### INFO411

Assignment 2 Report

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

**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

- **GitHub**: Jacques and Daniel

Contents

[1. Findings: EDA 2](#_Toc147660962)

[2. Findings: Data imputation 6](#_Toc147660963)

[3. Findings: Modelling 11](#_Toc147660964)

[4. Findings: Dashboards 11](#_Toc147660965)

[5. Conclusion 11](#_Toc147660966)

# 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.

## Statistical overview

Learning about what 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. The plots and correlogram are shown following this paragraph.

The figure below shows frequencies of males and females according to various measurements:

* Chest Pain Type: Most participants reported chest pain measurements of 4, which is asymptomatic.
* Exercise Induced Angina: Most participants did not experience exercise-induced angina.
* Fasting Blood Sugar: Most participants did not have high fasting blood sugar, above 120 mg/dl.
* Resting Electrocardiographic Results: There was an almost even split between participants who had normal results (0) and possible or definite left ventricular hypertrophy by Estes' criteria (2).

The pattern between males and females remains very similar across all plots.

A screenshot of a graph

Description automatically generated

Figure 1: Male and female statistics from the statlog-heart.data file

The figure that follows dives into the presence of heart disease according to several features. The following can be gathered from the figures as an indication of the presence of heart disease:

* Higher resting blood pressure with lower maximum heart rate achievement.
* Higher resting blood pressure with higher serum cholesterol.
* Lower age and lower maximum heart rate achieved.

There is not a clear indication of how serum cholesterol at different ages affects your chances of getting heart disease. It appears that the factors out of the ones below indicate that resting blood pressure and maximum heart rate achieved during exercise could be best used for determining the presence of heart disease.

A collage of different colored dots

Description automatically generated

Figure 2: Heart disease presence according to various parameters in the statlog-heart.data file

## Key findings

Seeing how much Exercise Induced ST 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 ST depression, and how the brain acts and reacts physiologically if we are to prevent heart disease in the future.

The correlogram that follows investigates the correlation each variable has with each other. The following is a summary on what can be discerned.

Low correlation:

* Resting Blood Pressure and Maximum Heart Rate.
* Maximum Heart Rate and Serum Cholesterol.
* Serum Cholesterol and Exercise Induced ST Depression.

Highest correlation:

* Resting Blood Pressure and Age.
* Exercise Induced ST Depression and Resting Blood Pressure.
* Serum Cholesterol and Age.
* Exercise Induced ST Depression and Age.
* Maximum Heart Rate and Age.
* Exercise Induced ST Depression and Maximum Heart Rate.

What can be inferred from this is that, if we understand that Exercise Induced ST Depression might be a strong determining factor for the presence of heart disease, the following variables should be avoided in favour of using ST Depression as a variable for a predictive model:

* Maximum Heart Rate.
* Age.
* Resting Blood Pressure.

The best other determining factor that should be considered is Serum Cholesterol.

A screenshot of a graph

Description automatically generated

Figure 3: A correlogram of select features

The figures below are density plots of age versus resting blood pressure and ST depression induced by exercise relative to rest. The density plots are separated between those with and without heart disease.

In both instances, age does not seem to influence the presence of heart disease much. On the other hand:

* Resting blood pressure looks to be somewhat positively correlated with the presence of heart disease.
* High measures of ST depression induced by exercise relative to rest look to be very strongly correlated with the presence of heart disease.

A screenshot of a graph

Description automatically generated

Figure 4: Heart disease vs age, resting blood pressure and ST depression

The plots below illustrate the frequency of males and females who have and do not have heart disease according to age. The following can be drawn from these figures in the scope of the test study group:

* There are more males who do not have heart disease than those who do not.
* There are more females who have heart disease than those who do not.
* Most of the males and females who do not have heart disease are around 60 years of age.
* The population of females who have heart disease are evenly spread across the ages.

What could be inferred from this is that there are more active 60-year-old females than at any other age, in this study group.

A screenshot of a graph

Description automatically generated

Figure 5: Disease-free populations in males and females

# 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.

As a reference, a heatmap was generated to show the missing data before and after imputation. Blank spaces in the heatmaps correspond to missing values. This is a visual guide to be used for targeted data cleaning. In all cases, variable 12 is significantly unpopulated, with variable 13 and 11 coming in second.

These variables are:

* 11: Slope of the Peak Exercise ST Segment.
* 12: Number of Major Vessels colored by fluoroscopy.
* 13: Thal: 3 = Normal; 6 = Fixed Defect; 7 = Reversable Defect.

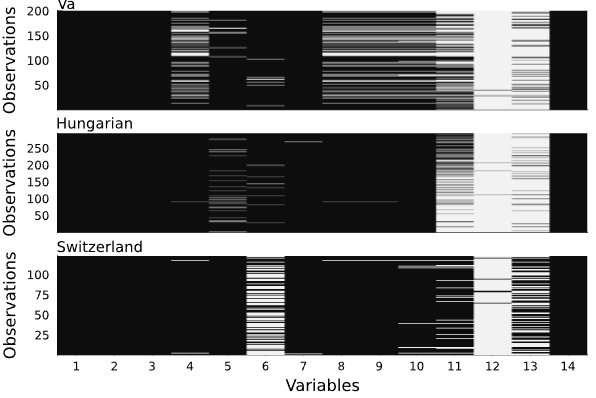


Figure 6: DS2 data sets before imputation

A black rectangular object with white text

Description automatically generated

Figure 7: Data sets after imputation

## 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.

The figure below shows that the Hungarian population represents the youngest age profile, VA the oldest, and the Swiss in the middle. The VA profile appeared to have far fewer participants, with a very small number of females which are concentrated in the 50–65-year range. The Hungarian profile seems to have a more even distribution for males and females compared to the other two datasets. The females in all cases have smaller age ranges than the males and appear to be smaller in number.

A diagram of different shapes

Description automatically generated

Figure 8: DS2 data set age and sex profiles across locations

For the second figure showing a histogram of Chest Pain Type, a ranking of 4 was most common across all datasets. The frequencies decrease as the ranking moves from 4 to 1, except in the case of Hungary, where rank 2 has the second highest number of reports.

A graph of different types of pain

Description automatically generated

Figure 9: DS2 data set chest pain type across locations

## Key findings

The following figures will explore more measures, how they contribute to the presence of heart disease, and compare. This will be on a basis of Age and Resting Blood pressure, across the three locations. For the first location, we have VA. Here, we can see that older people with higher resting blood pressure more than not have heart disease.



Figure 10: Heart disease in VA

For the second location, we have Switzerland. Here, we can see an interesting relationship. The density of people without heart disease is concentrated, but those without are not.

A graph with blue and orange lines

Description automatically generated

Figure 11: Heart disease in Switzerland

For the third location, we have Hungary. Here, there is no clear relationship between age, resting blood pressure and presence of heart disease.

A graph with blue and orange lines

Description automatically generated

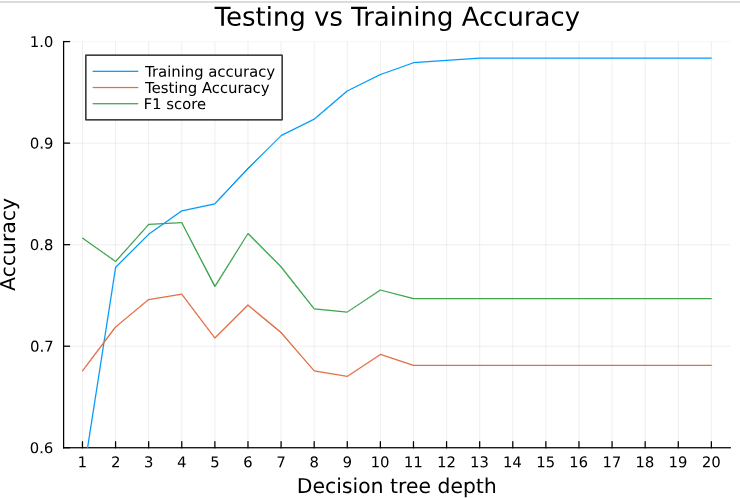
Figure 12: Heart disease in Hungary

# Findings: Modelling

To begin code was used to decide which models were compatible with modeling all the data (all 14 variables). This resulted in 5 models, two of which were used, a decision tree classifier and a random forest classifier. Then the data was divided into subsections with the continuous data subset allowing for more compatible models which will be discussed future down.

F1 value was decided on as a primary measure for the models as it provides a balanced measure of the model's performance that considers both the true positive and false positive rates of the model. Other measures such as accuracy and AUC were also recorded throughout the notebook.

A decision tree was initially implemented using the BetaML package. A plain tree was trained which resulted in a 10 fold CV F1 value of 0.7554, not bad for a first attempt. To gain an idea for the depth and how the resulting F1 score and accuracy was determined a test vs training plot was created for depths 1:20 as shown below.



As seen above the decision tree of depth 3 appears to have the best F1 score, and Testing accuracy is also good with training accuracy is not too much higher, this is showing a good ability to perform on new data for the model with depth 3.

Then the function TunedModel() was used to tune the model, using min\_gain, max\_depth and max\_features as parameters. Best features found were max\_depth = 3, min\_gain = 0.01, max\_features = 8, min\_records = 2, splitting\_criterion = gini.

This resulted in improved model performance and an F1 score of 0.824 along with increased accuracy. Mann-Whitney U test and two sample t test all showed difference between the two models indicating statistical significance.

Next a random forest was built. Unsurprisingly it outperformed the decision tree. Right off the bat a F1 score of 0.835 was given, parameters were then found and these were tuned (please see heatmaps.jl for tuning of the random forest). TunedModel was again used, max\_depth, bagging\_fraction and max\_features were all tuned as seen below.

A yellow and red gradient

Description automatically generated with medium confidence

A chart showing different colors of the same color

Description automatically generatedThis resulted in the following parameters, max\_depth = 2, max\_features = 4 and bagging\_fraction = 0.1. This improved the F1 value bringing it up to 0.855, accuracy was also improved. The roc curve can be seen below contrasting the tuned model vs the untuned model.

A graph of a curve

Description automatically generated with medium confidence

As evidenced by the Roc curve and the resulting two sample T test and Mann-Whitney U test the two models did not show difference that was statistically significant with the untuned model outperforming the tuned model on a selected fold, although as an average the tuned model appears to sit closer to the left corner of the graph, indicating high AUC and better performance, though not completely dominating in performance.

Given the two models that were trained on the entire data set, a Decision tree and the random forest, the tuned random forest performed the best and the f1 scores returned show statistical significance.

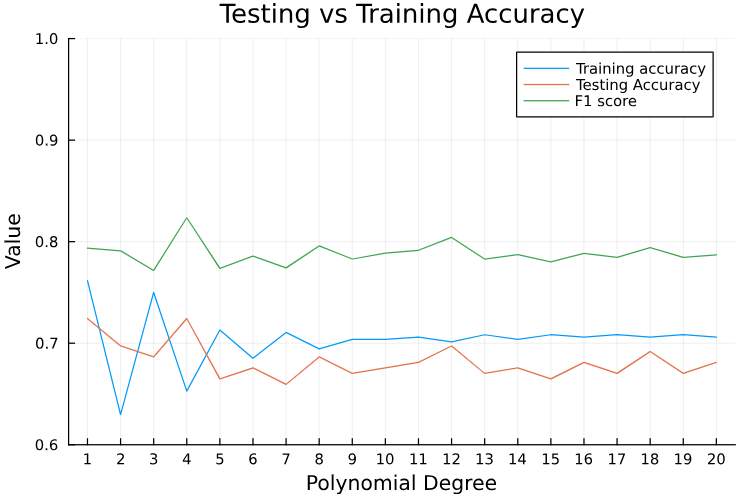
Feature importance using shapely values was also undertaken, shapely values are the weighted average of contribution to the model. As seen with the bar chart below there are a few leaders that are useful in determining heart disease status.

A graph with blue bars

Description automatically generated

From here a subset of data was created. The compatible models were then tested, with 54 models now compatible to the subset.

A support vector classifier was the first model to be created. The untrained model resulted in a F1 score of 0.768. This model was then plotted from 1:20 for polynomial degree with a model of degree 4 appearing to perform the best. This was not the case when cross validation was then used to test the models F1 score with a polynomial of 1 performing the best. This is not too supprising as the chart below also shows the degree = 1 performing well.



Overall, a F1 score of 0.797 was obtained. Unfortunately, the untuned and tuned f1 scores do not show any statistical significance, failing to reject the null hypothesis and very similar mean f1 scores. This model was outperformed by the decision tree.

Next, we trained a K nearest neighbor model. Untrained, its F1 performance was at 0.755. A weighted KNN was also trained. Then K was varied from 1:20 then 1:100 and tested as seen in the charts below.

A graph of a number of individuals

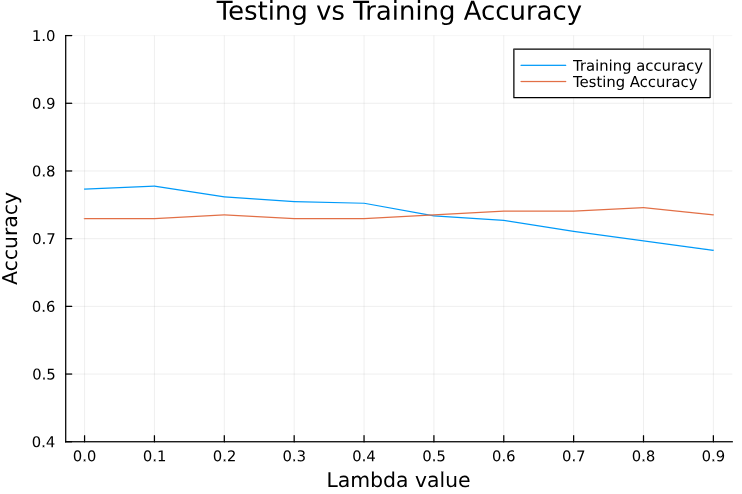
Description automatically generated

A graph of a number of individuals

Description automatically generated

Seen here K = 20 using a weighted (Inverse()) KNN model appears to perform the best. Different distance metrics were also tested, with Euclidean performing the best. This left us with the following results: Untuned KNN F1 Value = 0.755, Tuned Weighted KNN F1 Value = 0.788.   
These two models did not show any statistical difference in results when using the two sided T test and Mann-Whitney U test, although the p value was relatively close to 0.05, this meaning we are not 95% confident that the two models show statistical difference, but are more than 75% according to the tests. The random forest clearly outperforms the weighted KNN and this is backed up by the two tests.

Then a logistic classifier was created. Untrained with a F1 score of 0.797 this model performed surprisingly well. Lambda value was varied from 0.0:1 as shown below with not much effect, the untrained model was selected.



Regardless of the initial results the random forest outperformed the logistic classifier, and this was confirmed using the two statistical tests.

Finally, the last new model was built. This was the ridge classifier. Out of the box and untrained the model performed with an F1 score of 0.798. This model was also outperformed by the random forest. The model was tweaked while being build to trial a variety of settings but no score above 0.798 could be obtained, therefore base settings were kept.

PCA was then explored to see if it improved performance. PCA data was obtained from continuous data, then combined with categorical data and used in a random forest, this model did not outperform the original model as seen here:

Tuned Random Forest F1 Value = 0.852  
Tuned PCA Random Forest F1 Value = 0.848

This is not too surprising as random forests are robust and can handle high dimensional data well, although it did speed up the running of the model which may be beneficial in future use cases.   
A KNN PCA was then trained, with Minkowski(1.0) now in the best model parameters, although there was no improvement in results.

In conclusion, the random tuned random forest performed the best. F1 score of 0.852 Testing accuracy of 0.773 and training accuracy of 0.826.

# Findings: Dashboards

## The usefulness of dashboards

Using dashboards to explore some of the data allowed for some interesting observations. With density plots, peaks were formed based on how many layers of density were designated. This showed hotspots of the presence of heart disease, and also allowed for some sort of guesswork when it came to deciding how many layers were necessary.

The figures in the following sections show a comparison of different levels of density plot for the same figure, showing peak formation.

## The presence of heart disease versus ST depression and age (DS1)

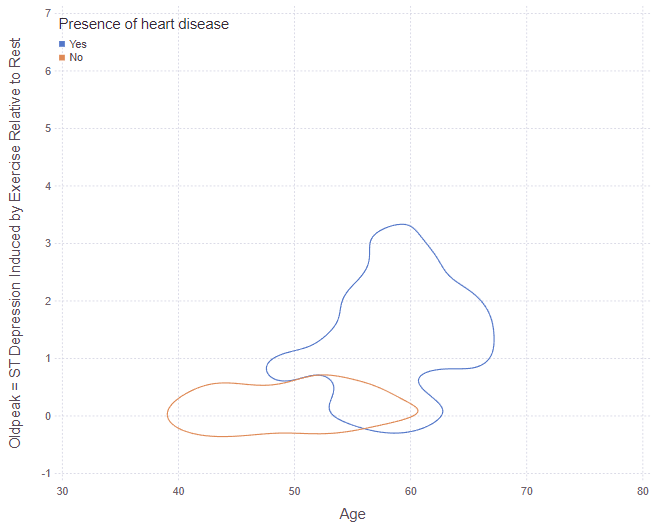


Figure 13: Heart disease, ST depression and age (Density layer level = 1)

A graph of a diagram

Description automatically generated with medium confidence

Figure 14: Heart disease, ST depression and age (Density layer level = 10)

In the figure where the density layer level = 1, it is shown that those with high ST depression induced by exercise values tend to have a heart disease more often than those with lower.

In the figure with the density layer level = 10, various hotspots can be seen at the ages of approximately 44, 52, 58 and 65, indicating that there may be some sort of age relation.

## The presence of heart disease versus resting blood pressure and age (DS2)

A graph with blue and orange lines

Description automatically generated

Figure 15: Heart disease, resting blood pressure, and age in Hungary (Density layer level = 1)

A graph showing a diagram of a person's age

Description automatically generated with medium confidence

Figure 16: Heart disease, resting blood pressure, and age in Hungary (Density layer level = 10)

In the figure where the density layer level = 1, it is shown that those with higher resting blood pressure values tend to have a heart disease more than those with lower.

In the figure with the density layer level = 10, various hotspots can be seen across many different ages, indicating that there may not be an age relation.

# Conclusion

The EDA of DS1 allowed for some interesting information to be extracted. Observing the correlations between features was important to see which should and should not be used for determining heart disease presence in a population. Observing the different statistical profiles between populations of different countries was equally as interesting, as it suggests people live wholly different lives. Observing that what appeared to be the most determinantal feature for heart disease presence was ST depression from induced exercise suggests that heart disease presence could be better predicted if people actively went for medical evaluations.