COMP 7006 Data Science

Computer Based Assignment - Data Analysis Task 2023

Due: Week 10, Tuesday November 21st Midnight

Submit via vUWS using Turnitin link

Students are expected to use your own data set from an application of interest. We recommend you use a real data set which will give you an opportunity to start on a publication. **Do not use time series data**. You should try to get a structured data set with **over 6 variables** and with **over 100 observations**. Students who are unable to find a good data set will be provided with a data set. Finding a suitable data set is also a part of the exercise. Students should have a **suitable data set ready for analysis** and **research question** by **week 7 (31st October**).

In this assignment you will look at various aspects of the data and write a report. You are expected to write the report as a publication to submit to a hypothetical conference. Use the format as the given sample paper (provided in vUWS) with additional information given as appendix. Submit the document online via the Turnitin link in vUWS. (Include a cover sheet provided in vUWS)

The report should explain the steps taken and conclusions drawn from the data and should include research questions stated in statistical language, analysis justification and assumptions of the analysis. Explain your data analysis process logically and connectively.

Students are expected to carry out the following:

1. <u>Supervised Learning (any 2)</u>

   Either Two methods for Regression analysis OR Two methods for Classification analysis OR One method for Regression analysis AND One method for Classification analysis.

   PLUS: Model validation, evaluation and testing must be carried out.

2. <u>Unsupervised Learning (any 1)</u>

   Perform Clustering OR Principal Component Analysis

The Report Headings:

- Abstract
- Introduction
- Data Description, Explore and Pre-processing
- Aims and Objectives
- Method 1: (Should include Model building, validation, evaluation and testing)
- Method 2: (Should include Model building, validation, evaluation and testing)

- Model Comparison
- Conclusions and Recommendations
- References

Use the paper format provided and **no more than 8 pages** plus Appendix. R Code and other explanatory material should be included as an appendix.

Finally – Project Presentation

4 Power Point Presentation slides and 3 minutes Zoom Video recording done over the Power Point slides to be submitted via Turnitin before 25th November (Saturday, midnight).

PLEASE NOTE: The marking scheme for Project is set according to using R and Lectures and Lab work done in this subject. So, using Python or any other language will be a disadvantage. To maintain the integrity of the project submission please demonstrate understanding of what you learned in this unit and any other content provided will not score marks.