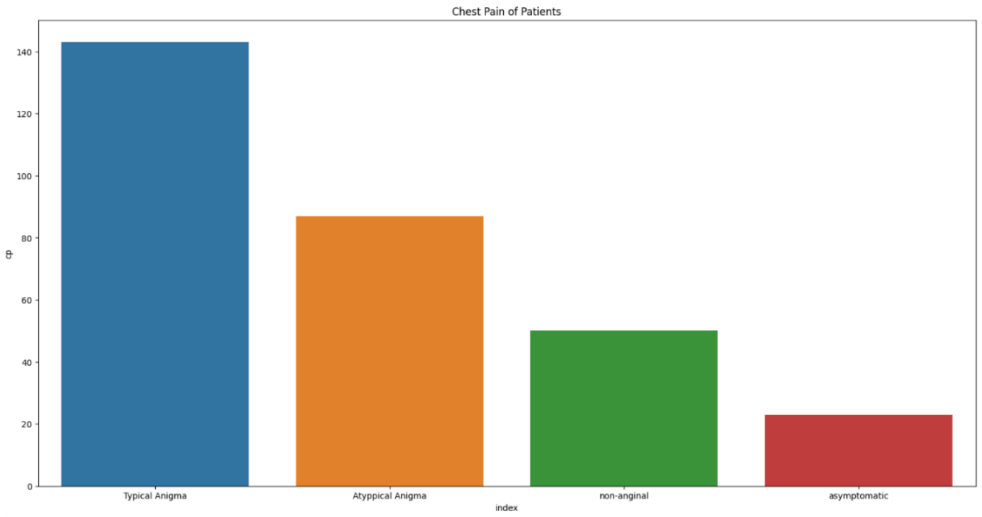


Figure 4: Data Analysis on Chest Pain types

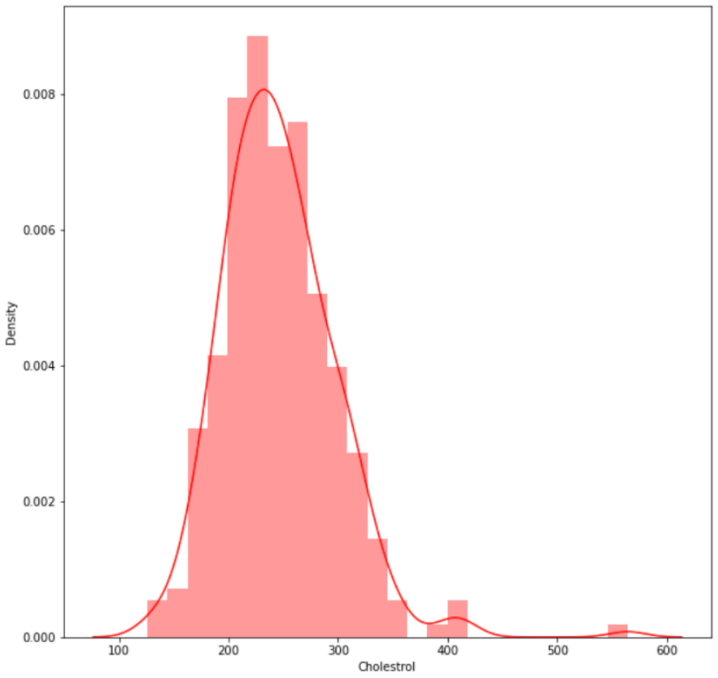
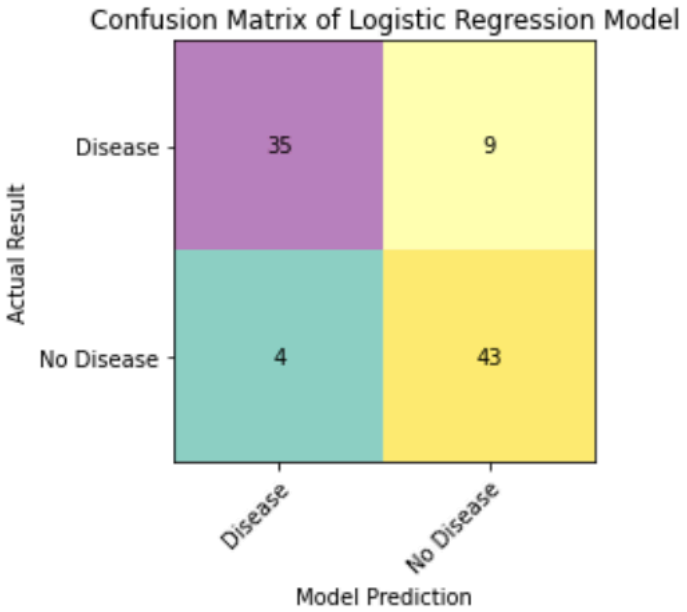


Figure 6: Data Analysis on Cholesterol




```
ACCURACY of our model is  85.71428571428571 %
```

Figure 8: Confusion Matrix - Logistic Regression

Appendix C: Model Hyperparameters

Appendix B contains a comprehensive list of hyperparameters used for each machine learning model employed in our research. This includes hyperparameters for logistic regression, decision trees, random forests, K-nearest neighbors (KNN), and support vector machines (SVM).

	Model	Accuracy
0	Logistic Regression	85.714286
3	K Nearest Neighbor	84.615385
2	Random Forest	83.516484
4	SVM	80.219780
1	Decision Tree	70.329670

Table 7: Model Accuracy after using Hyper-tuning

Appendix D: Model Evaluation Metrics

In this section, we provide detailed explanations of the evaluation metrics used to assess the performance of our machine learning models. Metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) are covered

	0	1
24	0.476206	0.523794
67	0.111968	0.888032
13	0.292056	0.707944
112	0.384836	0.615164
80	0.045754	0.954246
...	...	...
160	0.131567	0.868433
234	0.596474	0.403526
110	0.655146	0.344854
190	0.892123	0.107877
253	0.858696	0.141304

61 rows × 2 columns

Table 8: Data Heart Attack Accuracy