Data Mining and Decision Systems  
600092  
Assigned Coursework Report

Student ID: 561438  
Date: 18 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

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.

**Data Understanding**

**Data Cleaning**

In order to start creating models, it is crucial to make sure there are no invalid entries in the data set to avoid inaccurate results when making decisions.

When reading the .csv document, we need to make sure that each white space is read as a ‘NaN’ value as otherwise it will be counted as an actual entry.

To start with the data cleaning, we first need to find the unique values for each column to identify invalid entries, duplicate entries and even spelling mistakes. Next step is to count how many records have ‘NaN values in order to calculate the impact of those records in order to make a decision to whether replace the ‘NaN’ values or to drop the records that contain them.

When we identify all the unique values from different columns, we also find two ‘Unknown’ entries in the ‘label’ column. Since this entry does not relate to the rest of the entries, it is better to replace the value to a ‘NaN’ value just as we replaced the white space value.

It is also worth noticing that there is a spelling mistake in the ‘Indication’ column. This column contains 5 unique values (excluding the ‘NaN’ values) for which ‘ASx’ and ‘Asx’ are repeated. Since the values are the same but entered in different capitals, it is best to opt for a replacement to all capital ‘ASX’ just as the rest of the values in the ‘Indication’ column.

After replacing all the invalid values to ‘NaN’ values, it is now time to make a decision to what to do with records that contain that value. For this, we need to compare the valid records with the invalid records containing the ‘NaN’ and calculate how much it will impact in the final decision. Records containing ‘NaN’ values are equivalent to 1.3% of the whole record, for which it is better to drop those records as it will not make a big impact on the final decision.

Another issue with the data is that the ‘Random’ column is not unique. The random column was supposed to represent a unique record number for each record. But, the whole column does not

# 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.