Predicting Heart Attack Risk Using Machine Learning: A Comparative Study of Logistic Regression and Random Forest Algorithms

Predicting Heart Attack Risks

Comparative analysis of Logistic Regression and Random Forest for improved predictive analytics in healthcare.

b

bhagya lakshmi

Presenter

HEART ATTACK STUDY

Predicting Heart Attack Risk

Kusumanjali, MBA Student & Dr K BhagyaLakshmi, Assistant Professor, Chaitanya Bharathi Institute of Technology, Hyderabad

Heart Attack Prediction Using Machine Learning

Key Findings and Methodologies

Predictive Models

The study employs Logistic Regression and Random Forest algorithms to predict heart attacks, enhancing early detection and preventive strategies.

Logistic Regression

Used for analyzing linear relationships between predictors and heart attack likelihood, providing interpretable coefficients for feature significance.

Random Forest

Captures complex, non-linear interactions among variables to improve predictive accuracy, important for identifying high-risk individuals.

Key Medical Parameters

Includes 12 parameters such as age, sex, chest pain type, resting blood pressure, cholesterol levels, and more, critical for model training.

Findings

Demonstrates the potential of machine learning techniques in enhancing heart disease risk prediction, aiding clinical decision-making and preventive care.

Introduction to Cardiovascular Diseases and Machine Learning

Exploring the Intersection of Healthcare and Technology

Global Impact of Cardiovascular Diseases

Cardiovascular diseases rank among the most critical health challenges globally, significantly contributing to morbidity and mortality rates.

Complexity in Predicting Heart Attacks

Despite advances in medical science, predicting heart attack susceptibility remains complex due to multifaceted risk factors.

Role of Machine Learning in Healthcare

Machine learning is revolutionizing healthcare by analyzing large datasets to enhance disease understanding and clinical decision-making.

Objectives of the Study

The study aims to evaluate various machine learning techniques to identify the most effective models for predicting heart attack susceptibility.

Bridging Research with Clinical Application

The research seeks to bridge the gap between cutting-edge machine learning research and its practical application in clinical settings.

Review of Literature on Heart Disease Prediction

Exploring advanced techniques and models in heart disease prediction research

Hybrid Models for Heart Disease Prediction

T. Saravanan (2024) and MaahamMunsif et al. (2024) developed hybrid models combining deep learning with ensemble techniques, achieving high accuracies.

Significant Predictors Identified

Xinyi Liu et al. (2023) identified chest pain type and maximum heart rate as significant predictors using logistic regression and random forest models.

Advanced Feature Extraction Techniques

P. Morris and L. Collins (2023) enhanced model performance with wavelet transforms, providing deeper insights into predictive accuracy.

Explainable AI and Feature Importance

A. Price and K. Scott (2022) found SHAP provides reliable insights into feature importance, aiding model interpretation.

Multimodal Data Integration

J. Stewart and M. Rivera (2022) improved prediction accuracy by integrating genetic, clinical, and imaging data.

Federated Learning and Privacy

V. Scott and S. Green (2022) demonstrated that federated learning maintains predictive accuracy while ensuring patient privacy.

Time-Series Forecasting Models

T. Nguyen and Q. Tran (2021) showed that LSTM models outperform ARIMA in capturing complex temporal patterns.

Research Problem

Exploring Machine Learning for Heart Attack Risk Prediction

- **Heart attacks as a leading cause of mortality**

  Emphasizes the global significance and the necessity for accurate prediction methods to improve patient outcomes.

- **Machine Learning Algorithms**

  The study evaluates logistic regression and random forest models in predicting heart attack risk.

- **Objective: Compare Accuracy**

  Compare the accuracy and reliability of logistic regression and random forest algorithms for predicting heart attack risk.

- **Objective: Identify Predictors**

  Identify significant predictors of heart attack susceptibility across the two models.

- **Objective: Evaluate Generalizability**

  Use cross-validation techniques to evaluate the generalizability of these machine learning models.

- **Objective: Recommend Implementation**

  Recommend the best-performing model for clinical practice to improve risk prediction and preventive care.

Data Sources: Cleveland Heart Disease Study

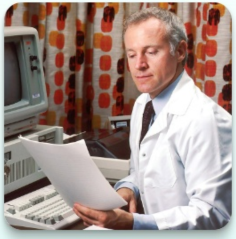
Comprehensive Data Analysis for Predicting Heart Disease

Primary Data Source

The study utilizes the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, providing comprehensive clinical and demographic information essential for cardiovascular health analysis.

Sample Size

The study will analyze the entire Cleveland Heart Disease dataset, comprising a total of insert number instances.



Clinical Variables Overview

Key clinical variables include blood pressure, cholesterol levels, fasting blood sugar, and more, crucial for predicting heart disease.

Detailed Clinical Variables

The dataset includes electrocardiographic results, maximum heart rate, exercise-induced angina, old peak, slope of the ST segment, vessel count by fluoroscopy, and heart disease presence.

Tools and Techniques for Heart Disease Prediction

Comprehensive Overview of Tools and Techniques

Data Processing Tools

Utilize Python libraries such as Pandas for efficient data manipulation and cleaning, and NumPy for numerical operations and handling arrays.

Machine Learning Algorithms

Apply Logistic Regression for binary classification of heart disease presence, and Random Forest for ensemble learning to improve prediction accuracy.

Evaluation Metrics

Evaluate model performance using Accuracy, Precision, Recall, F1 score, and ROC-AUC to ensure reliable predictions.

Data Collection for Research

The dataset from Kaggle, comprising 12 medical features and 1,000 patient records, is used for predicting heart disease risk. Only the first 20 rows are shown for illustration.

Key Features in the Dataset

Detailed Exploration of Dataset Features

Age

Patient's age, an important factor in analyzing health trends and conditions.

Sex

Gender of the patient, where 1 indicates male and 0 indicates female.

Chest Pain Type (cp)

Categorical variable with four types: 0 for Typical angina, 1 for Atypical angina, 2 for Non-anginal pain, and 3 for Asymptomatic.

Resting Blood Pressure (trestbps)

Measured in mm Hg, indicating the blood pressure of a patient at rest.

Serum Cholesterol (chol)

Cholesterol level measured in mg/dl, providing insight into the patient's lipid profile.

Fasting Blood Sugar (fbs)

Indicates if fasting blood sugar is greater than 120 mg/dl, where 1 means true and 0 means false.

Resting Electrocardiographic Results (restecg)

Categorical variable with options: 0 for Normal, 1 for ST-T wave abnormality, and 2 for Left ventricular hypertrophy.

Maximum Heart Rate Achieved (thalach)

Maximum heart rate achieved during a stress test, crucial for assessing cardiac function.

Exercise-Induced Angina (exang)

Presence of angina during exercise, coded as 1 for yes and 0 for no.

Oldpeak

ST depression during exercise relative to rest, indicating potential cardiac issues.

Slope of Peak Exercise ST Segment (slope)

Categorical variable with values: 0 for Upsloping, 1 for Flat, and 2 for Downsloping.

Number of Major Vessels Colored by Fluoroscopy (ca)

Numeric variable ranging from 0 to 3, indicating the number of major vessels colored by fluoroscopy.

Heart Attack Risk Prediction Models

Logistic Regression vs Random Forest

Logistic Regression vs Random Forest

- Accuracy: 86% vs 94%
- Precision: 84% vs 92%
- Recall: 85% vs 96%
- Specificity: 87% vs 93%
- Confusion Matrix: TP: 90, FP: 17, FN: 16, TN: 115 vs TP: 98, FP: 9, FN:
- Strengths: Interpretable coefficients vs Superior performance metrics
- Weaknesses: Limited in capturing non-linear relationships vs Higher

Vs

Logistic Regression vs Random Forest

- Accuracy: 86% vs 94%
- Precision: 84% vs 92%
- Recall: 85% vs 96%
- Specificity: 87% vs 93%
- Confusion Matrix: TP: 90, FP: 17, FN: 16, TN: 115 vs TP: 98, FP: 9, FN:
- Strengths: Interpretable coefficients vs Superior performance metrics
- Weaknesses: Limited in capturing non-linear relationships vs Higher

Clinical Implications

Early Detection

The Random Forest model shows promise for early diagnosis of heart disease, enabling timely interventions and potentially reducing the severity of outcomes.

Decision Support

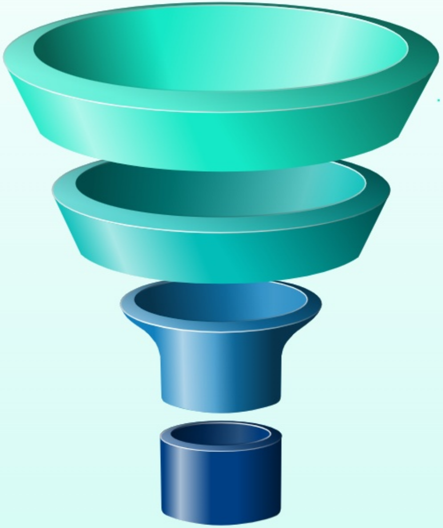
Integrating these models into clinical practice can enhance decision-making processes, improving diagnostic accuracy and leading to better patient outcomes.

Resource Allocation

Predictive analytics aids in prioritizing high-risk patients, ensuring that healthcare resources are allocated efficiently and effectively.

Recommendations

Enhancing Predictive Healthcare Models



Integration with EHR

Incorporate the models into electronic health records for real-time prediction and decision-making.

Model Refinement

Explore advanced techniques such as deep learning for further enhancement of predictive performance.

External Validation

Validate the models on external datasets to ensure broader applicability.

Personalized Predictions

Develop tailored models that account for individual patient history and lifestyle factors.

Key Findings on Heart Attack Risk Prediction

Comparative Analysis of Random Forest and Logistic Regression

01

Superior Performance of Random Forest

Random Forest outperformed Logistic Regression in predicting heart attack risk, demonstrating higher accuracy and reliability.

02

Alignment with Previous Research

Findings are consistent with studies by Xinyi Liu et al. (2023) and G. Wilson and B. Thomas (2020), affirming the effectiveness of ensemble methods.

03

Importance of Feature Engineering and Explainable AI

The study underscores the critical role of feature engineering and techniques like SHAP in enhancing model interpretability and accuracy.

04

Implications for Clinical Decision-Making

Random Forest provides a robust tool for early intervention and clinical decisions, supporting advanced machine learning integration in healthcare.

Predicting Heart Attack Risk Using Machine Learning: A Comparative Study of Logistic Regression and Random Forest Algorithms