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.

**Methodology**

Cross-industry standard process for data mining, also known as CRISP-DM, is a methodology commonly used by data scientists when analysing big amounts of data.

This methodology has 6 steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. Each one of these steps help gathering the right information for a specific purpose. During the first 4 steps, it is quite common to backtrack between them in order to understand the data and be able to take right action in order to get the right results for the models.

Business Understanding and Data Understanding allow the data scientist to understand the final objectives and the business perspective. Once having gathered the right raw data it is time to understand and get familiarized with it.

During data preparation, data scientists study the data more in depth and find any anomalies that the data contains. These anomalies can be any incorrect values that do not fit in the data, such as NaN values, misspelt values or characters that do not correlate with the rest of the values in the data. It is also helpful to rename certain labels to relate them to the data set they are given or even drop certain information if it does not contribute towards the data modelling step.

Next step is Data Modelling. During this step it is common to revisit some of the previous steps in order to adapt the data to make the required models. Some models such as decision trees and confusion matrixes require the variables in the data to be numeric or binary, for which its necessary to go back one step (to Data preparation) to convert the data entries into binary.

Once finished with the Data Modelling, it is now time for the Evaluation. Now we need to look at the model see if their predictions are similar and if every aspect of the business has been considered during the whole process.

The final step is Deployment. In this step, all the models are

Scrum Agile is another methodology used in software development which is helpful at reducing time from the beginning of the development to the final deployment in the real world. When compared to the CRISP-DM methodology, Agile can be more flexible as it does not have a specific development order, where as CRISP-DM is much stricter to its steps in order to achieve the outcome. On the other hand, when using the CRISP-DM methodology, it is possible to step back from first 4 phases if needed, where as it is not as easy when following an Agile methodology.

**Business understanding**

The data we are going to analyse is about medical records. This data comes from 1520 patients from the real world. There are 8 columns representing different diseases that patients may or may not have and one column representing whether the patient is on risk or no risk. The Random column represents each of the records for every visit of the patient. The Id column represents each patient which can be repeated. This is because a patient can revisit a clinic and have one more record with new symptoms which would become a new record with the new values. This can be helpful as this will show whether the new record would make an impact on the risk.

Once we understand the data and we know what each column represents, we can proceed to the next step, which is data cleaning.

**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 models.

First of all, we need to read the .csv file and make sure that any entry that only contains a space is read as a ‘NaN’ value. Then we need to make sure that the original document does not get affected by the changes we are going to make during the data cleaning process. For this, it is best to make a deep copy. This will create a parallel copy in memory, so when we make changes, it does not affect the original document.

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 is equivalent a ‘NaN’ entry, it is better to replace it to a ‘NaN’ value just as we replaced the white space value. Another issue with the ‘label’ column is the name. The name has no relation to the data, for which is better to rename it to ‘Risk’, since the values that the column contains are ‘Risk’ and ‘NoRisk’.

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.

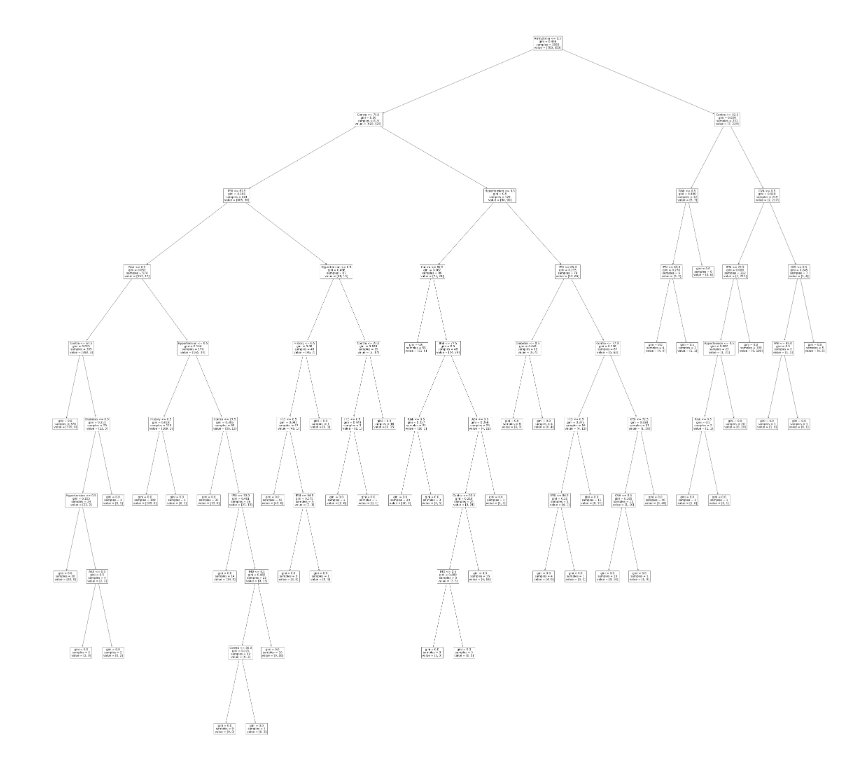
After replacing all the invalid values to ‘NaN’ values, it is now time to make a decision about what to do with records that do not contain a value. For this, we need to compare the valid records with the invalid records containing the ‘NaN’ and calculate how much it will impact in the models. Records containing ‘NaN’ values are equivalent to 1.3% ***(Appendix, Section 1.0)*** of the whole record, for which it is better to drop those records as it will not make a big impact on the models.

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. If we count the amount of unique values in the ‘Random’ column, we get 84.4% unique and 19.6% of duplicate values. After knowing that there are about 20% repeated values, this defeats the purpose of the entire column. The ‘Id’ column has a similar purpose, which represents every patient’s unique id. However, having an identifier for each record and for each patient does not influence in the result when making the models. Since these columns do not help in getting a result, it is better to drop them.

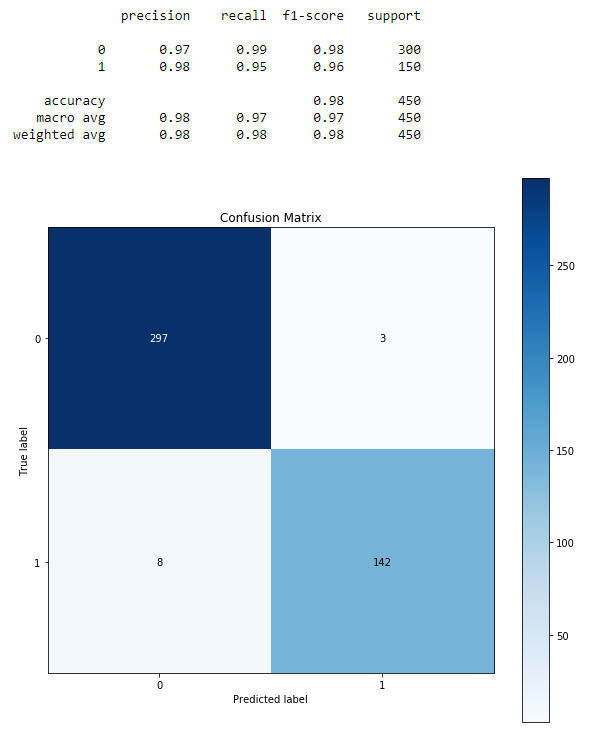
Now that we have all the records with ‘NaN’ values, the ‘Random’ column which contains repeated values and the ‘Id’ column which serve no purpose, we will proceed to drop them all and also replace the name of the ‘label’ column to ‘Risk’ to understand better the data included in that column.

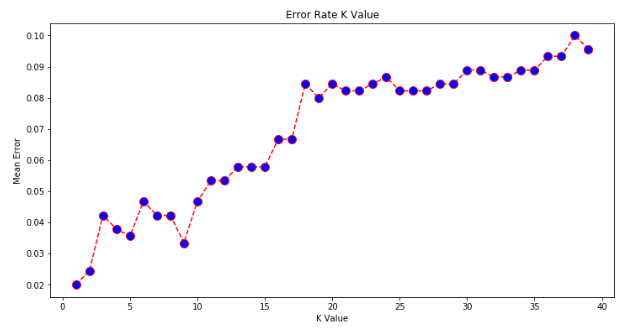
**Modelling**

# Results

***Figure 1. Decision Tree***

***Figure 2. Confusion Matrix based on the Decision Tree results***

******

***Figure 3. K-Nearest Neighbours Algorithm (KNN)***

# 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

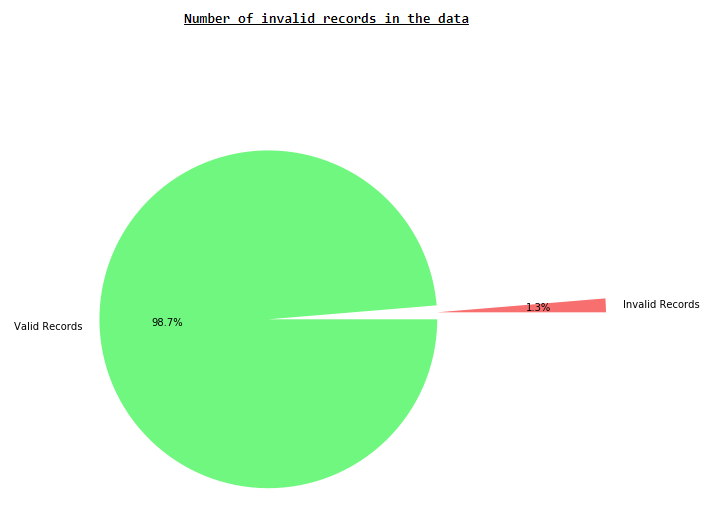
Any references used throughout the report should be included here in Hull Harvard Style. If no references used, remove this section.

<https://www.sv-europe.com/crisp-dm-methodology/>

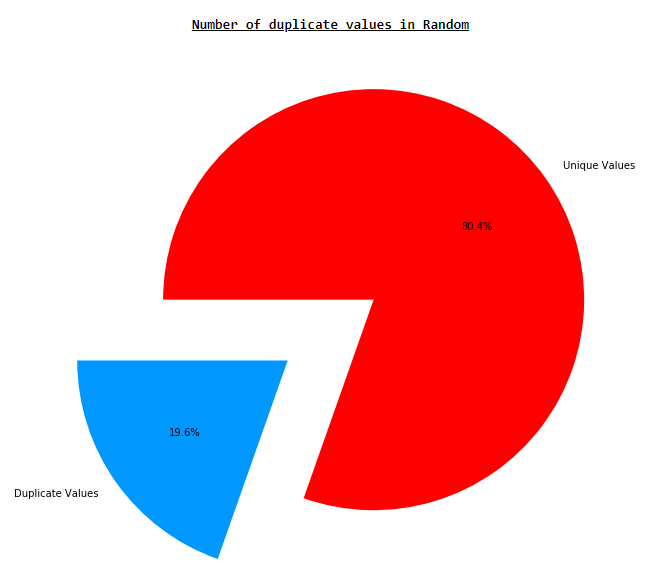
<https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome>

# Appendix

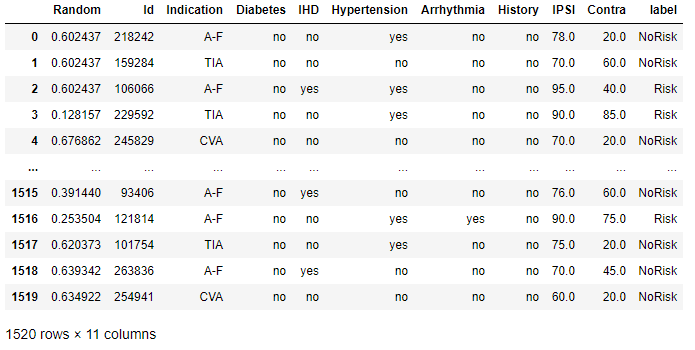
***1.0 - Pie chart representing Valid records with invalid records containing ‘NaNs’***



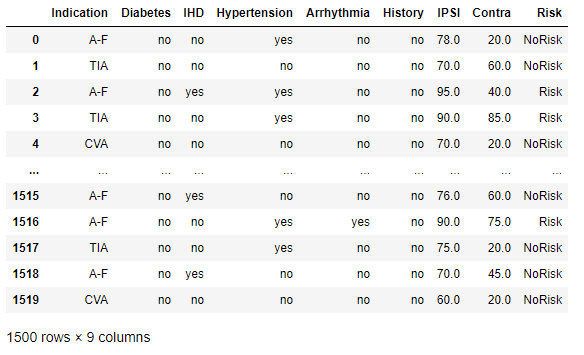
***1.1 - Pie chart representing duplicate values in***



***1.2 - Table with raw data (Raw data – Jupyter Notebook)***

******

***1.3 - Table after Data Cleaning (Table after Data Cleaning – Jupyter Notebook)***



***1.4 - Table after making all variables into binary***

