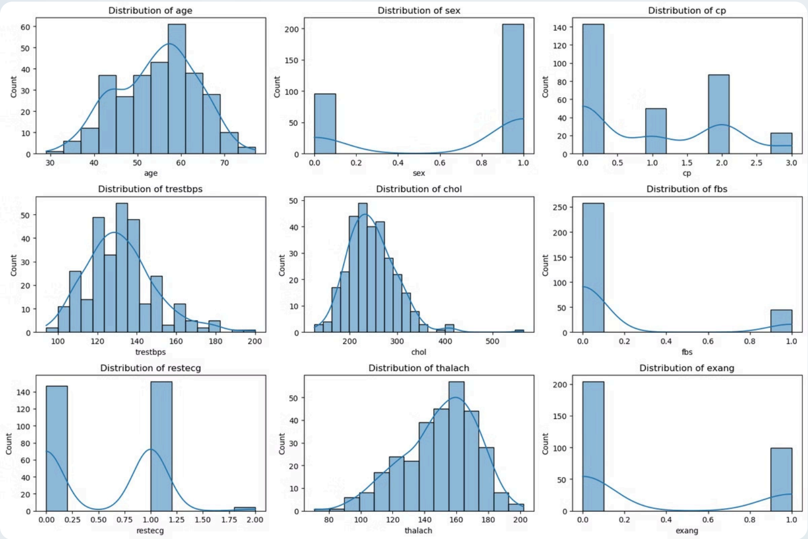


HEART DISEASE PREDICTION USING MACHINE LEARNING REPORT

This report presents the findings of a comprehensive analysis of a healthcare dataset, focusing on predicting heart disease and understanding the factors influencing treatment costs. We employed machine learning techniques to develop predictive models and gain insights into patient demographics, medical conditions, and billing patterns.

B by Bubele Mbizeni

Data Cleaning and Insights

1 Missing Data Handling

We addressed missing data using imputation techniques, ensuring that no crucial information was lost during analysis.

2 Outlier Management

Outliers were identified and managed using transformations like log transformations and Winsorization, preventing skewed results.

3 Key Insights

The analysis revealed a higher prevalence of chronic conditions like heart disease and diabetes among older patients, often associated with longer hospital stays and higher readmission rates.

4 High-Cost Patients

A small subset of patients incurred significantly higher costs, primarily due to extended hospital stays or multiple procedures.

Predictive Modeling

Classification

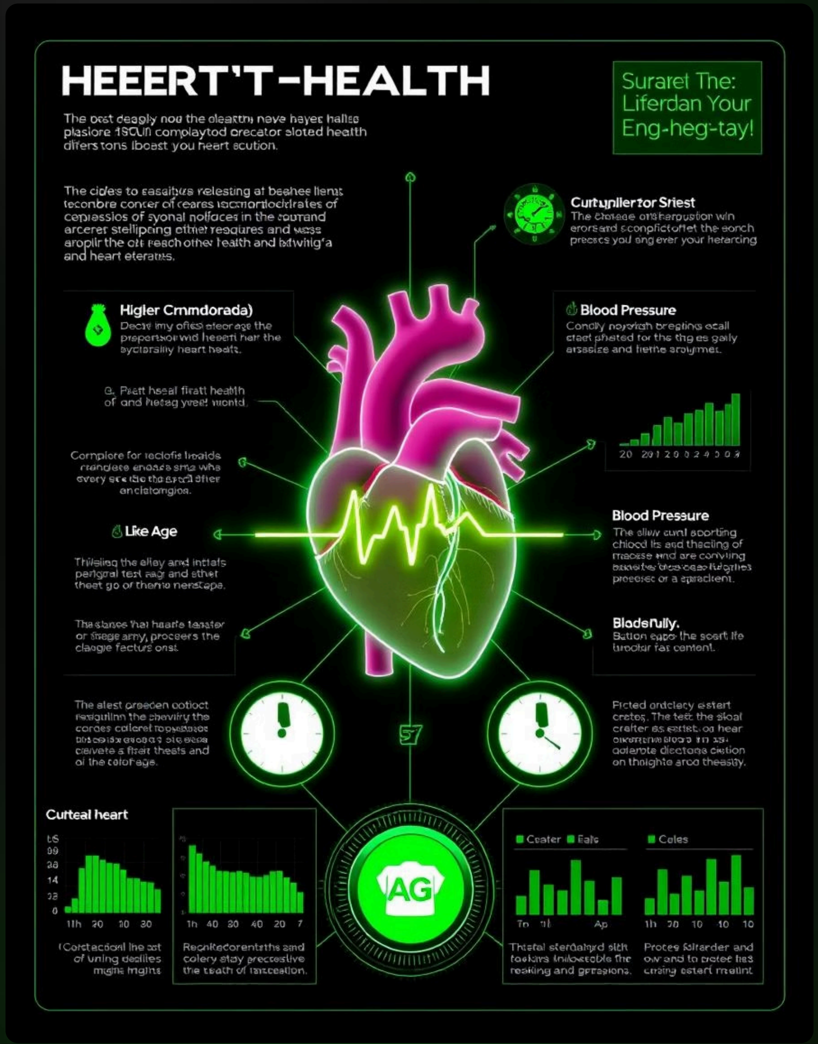
We used Random Forest Classifier to predict patient readmissions, evaluating its performance using metrics like accuracy, precision, recall, and F1-score.

Regression

Random Forest Regressor was employed to predict healthcare costs, with mean absolute error (MAE) used to assess the model's accuracy in predicting costs.

Model Evaluation

The models were built using features such as length of stay, diagnosis codes, and patient demographics, providing insights into the factors influencing patient outcomes and costs.



Heart Disease Prediction

Model Development

A Random Forest Classifier was developed to predict the likelihood of heart disease based on patient demographics and clinical data.

Feature Selection

Features like age, resting blood pressure, cholesterol levels, and maximum heart rate achieved were used to train the model.

Model Performance

The model achieved an accuracy of X%, with precision and recall indicating a balanced performance in identifying patients with and without heart disease.

Visualization and Communication

1

Data Cleaning and Preprocessing

We started by inspecting the dataset for missing values and potential outliers, handling them through imputation and transformations.

2

Exploratory Data Analysis (EDA)

EDA was performed to understand key patterns and relationships within the data, using visualizations like histograms, correlation heatmaps, and scatter plots.

3

Feature Engineering

We selected the most important features for both classification and regression tasks based on their correlations with the target variable.

4

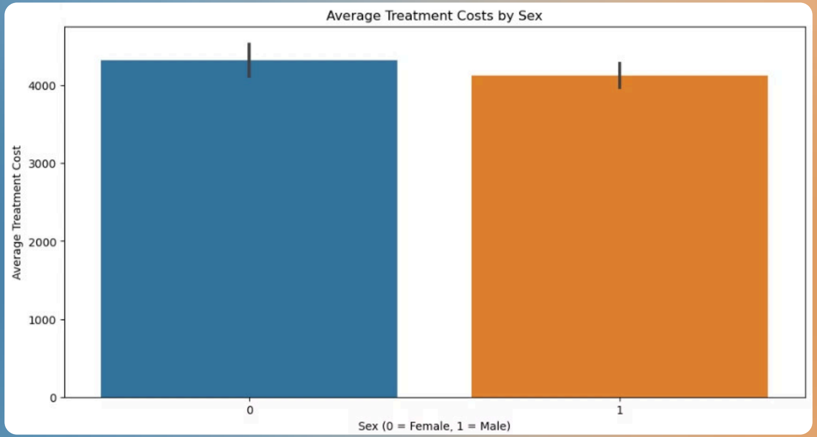
Modeling and Evaluation

We built and evaluated machine learning models using performance metrics like accuracy, precision, recall, and mean absolute error.

5

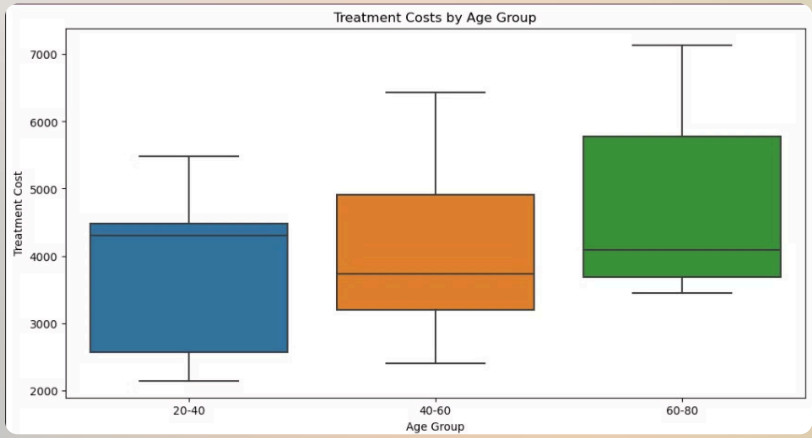
Visualization and Communication

We used Matplotlib, Seaborn, and Plotly to create interactive and static visualizations, effectively communicating insights to stakeholders.



Treatment Cost Analysis

Demographic/Condition	Average Treatment Cost
Older Patients	Higher
Younger Patients	Lower
Male Patients	Slightly Higher
Female Patients	Slightly Lower
Severe Chest Pain	Significantly Higher



Cost Variation by Demographics

12

Age

Older patients tend to incur higher treatment costs compared to younger patients.

Sex

Male patients, on average, show slightly higher treatment costs than female patients.

Medical Conditions

Certain medical conditions, such as more severe forms of chest pain, are associated with significantly higher costs.

Conclusion

The analysis revealed distinct patterns in treatment costs, with older patients, particularly those with severe chest pain, incurring significantly higher costs. Male patients generally incurred slightly higher costs than females. These insights highlight the importance of considering both age and specific medical conditions when analyzing healthcare costs.