Assignment 2: Analytics Report

Overview

* The purpose of this task is to provide students with practical experience in writing a data analytical report to provide useful insights, pattern and trends in a chosen dataset in the light of a set of tasks required within this document. This dataset will be chosen from the UC Irvine Machine Learning Repository[[1]](#footnote-1). This activity will give students the opportunity to show innovation and creativity in applying the WEKA data mining software, and designing useful visualization and data mining solutions presented as an analytics report.

Timelines and Expectations

Percentage Value of Task: 35%

Due: Week 11

Minimum time expectation: Preparation for this task will take approximately 40 hours

**Project Details**

You will use an analytical tool (i.e. WEKA) to explore, analyse and visualise a dataset of your choosing. An important part of this work is preparing a good quality report, which details your choices, content, and analysis, and that is of an appropriate style.

The dataset should be chosen from the following repository:

UC Irvine Machine Learning Repository

<https://archive.ics.uci.edu/ml/index.php>

The aim is to use the data set allocated to provide interesting insights, trends and patterns amongst the data. Your intended audience is the CEO and middle management of the Company for whom you are employed, and who have tasked you with this analysis.

Tasks

**Task 1 - Data choice.** Choose any dataset from the repository that has at least five attributes, and for which the default task is classification. Transform this dataset into the ARFF format required by WEKA.

**Task 2 - Background information.** Write a description of the dataset and project, and its importance for the organization. Provide an overview of what the dataset is about, including from where and how it has been gathered, and for what purpose. Discuss the main benefits of using data mining to explore datasets such as this. This discussion should be suitable for a general audience. Information must come from at least two appropriate sources be appropriately referenced.

**Task 3 - Data description.** Describe how many instances does the dataset contain, how many attributes there are in the dataset, their names, and include which is the class attribute. Include in your description details of any missing values, and any other relevant characteristics. For at least 5 attributes, describe what is the range of possible values of the attributes, and visualise these in a graphical format.

**Task 4 – Data preprocessing.** Preprocess the dataset attributes using WEKA's filters. Useful techniques will include remove certain attributes, exploring different ways of discretizing continuous attributes and replacing missing values. Discretizing is the conversion of numeric attributes into "nominal" ones by binning numeric values into intervals[[2]](#footnote-2). Missing values in ARFF files are represented with the character "?"[[3]](#footnote-3). If you replaced missing values explain what strategy you used to select a replacement of the missing values. Use and describe at least three different preprocessing techniques.

**Task 5 – Data mining.** Compare and contrast at least three different data mining algorithms on your data, for instance:. *k*-nearest neighbour, Apriori association rules, decision tree induction. For each experiment you ran describe: the data you used for the experiments, that is, did you use the entire dataset of just a subset of it. You must include screenshots and results from the techniques you employ.

**Task 6 – Discussion of findings.** Explain your results and include the usefulness of the approaches for the purpose of the analysis. Include any assumptions that you may have made about the analysis. In this discussion you should explain what each algorithm provides to the overall analysis task. Summarize your main findings.

**Task 7 – Report writing.** Present your work in the form of an analytics report.

**Submission**

The assignment is to be submitted via the Assignment submission box in Moodle. This can be found in the Assessments section of the course Moodle shell. Your report file will be submitted as either a MS word file or a PDF. If you are using MacOS, please submit as a PDF.

Your report will include the following in the order provided below:

A cover page with your name and student ID

Table of Contents

Table of Figures / Tables

Data choice

Background information

Data description

Data preprocessing

Data mining

Discussion of findings

References

Appendices

Your references should use the APAreferencing style; information is available here:

<https://federation.edu.au/library/student-resources/help-with-referencing>

<https://federation.edu.au/library/student-resources/fedcite>

Identify all sources of information used. You are reminded to read the “Plagiarism” section of the course description.

A passing grade will be awarded to assignments adequately addressing all assessment criteria. Higher grades require better quality and more effort. For example, a minimum is set on the wider reading required. A student reading vastly more than this minimum will be better prepared to discuss the issues in depth and consequently their report is likely to be of a higher quality. So before submitting, please read through the assessment criteria very carefully.

Marking Criteria/Rubric

|  |  |  |  |
| --- | --- | --- | --- |
| **Tasks** | **Marks** | **Awarded** | **Comments** |
| **1 - Data choice**   1. Data correctly transformed into the ARFF format. | **10** |  |  |
| **2 - Background information**   1. Description of the dataset including what the dataset is about, including from where and how it has been gathered, and for what purpose 2. Description of project, and its importance for the organization. 3. Main benefits of using data mining to explore datasets such as this. 4. At least two references from peer reviewed sources. | 4 x 5  = **20** |  |  |
| **3 - Data description**   1. General details of dataset 2. Detailed description of five attributes | 5 + 15  = **20** |  |  |
| **4 - Data preprocessing**   1. Implemented and described three different preprocessing techniques. | 3 x 10  = **30** |  |  |
| **5 – Data mining**   * Three different data mining algorithms used * Description of techniques with screenshots and results | 3 x 10  = **30** |  |  |
| **6 – Discussion of findings.**   * Explanation of results * Usefulness of the data mining algorithms * Summary of main findings. | 3 x 10  = **30** |  |  |
| **7 - Presentation of Report.**   * Report is well-written and presented professionally, containing all required sections. | **10** |  |  |
| **Total Marks** | **150** |  |  |
| **Total Marks out of 35** | **35%** |  |  |

Feedback

Feedback and marks will be provided in Moodle. Marks will also be available in FDL Marks.

Plagiarism

Plagiarism is the presentation of the expressed thought or work of another person as though it is one's own without properly acknowledging that person. You must not allow other students to copy your work and must take care to safeguard against this happening. More information about the plagiarism policy and procedure for the university can be found at <http://federation.edu.au/students/learning-and-study/online-help-with/plagiarism>

Please refer to the *Course Description* for information regarding late assignments, extensions, and special consideration. A reminder all academic regulations can be accessed via the university’s website, see: <http://federation.edu.au/staff/governance/legal/feduni-legislation>

1. <https://archive.ics.uci.edu/ml/index.php> [↑](#footnote-ref-1)
2. See: *weka.filter.DiscretizeFilter* [↑](#footnote-ref-2)
3. See: *weka.filter.ReplaceMissingValuesFilter* [↑](#footnote-ref-3)