Data Mining COMP5009 Assignment

Due Date

The assignment is due Friday 26th of September at 5pm.

Late marks will be applied at a rate of 10 marks per day or part thereof, unless you have an extension. Extensions before the deadline will be more generously received than those after the deadline. Please submit an extension application as soon as you know you need one. All assessment extensions are to be submitted via student oasis. If you have a CAP that allows for an extension to assignments, you still need to indicate that you need the extra time (but no additional justification is needed).

Weight

The assignment has a total mark of 100, which will count 25% to the total grade for this course.

Changes

If changes need to be made to any part of this assignment, they will be noted on blackboard and a revised version of this document may be created.

Conditions

This is an **individual assignment**, and all work submitted must be your own. While you're encouraged to discuss general data mining concepts and techniques with your peers to support your learning, you must ensure that your specific approach, code, and written report are developed independently. There are many valid ways to approach the task, and as a result, submissions should reflect your unique understanding and decision-making. Identical or highly similar submissions are extremely unlikely to occur by chance and will be investigated under the university’s academic misconduct policy. To protect the integrity of your work, please do not share your files with others, and ensure your assignment is stored securely. If you're ever unsure about what level of collaboration is appropriate, ask your lecturer for guidance.

Copying material (from other students, websites or other sources) and presenting it as your own work is plagiarism. Even with your own (possibly extensive) modifications, it is still plagiarism. Exchanging assignment solutions, or parts thereof, with other students is collusion. Engaging in such activities may lead to a grade of ANN (Result Annulled Due to Academic Misconduct) being awarded for the unit, or other penalties. Serious or repeated offences may result in termination or expulsion. You are expected to always understand this, across all your university studies, with or without warnings like this.

Use of AI

You are permitted to use large language models such as **ChatGPT** (or similar AI tools) to support your learning and problem-solving in this unit. These tools can be valuable for helping you develop your skills but must be used responsibly and with transparency.

Permitted uses include:

- Debugging your code or helping to understand error messages
- Suggesting possible approaches or solutions to coding problems you describe
- Helping you outline or structure your assignment report (e.g., organizing headings, suggesting sections)

Prohibited uses include:

- Generating written content to include in your assignment report
- Rewriting, rephrasing, or editing your own written work using AI tools

If you choose to use an AI language model at any stage of your assignment process, you must include a brief statement at the beginning of your report that clearly explains **where and how** the tool was used. This ensures transparency and academic integrity.

Important:
Using AI tools to generate report content, or failing to declare your use of such tools, will be treated as a breach of academic integrity and may result in penalties, including the annulment of your results. When in doubt, always ask your lecturer for clarification.

Overview

In this assignment, you will take on the role of a data scientist solving practical, real-world data mining problems. The task is designed to reflect the kinds of challenges faced in industry, where success depends not only on technical accuracy but also on your ability to make informed decisions, document your process, and clearly communicate your findings.

To complete this assignment, you will apply the theory introduced in workshops, build on the skills developed through practical exercises, and conduct independent research to understand the problem context and data. This will involve:

- Accessing and preparing real-world data
- Cleaning and transforming data to ensure quality and usability
- Selecting relevant features and designing effective inputs
- Choosing and training appropriate models for the task
- Evaluating model performance and refining your approach

The final deliverables will include both a set of predictions and a professionally written report. Your report should clearly document your end-to-end workflow, including justifications for your decisions at each stage.

This assignment is not just about getting the "right" answer - it's about demonstrating your ability to think like a data professional: to analyze a problem, design a solution, and communicate your results in a way that others can understand and build upon.

Problem Description: Vegetation Condition Monitoring on Wadjuk Noongar Country

Western Australia's southwest is home to one of the world's biodiversity hotspots and is the traditional land of the **Wadjuk Noongar people**. For tens of thousands of years, Noongar communities have practiced **sustainable land management** grounded in deep ecological knowledge, seasonal cycles, and cultural connection to Country.

In recent years, environmental scientists and traditional custodians have been working together to monitor **vegetation health and land condition** using a combination of modern sensor data and traditional knowledge. Understanding the condition of vegetation is essential for maintaining biodiversity, managing fire risk, guiding cultural burns, and protecting culturally significant areas.

The Task

You have been provided with a simulated dataset representing field sensor and satellite measurements across a range of land parcels on **Wadjuk Noongar country**. Each record contains various indicators of environmental condition, such as vegetation health, soil properties, moisture availability, proximity to water, and signs of human or invasive species disturbance.

Your goal is to build a **classification model** that can accurately predict the **vegetation condition** of each land parcel into one of the following categories:

- **Healthy** — Native vegetation appears robust and undisturbed.
- **Degraded** — Vegetation shows signs of stress, disturbance, or invasive species.
- **At Risk** — Vulnerable areas, often near sacred sites, under threat from human impact or ecosystem imbalance.

These categories are **imbalanced**, reflecting real-world monitoring challenges. For instance, truly “At Risk” areas are rare but critical to identify, while “Degraded” zones are unfortunately more common due to land use pressures.

The Data

You are provided an sqlite3 database file (`Assignment2025S2.sqlite`) which contains a total of 5500 samples across two tables. Table “train” contains 5000 samples have already been labels. Each row represents a land parcel sampled on Wadjuk Noongar country, with various ecological, geographic, and disturbance-related attributes .

This dataset is **simulated**, but it is inspired by real-world scenarios where **cultural sensitivity**, **data sovereignty**, and **collaborative stewardship** are paramount. You are encouraged to think critically about how data mining methods should be applied respectfully when working with or near Indigenous knowledge and lands.

This dataset has been **intentionally designed to reflect the kinds of data problems** that arise in real-world environmental monitoring, including but not limited to:

- Missing values in some columns,
- Outliers and different data distributions,
- Class imbalance,
- Noisy or redundant features,
- Duplicated instances.

Tasks

Data Preparation

In this first task, you will examine all data attributes and identify issues present in the data. For each of the issues you have identified, choose and perform necessary actions to address it. Note that you will need to apply these actions to both the training and test data at the same time. At the end of this phase, you will have two data sets: one for training and one for the final testing task. Your marks for this task will depend on how well you identify the issues and address them. Below is a list of data preparation issues that you need to address

- Identify and remove irrelevant attributes.
- Detect and handle missing entries.
- Detect and handle duplicates (both instances and attributes).
- Select suitable data types for attributes.
- Perform data transformation (such as scaling/standardization) if needed.
- Perform other data preparation operations (This is optional but may help improve your accuracy).

For each of the above issues your report should:

- Describe the relevant issue in your own words and explain why it is important to address it. Your explanation must consider the classification task that you will undertake subsequently.
- Demonstrate clearly that such an issue exists in the data with suitable illustration/evidence.
- Clearly state and explain your choice of action to address such an issue.
- Demonstrate convincingly that your action has addressed the issue satisfactorily.

Where applicable, you should provide references to support your arguments.

Data Classification

For this task, you will demonstrate convincingly how you select, train, and fine tune your predictive models to predict the missing labels. You must use **at least** the three (3) classifiers that have been discussed in the workshops, namely k-NN, Naive Bayes, and Decision Trees. You can also select additional classifiers (both base classifiers and meta-classifiers). Attempt and report the following:

- **Class imbalance**: the original labelled data is not equally distributed between the three classes. You need to demonstrate that such an issue exists within the data, explain the importance of this issue, and describe how you address this problem.
- **Model training and tuning**: Every classifier typically has hyperparameters to tune in order. For each classifier, you need to select (at least one) and explain the tuning hyperparameters of your choice. You must select and describe a suitable cross-validation/validation scheme that can measure the performance of your model on labelled data well and can address the class imbalance issue. Then you will need to conduct the actual tuning of your model and report the tuning results in detail. You are expected to look at several classification performance metrics and make comments on the classification performance of each model. Finally, you will need to clearly indicate and justify the selected values of the tuning hyperparameters of each model.

- **Model comparison**: Once you have finished tuning all models, you will need to compare them and explain how you select the best two models for producing the prediction on **all** of the test samples.
- **Prediction**:
    - Use the best two (2) models identified in the previous step to predict missing class labels of the test samples. Clearly explain in detail how you arrive at the prediction.
    - Produce a classification table as a .csv file that contains your prediction in the format: the first column is the index corresponding to the 'test' table, the second and third columns are the predicted class labels. This file must be submitted electronically with the electronic copy of the report via Blackboard. An example of such a file is given below:

      ```
      index,Predict1,Predict2
      5000, Healthy, Healthy
      5001, Healthy, Degraded
      5002, Degraded, AtRisk
      ...
      5499, AtRisk, Degraded
      ```

    - You must also indicate clearly in the report your estimated prediction accuracy for each selected model and explain how you arrive at these estimates.

Reporting

You are required to submit a **written report** that documents your approach to the environmental monitoring classification task. Your report will be assessed based on how clearly it communicates your understanding of the problem, your analytical process, and the decisions you made during the task.

Your report should:

- Demonstrate your understanding of the task and its broader context.
- Showcase your ability to conduct data-driven analysis, including research, data preparation, and model development.
- Remain within the **page limit of 20 pages**. Content beyond this limit will **not** be considered for marking.

Please organize your report with the following sections:

1. Cover Page - Include your name and student ID
2. Summary - Include an overview of the work including the performance of your best models.
3. Methodology - Organize this section into two parts:
    a. Data Preparation - Justify and describe the actions you performed.
    b. Classification – Describe your approach to model training and selection.
4. Conclusion - Summarize your results and observations, reflect on limitations, and identify future work.
5. References - List any materials you cited (not including the lecture materials).

Visual illustration to support your analysis which may include tables, figures, plots, diagrams, and screenshots. **Do not include code snippets** - they are not required and will not be assessed.

Submission

The assignment is submitted in two parts:

- The main **report** in PDF format must be submitted through Turnitin. A submission link will be provided on Blackboard. You should name the report file using the following naming convention report_surname_studentID.pdf, for example report_hancock_12345678.pdf
- Your **code** and **predictions** must be submitted through a separate assignment submission link. You must put the following files in a single zip file using your surname and student ID as the name of the zip file (for example hancock_12345678.zip):
    - Answers_<sid>.csv
    - Notebook_<sid>.ipynb

Source Code

In addition to the main report which details your analysis of the assignment tasks, you will need to submit a Jupyter Notebook that will reproduce your prediction. The notebook will:

- Run without error on **Google Colaboratory** when you select “restart and run all”. You should assume that the data file for this assignment is in your root directory.
- Contain a combination of Text and Code cells that describe the tasks that you are performing. See the practicals for examples of how to do this.
- Produce the Answers_<sid>.csv file without need for further modification.
- Be named notebook_SID.ipynb (eg notebook_12345678.ipynb)

Once you have completed your notebook please go to “Edit -> Clear all outputs”, and then “File -> Download -> Download .ipynb”.

Any notebooks that fail to execute completely, or do not reproduce the submitted prediction file, will lose marks.

Please note that the Jupyter notebook will be run to ensure that it will produce your output file; however, it will not be read. **The markers will not be reading your notebook when marking your assignment, so please ensure all the important information and analysis that you do is in your report.**

Mark Allocation

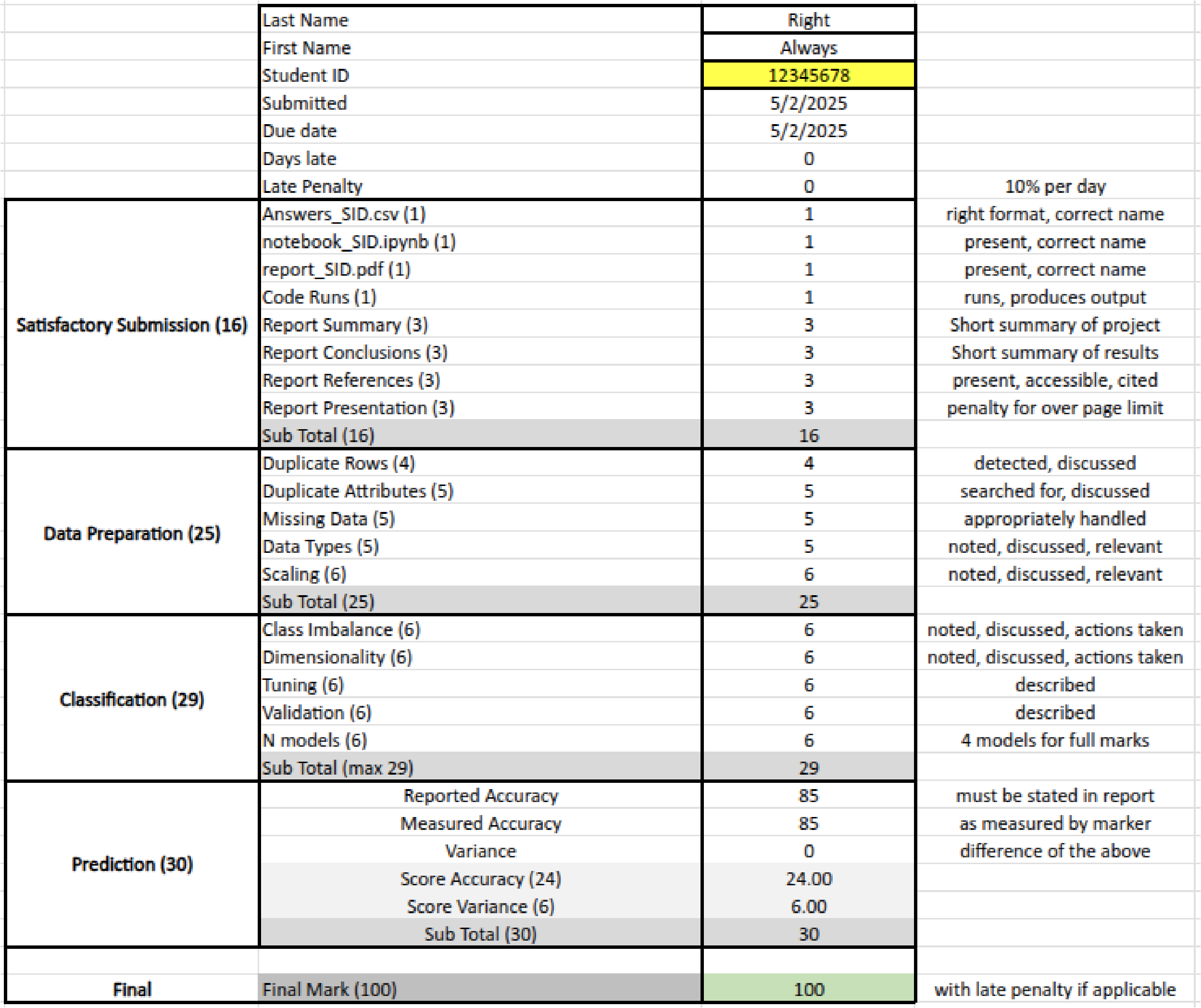
The total mark of this assignment is 100, and it is distributed as follows

- **Satisfactory submission**: 16 marks. This is based on
  - All required files are submitted with the correct name and format.
  - Your source code: your code must run without errors and produce the same prediction that you submitted.
  - Summary, conclusion, and references in the report.
  - The overall presentation of the report.
- **Data Preparation**: 25 marks. This is based on how well you identify and address data preparation issues in the report. This includes irrelevant attributes, duplicates, missing entries, data types, and scaling/standardization.
- **Data Classification**: 29 Marks. This is based on how well you present the class imbalance, training, tuning, validation, comparison of different models, and how you arrive at the prediction as described in the report.
- **Prediction**: 30 Marks. This is based on two factors: actual prediction accuracy (maximum 24 marks) and your estimate of the prediction accuracy (maximum 6 marks). For the actual and estimated prediction accuracy, the allocation is as follows:

Accuracy	Marks	Estimate of Accuracy	Marks
<55%	0	Within ± 2%	6
55%	1	Within ± 3%	5
56%	2	Within ± 4%	4
57%	3	Within ± 5%	3
58%	4	Within ± 6%	2
59%	5	Within ± 7%	1
60%	6	Outside± 7%	0
61%	7		
62%	8		
63%	9		
64%	10		
65%	11		
66%	12		
67%	13		
68%	14		
69%	15		
70%	18		
71%-74%	20		
≥75%	24		

Marking Rubric

Below is an example of the marking rubric used to grade your assignment. You should refer to this when completing your report.



	Last Name	Right	
	First Name	Always	
	Student ID	12345678	
	Submitted	5/2/2025	
	Due date	5/2/2025	
	Days late	0	
	Late Penalty	0	10% per day
Satisfactory Submission (16)	Answers_SID.csv (1)	1	right format, correct name
	notebook_SID.ipynb (1)	1	present, correct name
	report_SID.pdf (1)	1	present, correct name
	Code Runs (1)	1	runs, produces output
	Report Summary (3)	3	Short summary of project
	Report Conclusions (3)	3	Short summary of results
	Report References (3)	3	present, accessible, cited
	Report Presentation (3)	3	penalty for over page limit
	Sub Total (16)	16	
Data Preparation (25)	Duplicate Rows (4)	4	detected, discussed
	Duplicate Attributes (5)	5	searched for, discussed
	Missing Data (5)	5	appropriately handled
	Data Types (5)	5	noted, discussed, relevant
	Scaling (6)	6	noted, discussed, relevant
	Sub Total (25)	25	
Classification (29)	Class Imbalance (6)	6	noted, discussed, actions taken
	Dimensionality (6)	6	noted, discussed, actions taken
	Tuning (6)	6	described
	Validation (6)	6	described
	N models (6)	6	4 models for full marks
	Sub Total (max 29)	29	
Prediction (30)	Reported Accuracy	85	must be stated in report
	Measured Accuracy	85	as measured by marker
	Variance	0	difference of the above
	Score Accuracy (24)	24.00	
	Score Variance (6)	6.00	
	Sub Total (30)	30	
Final	Final Mark (100)	100	with late penalty if applicable