Data Mining and Decision Systems  
600092  
Assigned Coursework Report

Student ID: 201601628  
Date: 07 October 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

**Introduction**

This project looked to solve a toy classification problem, using patient medical data to classify patients as “Risk” or “NoRisk”.

A context-adapted CRISP-DM methodology was followed, limiting the focus on the, “Business Understanding” and “Deployment” phases, but still utilising its iterative nature. This document outlines the processes used whilst adhering to this framework, the results achieved, and a discussion of findings.

**Business Understanding**

This phase sought to understand and summarise the problem at hand, giving due consideration to the domain based on the project specification and personal domain knowledge. This is summarised below:

* **Domain:** Cardio-vascular medicine / healthcare.
* As a healthcare dataset it may be "natural", anonymised patient data, study data (e.g. clinical trial), or an aggregation of many different datasets (“warehoused”).
* There is a chance there is "control" data (healthy cohorts) within the dataset or, similarly, focus groups that consist of unhealthy cohorts.
* Due to the largely subjective nature of clinical diagnosis (i.e. different doctors with varying levels of experience make the diagnoses), it's entirely possible that some data is mislabelled (has the wrong classification).
* It is also possible that some diagnoses or features are self-certified or derived from incorrect patient assumptions (e.g. "Yes, I have a history of...").
* **Inputs:** Tabulated patient data; up-to 1520 records of 11 features.
* **Outputs:** “Risk” or “NoRisk”.
* **Problem type:** (Binary) Classification.

**Data Understanding**

This phase focused on gaining a high-level understanding of data correctness and patterns in the data. Most time was spent in this phase since a thorough understanding of the data allows for more effective choice of tools and quicker reasoning in the face of poor results.

One problem with this phase is that the issues found affect all other details being inspected. It’s possible to conduct infinite inspections on possible combinations of the data. For this reason, each step in the data understanding was carried out on a copy of the raw data, with the awareness that some findings might affect others at the cleaning stage (e.g. some null values may be irrelevant because the decision to drop the feature with the missing value).

1. Dictionary contradictions
2. Missing values
3. Duplicates
4. Distribution
   1. With corrections
   2. Without corrections
5. Outliers

**Data Preparation**

1. Raw data
2. Cleaned data (removed discrepancies)
3. Data with imputed values
4. Transformation – for all
5. Stratification (TTS) – for alll

**Modelling**

**Evaluation**

**Deployment**

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

A

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

A

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.