ASSESSMENT 2

Deadline:	Hand in by midnight 7 May 2023
Evaluation:	15% of your final course grade.
Late Submission*:*	Refer to the course guide.
Work	This assignment is to be done **individually**.
Purpose:	Implement the entire data science/analytics workflow. Use regression techniques to solve real-world problems. Gain skills in extracting data from the web using APIs and web scraping. Build on the data wrangling, data visualization and introductory data analysis skills gained up to this point as well as problem formulation and presentation of findings. Gain skills in kNN regression modelling and supervised and unsupervised learning. Learning outcomes 1 - 5 from the course outline.

These assignments will take longer than you think, so…

Do not leave starting the assignment until the last minute. **Start now**.

IMPORTANT

You have not been equipped to complete the entire assignment yet. But you can start on several tasks now and complete the rest as we cover the relevant material in class

- TASKS 1, 2 and a part of TASK 3 can be started from Week 4.
- TASK 3.2 can be started from Week 6, but you will acquire suitable evaluation skills from Week 7.
- TASKS 3.3 and 4 can be fully completed in Week 8.

You are of course welcome to work ahead of the class if you wish and watch the prerecorded materials if you wish to move at a faster pace and complete the assignments earlier. All the relevant course content is up on Stream now.

PROJECT DESCRIPTION OVERVIEW: TENNIS DATA ANALYTICS (PART 2) WITH REGRESSION, KNN AND CLUSTERING MODELS

In this project, you will build upon the work you completed in Project 1 by performing analytics involving prediction and clustering on the tennis dataset. You will also have the opportunity to incorporate additional data of your choice, generate new features, and perform regression and kNN modelling, as well as clustering.

The project aims to enable you to explore and understand the tennis data deeper, and use data science techniques to uncover insights and develop models to predict outcomes and cluster similar data points. You will work with real-world data, and use data science tools and techniques to develop solutions to various research questions.

The project will be broken down into several phases, including:

1. **Data acquisition, preparation and cleaning**: In this phase, you will acquire, clean and preprocess the tennis dataset used in Project 1. You will also identify and acquire additional data of your choice to be incorporated into the analysis.

2. **Feature engineering**: In this phase, you will generate new features based on the data, using domain knowledge and data science techniques.

3. **Model development and evaluation**: In this phase, you will develop regression and kNN regressor models to predict outcomes based on the data. Additionally, you will also start to experiment with classification. You will also perform clustering analysis to group similar data points together. You will evaluate the performance of the developed models using appropriate metrics and techniques.

4. **Presentation and report**: In this final phase, you will present your findings and report on the results of the project, including any insights or conclusions drawn from the analysis.

TASK 1: DATA ACQUISITION

Here are some steps you should follow in this phase:

- Identify and acquire additional data: Search for additional data that can be incorporated into the analysis. This could include player-specific information, such as age, height, weight, handedness, and past performance. You can obtain this data from online sources such as tennis websites which you may be permitted to scrape, downloadable datasets or APIs. Some examples:

  - http://www.tennisabstract.com/
  - https://www.atptour.com/en/stats
  - https://datahub.io/sports-data/atp-world-tour-tennis-data#data
  - https://www.ultimatetennisstatistics.com/

- Clean and integrate additional data: Once you have obtained the additional data, clean and preprocess it, and integrate it into the original dataset. This can be done by merging the datasets based on common fields such as player names or match dates. Validate the data. Check that the combined dataset is free of errors and inconsistencies and that it has the necessary variables for your analysis.

TASK 2: FEATURE ENGINEERING

Based on all the data you have managed to gather, here you will start considering what research questions you can formulate and answer with the data you have, using regression and classification techniques.

Here are the specific steps you should follow in this phase:

1. Identify research questions: Start by identifying research questions that you would like to answer using the dataset. Consider what values you can predict in the dataset, and what kinds of insights you can gain from the data. Formulate 4-6 research questions that can guide your analysis.

2. Feature engineering: Once you have identified research questions, think about what new features you can generate from the data to help answer them. This could include transforming existing variables, creating new variables from existing ones, or perhaps dummy variables, and the task is also likely to include creating new dependent variables you'd like to predict.

TASK 3: MODELLING

Based on all the data you have managed to gather, here you will start considering what research questions you can formulate and answer with the data you have, using regression and classification techniques.

1. Develop regression models: Using the enriched dataset, develop regression models to predict values that either already exist in the original dataset, or ones you have engineered.

    a. Try simple linear regression polynomial and multiple regression as well as kNN regression with different parameters. Experiment with different independent and dependent variables, and use the newly generated features in your models. Make plenty of models and compare them. Reason about which ones are better than others.

2. Experiment with classification: After developing regression models, experiment with classification techniques to predict categorical variables in the dataset. This could include kNN, or you can use simple regression techniques for predicting two labels like 0 or 1.

    a. Use appropriate evaluation metrics covered in Week 7 material.

3. Perform clustering analysis: Finally, perform clustering analysis to group similar data points together. Explore if there is some structure to the data and if some interesting patterns appear. Be curious and creative.

    a. Try out different configurations of k-Means (or other clustering algorithms if you wish) and use the evaluation measures covered in Week 8. Try also to visualise the results. Contribute your interpretation of the results.

TASK 4: REPORT-'IZE' YOUR WORK

Go back through what you have done I the previous sections and convert a part of your analysis into something that looks like a report that you could hand to a client (a technically savvy client as you still need to include your scripting for marking). Follow the Jupyter Notebook template as an example, so include a brief introduction, that describes the selected modelling problem you formulated and a brief description of the datasets that you use with chosen research questions and an executive summary. Decide on the structure of the body of your report, but also make sure you have a conclusion. Clear out any unnecessary code and outputs that clutter your work. Run your text through a spell-checker extension.

Component	**Marks**	**Requirements and expectations**
Task 1	15	Acquisition and integration of additional dataset(s), quality of EDA
Task 2	15	Formulation of research questions for modelling, creativity in generating new features to support the modelling process
Task 3	60	Regression (25 marks) Classification (25 marks) Clustering (10 marks) Multiple models should be developed for each research question and contrasted, models need to be robustly contrasted and evaluated and the best ones identified, quality of experimentation, analysis, interpretation of results and conclusions. Appropriate use of visualisations in evaluations where possible.
Task 4	10	Report structure, presentation of findings and interpretation, emphasis on what is central to the message being presented, tidiness of code and outputs

HAND-IN:

Zip-up all you notebook, html notebook, python files and dataset(s) into a single file. Submit this file via stream. Make sure that your jupyter notebook has been run with all outputs visible. Download an HTML version of your notebook (with outputs showing) and include this in your zip file.

****

*** Plagiarism ***

****

It is mandatory that any assessment items that you submit during your University study are your own work. Massey University takes a firm stance on academic misconduct, such as plagiarism and any form of cheating.

Plagiarism is the copying or paraphrasing of another person’s work, whether published or unpublished, without clearly acknowledging it. It includes copying the work of other students and reusing work previously submitted by yourself for another course. **It also includes the copying of code from unacknowledged sources**.

Academic integrity breaches impact on students as it disadvantages honest students and undermines the credibility of your qualification. Plagiarism, and cheating in tests and exams will be penalised; it is likely to lead to loss of marks for that item of assessment and may lead to an automatic failing grade for the course and/or exclusion from reenrolment at the University.

Please see the Academic Integrity Guide for Students on the University website for more information. The Guide steps you through the University Academic Integrity Policy and Procedures. For example, you will find definitions of academic integrity misconduct, such as plagiarism; how misconduct is determined and managed; and where to find resources and assistance to help develop the skills of academic writing, exam preparation and time management. These skills will help you approach university study with academic integrity.