COMP 309/AIML421 — *Machine Learning Tools and Techniques*

Assignment 2: Data Exploration, Manipulation and Modelling

12% of Final Mark — Due: 11:59pm Friday 20 August 2021

1 Objectives

The goal of this assignment is to help you understand data manipulation and visualisation tools for machine learning. The purpose is to implement common data handling methods on real-world observations. To validate the effectiveness of the implemented methods, you are also required to perform data analysis tasks to draw useful conclusions. In particular, the following topics should be reviewed:

- Cross Industry Standard Process for Data Mining (CRISP-DM)
- Exploratory Data Analysis (EDA)
- Data Preparation
- Feature Manipulation

These topics are (to be) covered in weeks 4-6, but will also involve content from previous weeks. Research into online resources for AI and machine learning is encouraged. You are required to complete the following questions using data mining/machine learning tools introduced in the lectures - *Python* and/or *Orange* (https://orangedatamining.com/). For each part, make sure you finish reading all the questions before you start working on it, and your report for the whole assignment should *not exceed **10** pages* (note that this is a *maximum*, not a goal/target) with font size no smaller than 10.

2 Question Description

Happiness plays an important role in human emotion and personal growth. The *World Happiness Report* is an annual publication of the United Nations Sustainable Development Solutions Network, which may be a point of interest survey of the state of worldwide bliss. The report and related data are publicly available in their website (https://worldhappiness.report). The report contains articles and rankings of national happiness based on respondent ratings of their own lives. The data is adopted from the Gallup World Poll. The poll asked living evaluation questions, known as the *Cantril Ladder*, which asks respondents to rate their lives in a range of 0 to 10.

What makes the world's happiest countries so happy? In this assignment, let's try to answer this question by finding the most important factors that affect the happiness score in the past several years (Part 1 uses data from 2017 to 2019, and Part 2 uses data from 2021). The columns following the happiness score ("Ladder Score") estimate the extent to which each of the six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in one country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. Your tasks in this assignment is to use data from years to discern the relationship between the happiness score ("Ladder Score") with the six factors and residual (the last feature in the data). The data are available at https://ecs.wgtn.ac.nz/Courses/COMP309_2021T2/Assignments.

2.1 Part 1: Understanding Data From World Happiness Report [40 marks]

The first part of this assignment is to explore the data. The task is to use CRISP-DM, EDA and data manipulation to define the machine learning tasks, understand, and prepare the data.

You should:

1. (15 marks) Perform an **initial EDA** on the given data to gain an understanding of the data. The analyses should explore the data from three different aspects including:

   (a) the correlation between variables, check any important pattern.

   (b) figure out any countries consistently scoring high or low across the variables,

   (c) any happiness discrepancy across regions and years.

   Provide answers to these three questions. Report your EDA methods and results. Identify how EDA helps you answer the questions. Submit your code if you use Python or the workflow (your ows file) if you use Orange.

2. (10 marks) Investigate the business understanding questions (or business objectives) based on your exploration of the data. It is noted that a key question to ask is "which factors contributes most to the happiness score?"

   (a) Identify and describe two other business questions that could be asked of the happiness data as well.

   (b) For these three business understanding questions, decide which machine learning techniques can be applied to meet them and provide your justifications in the report.

3. (5 marks) Use suitable techniques to merge the datasets to form a new dataset. *In the report, describe how this merging is achieved*. Note that appending datasets with identical features but different instances is different from merging distinct data sources that consider different features related to the same task/problem. Determine and describe how to process the given data to form an initial dataset for machine learning tasks. Submit *a copy of the established new dataset* (in CSV format) and give a brief summary of the new dataset. Submit your code or workflow.

4. (10 marks) EDA using clustering is very useful for understanding the important characteristics of the data. Provide a **further EDA** on the establish dataset using Hierarchical clustering to answer the question — "is there a noticeable geographical pattern on the happiness score across years?". Report the output dendrogram and show how it helps you to answer the question.

2.2 Part 2: Data Preparation and Machine Learning on Happiness Report Data [60 marks]

Address the business question of "what makes the world's happiest countries so happy?" using the provided dataset in Part 2. You should:

1. (5 marks) Translate the given business question into a data mining goal and select two machine learning techniques, e.g. classification, regression and so forth, that can help you achieve this goal. Provide justifications of your decision.

2. (10 marks) Determine and describe the data preprocessing steps applied to the provided dataset, e.g. handle missing data, normalise the data if necessary, and/or remove any unnecessary instances, these could be redundant instances, outliers or non-effective instances and so forth. Submit *a copy of the processed dataset* (in CSV format). Submit your code or workflow.

3. (10 marks) Utilise domain knowledge and dimensionality reduction technique(s) to identify which features are irrelevant and/or redundant to the performance of selected machine learning techniques. Report the dimension reduction process and remove redundant/irrelevant data. Submit your code or workflow.

4. (10 marks) After the previous steps only a subset of the original features and/or instances should exist. Use suitable techniques to identify which features are important to the selected machine learning techniques, e.g. which features are near the start of the decision tree, or which features have the highest weights in regression and so forth.

5. (25 marks) Now approach your data mining goal on your preprocessed data using machine learning methods.

   (a) Analyse the data using the two machine learning methods/techniques you choose in Question 1 in this part to approach your data mining goal. Identify how the two techniques are different, e.g. how do the expected results of clustering differ from classification/regression techniques. Submit your code or workflow

   (b) Discuss the results of each technique used on the dataset. Note that the most appropriate form of results may differ between each technique. Exercise your skill and judgement to decide how the results should be communicated.

6. (20 marks) **(For AIML421 students only)** Use three different regression techniques to build prediction models for the happiness score. Discuss the results, and identify which technique is more suitable for this question and provide your justifications.

3 Relevant Data Files and Program Files

A soft copy of this assignment, the relevant data and program files are available from the course homepage: http://ecs.victoria.ac.nz/Courses/COMP309_2021T2/Assignments.

4 Assessment

We will endeavour to mark your work and return it to you as soon as possible, hopefully in 2 weeks. The tutor(s) will run a number of helpdesks to provide assistance to answer any questions regarding what is required.

5 Submission Guidelines

5.1 Submission Requirements

1. Programs for all individual parts. To avoid confusion, all the individual parts should use directories `part1/`, `part2/`, `...` and all programs should be stored in their corresponding directories. Within each directory, please provide a `readme file` that specifies how to run your Python programs on the ECS School machines. If you use Orange, you can put your workflows for each part (Part 1/Part 2) in one ows file. Make sure well organise the workflows and make them tidy. A script file called `sampleoutput.txt` should also be provided to show how your program run properly. If you programs cannot run properly, you should provide a `buglist` file.

2. A document that consists of the reports of all the individual parts. The document should mark each part clearly. The document must be submitted as a PDF.

5.2 Submission Method

You are required to submit both the program *code* and the PDF version of the *report*, which should be submitted through the web submission system from the COMP309 course web site **by the due time**.
There is NO required hard copy of the documents. Problems with personal PCs, internet connections and lost files, which although eliciting sympathies, will not result in extensions for missed deadlines. There will be *no* extension for minor/routine situations — this is what your late days are for.

5.3 Late Penalties

The assignment must be submitted on time unless you have made a prior arrangement with the course coordinator or have a valid medical excuse (for minor illnesses it is sufficient to discuss this with the course co-ordinator). The penalty for assignments that are handed in late without prior arrangement is one grade reduction per day. Assignments that are more than one week late will not be marked.

5.4 Plagiarism

Plagiarism in programming (copying someone else's code) is just as serious as written plagiarism, and is treated accordingly. Make sure you explicitly write down where you got code from (and how much of it) if you use any other resources asides from the course material. Using excessive amounts of others' code may result in the loss of marks, but plagiarism could result in zero marks!