

#### Electrical & Computer Engineering & Computer Science (ECECS)

### Technical Report SPRING 23



**1**

**CONTENTS**

[HEART DISEASE PREDICTION 2](#_Toc130380066)

[Team Working On it 3](#_Toc130380067)

[Abstract 4](#_Toc130380068)

[Data Collection 5](#_Toc130380069)

[Data Cleaning 6](#_Toc130380070)

[Data Understanding 7](#_Toc130380071)

[Data Preparation 9](#_Toc130380072)

[Methodology 7](#_Toc130380073)

[Contributions/References 8](#_Toc130380074)



University of NewHaven Tagliatela college of Engineering

Distributed and Scalable engineering

Spring 2023

## HEART DISEASE PREDICTION

## Team Working On it

Sravan Kumar Reddy Pebbeti (Data Analyst)

Rakshala Mukka (Data Engineer)

Keerthana Krishnamoorthy (Machine Learning Engineer)

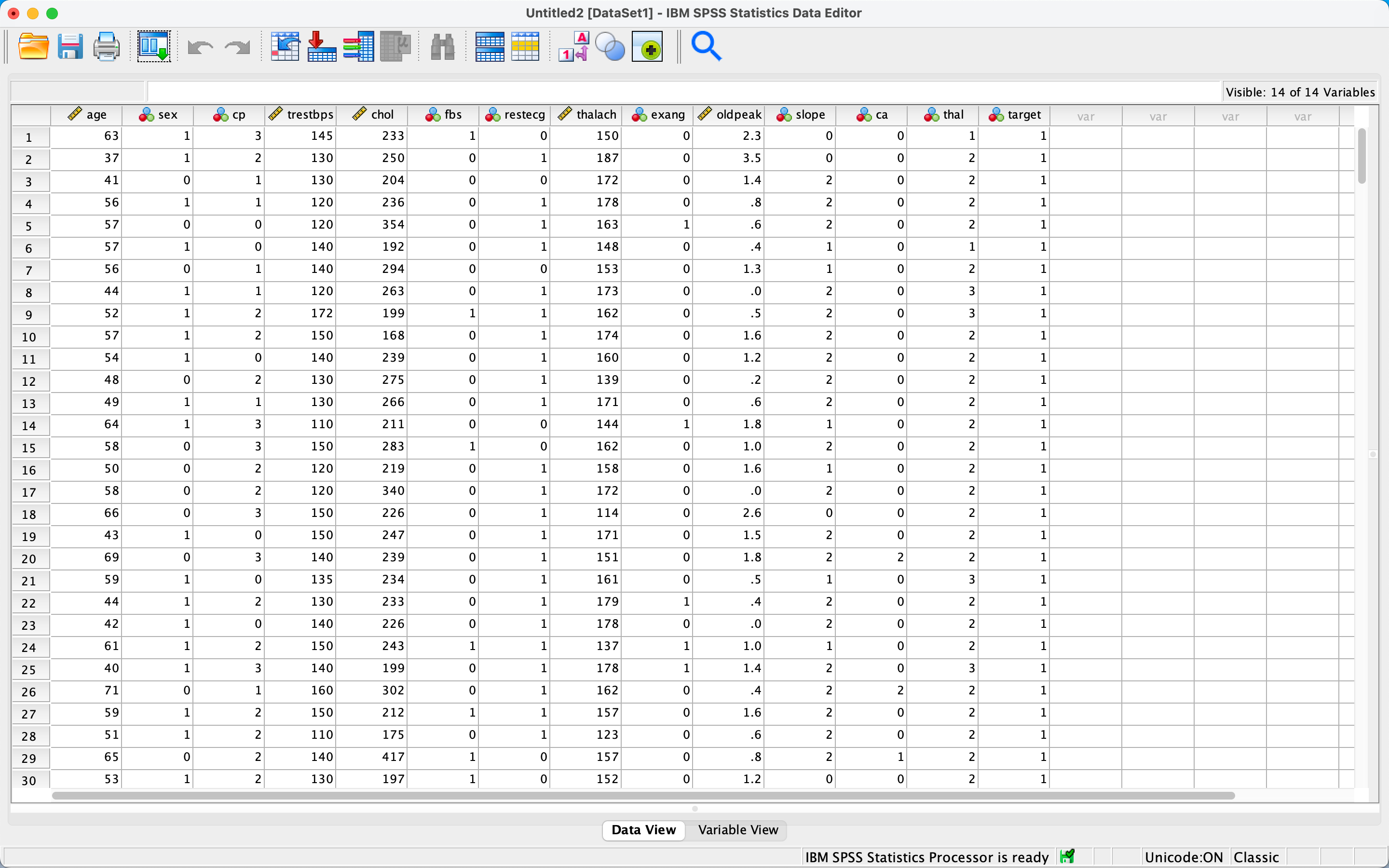
Svachuta Siva Sai Krishna Prasad Gollavilli (Lead Data Scientist)

## Abstract

The objective of this project is to determine if a patient is suffering from heart disease by employing various detection techniques. To achieve this, the study utilized different models such as Logistic Regression, Random Forest, and Decision Tree Classifier, which were compared to obtain precise predictions of the patients' conditions. This intelligent system leverages artificial intelligence (AI) strategies to make the most accurate disease predictions based on patient information. Patients can then seek advice from medical experts based on the system's output. The study employed a dataset that categorized patients as having or not having heart disease based on specific characteristics. The aim was to use this data to develop a model that can predict whether a patient has the disease or not. The findings demonstrate that all four models can be utilized to detect heart diseases in patients; however, the study identifies the most effective model by comparing the accuracy.

## Data Collection

Our data collection involved systematically gathering and measuring information on specific variables of interest, which enabled us to address important questions and make informed decisions about the future. We took great care to ensure that the data we obtained was both relevant and validated. To ensure the quality and validity of the data, we opted to use Kaggle as our source site.



## Data Cleaning

In order to improve the quality of our information and make more accurate decisions, we carried out a process known as Data Cleaning on our raw data. Here are the steps we took to achieve a cleaner dataset:

1. We eliminated any duplicate values that arose from combining multiple datasets and removed irrelevant observations that were not relevant to the problem at hand.
2. We addressed missing values using a variety of techniques, such as Imputation, dropping features or observations with missing values.
3. We reformatted the data types, such as Boolean, numeric, and Datetime, as needed.
4. Finally, we restructured and validated the data by reformatting strings and ensuring data consistency.Top of Form

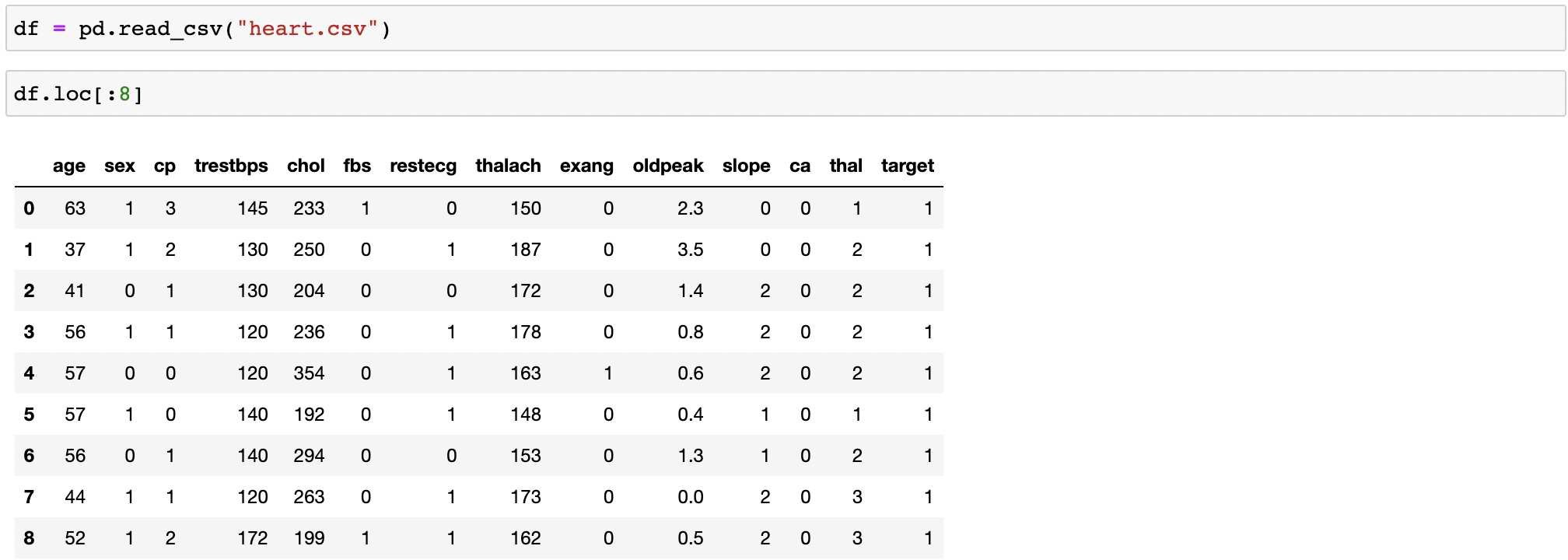
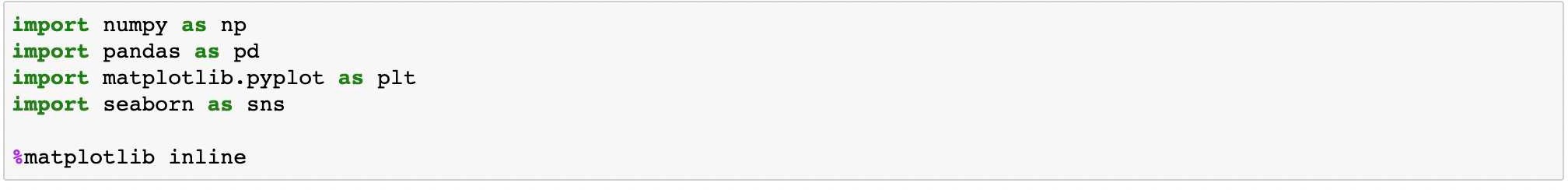
## Data Understanding

We have imported the data from the data set using different libraries like numpy, pandas, matplotlib and seaborn. After loading the data, we read the dataset. These are the parts of the data loading. Our dataset consists of fourteen columns which are

1. Age: patient’s age
2. Sex: Gender of the patient (Male = 1, Female = 0),
3. Chest pain type(CP): Includes pressure, fullness, burning or tightness in the chest .
4. Trestbps: A person’s resting blood pressure >120.
5. Chol: cholesterol levels. Good = less than 200, Moderate = 200-239, High = >240.
6. Fasting blood sugar(FBS): Measures the blood sugar after the overnight fast. 0 = normal, 1 = abnormal.
7. Resting electro cartographic results(rest ECG): Normal heart rate beats 60-100 per minute. 0 = normal, 1 = abnormal.
8. Thalach: Persons maximum heart rate achieved.
9. Exercise included angina(exang): The pain or stress that occurs after exercise. 0 = normal, 1 = abnormal.
10. Oldpeak: Exercise relative to rest. Records the slope of the peak exercise whether up, flat or down. Ranges between 0 to 5.
11. Slope: It is ST segment shift relative to the exercise-induced increments in patients heart rate.
12. CA: Coronary artery disease. Patients records of having CA. 0 no records, 1 or 2 had the cardiac attacks in the past.
13. Thal: A blood disorder called thalassemia.
14. Target.

By providing all the fields required we can predict whether the patient has the heart related diseases or not. If our end result is the integer valued 0 then that particular patient has no heart disease and if the result is integer 1 then the patient has heart disease. We used info, dtypes, shape, describe to understand our dataset.

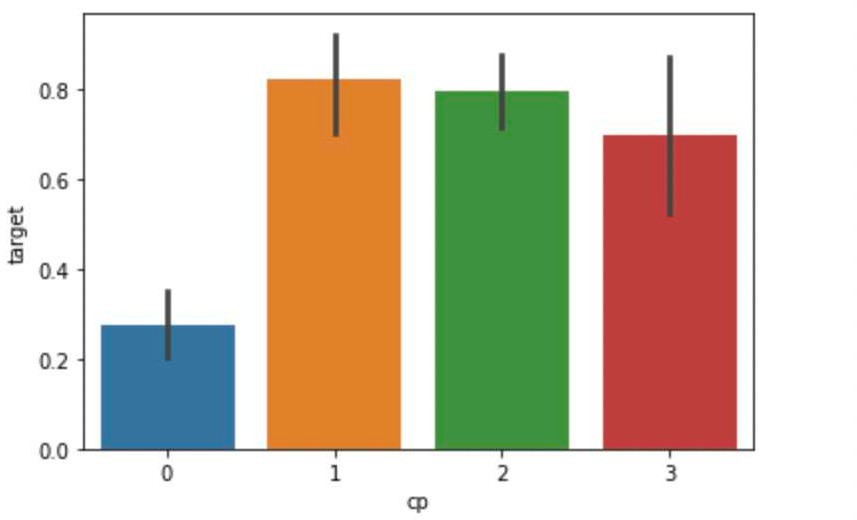
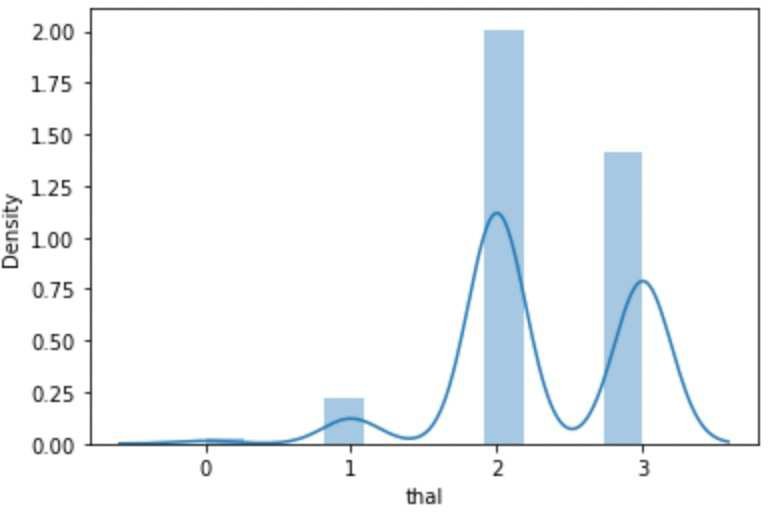
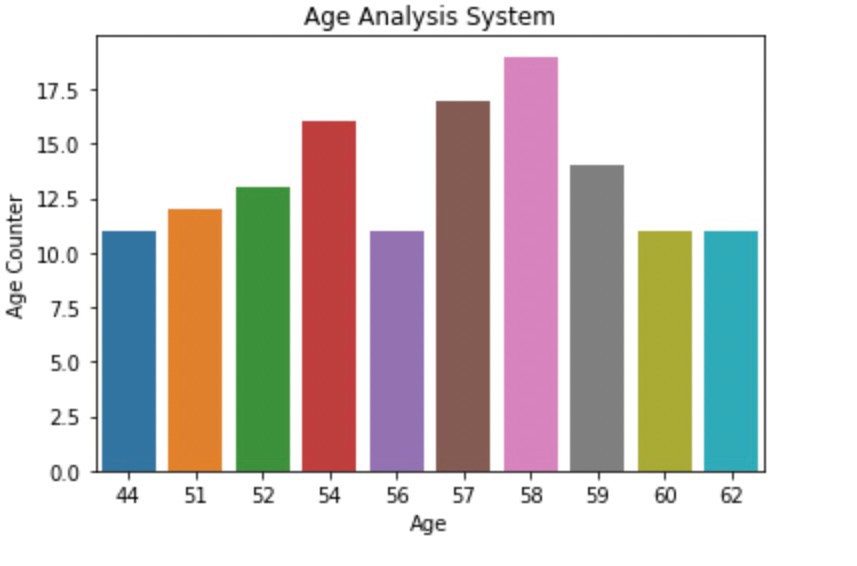
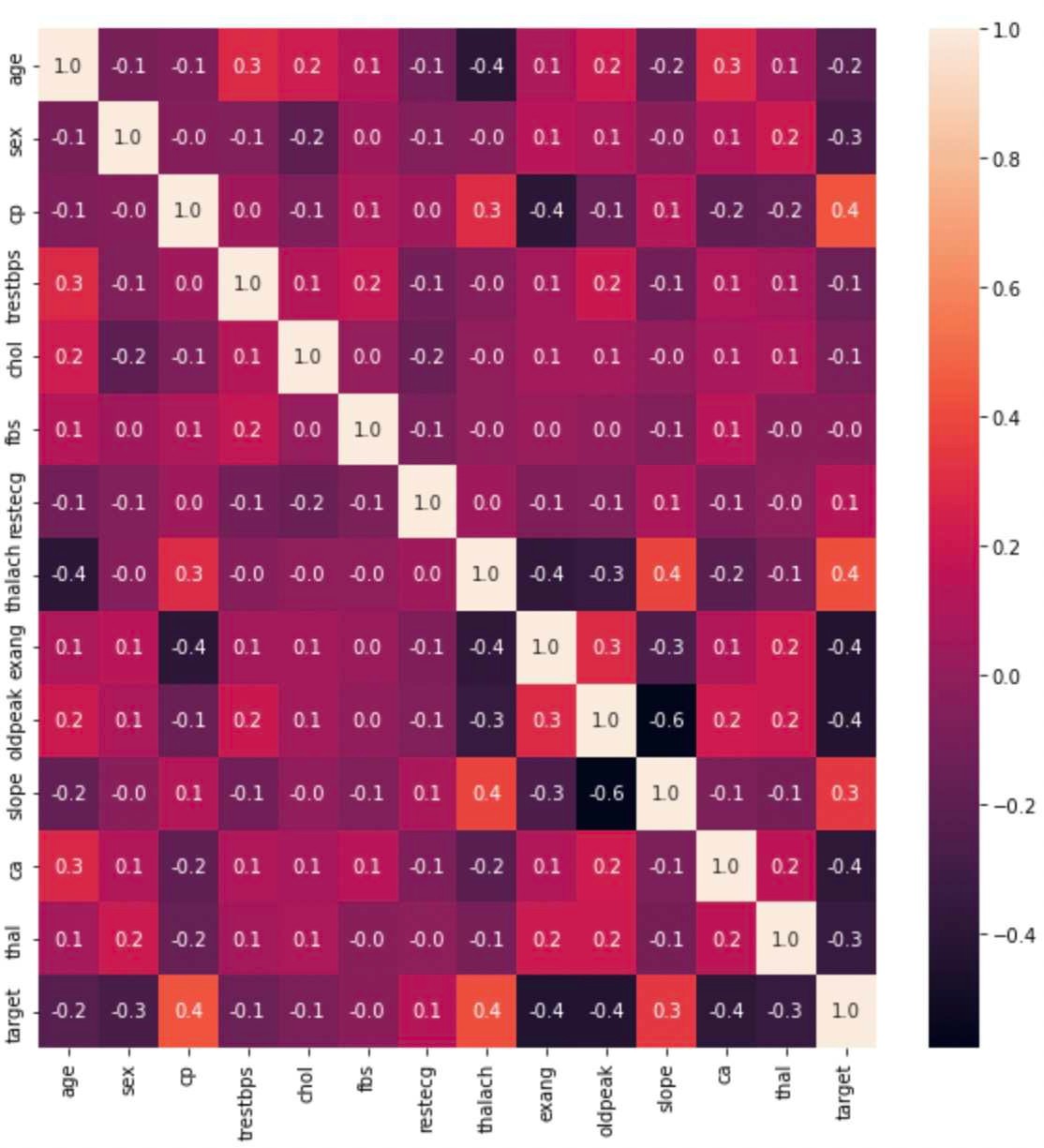
**6**



## Data Preparation

We conducted Exploratory Data Analysis (EDA) to identify any underlying patterns in the data, detect outliers, and test assumptions in order to find a model that best fits the data. EDA generally includes four main types of analysis, which are:

1. Univariate non-graphical analysis: This involves observing the population and analyzing the distribution of a single variable, such as measures of central tendency, measures of spread, and outlier detection.
2. Univariate graphical analysis: This involves using graphs to analyze a single variable, such as histograms, box plots, and stem-and-leaf plots.
3. Multivariate non-graphical analysis: This involves using techniques to analyze the relationship between two or more variables, such as covariance and correlation.
4. Multivariate graphical analysis: This involves using graphical methods to analyze the relationship between two or more variables, such as bar plots and scatterplots



## Methodology

In In this project, we performed data modeling to develop the most effective approach for obtaining accurate results using three different algorithms: Random Forest Classification, Logistic Regression, and Decision Tree. We determined the best model for predicting heart disease in patients based on the highest accuracy rate achieved by the algorithms.

To evaluate the accuracy of our model, we utilized the train and test approach, which involved splitting the data into two sets: an 80% training set and a 20% testing set. We trained our model using the training set and then tested the model using the testing set. The accuracy rates of the algorithms were compared to determine which model was the most effective for predicting heart disease in patients.

## Contributions/References

* [https://www.kaggle.com/kralmachine/analyzing](http://www.kaggle.com/kralmachine/analyzing-the-heart-disease)-the-[heart](http://www.kaggle.com/kralmachine/analyzing-the-heart-disease)-[disease](http://www.kaggle.com/kralmachine/analyzing-the-heart-disease)
* https://towardsdatascience.com/simple-linear-regression-model-using-python- machine-learning-eab7924d18b4
* https://github.com/jupyterlab/jupyterlab
* [https://www.atmosera.com/blog/creating](http://www.atmosera.com/blog/creating-a-simple-linear-regression-machine-)-a-[simple](http://www.atmosera.com/blog/creating-a-simple-linear-regression-machine-)-li[near](http://www.atmosera.com/blog/creating-a-simple-linear-regression-machine-)-[regres](http://www.atmosera.com/blog/creating-a-simple-linear-regression-machine-)s[ion-machine-](http://www.atmosera.com/blog/creating-a-simple-linear-regression-machine-) learning-model-with-scikit-learn/
* [https://www.ibm.com/support/pages/producing](http://www.ibm.com/support/pages/producing-line-graph-dual-y-axes-spss-statistics)-line-[graph](http://www.ibm.com/support/pages/producing-line-graph-dual-y-axes-spss-statistics)-[dual](http://www.ibm.com/support/pages/producing-line-graph-dual-y-axes-spss-statistics)-[y-axes-spss-statistics](http://www.ibm.com/support/pages/producing-line-graph-dual-y-axes-spss-statistics)
* https://[www.open.ac.uk/socialsciences/spsstutorial/files/tutorials/graphs.pdf](http://www.open.ac.uk/socialsciences/spsstutorial/files/tutorials/graphs.pdf)

https://jakevdp.github.io/PythonDataScienceHandbook

Thanking You