Data Mining and Decision Systems  
600092  
Assigned Coursework Report

Student ID: 201707824  
Date: 15th November 2019

## Due Date: 12 December 2019

**Report must be within 8 page maximum. Strict page limits will be enforced. Any extra pages will be ignored and no marks awarded for any work on these. Exclusions to this limit are the front page, the references section, and any appendices. Please keep to the given section headings and format; subsections are permitted.**

# Methodology

The methodology followed for this report, will be a slightly modified version of the popular CRISP-DM methodology. In this instance, we do not have the first and last stages of the methodology, which include the **business understanding phase** and the **deployment phase**, so these will be included in a hypothetical sense rather than a more general sense.

The task is to create a classification model with a cardiovascular dataset supplied to us. The aim of this is to identify if a patient is at risk or not at risk. The model will be of a binary nature as we are working with an if or else classification, also known as a 0 or 1 classification mode (binary classifier).

### Data Understanding

The first phase is the Data Understanding phase. The first step is to load the data into a pandas data frame and then use the describe function to look at the columns, which can be seen at output lines [2] to [6] of the code. The first task is to check what is in the dataset, and how closely it matches the expected values found within the data description provided.

There are observed differences, the most notable being the attribute ‘Contra’ not being a numeric value, instead being an object, and the total values being 1520, with some attributes having less records than the max, indicating missing or null values. There is also the attribute Indication, with 5 unique values, but 4 being expected, and the attribute label having 3 unique values, rather than the 2 that are expected.

The target attribute of the classification model will be label. This is due to all other attributes indicating if the patient involved has a risk of mortality or not. The data description is also helpful, as the description for the other attributes indicate steps taken to discover the risk of mortality of the patient, therefore a suitable target for a binary classification model. The data appears to have slight differences from what is expected through the data description, but the differences are small and will not require much to be changed.

### Data Preparation

Code blocks [7] to [23] of the notebook code show the data cleaning steps.

For the task, all features will be used besides Random and ID. This is because, while useful for indicating individual patients and their scores, they don’t have suitability towards the end goal, and would be the only columns I would consider excluding. As the model will be of a binary nature, these columns indicate high numerical values and can thus cause incorrect classification later

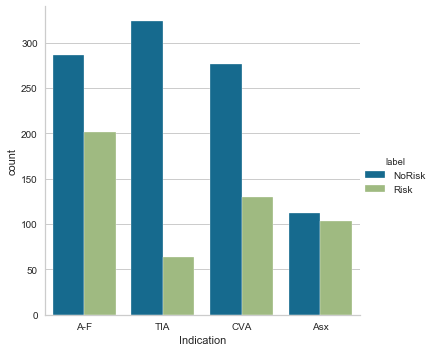
The first and most important step in the early stages of CRISP-DM is cleaning the data. The first task is to check the unique values that were found before that don’t add up with the data description starting with label. As shown in code blocks [7] to [9], there is a third categorical type in the label attribute of Unknown, which will be changed into a numpy value of NaN, which stands for not a number.

Once this is achieved, the feature columns of Random and ID are dropped from the overall dataset (code block [11]), and then the feature Contra is changed to be of a numerical type float64 using pandas at code block [13].

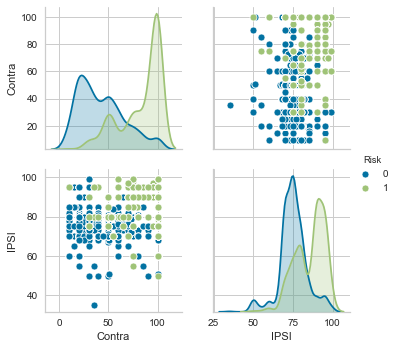
There is always the argument of imputing missing data or simply removing it. This depends on the amount of data loss, however. In code blocks [13] and [14] the null data is explored and is found to equal 20 values. Considering we have 1520 expected values and only 20 contain a null value, I decided to drop these values. There is an argument for imputation here, but due to this data being legacy, and the percentage of values missing being low (code block [14]), the data was removed instead.

The Indication is corrected later at code block [17], which is discovered to be a duplicate name, one which has an uppercase and the other a lowercase spelling causing a split in attributes of the column. This is corrected by simply renaming one to be equal to the other.

#### Visualizing the data

An important aspect of data preparation is data visualization. The graph to the right shows the indication value count against the classification target.

This count is useful, indication is a nominal value that describes what type of event the patient had prior to hospitalization. This is important towards the end goal, because the indication has a bearing on the outcome, as it defines the seriousness of the incident.

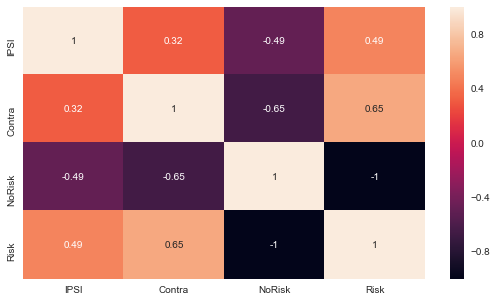
Let’s now look at the two predominant numerical values of IPSI and Contra from the dataset.

The graphs here show that Contra values have a higher range across the dataset, but the higher numeric values indicate risk more than the lower values do. However, IPSI values are generally right skewed as a whole, with the distribution of risk and no risk patients generally falling between 70 and 100. What is important is not all high value IPSI patients have a risk outcome.

The plots showing IPSI vs Contra in the top right and bottom left, show that there is no true correlation between the two, this can be seen more on the following page.

But what is correlation exactly?

Correlation is a measure of the linear relationship or connection between two or more things, better thought of as, if one value increases, what does the other value do? Continuing with this trend towards IPSI and Contra, it is important to consider the measure of Correlation.



The correlation heatmap shows a visual look at this. The important things to look at here, are the value of 0.65 for Contra against Risk, and 0.49 for IPSI against Risk. What this means is that, an increase in Contra has a 65% correlation with a risk classification, whilst IPSI is better with a 49% increase. No Risk against these two valuations has the inverse effect, causing a decrease in the same numerical measure independent of the other.

The score for IPSI against Contra is 0.32, indicating a weak correlation between the two, which was also seen the plot above. The Distribution plots for IPSI and Contra can be found either in the code at blocks [28] and [29] or within the appendix. These plots aim to better distinguish the spread of the data based on these values.

Provide details on the methodology applied towards the data mining analysis undertaken, providing rationale for these steps.

This should detail how you went from the raw data provided to the chosen model(s), choice of model, and how this methodology helps address the problem domain.

Evidence to support the following of this methodology should be presented, especially any cases which required moving backwards in the process to readdress issues.

# Results

Results should include tables showing model performance with appropriately selected metrics. No rationale should be provided for this section - simply results of evaluative processes.

If using modified variants of the dataset, these should be clearly identified in the tables with appropriate naming. The justification and description of modification is not for this section.

Additional figures may be used as appropriate, in support of discussion points in the Evaluation & Discussion section, or as evidence for methodology following above.

# Evaluation & Discussion

Evaluation methodology used for generating the results provided in the previous section. How were these evaluated? Why was this selected? What metrics were used and why?

Discussion of the results should be presented with appropriate evidence and rationale. E.g Which is the best model, and why?

Consider each stage in the methodology, and reflect on any improvements which could have been made. Could any techniques have been used which may have improved performance? Why?

# References

Any references used throughout the report should be included here in Hull Harvard Style. If no references used, remove this section.

# Appendix

