Data Mining and Decision Systems  
600092  
Assigned Coursework Report

Student ID: 201707408  
Date: 07 October 2019

# Methodology

## Business understanding

The data will be analyzed using the CRISP-DM. The data consists of visits to a hospital due to cardiovascular complications. The model being developed will be used to identify patents that are at risk, and therefore there are significant ramifications if the model is malformed therefore false negatives must be avoided when selecting a model. Data protection is also a significant concern however as we only have access to numerical identifiers and therefore identifying individuals would be extremely difficult. The model itself will be used to classify patients to risk or no risk depending on a variety of features, this however may have to be overseen by a trained medical professional due to regulations in the industry.

## Data understanding

Reading the data description gave the data scientist insight into the domains of the columns and suggested. Comments were also provided specifying what each column means along with non-clinical descriptions. This allowed the data scientist to research how certain features could affect risk. Certain medical were harder to decipher such as ischemic and contralateral. Contralateral means opposite side of the body and Ischemic is defined as lack of oxygen leading to necrosis or cell death. Therefore, this means IPSI and Contra are the percentage of oxygen lacking lesions for each side of the body, and the closer they are to 100% the more risk that is involved. This would be verified with a medical professional or the data owner however this is one of the issues working with legacy data.

Looking at the data types and using the describe function in pandas allows the data scientist to get a good overview. Session ID did not seem to be provided and random seemed to have duplicates which did not match the data description which implied that either the data description or data is wrong. The unique values per column showed that many columns contained invalid values as many columns could only have binary values or certain other sets as is the case for the Indication feature.

## Data prep

Due to the data containing invalid or incomplete values, cleaning was required. Removing or imputing these data values would allow for more consistent and accurate visualizing and models.

Using the data description as a guide to ensure data integrity, the data scientist checked the data types as this would aid with finding the domain of the columns, allowing him/her to spot and invalid values.

Firstly, the entire random column seemed to be to contain duplicates, therefore having a closer look at these data points suggested that these we not identical patents as they had varying features even if random was identical.

Setting an Index would be useful to call individual records so the data scientist can pull records by ID and could review them individually. ID was chosen as it was already unique and therefore could also be integrated into other legacy systems when deployed.

Contra was detected as an object, this made numerical analysis of the feature difficult. Forcing the column to numeric would make invalid values into NaN values which could be then be imputed or dropped.

Checking each feature’s unique values showed ‘Unknown’ in the label column, as the model being trained would rely on the label column keeping these records would be meaningless. ASX had 2 variants and with further research it was found there was only one ASX and therefore the capitalization was fixed to be inline with the other indication values.

Listing all the null values allowed the data analyst to decide whether to impute or drop the records containing null values. However, there were only 20 null records in total which made up only 1.32% of all the data and therefore removing them would not be detrimental to the models.

Random and ID did not affect the label and therefore was dropped as this would just interferer with the model training process.

### Visualization

This not a part of crisp - DM however visualization would be helpful in identifying key features that affect risk.

Encoding binary values to numbers allowed the models to be trained more effectively as machine understand 1 and 0 better than yes or no. This was due using label encoder as this would replace values as opposed to creating new columns like get\_dummies.

## Models

### Logistic Regression

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.