Data Mining and Decision Systems  
600092  
Assigned Coursework Report

Student ID: 201602395  
Date: 07 October 2019

## Due Date: 12 December 2019

# Methodology

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.

A large set of legacy data from a domain of cardio-vascular medicine has been provided to be understood and analysed.

CRISP-DM methodology has been used to complete this task by filtering, cleaning, and transforming the data as appropriate such that it can be used to produce optimal classification for patient risk of dying.

UNDERSTANDING OF DATA (BUSINESS)

The data provided consists of over 1500 patient records each containing 11 attributes describing factors of the patients health and numerical data such as ID. The first attribute ‘Random’ displays a randomly generated number. This is used to help in randomly sorting the data. Next is the ‘ID’ column which is a unique identifier for each patient. The third attribute ‘Indication’ is a categorical value expressing the cardiovascular event which triggered the hospitalisation. This

Data set consists of over 1500 patient records each containing 11 attributes describing factors of their health.

* Random number provided to help in randomly sorting the data records.
* Unique ID to distinguish patients
* Indication indicating why the patient has been hospitalised using four values. “A-F”,”ASX”,”CVA”,”TIA”.
* A series of Yes or No attributes including Diabetes, IHD (Ischemic Heart Disease, or otherwise known as CHD (Coronary heart disease), Hypertension, Arrhythmia and History
* IPSI
* Contra
* Label

UNDERSTANDING OF DATA (DATA)

DATA PREPERATION

MODELING

# 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.