Predictive Analysis of Heart Attack

Submitted To :Dr Awais Hasan

Submitted By :Fatima Rauf

STRATEGIES FOR SUCCESS IN HEART ATTACK PREDICTION

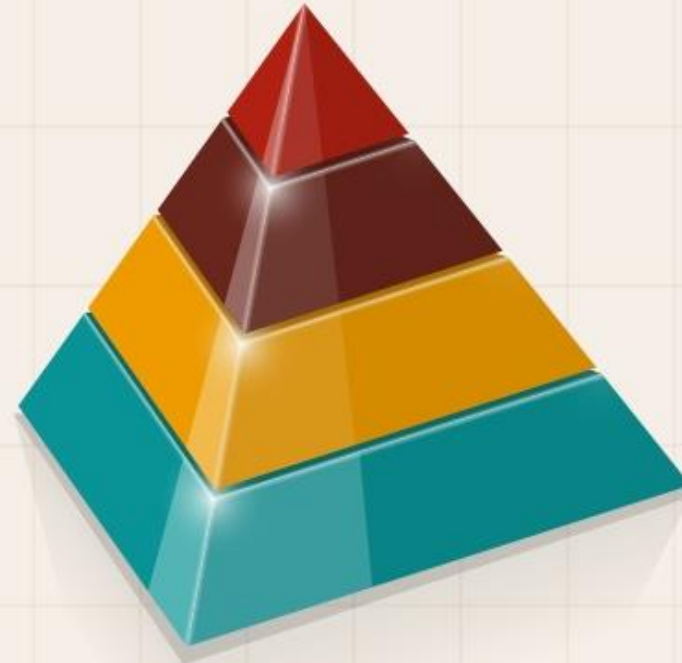Exploring effective strategies for heart attack risk modeling

DATASET OVERVIEW

We will introduce the Heart Attack Prediction Dataset, detailing its significance in predicting health outcomes.

KEY STATISTICS

Explore critical statistics within the dataset, including patient demographics and health indicators.

DATA PREPARATION STEPS

Discuss the necessary steps for data cleaning and preparation to ensure accurate modeling.

MODELING STRATEGIES

Outline the modeling strategies to predict heart attack risks, focusing on various algorithms and techniques.

HEART ATTACK PREDICTION DATASET

Key Features and Insights of the Dataset

PURPOSE OF THE DATASET

The dataset aims to analyze and predict heart attack risks based on patient features.

DATASET SIZE

It consists of 303 rows, providing a robust sample for analysis.

NUMBER OF FEATURES

The dataset includes 14 columns representing various patient data points.

SIGNIFICANCE

This dataset is crucial for understanding heart attack risks and contributing to preventive measures.

DATASET FEATURES OVERVIEW

Detailed Overview of Dataset Attributes and Types

COLUMN NAME	DATA TYPE	DESCRIPTION
age	Numerical	Age of the patient in years
sex	Categorical	Gender (1 = Male, 0 = Female)
cp	Categorical	Chest pain type (0: Typical angina, 1: Atypical angina)
trestbps	Numerical	Resting blood pressure in mm Hg
chol	Numerical	Serum cholesterol level in mg/dl
fbs	Categorical	Fasting blood sugar (>120 mg/dl, 1 = True, 0 = False)
restecg	Categorical	Resting electrocardiographic results

SUMMARY STATISTICS OF NUMERICAL AND CATEGORICAL FEATURES

Comprehensive overview of statistical metrics

COLUMN NAME	MEAN	MEDIAN	STD. DEV.	MIN	MAX
age	value	value	value	value	value
chol	value	value	value	value	value
trestbps	value	value	value	value	value

HANDLING MISSING VALUES IN DATA

Strategies for Imputing Missing Data Effectively

COLUMN NAME	MISSING COUNT	PERCENTAGE MISSING
chol	5	1.65%
thal	3	0.99%

EXPLORING CORRELATION

Key Insights on Heart Attack Risk

UNDERSTANDING CORRELATION MATRIX

A correlation matrix is a table showing correlation coefficients between variables, highlighting their relationships.

STRONG POSITIVE CORRELATION

`thalach` (Maximum heart rate) exhibits a strong positive correlation with `target` (heart attack risk), indicating higher rates link to increased risks.
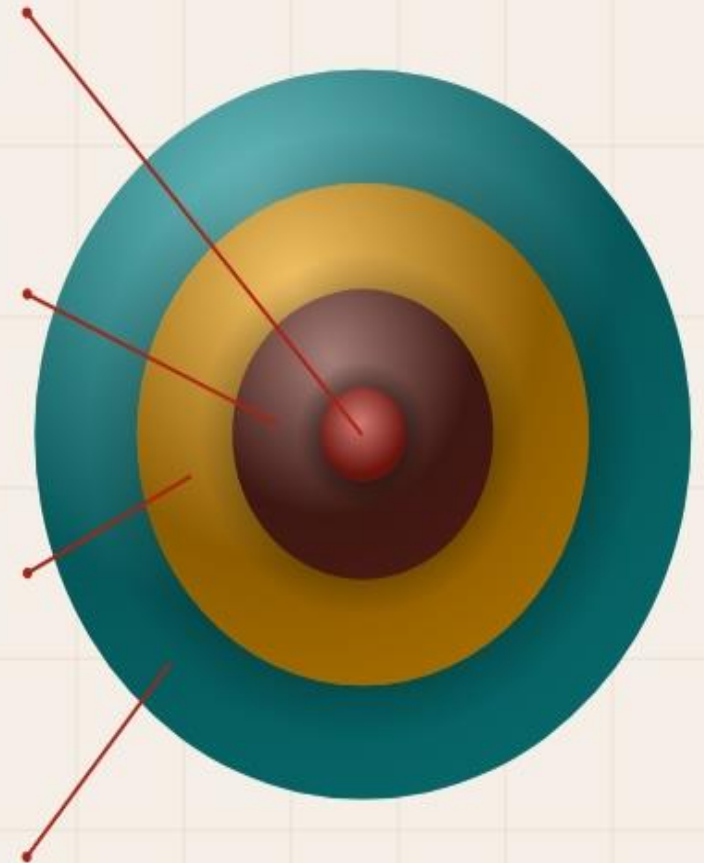
NEGATIVE CORRELATION INSIGHT

`oldpeak` (ST depression) reveals a negative correlation with `target`, suggesting lower ST depression relates to higher heart attack risks.

IMPORTANCE OF VISUALIZATION

Visualizing the correlation through a heatmap is essential for understanding relationships and guiding feature selection.

DISTRIBUTION OF KEY FEATURES

Analyzing Features for Model Training

HISTOGRAMS FOR NUMERICAL DATA

Examine the distribution of numerical features like Age, Cholesterol, and Resting Blood Pressure using histograms.

COUNT PLOTS FOR CATEGORICAL DATA

Utilize count plots to analyze the frequency distributions of categorical features such as Gender, Chest Pain Types, and Thalassemia.

FEATURE: AGE ANALYSIS

Assess how age is distributed across the dataset, identifying trends and patterns crucial for model training.

FEATURE: CHOLESTEROL LEVELS

Investigate cholesterol levels to understand their impact on health and model predictions.

FEATURE: BLOOD PRESSURE INSIGHTS

Explore resting blood pressure data to uncover relationships with other health metrics.

IMPORTANCE OF GENDER DISTRIBUTION

Evaluate the gender distribution in the dataset to ensure balanced model training and avoid bias.

UNDERSTANDING CHEST PAIN TYPES

Analyze the frequency of different chest pain types to identify common patterns in the data.

THALASSEMIA FREQUENCY ANALYSIS

Examine the frequency of thalassemia occurrences within the dataset for better predictive modeling.

IDENTIFYING ANOMALIES

Utilize these distributions to pinpoint anomalies that may skew model training or predictions.

PREPARATION FOR MODEL TRAINING

Understanding feature distributions is crucial for preparing effective machine learning models.

KEY OBSERVATIONS FROM THE DATASET

Understanding critical factors influencing dataset analysis

01 GENDER IMBALANCE

The dataset exhibits a significant gender imbalance with more males than females, potentially skewing model predictions.

02 CHOLESTEROL LEVELS

Cholesterol levels show a broad range, suggesting the need for normalization to ensure model accuracy.

03 LOW VARIANCE FEATURES

Features like 'ca' (major vessels) and 'thal' show low variance, indicating areas requiring further analysis.

04 IMPLICATIONS FOR PREPROCESSING

These observations will guide our data preprocessing and feature engineering strategies for improved model performance.

NEXT STEPS IN DATA PREPARATION

Essential Steps to Prepare Data for Analysis

HANDLE MISSING VALUES

Impute missing values in the 'chol' and 'thal' columns to maintain data integrity.

SCALE DATA

Normalize numerical features such as 'chol', 'trestbps', and 'thalach' for consistent model training.

FEATURE SELECTION

Analyze importance of features like 'cp', 'ca', and 'thal' to focus on impactful variables.

MODEL BUILDING

Utilize algorithms such as Logistic Regression, Decision Trees, and Neural Networks based on data characteristics.