Data Analytics Coursework 2 - [SET09120](https://moodle.napier.ac.uk/course/view.php?id=35533)

**Davide Pollicino**

40401270@live.napier.ac.uk

Edinburgh Napier University, Edinburgh (UK) - [SET09120](https://moodle.napier.ac.uk/course/view.php?id=35533)

1. Introduction

The aim of this coursework is to demonstrate, understand and use OpenRefine and Weka to conduct an exploration of a given data provided and create a report about my discovery in terms of rules found and data mining. OpenRefine, is a standalone web app created by Google used for data cleaning and data preparation. Weka is an extremely popular suite of Machine Learning algorithm developed at the University of Waikato, New Zeeland.

Keywords: Open Refine, weka, data preparation, data cleaning, python, regression, classification, j48.

2. Data Preparation

2.1. Data Cleaning: using OpenRefine I have been able to clean our dataset. The first step has been upload the excel file and create the project. Analyzing the values contained in each field (or columns), using the text or numerical facet, I have been able to see analyze all the values and their frequencies in each column , using **edit cell** functionality.

The following data outliers / anomalies has been found during the data cleaning process:

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Initial state | After Been updated | Description |
| purpose |  |  |  |
|  | ather (1 record) | other | Misspelled |
|  | Education (2 records) | education | Misspelled |
|  | Radio/Tv (2 records) | ratio/tv | Lowe case |
|  | busines (1 record) | business | Misspelled |
|  | busness ( 1 record) | business | Misspelled |
|  |  |  |  |
| job |  |  |  |
|  | yes | skilled | Suppose that “yes” would be that the person is skilled, “no” otherwise. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Data conversion

Data analytics

3.1 Classification

3.2 Regression

Regression is a ML technique that allows to predict a numerical value for a given set of input values. For example, a cupof coffee in the next 2 years, considering the cost, logistic and selling costs registered in the last 60 years in a specific nation, considering the same coffee brand.

3.3. Association

3.4 Clustering

2.2. Data conversion Describe how the cleaned data is converted to data sets, which can be analyzed by algorithms in Weka.

After have updated, cleaned and converted data in order to be bale to use them in Weka, the dataset needed to be exported in .arff format, a format used to upload our dataset in weka.

To export the dataset as .arff file, I had to: run OpenRefine, load the excel file from the weka explorer and export it as .csv. Once exported as .csv, I had to upload this .csv file in weka using the Weka Explorer, and export it in .arff format.

1. Data Analytics

Speack about differences between classification,r egression, association e clustering, speack about the algorithm used and where they can be applied and solutions found form this approach.

3.1. Classification

3.2. Regression

3.3. Association

3.4. Clustering

Note: • Three of the above 4 sections will be fine. For each section, apply at least one of the algorithms in that category. Settings, datasets used, attributes involved for each run should be described. Patterns/rules need to be retrieved, analysed and interpreted. About 6 patterns/rules for each algorithm used are expected.

4. Summary/Conclusion References are expected