TU59/TU60 - Data Mining Assignment 1 (CA1).

Due Before: Sunday 10th March @23:59

Group Assignment: This assignment should be completed in groups of 2 people. **Students are responsible for the formation of groups.**

Introduction & Overview

This data mining assignment requires you to analyse a data set, identify data insights, build and evaluate several data mining models and compare the performance of these models with previously published work on the same dataset.

The dataset and problem are from a well-known (and discussed) problem, and the dataset is available for download from the UCI ML Repository called Census Income Dataset: https://archive.ics.uci.edu/ml/datasets/census+income

The prediction task is determining whether a person makes over 50K a year. The purpose of this data mining project is to identify the profile of people who are making more than 50k per year. Also, create different groups and categories and see which ones are less likely to make 50k per year.

The website (given above) contains the data files and attribute descriptions of the dataset.

The dataset contains two files. The training set contains 32,561 instances (2/3), and the testing set contains 16,281 instances (1/3). Some of them have missing values. There is a file called “adult.names” that contains a description of the different variables. It also gives some information on the performance of different models.

The dataset is well-known in the literature, and there is a large number of papers that have used this dataset over the years. These papers are listed at the button on the provided website. These papers can give you more information on the dataset, but you should also be able to find more information online.

You will be required to build several data mining models using Orange data mining (maybe Python for some tasks) and compare the results that you generate with those that the original researchers generated.

Required Tasks

You are required to produce a report detailing your work investigating the data, building classification models, analysing the results, and comparing your results with the original findings.

The first task you should complete is a data investigation exercise, where you will document the characteristics and other information that you can determine about each Feature. Identify any data insights discovered and detail all data preparation tasks and any decisions made. This work should be completed using Orange DM software. Any additional data preparation not possible in Orange DM EM can be completed using Python. But this should be minimal compared to the use of Orange DM EM.

You will need to work through/develop several classification models. To do this you need to use the data mining tool used in class (Orange data mining). In this tool, you can have several different classification techniques and within each of these, you can modify the various parameter settings.

You will need to evaluate the results from each of the models to determine which of the models gives the best results for you. You can then compare your results with the original research and discuss the outcomes.

Since this data represents income in society and one of the goals in society is more equality. In the sense that there are no big gaps based on features like gender, origin, etc. Give some actions based on the analysed data you think can help the government to address this issue.

The original research project used a certain number of data mining algorithms. For your assignment, you should use the algorithms that are available to you in Orange data mining.

Deliverables

You will be required to document your approach to solving and evaluating this classification problem, based on the CRISP-DM process and documentation template guide attached with this CA.

- Your report will probably be a maximum of 10 pages.
- Only include details and images which are important for your findings and narrative. Do not fill your report with reports/charts/etc which do not add to your discussion.

The report should clearly show your work in the following areas (similar to CRISP-DM):

- Definition of the problem
- Data Exploration, Descriptive Analytics, and Identification of data insights are important − Details of any additional data preparation (cleaning, transformations, etc), data enrichment, feature engineering, feature reduction, etc
- Details on which variables are more important to determining the output of the model.
- Details of each data mining algorithm used, the configuration settings used, etc
- Details of the evaluation and performance measures from your data mining models. Examine which one performed best, why this might have been the case and how the results compare across all the models
- Discussion of how your results compared to the results from the original research and other researchers [minimum=3, maximum=5], and any conclusions that you can draw from this comparison
- List of actions a government can take to improve the development across different groups in society.

Submission Details

You should create one document/report containing all the material for each part of the assignment. Convert this document into a PDF. It is this PDF document that should be submitted. All images should be embedded in this document.

The maximum page limit of 10 pages for your report. However, you can create an appendix section to include more material if it is necessary.

You will need to submit your assignment on **BrightSpace**. You cannot submit your assignment via email.

Marking Scheme

The marking scheme for this assignment is:

- 20% Problem Definition, Descriptive Analytics, Data Insights, etc & summary of initial findings/insights
- 10% Details of any additional data preparation, data enrichment, feature engineering, feature reduction, etc
- 15% Details of each data mining algorithm used, the configuration settings used, etc

- 25% Details & Discussion of the evaluation and performance measures from your data mining models.
- 10% Discussion and comparison of your results with original and other researchers.
- 20% Provide a list of actions based on your insights the government can take to improve the development across all different groups in society

The documentation for your assignment must contain the name, student number, class, course (**TU??**) and year information for each student in the group. **Failure to give this information will incur a 10% penalty.**

This assignment should be completed in groups of 2 people. **Students are responsible for the formation of groups.**

Each submission must be original work, as plagiarism will result in **zero marks** (0%). This includes any text used, generated and applied to your work by/from ChatGPT and any other Large Language Models (LLMs).
There will be a 10% penalty deduction will be applied for each day the assignment is late.
There is no penalty for submitting early.

TUDublin Plagiarism Policy :
https://tudublin.libguides.com/c.php?g=674049&p=4794713
https://www.tudublinsu.ie/advice/exams/breachesofregulations/

Assignment Feedback
I will endeavour to mark the assignments and provide feedback via Brightspace VLE. This will consist of a mark for your assignment and a short comment on the assignment. I hope to provide this feedback before you submit Assignment 2 (CA2), or shortly after your submission for this assignment.