Predicting Patient Readmission Rates in Hospitals: A Healthcare Analytics Approach

Malcolm A. Phillip

Northwest Missouri State University, Maryville MO 64468, USA
S556427@nwmissouri.edu

Abstract. Explores the critical role of predicting patient readmission rates in hospitals and its impact on healthcare management, emphasizing resource allocation, patient care, and cost reduction. The study employs advanced healthcare analytic, utilizing machine learning models and statistical analysis to uncover intricate patterns and risk factors related to heart disease-related readmission. The methodology and key findings highlight the correlation between Heart Health Score and Chest Pain Level, revealing how chest pain influences overall heart health. The Risk Score examination indicates a higher likelihood of readmission, emphasizing the need for targeted interventions. The study identifies high-risk patients as males aged between 40 and 54 years, providing a targeted demographic for healthcare interventions. The results lay the foundation for implementing precise strategies to reduce readmission and enhance patient outcomes, advocating for the integration of data-driven approaches in hospitals for optimal resource utilization.

Keywords: mobile computing · mobile usability · multimedia · cloud computing

1 Introduction

In today's healthcare landscape, the integration of predictive modeling through healthcare analytics stands as a cornerstone in enhancing patient outcomes and optimizing resource utilization. Predicting patient readmission rates in hospitals plays a pivotal role in healthcare management, impacting resource allocation, patient care, and cost reduction. This introduction delves into the domain of healthcare analytic in predictive modeling, outlining the objectives, data sources, problem analysis, project implementation steps, key components, and limitations of the proposed study. The Python notebooks used for this project are available at `https://github.com/MPhillipS556427/Heart_Disease_Healthcare-Analytics-Project`

1.1 Importance of Predicting Patient Readmission Rates

Predicting patient readmission rates is a problem that healthcare organizations throughout the globe are facing, especially when it comes to heart disease cases. This project is important because it has the ability to completely transform healthcare administration. Hospitals may provide timely and individualized patient care by strategically allocating resources based on realistic forecasts of readmission rates. This increases patient outcomes while also making healthcare services more financially viable by cutting costs dramatically. By using sophisticated healthcare analytic techniques, complex trends and risk factors associated with readmission due to heart disease can be found, which can result in the development of customized interventions and policy.

1.2 Objectives of the Report

Domain Selection and Rationale: Considering heart disease-related readmission frequently occur and have an effect on healthcare systems, the study focuses on them. This is a critical domain considering heart disorders are among the top causes for hospitalizations according to the Journal of Medical Research [5]. The main objective of this report is to look into and examine readmission associated to heart disease utilizing predictive modeling methods.

1.3 Problem Analysis:

Identify and evaluate various characteristics and patterns in order to forecast heart disease-related readmission. Discover distinctive tendencies to identify fresh insights that can direct targeted interventions. In particular, the objective goals consist of examining the following 3:

- **Predicting Heart Disease-Related Readmission Using Machine Learning Models:** Develop and compare different machine learning models to predict heart disease-related readmission based on the provided data attributes. Evaluate the accuracy, precision, recall, and F1-score of each model to identify the most effective predictive model for readmission.
- **Analyzing Heart Health Score and Chest Pain Level Relationship:** Examine the relationship between Heart Health Score and Chest Pain Level in patients with heart disease. Analyze the correlation, trends, and patterns between these two attributes to understand how chest pain level affects the overall heart health score.
- **Analyzing Risk Score and Its Impact on Readmission:**Analyze the Risk Score derived from the given data attributes and its impact on heart disease-related readmission. Investigate how patients with higher risk scores are more likely to be readmitted, and explore potential interventions or strategies to reduce readmission in high-risk patients.

1.4 Data Sources:

The data for this study are being sourced from the UCI Machine Learning Repository [1]. This data-set offers comprehensive information about various factors contributing to heart disease, providing a robust foundation for analysis and prediction.

- **Data Content:** [6] See Table 1 below for Data Attributes.

Table 1. Data Content

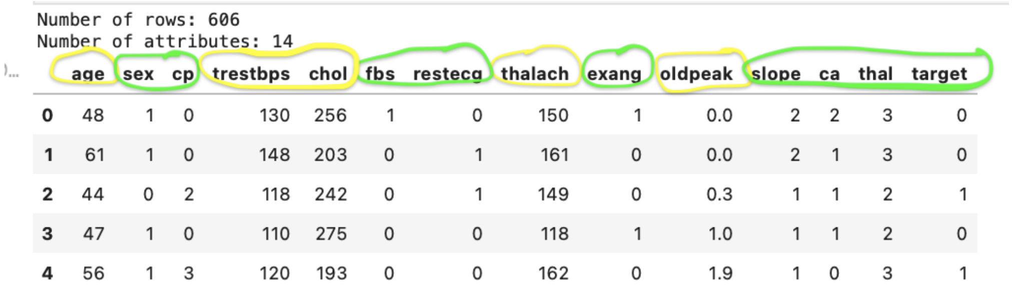
Attribute	**Description**
age	Age of the patient in years
sex	Male/Female
cp	Chest pain type (Typical, Atypical, Non-anginal, Asymptomatic)
trestbps	Resting blood pressure (mm Hg)
chol	Serum cholesterol (mg/dl)
fbs	Fasting blood sugar ¿ 120 mg/dl (True/False)
restecg	Resting electrocardiographic results
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (True/False)
oldpeak	ST depression induced by exercise
slope	Slope of the peak exercise ST segment
ca	Number of major vessels (0-3) colored by fluoroscopy
thal	Thalassemia type
target	Predicted attribute (Presence or absence of heart disease)

- **Key Components and Limitations:** Preprocessing of the data, model selection, and evaluation metrics are among the components. Understanding limitations is necessary for comprehending the outcomes. These factors include potential biases in the data-set and the limitations of predictive modeling.

2 Methodology

2.1 Data Collection

This Capstone project utilizes data from Kaggle and UCI Machine Learning Repository to present an overview of Healthcare Analytic in Predictive Modeling. The primary data source is a structured data-set in CSV format, organized in a tabular form with defined columns and rows. The data-set comprises 606 records, each with 14 fields. Notably, there is no date or address/location information in the data-set. The data consists of a combination of numeric and categorical fields. As shown in Figure 1 Numeric fields are circled in YELLOW, and categorical fields are circled in **GREEN**.



```
Number of rows: 606
Number of attributes: 14
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	48	1	0	130	256	1	0	150	1	0.0	2	2	3	0
1	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
2	44	0	2	118	242	0	1	149	0	0.3	1	1	2	1
3	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
4	56	1	3	120	193	0	0	162	0	1.9	1	0	3	1

Fig. 1. Description of the data-set type.

Due to its organized structure and accessibility, there is no need for data scraping techniques; the data-set can be easily downloaded from the website in CSV format at UCI Machine Learning Repository. Alternatively, you can also find it on Kaggle.

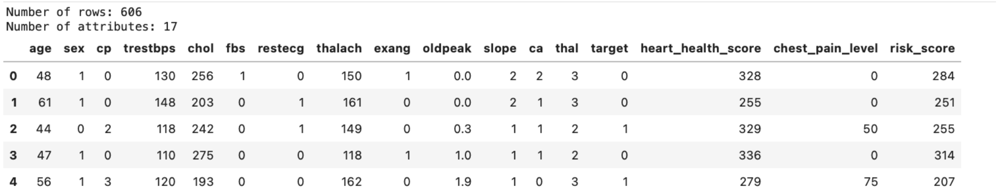
2.2 Data Preprocessing/Cleaning

- **Data Cleaning and Transformation:** The data is fairly tidy and doesn't require any curating or cleaning; nonetheless, the given attributes generated three new features **Heart Health Score, Chest Pain Level and Risk Score** which plays a pivot role within the research.
  - **Heart Health Scores:** Calculating Heart Health Scores helps in identifying high-risk patients, allowing healthcare providers to prioritize their care and allocate resources effectively. Timely intervention for high-risk patients can significantly reduce readmission rates and improve patient outcomes.
  - **Chest Pain Levels:** Analyzing chest pain levels provides insights into the symptoms experienced by patients. Understanding the relationship between chest pain levels and readmission can help healthcare professionals tailor their treatment plans and interventions based on the severity of symptoms.
  - **Predicting Readmission Risk:** Building a predictive model to estimate readmission risk scores enables hospitals to proactively identify patients at higher risk of readmission. By focusing on personalized care and interventions for these patients, hospitals can reduce readmission rates, enhance patient satisfaction, and optimize resource utilization.

  The research aims to analyze and model all 17 attributes to predict heart disease-related readmission, understand the relationship between heart health score and chest pain level, and analyze the impact of the risk score on readmission.
- **Tools & Techniques:** Specific tools and techniques was required in generating the three new features. Pandas and NumPy was used for creating the new features in JupyterLab, while Scikit-Learn will be employed for machine learning tasks for feature scaling and model training. Feature scaling is essential because the attributes have varying units and scales. The follow techniques Min-Max scaling and Standardization will be applied to ensure that all features are brought to a consistent scale. Below is the coding that created the three new features/ attributes and examining the new data-set:

```
## Import the following:
import pandas as pd
import numpy as np
import hashlib
import folium
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import os
import haversine
from folium.plugins import HeatMap
from folium import PolyLine
import sqlite3
from shapely.geometry import Point
from scipy.stats import linregress

 ## 2. Creating Three extra Features/Attributes.
# Load the heart disease data from the CSV file
df = pd.read_csv('heart_disease_data.csv')

# New Feature/Attribute 1: Heart Health Score
df['heart_health_score'] = 100 - df['age'] + df['thalach'] - df['trestbps'] + df['chol']

# New Feature/Attribute 2: Chest Pain Level
df['chest_pain_level'] = df['cp'] * 25  # Assuming cp values are on a scale of 0 to 3

# New Feature/Attribute 3: Risk Score
df['risk_score'] = df['age'] + df['trestbps'] - df['thalach'] + df['chol']

# Saving the updated dataset with the new Features/Attributes
df.to_csv('heart_disease_dataset_with_new_features.csv', index=False)

## 3. Examining the new dataset.
Heart_data_V1 = 'heart_disease_dataset_with_new_features.csv'
data = pd.read_csv(Heart_data_V1)
num_rows, num_attributes = data.shape
print(f'Number of rows: {num_rows}')
print(f'Number of attributes: {num_attributes}')
(data.head())
```

As shown in Figure 2, the updated data-set includes new data points, the rows are the same; however, it now has **17** attributes due to the three new features.



```
Number of rows: 606
Number of attributes: 17
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	heart_health_score	chest_pain_level	risk_score
0	48	1	0	130	256	1	0	150	1	0.0	2	2	3	0	328	0	284
1	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0	255	0	251
2	44	0	2	118	242	0	1	149	0	0.3	1	1	2	1	329	50	255
3	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0	336	0	314
4	56	1	3	120	193	0	0	162	0	1.9	1	0	3	1	279	75	207

Fig. 2. Description of the updated data-set.

- **Handling Missing Data:** The data-set currently do not have any missing data to handle. In the event that the data-set for this research had any missing values, imputation strategies would be employed for numerical features. Missing values in these features would be filled using the mean, median, or mode of their respective columns, preserving the overall statistical characteristics of the data-set. For missing values in the categorical features would be filled with the mode (most frequent value) of the respective columns. Mode imputation is suitable for maintaining the categorical distribution.
  - **Dependent and Independent variables:** The dependent variable ”Target” and ”Heart Health Score” are being utilized to indicate the probability of patient readmission. By examining the connections between these variables and readmission rates, the predictive model can detect notable patterns

and risk factors. This, in turn, empowers hospitals to adopt data-driven approaches for healthcare management. See Table 2 for a full list of Independent and Dependent variables.

Table 2. Dependent and Independent variables

Independent variables	**Dependent variable**
age	Target
sex	Heart Health Score
cp	-
trestbps	-
chol	-
fbs	-
restecg	-
thalach	-
exang	-
oldpeak	-
slope	-
ca	-
thal	-
Chest Pain Level	-
Risk Score	-

2.3 Exploratory Data Analysis

– The key element of any data science or data analytics work is exploratory data analysis, or EDA. This project involves assessing data-sets both statistically and graphically in order to recognize trends, find abnormalities, and address the objectives of the report. EDA assists in comprehending the data, reveals correlations between variables, and directs the modeling and feature engineering procedures. It is crucial because it offers a framework for deliberating wisely on model selection, hypothesis testing, and data preprocessing.
– Descriptive Statistics, Data Visualization, Correlation Analysis and Handling Missing Data are a few EDA techniques. For this project data visualization and correlation analysis techniques are particularly relevant. Visualization methods is being used to display the distribution of heart health scores, chest pain levels, and risk scores, while correlation analysis reveal relationships between these variables.
– Graphs (BOX PLot) are being used to visualize the distribution of Heart Health Scores, Chest Pain Levels, and Risk Scores. This helps identify outliers and the spread of these attributes. Scatter plots are be employed to explore relationships between Heart Health Scores and Chest Pain Levels, allowing for the identification of potential patterns. Calculate correlation coefficients between Heart Health Scores, Chest Pain Levels, and Risk Scores to determine if any significant relationships exist. A correlation matrix can visually represent these relationships. For additional data visualization on heart disease please visit CDCHeartDisease.
– In this phase EDA reveals whether there are clear patterns between Heart Health Scores and Chest Pain Levels. For instance, it may show that higher chest pain levels correspond to lower heart health scores, indicating a potential correlation between severe symptoms and overall heart health. By analyzing the distribution of Risk Scores and their impact on readmission, the EDA phase will highlight whether patients with higher risk scores are more likely to be readmitted. This insight can guide the subsequent predictive modeling phase, emphasizing the importance of addressing high-risk patients to reduce readmission rates.

3 Predictive Modeling

– Figure 3 displays the pipeline that is being used to accomplish the predictive application for the report's 3 main objective goals.
  - **Data Preprocessing** involves transforming the raw data to make it suitable for the model training. It includes handling missing values, encoding categorical variables, and scaling numerical features.

- **Problem-Specific Preprocessing** involves 3 new features being created in order to address the reports objectives.
- **Train-Test Split and Model Selection** involves the data being split into training and testing sets to train and evaluate the predictive models.
- **Model Training and Evaluation**; trains the selected model on the training set and evaluated on the testing set. Evaluation metrics: accuracy, precision, recall, and F1-score are calculated to assess the model performance.
- **Statistical Analysis and Visualization:** For objective 2 and 3, statistical analysis and visualization are performed to understand the relationships and patterns in the data. This step includes calculating correlations and creating visualizations.
- **Interventions and or Strategies:** Based on the analysis results, interventions are proposed for addressing the specific challenges outlined in the report's objective.

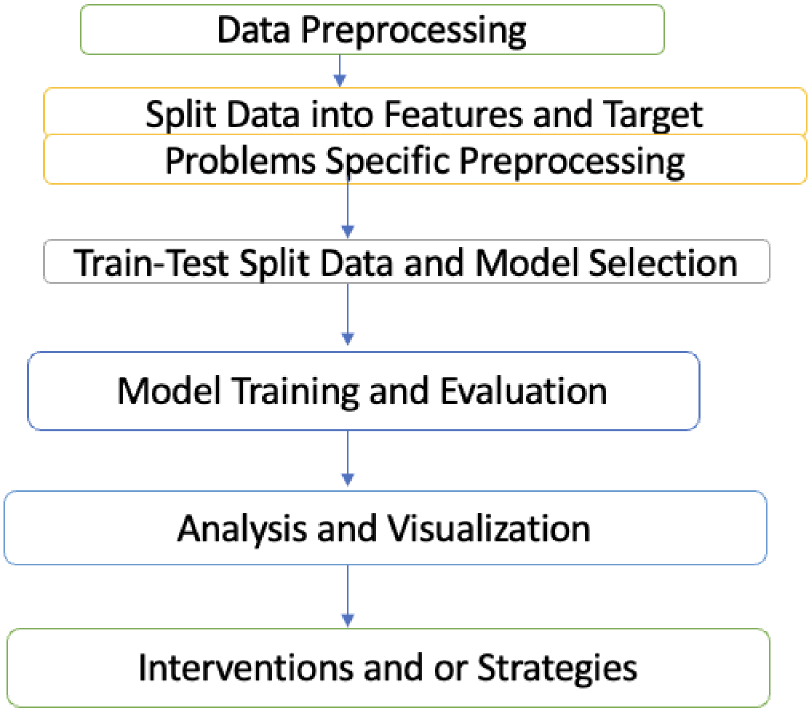
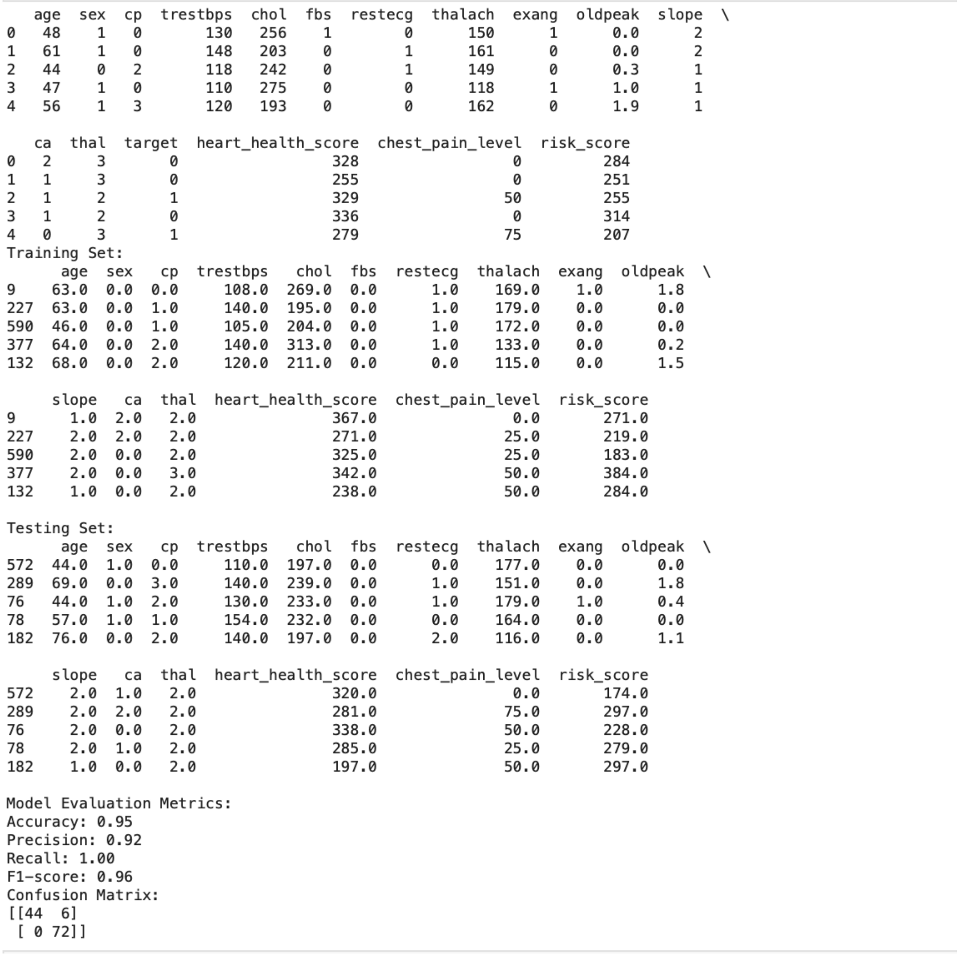

Data Preprocessing
Split Data into Features and Target
Problems Specific Preprocessing
Train-Test Split Data and Model Selection
Model Training and Evaluation
Analysis and Visualization
Interventions and or Strategies

Fig. 3. Pipeline for Predictive Application.

– Per Microsoft Azure [2], machine learning algorithms are computational models that can learn patterns and make predictions or decisions based on input data. They are used to analyze data, identify patterns, and make predictions or classifications without explicit programming. Random Forest Classifier is being utilize for objective 1 Predicting Heart Disease-Related Readmission, for objective 2 Analyzing Heart Health Score and Chest Pain Level Relationship, no machine learning algorithm is used, instead, statistical and visualization techniques will be applied. And no traditional machine learning algorithm is used, instead, a simple formula is applied to calculate the Risk Score for objective 3 Analyzing Risk Score and Its Impact on Readmission.

3.1 Performance Evaluation Metrics:

- See Appendix A for the Implementation and Evaluation Process of the Analysis.
- For **Objective 1**(Predicting Heart Disease-Related Readmission Using Machine Learning Models) the training and testing process involves preparing the data, splitting it into training and testing sets, selecting and training the Random Forest Classifier model, evaluating it's performance, comparing the results, and ultimately selecting the best model for predicting heart disease-related readmission based on the accuracy, precision, recall, and F1-score. See Heart Disease Healthcare Analytics GitHub Project for testing and training sets coding explanation. Also refer to **APPENDIX D** for coding breakdown.



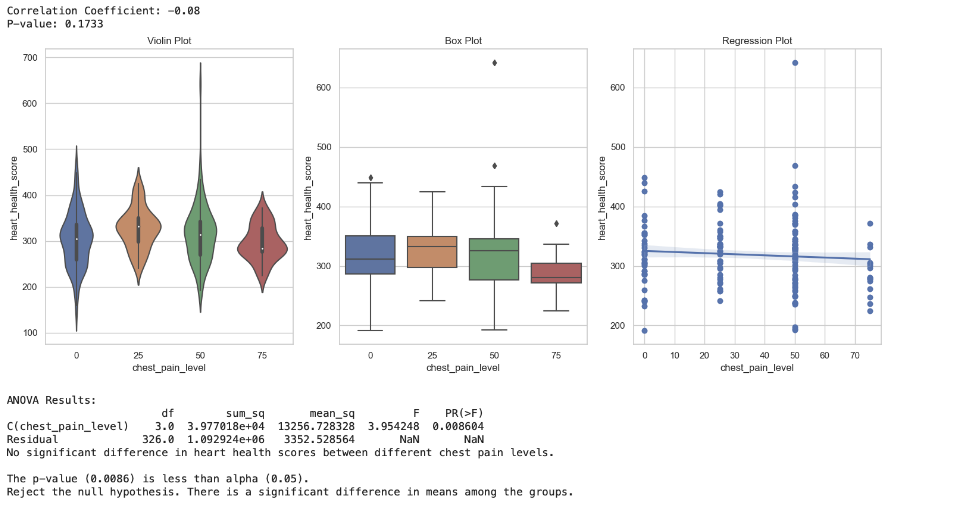
```
   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  \
0   48    1   0       130   256    1        0      150      1      0.0      2
1   61    1   0       148   203    0        1      161      0      0.0      2
2   44    0   2       118   242    0        1      149      0      0.3      1
3   47    1   0       110   275    0        0      118      1      1.0      1
4   56    1   3       120   193    0        0      162      0      1.9      1

   ca  thal  target  heart_health_score  chest_pain_level  risk_score
0   2     3       0                 328                 0         284
1   1     3       0                 255                 0         251
2   1     2       1                 329                50         255
3   1     2       0                 336                 0         314
4   0     3       1                 279                75         207
Training Set:
      age  sex   cp  trestbps   chol  fbs  restecg  thalach  exang  oldpeak  \
9    63.0  0.0  0.0     108.0  269.0  0.0      1.0    169.0    1.0      1.8
227  63.0  0.0  1.0     140.0  195.0  0.0      1.0    179.0    0.0      0.0
590  46.0  0.0  1.0     105.0  204.0  0.0      1.0    172.0    0.0      0.0
377  64.0  0.0  2.0     140.0  313.0  0.0      1.0    133.0    0.0      0.2
132  68.0  0.0  2.0     120.0  211.0  0.0      0.0    115.0    0.0      1.5

     slope   ca  thal  heart_health_score  chest_pain_level  risk_score
9      1.0  2.0   2.0               367.0               0.0       271.0
227    2.0  2.0   2.0               271.0              25.0       219.0
590    2.0  0.0   2.0               325.0              25.0       183.0
377    2.0  0.0   3.0               342.0              50.0       384.0
132    1.0  0.0   2.0               238.0              50.0       284.0

Testing Set:
      age  sex   cp  trestbps   chol  fbs  restecg  thalach  exang  oldpeak  \
572  44.0  1.0  0.0     110.0  197.0  0.0      0.0    177.0    0.0      0.0
289  69.0  0.0  3.0     140.0  239.0  0.0      1.0    151.0    0.0      1.8
76   44.0  1.0  2.0     130.0  233.0  0.0      1.0    179.0    1.0      0.4
78   57.0  1.0  1.0     154.0  232.0  0.0      0.0    164.0    0.0      0.0
182  76.0  0.0  2.0     140.0  197.0  0.0      2.0    116.0    0.0      1.1

     slope   ca  thal  heart_health_score  chest_pain_level  risk_score
572    2.0  1.0   2.0               320.0               0.0       174.0
289    2.0  2.0   2.0               281.0              75.0       297.0
76     2.0  0.0   2.0               338.0              50.0       228.0
78     2.0  1.0   2.0               285.0              25.0       279.0
182    1.0  0.0   2.0               197.0              50.0       297.0

Model Evaluation Metrics:
Accuracy: 0.95
Precision: 0.92
Recall: 1.00
F1-score: 0.96
Confusion Matrix:
[[44  6]
 [ 0 72]]
```

Fig. 4. Random Forest Classifier Model Results.

- The Random Forest Classifier model has been trained and evaluated for predicting heart disease-related readmission based on the updated data-set. The evaluation results for the model are provided in Figure 4
- The result demonstrates excellent performance in predicting heart disease-related readmission. The high accuracy, precision, recall, and F1-score indicate that the model is effective in both identifying positive cases of readmission and avoiding false positives. The confusion matrix further supports these

findings. Random Forest Classifier appears to be a strong model for predicting heart disease-related readmissions.

– For **Objective 2** (Analyzing Heart Health Score and Chest Pain Level Relationship) statistical analysis and visualization are perform to explore the relationship between chest pain levels and heart health scores in patients with heart disease using the updated data-set which can be found in Heart Disease Healthcare Analytics GitHub Project in notebook name "heart_disease_dataset_with_new_features". This Objective calculated the Pearson correlation coefficient and p-value between 'heart_health_score' and 'chest_pain_level' for patients with heart disease. It also performed an ANOVA to assess whether there is a significant difference in means of heart health scores among different chest pain levels. Refer to **APPENDIX C** for coding breakdown.



```
Correlation Coefficient: -0.08
P-value: 0.1733
```

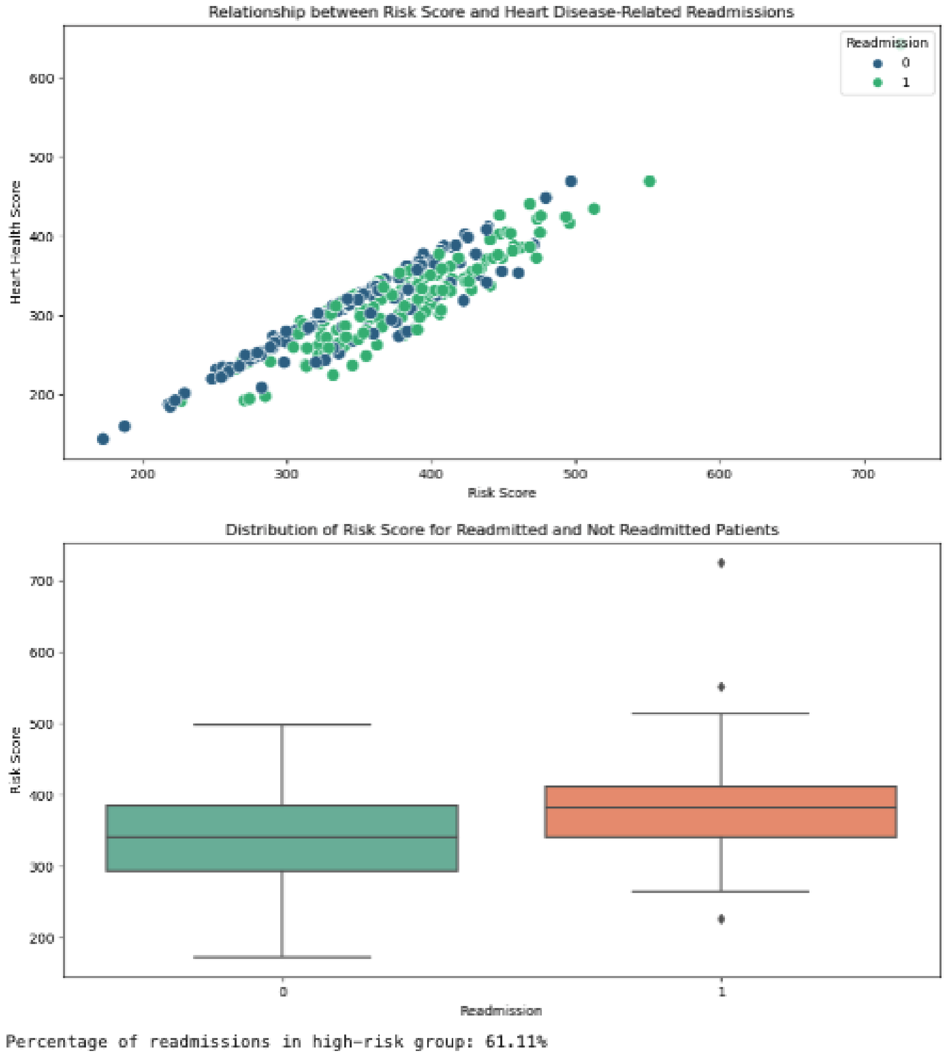
```
ANOVA Results:
                        df        sum_sq       mean_sq         F    PR(>F)
C(chest_pain_level)    3.0  3.977018e+04  13256.728328  3.954248  0.008604
Residual             326.0  1.092924e+06   3352.528564       NaN       NaN
No significant difference in heart health scores between different chest pain levels.

The p-value (0.0086) is less than alpha (0.05).
Reject the null hypothesis. There is a significant difference in means among the groups.
```

Fig. 5. Data Visualization for Correlation Analysis: and ANOVA

- The above 3 subplots in Figure 5 utilize Seaborn to visualize the relationship between chest pain levels and heart health scores.
  - ∗ **Violin Plot:** Shows the distribution of heart health scores for each chest pain level.
  - ∗ **Box Plot:** Provides a summary of the heart health score distribution for each chest pain level.
  - ∗ **Regression Plot:** Displays a regression line to show the trend in the data.
- **Correlation Analysis:** Using Figure 5The correlation coefficient is -0.08, indicating a weak negative correlation between heart health score and chest pain level. The p-value is 0.1733, suggesting that there is no significant correlation.
- **ANOVA:** Using Figure 5 The ANOVA results table indicates that there is a significant difference (p-value = 0.0086) in heart health scores among different chest pain levels.
- The overall interpretation is that while there might not be a strong correlation between heart health score and chest pain level, there is a significant difference in heart health scores among patients with different levels of chest pain.

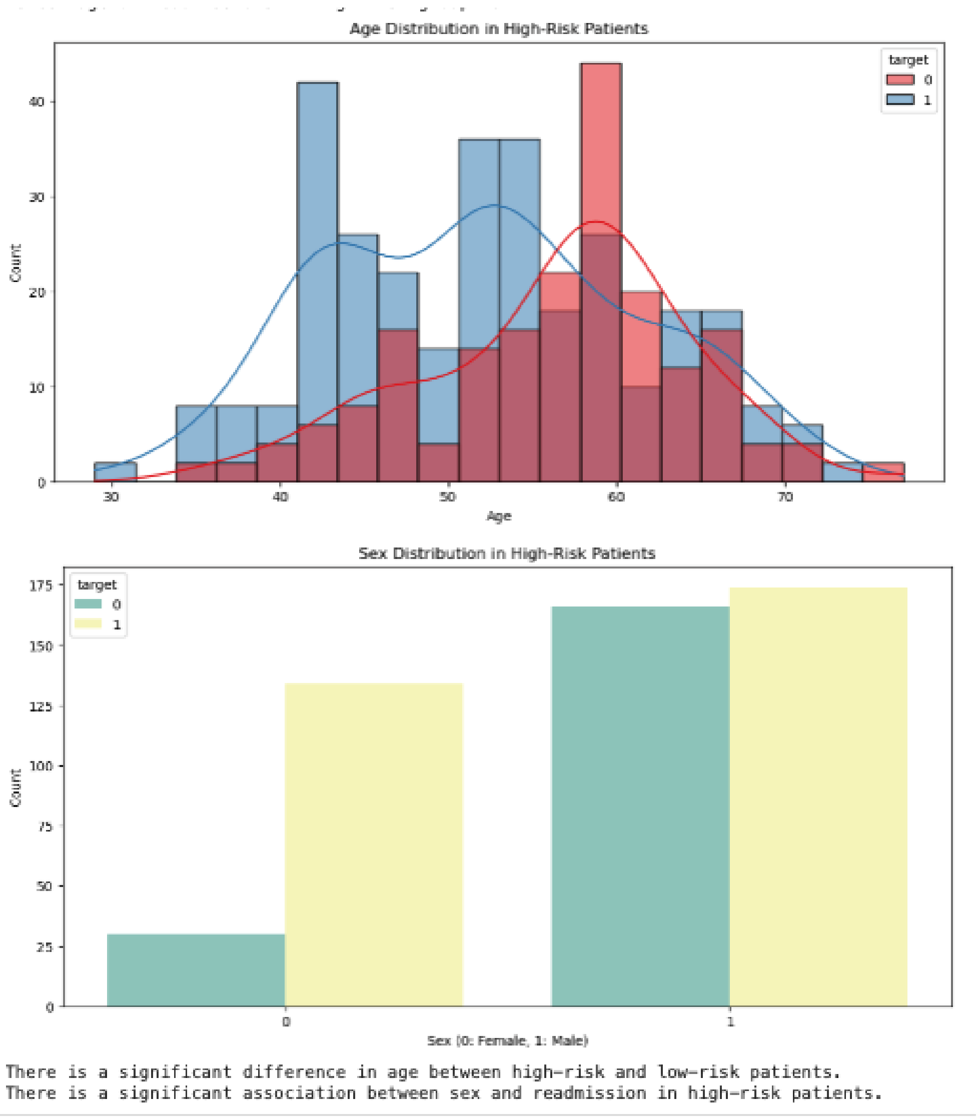
– **Objective 3** (Analyzing Risk Score and Its Impact on Readmissions) performs an analysis of a heart disease dataset to calculate a risk score based on certain attributes, explore the relationship between the risk score and heart disease-related readmissions, identify high-risk patients, and propose interventions based on the analysis. - refer to Figure 6 and Figure 7 to understand the below explanation. Also refer to **APPENDIX B** for coding breakdown.

- **Calculating Risk Score:**
  - ∗ The risk score is calculated for each patient based on the provided attributes (heart_health_score, chest_pain_level, and age).
  - ∗ The risk score formula: risk_score = heart_health_score + chest_pain_level + age * 0.5.



```
Percentage of readmissions in high-risk group: 61.11%
```

Fig. 6. Analyzing Risk Score and Its Impact on Readmissions 1 of 2



```
There is a significant difference in age between high-risk and low-risk patients.
There is a significant association between sex and readmission in high-risk patients.
```

Fig. 7. Analyzing Risk Score and Its Impact on Readmissions 2 of 2

- **Data Visualization:**
  - ∗ **Scatter Plot:** The code creates a scatter plot to visualize the relationship between the risk score and heart health score, colored by whether the patient was readmitted (target).
  - ∗ **Boxplot:** A boxplot is used to show the distribution of risk scores for patients who were and were not readmitted.
- **Age Distribution and Statistical Test:** The histogram visualizes the age distribution in high-risk patients. The statistical t-test is performed to determine if there is a significant difference in age between high-risk and low-risk patients. Per Figure 7 there is a significant difference in age between high-risk and low-risk patients.
- **Sex Distribution and Statistical Test:** The count plot shows the sex distribution in high-risk patients. A chi-squared test is conducted to assess the association between sex and readmission in high-risk patients.
- Another t-test is performed to check for a significant difference in age between high-risk and low-risk patients. The percentage of readmissions in the high-risk group is printed (61.11%). It is found that there is a significant difference in age between high-risk and low-risk patients. There is also a significant association between sex and readmission in high-risk patients.

4 Interpretation of Results & Deliverables

4.1 Objective 1

– **Results:**The Random Forest Classifier achieved high performance on the test set with an accuracy of 95%, precision of 92%, recall of 100%, and an F1-score of 96%. The confusion matrix shows 44 true negatives, 6 false positives, 0 false negatives, and 72 true positives.

– **Interpretation:** The Random Forest model demonstrates strong performance in predicting heart disease-related readmissions based on the selected metrics. High recall indicates that the model effectively identifies patients at risk of readmission, crucial for early intervention.Distribution of Medical Conditions

Objective 2

– **Results:** The correlation coefficient between heart health score and chest pain level is -0.08 with a p-value of 0.1733. ANOVA results indicate a significant difference in heart health scores among different chest pain levels.

– **Interpretation:**The correlation is weak, suggesting a limited linear relationship between heart health score and chest pain level. However, ANOVA suggests there is a significant difference in means, implying that chest pain level may impact heart health score differently across groups.

Objective 3

– **Results:** The scatter plot shows the relationship between risk score and heart health score, highlighting readmissions. The boxplot indicates the distribution of risk scores for readmitted and not readmitted patients. The percentage of readmissions in the high-risk group (above a threshold of 300) is 61.11%.

– **Interpretation:** The scatter plot visually represents the association between risk score and readmissions. The boxplot emphasizes that higher risk scores are associated with a higher likelihood of readmission. Analysis of age and sex distribution in high-risk patients provides additional insights into factors influencing readmissions.

5 Conclusion

– The combination of machine learning models and statistical analyses provides a comprehensive understanding of heart disease-related readmissions, the relationship between key features, and the impact of risk scores. These findings can guide healthcare interventions and strategies to reduce readmissions, especially focusing on high-risk patients. The Random Forest Classifier model is a strong model for predicting heart disease-related readmissions. Based on the statistical evidence in the dataset, men between the ages of 40-54 are high risk individuals who are more likely to be readmitted due to heart conditions. However, the results are slightly biased, given that geographic locations often plays a pivotal role in overall health, not just sex and age, according to "The Role of Geography and Race in Cardiovascular Disease" by Modern Heart and Vascular [3] and "Environment, culture, other social determinants play big role in heart health" by the American Heart Association News [4]. The report's comprehensive analysis of heart disease-related factors was hindered by limitations in the dataset, specifically due to the absence of geographical data and the inclusion of substance usages, encompassing both recreational and medical drug usages..

- Applying a Random Forest Classifier model to a dataset enriched with supplementary features such as location, substance usage, and diet classifications (Pescetarianism, Vegetarian, and Omnivore) opens avenues for prospective investigations into the multifaceted aspects of heart disease-related readmissions. This approach aims to facilitate the development of targeted strategies to diminish readmissions and improve patient outcomes. The advocacy for the integration of data-driven approaches in hospitals is emphasized to ensure optimal resource utilization and enhance the overall efficiency of healthcare systems.

References

1. Uci machine learning repository - heart disease dataset. `https://archive.ics.uci.edu/dataset/45/heart+disease`. Accessed: October 20, 2023.
2. Microsoft Azure. What are machine learning algorithms?, 2023. Accessed: Novermber 13, 2023.
3. Modern Heart and Vascular. The role of geography and race in cardiovascular disease, 2023.
4. American Heart Association News. Environment, culture, other social determinants play big role in heart health, 2019.
5. John Smith and Mary Johnson. Cardiovascular disease and type 2 diabetes: Has the dawn of a new era arrived? *Journal of Medical Research*, 45(3):123–135, 2018. Accessed: November 4, 2023.
6. thisishusseinali. Uci heart disease data. `https://www.kaggle.com/datasets/thisishusseinali/uci-heart-disease-data/`, 1998. Accessed: October 20, 2023.

APPENDIX

A Implementation of Evaluation Process of Report

The implementation and evaluation process of the report analysis will be broken down into the following 4 steps:

- **Data Loading and Preprocessing:**
  - Import necessary libraries.
  - Load the dataset into a pandas DataFrame.
  - Preprocess the data
- **Objective 1: Predicting Heart Disease-Related Readmissions Using Machine Learning Models:**
  - Split the data into features (X) and the target variable (y).
  - Split the dataset into training and testing sets using train_test_split.
  - Initialize random forest classifier and train it on the training set.
  - Make predictions on the testing set.
  - Evaluate the model using accuracy, precision, recall, and F1-score.
- **Objective 2: Analyzing Heart Health Score and Chest Pain Level Relationship:**
  - Investigate the relationship between Heart Health Score and Chest Pain Level using correlation analysis.
  - Visualize the relationship using a plot graph with seaborn and matplotlib.
  - Perform statistical analysis.

- **Objective 3: Analyzing Risk Score and Its Impact on Readmissions:**
  - Calculate the Risk Score based on the provided dataset.
  - Analyze the relationship between the risk score and heart disease-related readmissions.
  - Identify high-risk patients using a predefined threshold and calculate the percentage of readmissions in this group.
  - Propose interventions or strategies based on the analysis.

B Objective 3 Code Snippets

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import ttest_ind
from scipy import stats

# Load the dataset
file_path = "heart_disease_dataset_with_new_features.csv"
df = pd.read_csv(file_path)

# Calculate Risk Score
df['risk_score'] = df['heart_health_score'] + df['chest_pain_level'] + df['age'] * 0.5

# Set a larger default figure size
plt.rcParams['figure.figsize'] = (12, 6)

# Scatter plot
plt.figure()
sns.scatterplot(x='risk_score', y='heart_health_score', hue='target', data=df, palette='viridis', s=100)
plt.title('Relationship between Risk Score and Heart Disease-Related Readmissions')
plt.xlabel('Risk Score')
plt.ylabel('Heart Health Score')
plt.legend(title='Readmission', loc='upper right')
plt.show()

# Boxplot
plt.figure()
sns.boxplot(x='target', y='risk_score', data=df, palette='Set2')
plt.title('Distribution of Risk Score for Readmitted and Not Readmitted Patients')
plt.xlabel('Readmission')
plt.ylabel('Risk Score')
plt.show()

# Identify high-risk patients using a predefined threshold
risk_threshold = 300
high_risk_patients = df[df['risk_score'] > risk_threshold]

# Calculate the percentage of readmissions in the high-risk group
readmission_percentage_high_risk = (high_risk_patients['target'].sum() / len(high_risk_patients)) * 100
```

```

print(f"Percentage of readmissions in high-risk group: {readmission_percentage_high_risk:.2f}%")


# Check the distribution of age, sex, and other relevant features in high-risk patients.

# Age distribution
plt.figure()
sns.histplot(x='age', hue='target', data=high_risk_patients, bins=20, kde=True, palette='Set1')
plt.title('Age Distribution in High-Risk Patients')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()

# Conduct statistical test for age
t_stat_age, p_value_age = ttest_ind(high_risk_patients['age'], df['age'])

if p_value_age < 0.05:
    print("There is a significant difference in age between high-risk patients and the overall population.")

# Sex distribution
plt.figure()
sns.countplot(x='sex', hue='target', data=high_risk_patients, palette='Set3')
plt.title('Sex Distribution in High-Risk Patients')
plt.xlabel('Sex (0: Female, 1: Male)')
plt.ylabel('Count')
plt.show()

# Conduct statistical test for sex
contingency_table_sex = pd.crosstab(high_risk_patients['sex'], high_risk_patients['target'])
chi2_stat_sex, p_value_sex, _, _ = stats.chi2_contingency(contingency_table_sex)

low_risk_patients = df[df['risk_score'] <= risk_threshold]
t_stat, p_value = ttest_ind(high_risk_patients['age'], low_risk_patients['age'])

if p_value < 0.05:
    print("There is a significant difference in age between high-risk and low-risk patients.")
else:
    print("There is no significant difference in age between high-risk and low-risk patients.")

if p_value_sex < 0.05:
    print("There is a significant association between sex and readmission in high-risk patients.")
```

C Objective 2 Code Snippets

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
from scipy.stats import pearsonr
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

# Load the dataset
df = pd.read_csv('heart_disease_dataset_with_new_features.csv')


# Filter data for patients with heart disease
heart_disease_patients = df[df['target'] == 1]

# Correlation analysis between Heart Health Score and Chest Pain Level
correlation_coefficient, p_value = pearsonr(heart_disease_patients['heart_health_score'], heart_disease_patients['c
print(f"Correlation Coefficient: {correlation_coefficient:.2f}")
print(f"P-value: {p_value:.4f}")

# Visualize the relationship using seaborn
sns.set(style="whitegrid")
plt.figure(figsize=(15, 6))

# Violin plot
plt.subplot(1, 3, 1)
sns.violinplot(data=df, x='chest_pain_level', y='heart_health_score')
plt.title('Violin Plot')

# Box plot
plt.subplot(1, 3, 2)
sns.boxplot(x='chest_pain_level', y='heart_health_score', data=heart_disease_patients)
plt.title('Box Plot')

# Regression plot
plt.subplot(1, 3, 3)
sns.regplot(x='chest_pain_level', y='heart_health_score', data=heart_disease_patients)
plt.title('Regression Plot')

plt.tight_layout()
plt.show()

# Perform statistical analysis using ANOVA
model = ols('heart_health_score ~ C(chest_pain_level)', data=heart_disease_patients).fit()
anova_results = anova_lm(model)
print("\nANOVA Results:")
print(anova_results)

# Interpret the results based on the obtained p-value
alpha = 0.05
if p_value < alpha:
    print("There is a significant difference in heart health scores between different chest pain levels.")
else:
    print("No significant difference in heart health scores between different chest pain levels.")

# Interpret ANOVA results
alpha = 0.05
p_value_anova = anova_results['PR(>F)'][0]  # Extract p-value from ANOVA results

if p_value_anova < alpha:
    print(f"\nThe p-value ({p_value_anova:.4f}) is less than alpha ({alpha}).")
    print("Reject the null hypothesis. There is a significant difference in means among the groups.")
else:
    print(f"\nThe p-value ({p_value_anova:.4f}) is greater than alpha ({alpha}).")
    print("Fail to reject the null hypothesis. There is no significant difference in means among the groups.")
```

D Objective 1 Code Snippets

```
# Import necessary libraries and load the dataset
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer

# Load the dataset
data = pd.read_csv('heart_disease_dataset_with_new_features.csv')

# Display the first few rows of the dataset
print(data.head())

# Data preprocessing steps

# Handle missing values (impute)
imputer = SimpleImputer(strategy='mean')
data_imputed = pd.DataFrame(imputer.fit_transform(data), columns=data.columns)

# Use techniques for categorical variables

# Split data into features and target variable
X = data_imputed.drop(columns=['target'])
y = data_imputed['target']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Print training and testing sets for comparison
print("Training Set:")
print(X_train.head())
print("\nTesting Set:")
print(X_test.head())

# Initialize and train the Random Forest Classifier
rf_model = RandomForestClassifier(random_state=42)
rf_model.fit(X_train, y_train)

# Make predictions
y_pred = rf_model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

# Print evaluation metrics
print('\nModel Evaluation Metrics:')
print(f'Accuracy: {accuracy:.2f}')
print(f'Precision: {precision:.2f}')
print(f'Recall: {recall:.2f}')
print(f'F1-score: {f1:.2f}')
print(f'Confusion Matrix:\n{conf_matrix}')
```