INFO411

Assignment 2 Report

Submitted By: Jacques Klavs

Daniel Pienaar

Submission Date: 07/10/2023

Contributions:

* **Exploratory data analysis**: Jacques and Daniel
* **Data imputation**: Jacques and Daniel
* **Modelling**: Jacques
* **Deliverables**: Jacques and Daniel
* **Dashboards**: Jacques and Daniel
* **Report**: Daniel

|  |
| --- |
| An overview of the findings and actions throughout Assignment 2 |

# Findings: EDA

## Understanding the data

Understanding the data, its structure, the element type and how it was first gathered and used is essential. Knowing which variables are categorical (ordered and not) or numerical is key to performing great analyses. A combination of reading the published paper and studying the UCI ML website (and its Variables Table) guided us in understanding what each variable meant, and how it should be used. We believe this is the correct way of doing things, because it informs future procedures.

Much of the graphics developed in the EDA section of the assignment required an understanding of the type of data we were working with. This was of course designated, but only after cleaning it. The clean stalog-heart.data.txt dataset was great to work with to make sure graphing techniques could work, which helped figure out what could be done with the DS2 dataset, which had missing data.

## Statistical overview

Learning about what actually could be a determining factor for heart disease was eye-opening. Learning that some conditions may not contribute much to heart disease presence was just as eye-opening. Comparing plots and studying the correlogram helped discern what should and shouldn't be used.

## Key finding

Seeing how much Exercise Induced Depression correlates to the presence of heart disease was frightening. It is a contrast that is more significant than most relationships. This emphasises the need to understand mental health, and how the brain acts and reacts physiologically if we are to prevent heart disease in the future.

# Findings: Data imputation

## The DS2 dataset

We chose to use the Hungarian, Swiss and VA datasets. Much of the data was missing. Julia does make it very easy to rectify this, however.

After some cleaning that was performed in the same way as in the EDA section, a simple application of the Julia Impute package was performed. This filled in all the missing data. EDA could then immediately be performed.

## Results of the EDA on DS2

Statistical profiles, not surprisingly, varied significantly across the datasets. There were more people in the Hungarian dataset, but older people in the VA dataset. The Hungarian dataset appeared to be filled with healthier people - possibly due to age - but also had a great portion of people experience rank 2 chest pain, the second highest in rank. The Swiss and VA test group participants followed a similar descending profile to that of DS1 from rank 4 to 1.

Seeing how people of different ages in different locations fit within the scope of the presence of heart disease according to blood pressure was interesting: high concentrations of people 10 years apart do not have heart disease in Hungary, but low concentrations do in this same location spread out more evenly across the age groups. There appeared to be some sort of correlation between blood pressure and heart disease routinely.

# Findings: Modelling

# Findings: Dashboards

# Conclusion