Heart Failure Prediction and Analysis Project

Dharmik Shah

Contents

/01

Introduction

Project Overview

Objective of the Project

The project aims to analyze clinical records of heart failure patients to develop predictive insights, enabling early intervention and treatment by medical professionals.

Importance of Heart Failure Prediction

Early prediction of heart failure can significantly reduce mortality rates by allowing timely medical interventions and personalized treatment plans.

Target Audience

The target audience includes healthcare professionals, researchers, and hospital administrators who can utilize the insights for patient care and resource management.

Dataset Description

Dataset Overview

The dataset contains 299 patient records with 13 features, including clinical and lifestyle factors, to predict heart failure events.

Key Features

Key features include age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and death event.

Data Structure

The dataset is structured with each row representing a patient and columns representing clinical and lifestyle features, with a binary target variable indicating death events.

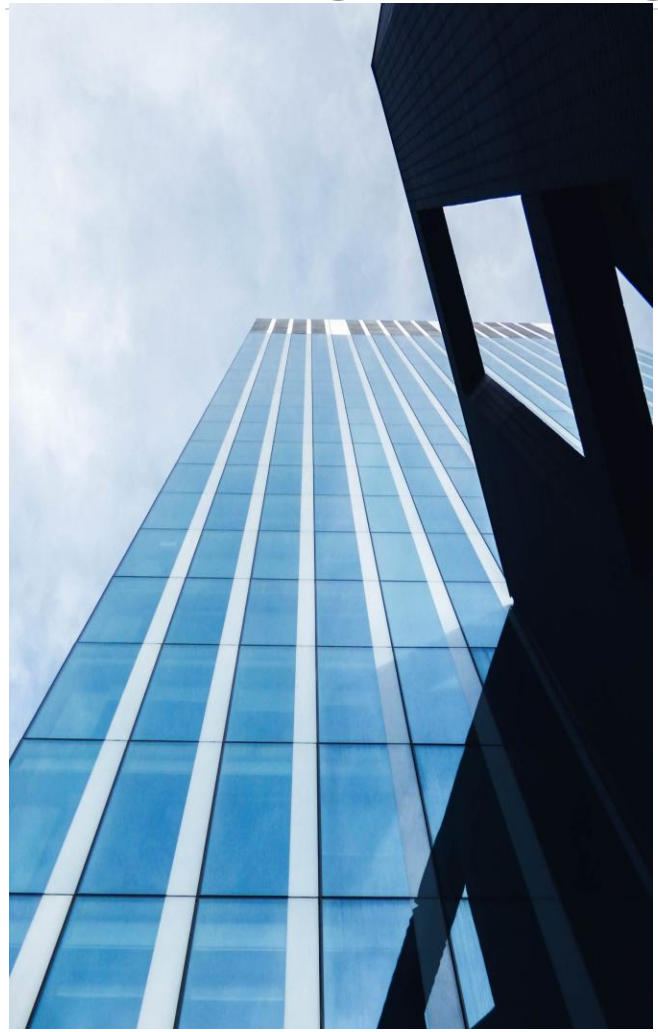
Data Sources

The data is sourced from clinical records, providing a comprehensive view of patient health and lifestyle factors influencing heart failure.

/02

Data Preprocessing

Handling Missing Values



01

Techniques for Missing Data

Techniques include imputation using mean/median for numerical data and mode for categorical data, or removing rows with significant missing values.

02

Impact on Analysis

Proper handling of missing values ensures the integrity of the dataset, preventing biased or inaccurate model predictions.

03

Example Cases

For example, missing serum creatinine levels can be imputed using the median value to maintain the dataset's statistical properties.

Feature Standardization

Standardization Techniques

Techniques include z-score normalization and min-max scaling to bring all numerical features to a common scale.

Benefits of Standardization

Standardization improves model performance by ensuring that all features contribute equally to the analysis, especially in distance-based algorithms.

Implementation in Python

Using Scikit-learn's StandardScaler or MinMaxScaler to standardize features like age, creatinine phosphokinase, and serum sodium.

Encoding Categorical Variables

Encoding Methods

Methods include one-hot encoding for nominal variables and label encoding for ordinal variables to convert categorical data into numerical formats.

Impact on Model Performance

Proper encoding ensures that categorical variables are correctly interpreted by machine learning models, improving accuracy and predictive power.

Example: Anaemia and Diabetes

Anaemia and diabetes are encoded as binary variables (0 or 1), allowing models to process these features effectively.

/03

Exploratory Data Analysis (EDA)

Statistical Analysis

Key Insights

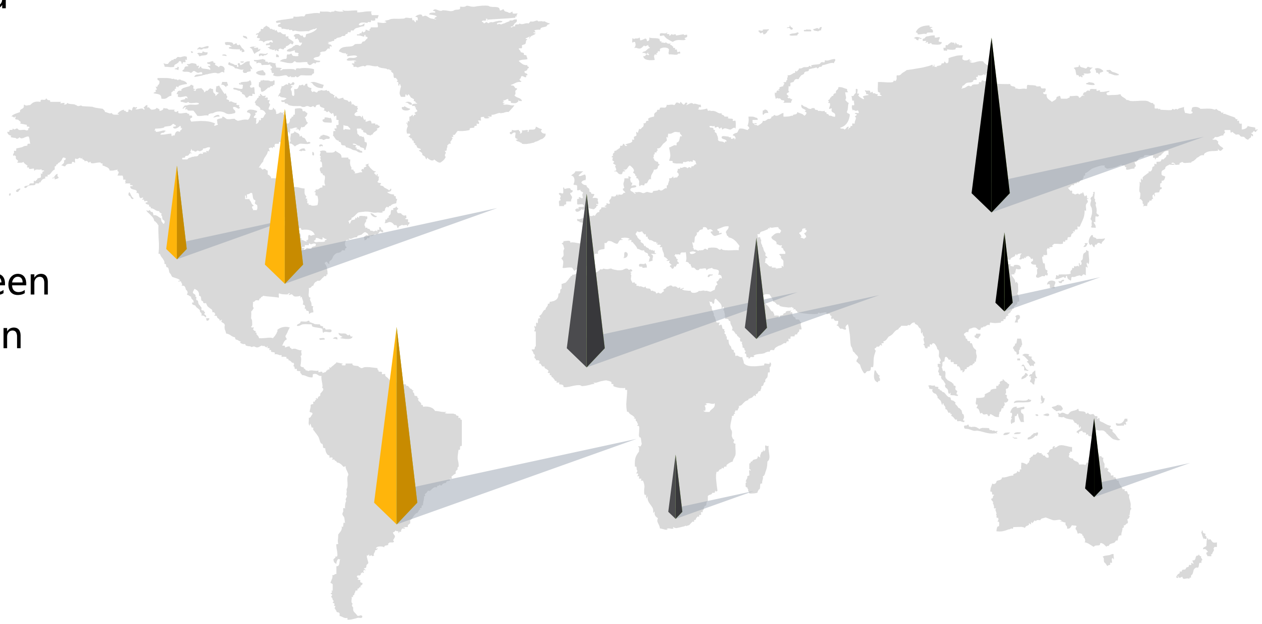
Key insights include identifying high-risk groups based on features like low ejection fraction and high serum creatinine levels.

Correlation Analysis

Correlation analysis reveals relationships between features, such as the strong correlation between serum creatinine levels and death events.

Descriptive Statistics

Descriptive statistics provide insights into the central tendency, dispersion, and distribution of features like age, ejection fraction, and serum creatinine.



Data Visualization

01

Histograms and Box Plots

Histograms show the distribution of numerical features like age, while box plots highlight outliers in features like creatinine phosphokinase.

02

Scatter Plots

Scatter plots visualize relationships between features, such as serum creatinine vs. ejection fraction, with color coding for death events.

03

Heatmaps

Heatmaps display correlation matrices, helping identify strong positive or negative relationships between clinical features and outcomes.

04

Insights from Visualizations

Visualizations reveal patterns, such as higher death rates in patients with elevated serum creatinine and low ejection fraction.

/04

Predictive Modeling

Model Selection

Logistic Regression

01

Logistic regression is used for binary classification, predicting the likelihood of heart failure based on clinical and lifestyle features.

Random Forest is employed for its ability to handle non-linear relationships and provide feature importance rankings.

02

Random Forest

Support Vector Machines

03

SVM is chosen for its effectiveness in high-dimensional spaces, making it suitable for datasets with multiple features.

Neural networks are utilized for their capacity to model complex relationships and improve prediction accuracy with sufficient data.

04

Neural Networks

Model Evaluation

01 Accuracy and Precision

Accuracy measures the overall correctness of predictions, while precision focuses on the proportion of true positive predictions among all positive predictions.

02 Recall and F1-Score

Recall measures the ability to identify all positive cases, and F1-score balances precision and recall, providing a single metric for model performance.

03 ROC-AUC Analysis

ROC-AUC evaluates the model's ability to distinguish between classes, with higher values indicating better performance.

Feature Importance

Random Forest Feature Importance

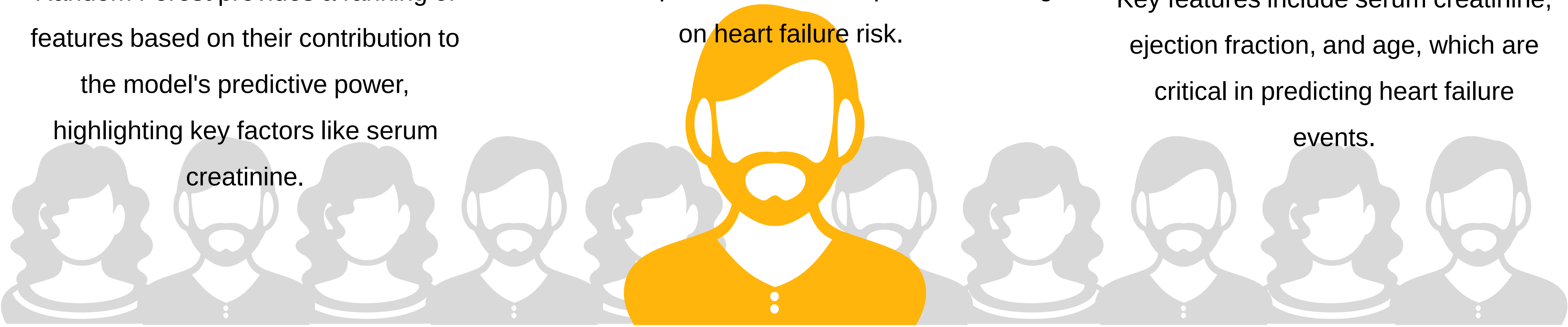
Random Forest provides a ranking of features based on their contribution to the model's predictive power, highlighting key factors like serum creatinine.

SHAP Values

SHAP values explain individual predictions, showing how each feature contributes to the model's output, such as the impact of smoking on heart failure risk.

Key Features Identified

Key features include serum creatinine, ejection fraction, and age, which are critical in predicting heart failure events.



/05

Dashboard Development

Dashboard Purpose

01 **Visual Interface Overview**

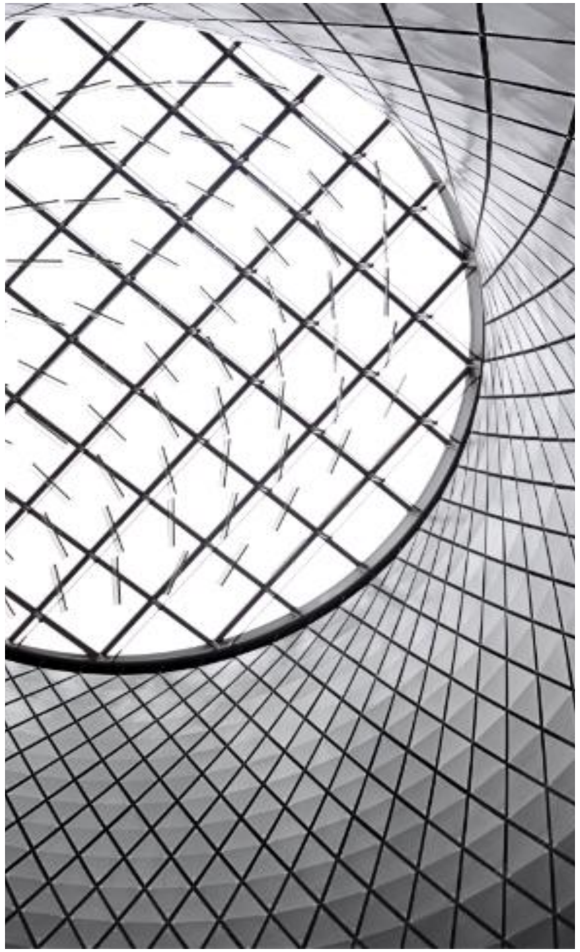
The dashboard provides a visual interface to explore key metrics and insights, making complex data accessible to non-technical stakeholders.

02 **Stakeholder Benefits**

Stakeholders benefit from real-time data access, enabling informed decision-making and improved patient care strategies.

03 **Real-Time Monitoring**

Real-time monitoring allows healthcare professionals to track patient outcomes and adjust treatments promptly based on the latest data.



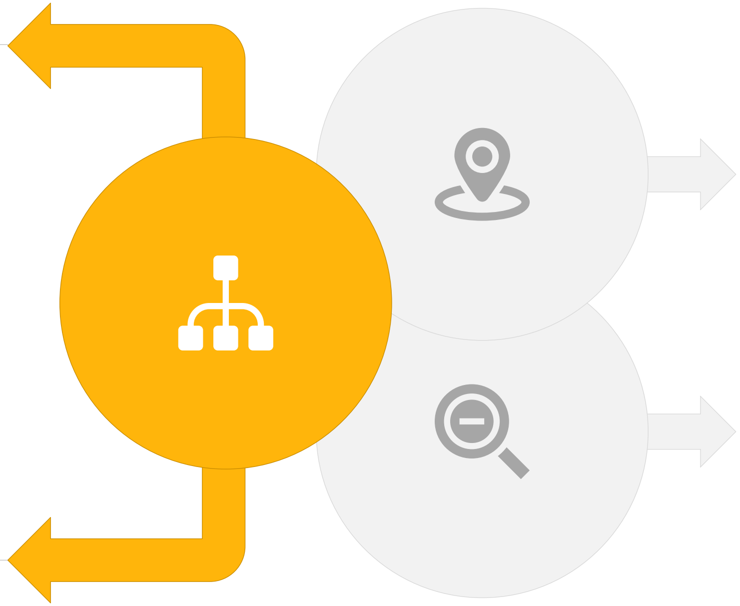
Dashboard Features

Patient Demographics

The dashboard includes visualizations of age distribution and gender analysis to identify demographic trends in heart failure cases.

Mortality Trends

Mortality trends are visualized to show the proportion of death events and identify high-risk periods during patient follow-up.



Clinical Metrics

Clinical metrics like serum creatinine and ejection fraction are displayed to highlight patients at risk of heart failure.

Lifestyle Factors

Lifestyle factors such as smoking habits and diabetes prevalence are analyzed to assess their impact on heart failure risk.

Interactive Elements

Filters and Customization

Interactive filters allow users to customize views by age, gender, or clinical metrics, enabling targeted analysis of specific patient groups.

Example Visualizations

Example visualizations include bar charts for death events by age, pie charts for gender distribution, and scatter plots for serum creatinine vs. ejection fraction.

Predictive Insights

The dashboard provides predictive insights, displaying personalized risk scores for heart failure based on patient data.

/06

Insights and Recommendations

Key Findings

01 **Correlation Analysis Results**

Correlation analysis reveals significant relationships between features like serum creatinine, ejection fraction, and death events.

02 **High-Risk Patient Groups**

High-risk groups include patients with elevated serum creatinine levels, low ejection fractions, and older age.

03 **Impact of Lifestyle Factors**

Lifestyle factors such as smoking and diabetes significantly increase the risk of heart failure, highlighting the need for targeted interventions.

Actionable Recommendations

1

Early Intervention Strategies

Early intervention strategies should focus on high-risk patients, with regular monitoring of serum creatinine and ejection fraction levels.

2

Resource Allocation

Resources should be allocated to high-risk groups, ensuring timely access to treatments and reducing the burden on healthcare systems.

3

Future Research Directions

Future research should explore additional data sources and advanced modeling techniques to enhance prediction accuracy and uncover new risk factors.

Next Steps

Data Enrichment

Enriching the dataset with more patient records and additional features will improve the robustness of the predictive models.

Model Enhancement

Model performance can be enhanced through hyperparameter tuning and the integration of additional clinical data.

Deployment Strategies

Deployment strategies include using Flask or Streamlit to create a web-based application for real-time heart failure risk predictions.

Thank you for listening.

Dharmik Shah