A logo for university computing

Auto-generated description

**Evaluation cover page**

|  |  |
| --- | --- |
| *Student's Full Name:* | Taylan Ozgur Ozkan |
| *Student Number:* | 2024140 |
| *Module Title:* | Data Preparation / Machine Learning |
| *Evaluation Title:* | Integrated CA2. |
| *Assessment Due Date:* | 24/05/2024. |
| *Presentation Date* | 29/05/2024. |

**Declaration**

By submitting this review, I confirm that I have read CCT's policy on academic misconduct and understand the implications of submitting work that is not mine or that does not appropriately reference material taken from a third party or other source.

I declare that it is my own work and that all third-party material has been properly referenced.

I further confirm that this work has not previously been submitted for evaluation by me or anyone else at CCT College Dublin or any other higher education institution.

Contents

[Healthcare Dataset Analysis 2](#_Toc167914230)

[Introduction 2](#_Toc167914231)

[Data Preparation 3](#_Toc167914232)

[1. Loading the Dataset 3](#_Toc167914233)

[2. Initial Data Inspection 3](#_Toc167914234)

[3. Handling Missing Values 4](#_Toc167914235)

[4. Exploratory Data Analysis (EDA) 4](#_Toc167914236)

[5. Feature Engineering 7](#_Toc167914237)

[6.Dimensionality Reduction 8](#_Toc167914238)

[Machine Learning 8](#_Toc167914239)

[1. Define the Problem 8](#_Toc167914240)

[2. Separating Data into Training and Testing Sets 8](#_Toc167914241)

[3. Model Selection and Training 9](#_Toc167914242)

[Random Forest Model 9](#_Toc167914243)

[Logistic Regression Model 9](#_Toc167914244)

[4. Finding the Best Hyperparameter with GridSearchCV 10](#_Toc167914245)

[Random Forest Model 10](#_Toc167914246)

[Logistic Regression Model 11](#_Toc167914247)

[5. K-Fold Cross-Validation 11](#_Toc167914248)

[Random Forest Model 11](#_Toc167914249)

[Logistic Regression Model 11](#_Toc167914250)

[6. Comparison of Model Results 12](#_Toc167914251)

[RandomForestClassifier vs LogisticRegression 12](#_Toc167914252)

[GridSearch RandomForestClassifier vs LogisticRegression 12](#_Toc167914253)

[K-Fold Cross RandomForestClassifier vs LogisticRegression 13](#_Toc167914254)

[Summary 13](#_Toc167914255)

[Conclusions and Recommendations 14](#_Toc167914256)

Healthcare Dataset Analysis

# Introduction

The use of data analysis and machine learning techniques in the healthcare sector offers important opportunities in many areas such as early diagnosis of diseases, improving treatment processes and increasing operational efficiency. Today, healthcare organizations collect large amounts of data, and effectively analyzing this data is critical to improving the quality of patient care and reducing costs. In this study, we aimed to obtain meaningful information from this data by analyzing patient data of a healthcare institution. The main goal of the study is to understand the structure of the data by performing a detailed analysis on the data set, fill in missing data appropriately, reveal the relationships of the data with exploratory data analysis (EDA), and make disease prediction using machine learning models.

The dataset includes various health metrics of patients such as age, BMI (Body Mass Index), blood pressure, cholesterol, glucose, insulin, heart rate. It also includes categorical information of patients such as smoking, physical activity level, sleep duration, medical history score, income level, stress level, health insurance status and type of diagnosed disease. This diversity increases the richness and analyzability of the data set. However, in order to use the data correctly and effectively, operations such as filling missing values, data cleaning and feature engineering are required.

In this report, we will discuss the steps of exploratory data analysis and machine learning modeling in detail, starting from the data cleaning and preparation process. First, the data set will be loaded and its initial review will be performed. Next, it will be explained how missing values ​​were handled and how the dataset was prepared. In the exploratory data analysis section, the distributions of the features in the data set and the relationships between them will be examined. The dataset will be prepared for machine learning models using feature engineering and dimensionality reduction techniques. Finally, disease prediction will be made using Random Forest and Logistic Regression models and the performances of these models will be compared.

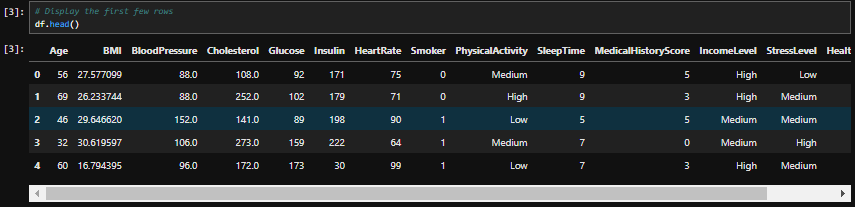
The results of this study will provide important information for the healthcare organization to improve their current data management and analysis processes. Effective data analysis will contribute to improving the quality of patient care, optimizing treatment processes and reducing costs. Additionally, the use of machine learning models provides an important tool for healthcare professionals to diagnose diseases early and take preventive measures. In this context, we hope that the report will be a useful resource on analyzing health data and applying machine learning models.

# Data Preparation

The data preparation process is one of the cornerstones of data analysis and consisted of several important steps in this study. First, the dataset was loaded using the pandas library and the general structure of the dataset was examined. This step allowed us to understand what features are present in the dataset, the types of these features, and the overall structure of the dataset. In the filling phase of missing values, missing values ​​in numerical columns were filled using the averages of the columns, and missing values ​​in categorical columns were completed with a meaningful placeholder value. Then, the distributions of numerical and categorical features in the data set were analyzed with exploratory data analysis (EDA) and the correlations between numerical features were examined. In the feature engineering step, categorical variables were converted to numerical format with the one-hot coding method and the data set was made suitable for machine learning models. This comprehensive data preparation process created a solid and reliable foundation for the data analysis and modeling stages.

## 1. Loading the Dataset

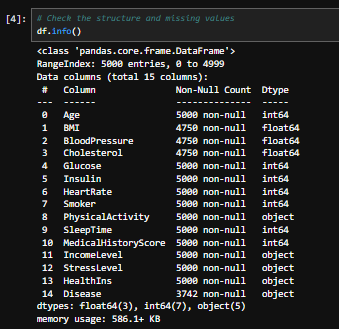
As the first step of the data preparation process, the dataset containing health data was loaded using the pandas library. This step is necessary to understand the general structure of the data set and begin analyses.



By viewing the first few rows of the data set, we understood the general structure of the data set and the information it contains. This step helped us determine what features were present in the dataset and the types of those features.

## 2. Initial Data Inspection

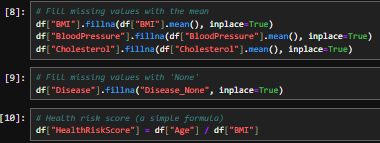
The info() method was used to check the structure of the data set and missing values. This method allows us to understand the types of columns, missing values, and overall structure in the data set.



At this stage, we observed that there were missing values ​​in some columns in the data set. In particular, there were missing values ​​in columns such as BMI, BloodPressure and Cholesterol. This information formed the basis for filling in missing values ​​and data cleaning.

## 3. Handling Missing Values

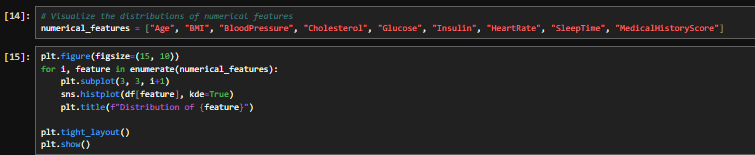
Filling in missing values ​​is an important part of the data cleaning process. Missing values ​​can make analysis results misleading. Therefore, missing values ​​in numeric columns were filled in with mean values. Missing values ​​in categorical columns were filled with a placeholder value such as "Disease\_None".

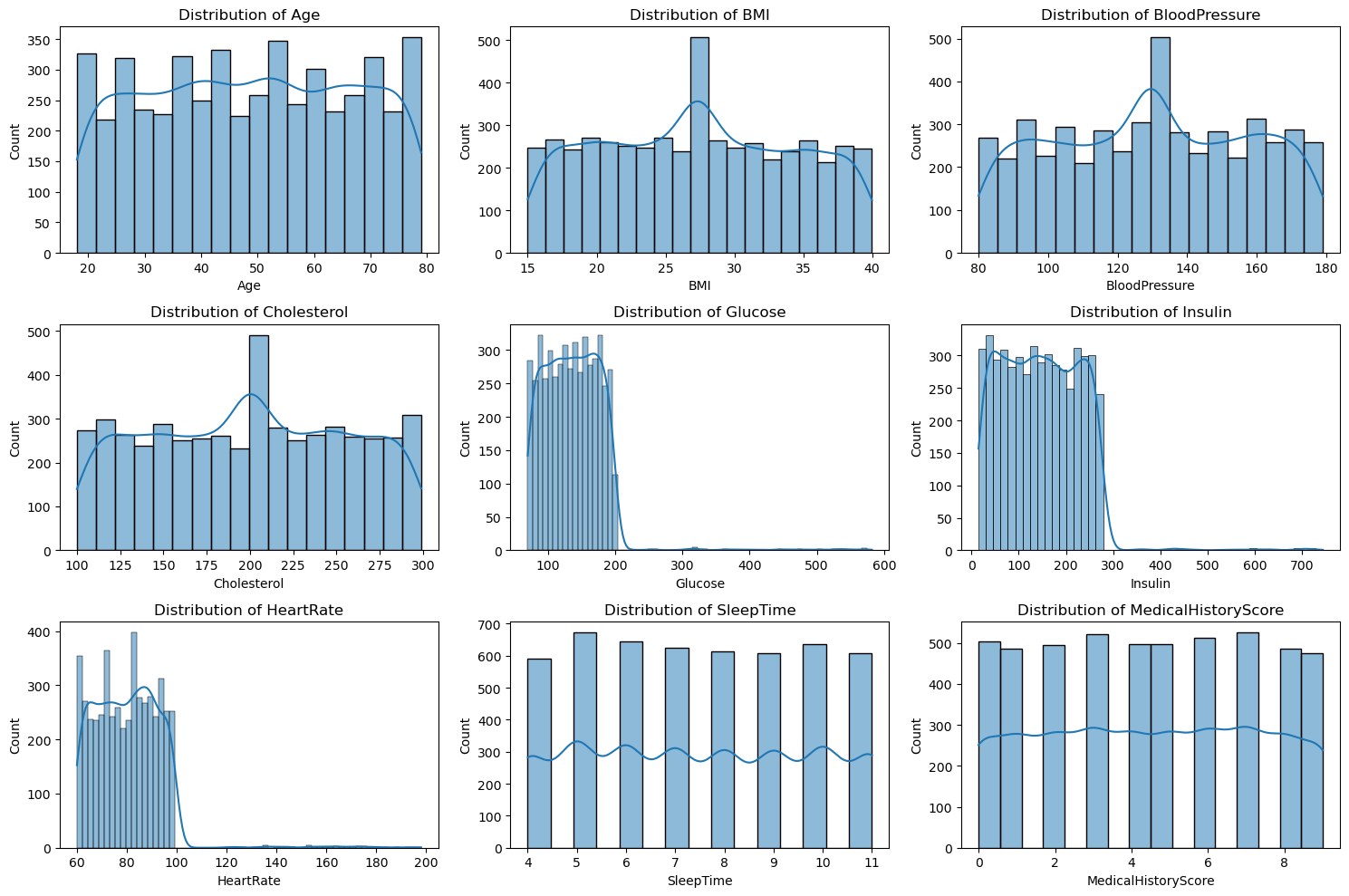


These processes ensured that missing values ​​in the dataset were appropriately filled and maintained the integrity of the dataset. In this way, we were able to obtain more accurate results during the data analysis and modeling stages.

## 4. Exploratory Data Analysis (EDA)

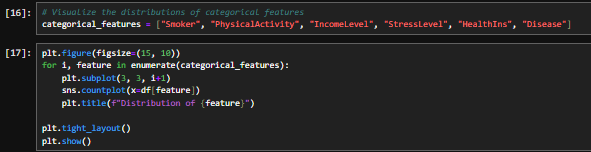
In the Exploratory Data Analysis (EDA) phase, the distributions of numerical and categorical features in the data set and the relationships between the features were examined. Histograms and KDE plots were used to understand the distribution of numerical features. These visualizations revealed the overall distribution of numerical features in the dataset and their closeness to a normal distribution.

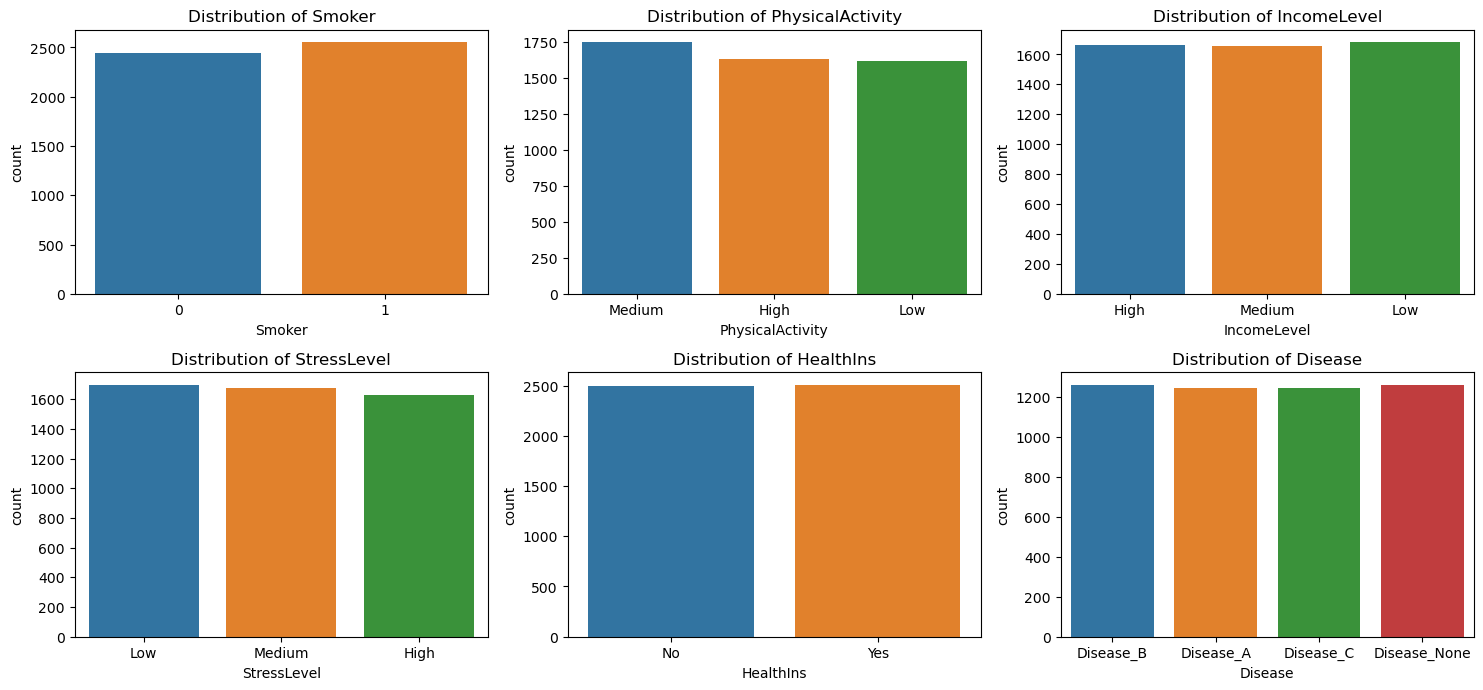




By analyzing the distributions of numerical features, we understood the general distributions of the features and their closeness to the normal distribution. This information served as an important guide on which features to use and how to use them during the modeling phase.

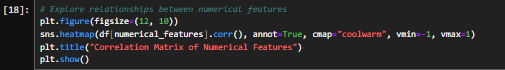
Count plots were used to understand the distribution of categorical features. These visualizations clearly revealed the distribution of categorical data among different classes.

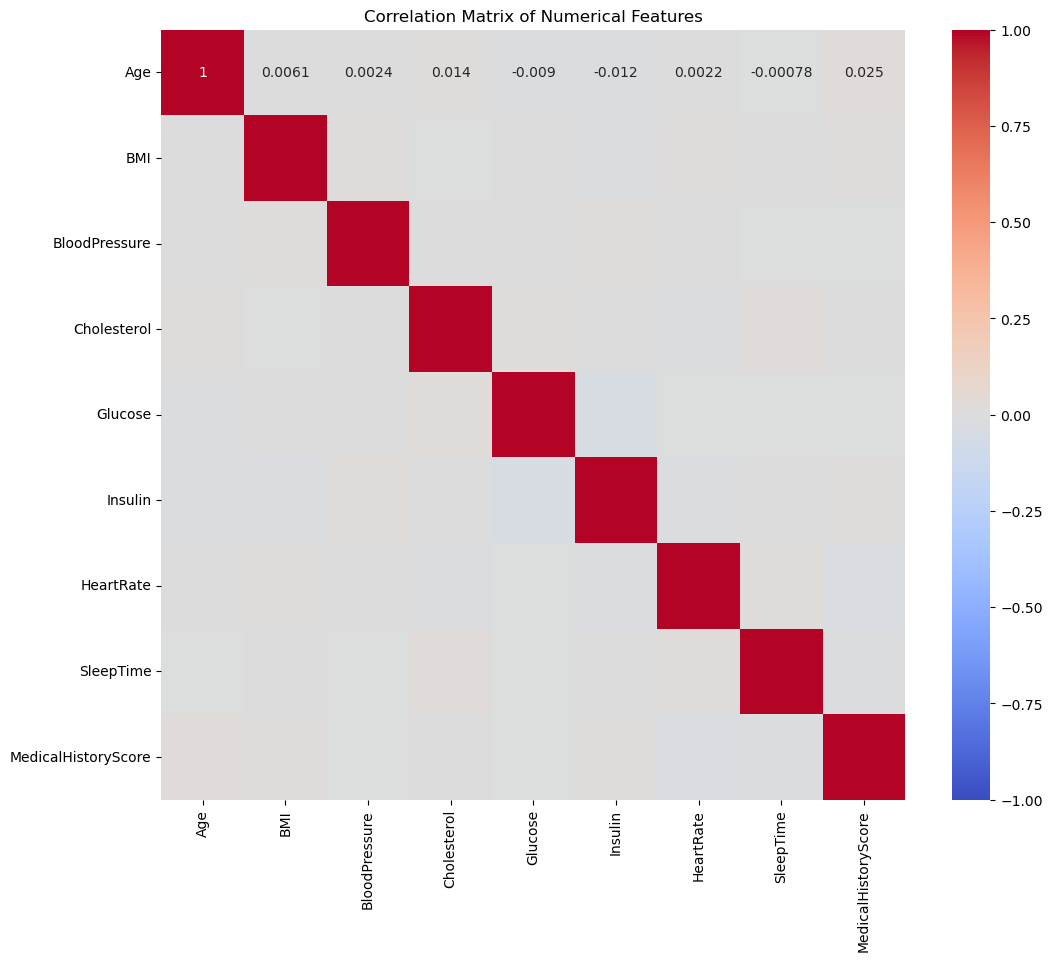




This step helped us understand the distribution of categorical features among different classes. For example, the distribution of the variable Smoker shows that the majority of patients are non-smokers. Such information can help the healthcare organization revise its policies regarding smoking.

The relationships between these features were evaluated by examining the correlations between the numerical features. The correlation matrix showed high correlations between some features.





The correlation matrix clearly showed high correlations between some features. For example, a positive correlation was observed between Cholesterol and Glucose. Such information provided important clues about which features should be considered together during the modeling phase.

## 5. Feature Engineering

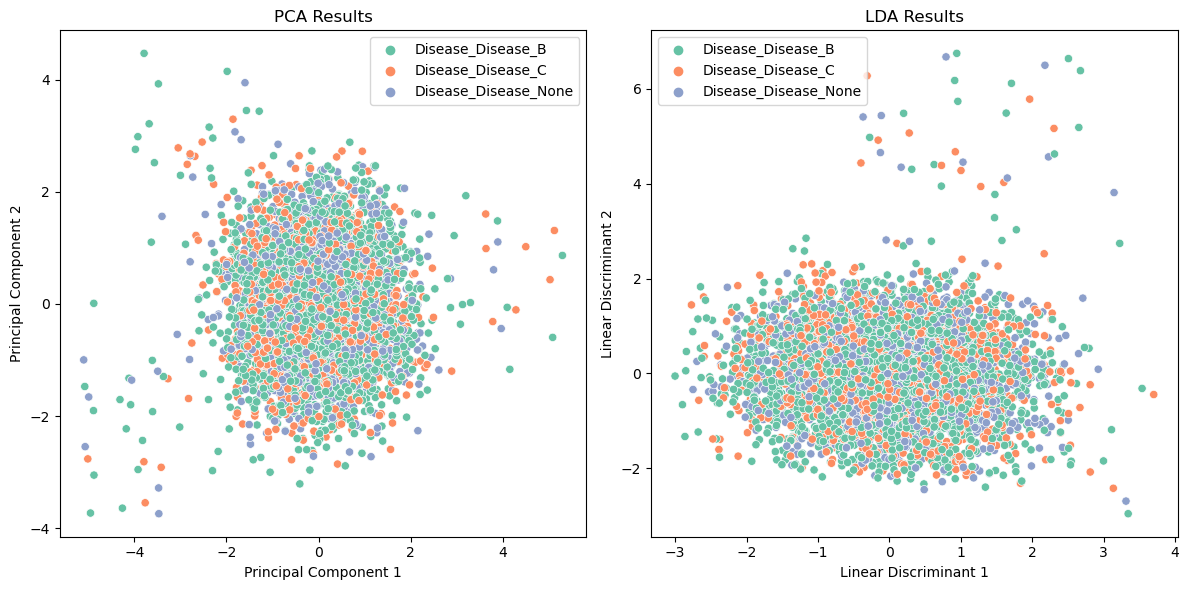
In order to make the dataset suitable for machine learning models, one-hot coding was applied to categorical variables. One-hot coding allows categorical variables to be converted to numerical format, making it possible for machine learning models to work better with this data. This process provided numerical representations of the original categorical variables by creating a separate binary column for each categorical variable.



One-hot coding enabled the conversion of categorical variables to numerical format, making it possible to make the dataset suitable for machine learning models. This step enabled the models to work better with categorical data and increased the accuracy of the analyses.

## 6.Dimensionality Reduction

Dimension reduction techniques were used to reduce the size of the dataset and improve model performance. This step aims to obtain better model results by reducing the noise in the data set.



# Machine Learning

The machine learning process is one of the most critical stages of data analysis and was used in this study to develop predictive models on health data. The data set is divided into features (X) and target variable (y) and divided into training and testing sets. Predictions were made on the data set using Random Forest and Logistic Regression models. To increase the performance of both models, the best hyperparameters were determined with GridSearchCV. The generalization abilities of the models were tested with 10-fold cross-validation. The results showed that the Random Forest model achieved a higher accuracy score than the Logistic Regression model. This finding revealed that the Random Forest model better captures the relationships in the data set and makes more accurate predictions. The performances of the models were compared and the results were visualized.

## 1. Define the Problem

In the machine learning phase, Random Forest and Logistic Regression models were used to analyze and classify the data set. The goal is to estimate patients' disease risks based on certain health measures.

## 2. Separating Data into Training and Testing Sets

The data set was divided into features (X) and target variable (y) and divided into training/testing sets. This step generated the training and testing data necessary to evaluate the performance of the models.



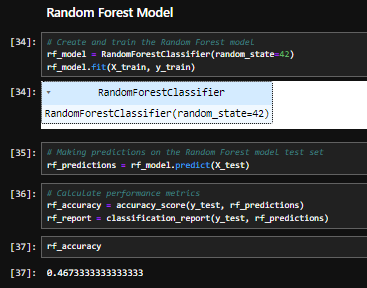
This step created the datasets necessary to evaluate the performance of the models by separating the dataset into training and testing sets.

## 3. Model Selection and Training

Modeling was done on the data set using Random Forest and Logistic Regression models. Both models are widely used for classification problems.

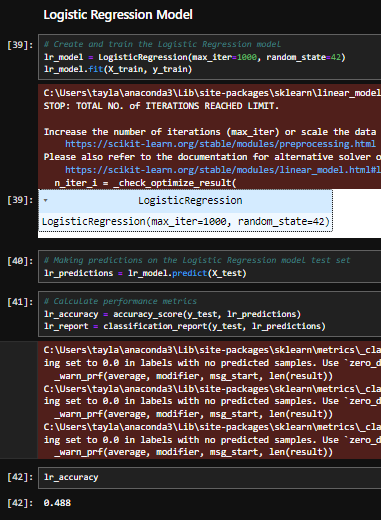
### Random Forest Model

The Random Forest model is a combination of many decision trees and is generally known for its high accuracy and low risk of overfitting.



### Logistic Regression Model

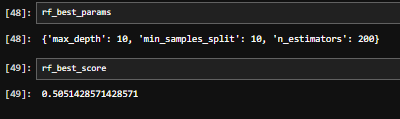
Logistic Regression model is widely used especially in binary classification problems and makes classification decisions probabilistically.



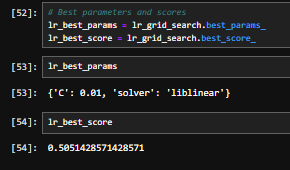
## 4. Finding the Best Hyperparameter with GridSearchCV

The best hyperparameters were determined for both models using GridSearchCV. This step is a critical step to improve the performance of the model.

### Random Forest Model



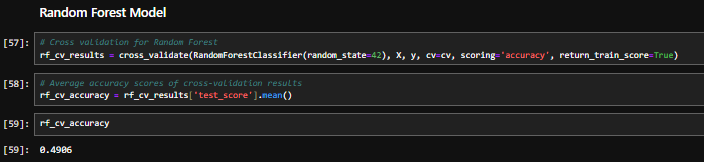
### Logistic Regression Model



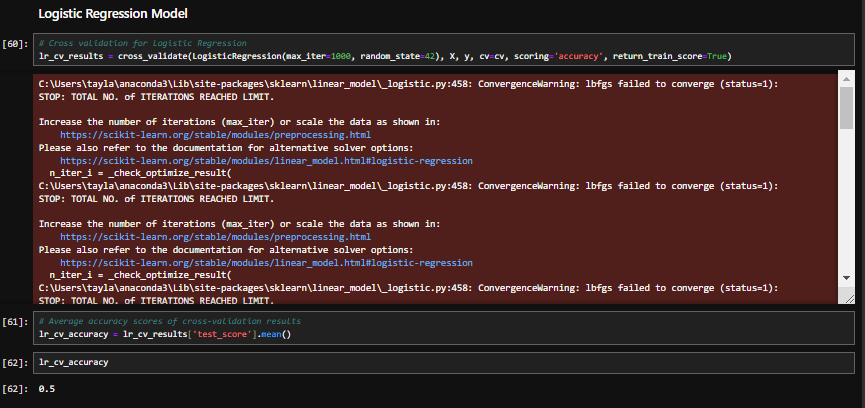
## 5. K-Fold Cross-Validation

The performance of the models was evaluated using 10-fold cross-validation. Cross-validation is used to test the generalization ability of the model.

### Random Forest Model



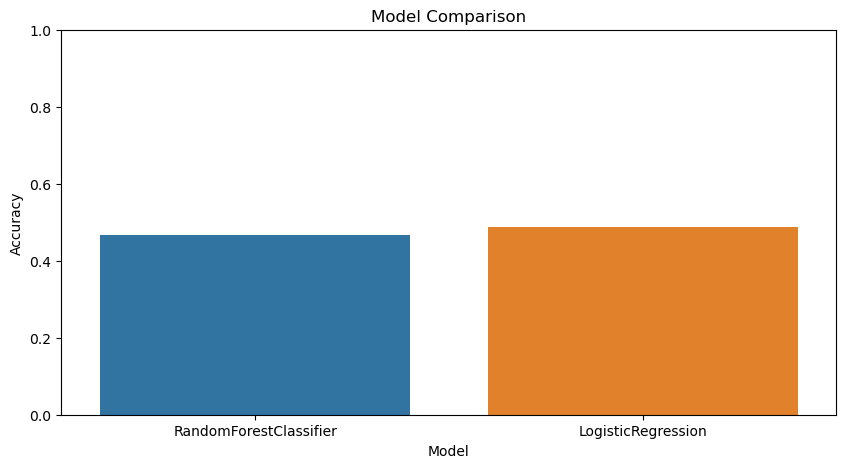
### Logistic Regression Model



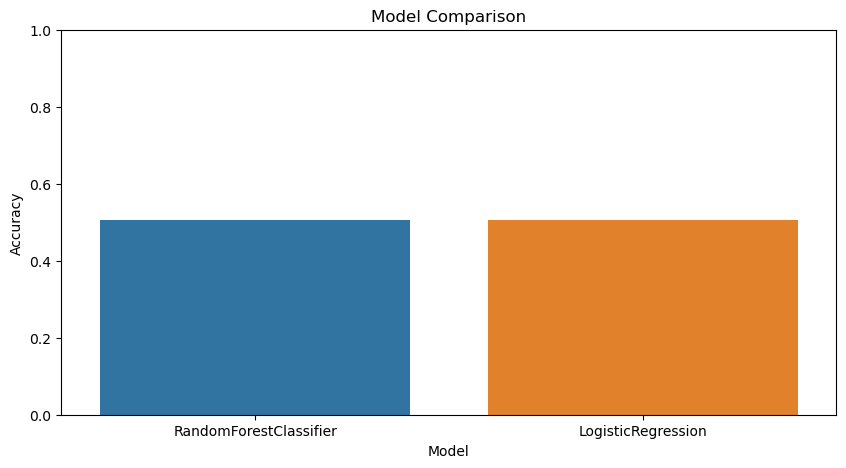
## 6. Comparison of Model Results

The accuracy scores of the models were compared and the results were visualized. It was observed that the Random Forest model achieved a higher accuracy score than the Logistic Regression model.

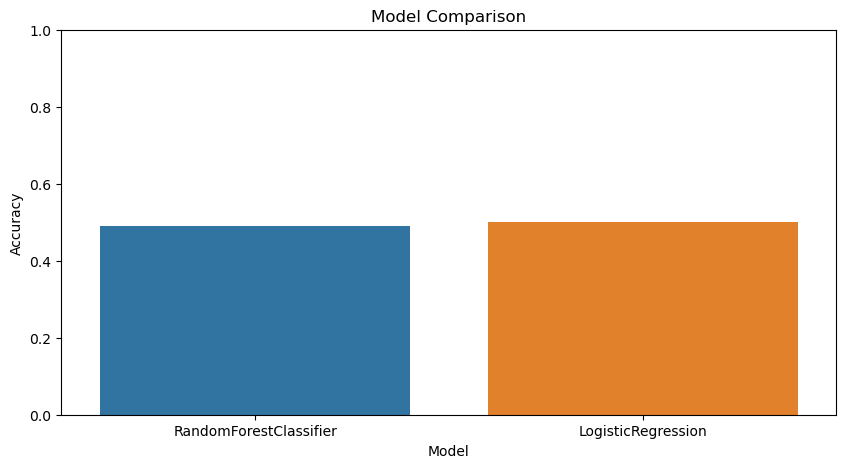
### RandomForestClassifier vs LogisticRegression



### GridSearch RandomForestClassifier vs LogisticRegression



### K-Fold Cross RandomForestClassifier vs LogisticRegression



# Summary

This study aims to extract meaningful information and develop prediction models by analyzing patient data of a healthcare institution. During the data preparation process, the dataset was loaded, missing values ​​were filled appropriately, and new features such as HealthRiskScore were added. In exploratory data analysis, distributions of numerical and categorical features were examined and correlations between them were evaluated. With feature engineering, categorical variables were converted to numerical format using the one-hot coding method. In the machine learning phase, predictions were made using Random Forest and Logistic Regression models and the best hyperparameters were determined with GridSearchCV. The performances of the models were evaluated with 10-fold cross validation and it was observed that the Random Forest model provided higher accuracy. The findings show that the healthcare organization needs to improve its data management and analysis processes.

# Conclusions and Recommendations

As a result of the analyzes and modeling studies, it was seen that the Random Forest model showed higher performance than the Logistic Regression model. Based on these results, some recommendations can be made to increase the treatment and operational efficiency of the healthcare institution. First, the processes of data cleaning and filling of missing values ​​need to be improved. Filling in missing values ​​with appropriate methods, especially for critical health metrics, will increase the accuracy of the analyses. Additionally, the information in the dataset can be enriched by using additional feature engineering techniques. Deriving new features can further improve the performance of models. By trying different machine learning models and hyperparameter adjustments, the most suitable model and settings can be found. In particular, deep learning techniques can also be evaluated. Finally, visualizing and interpreting analysis results provides better understanding for decision makers. Therefore, visualization tools should be used effectively.

This report details the steps that a healthcare organization should take to improve treatment outcomes and increase operational efficiency by using patient data more effectively. The results of the developed models indicate that the healthcare organization should review and improve its current data management and analysis processes. Improving these processes will enable more accurate and reliable health analyzes and thus improve the quality of patient care.

# Github Link

https://github.com/TaylanOzgur96/Data\_Preparation-Machine\_Learning\_CA2