

# Cover sheet for submission of work for assessment

## UNIT DETAILS

Unit name	Data Science Principles	Class day/time	COS10022.1	Office use only
Unit code	COS10022	Assignment no.	01	Due date
Name of lecturer/teacher	Mr. Minh Anh			
Tutor/marker's name				Faculty or school date stamp

## STUDENT(S)

	Family Name(s)	Given Name(s)	Student ID Number(s)
(1)	Nguyen	Minh Duy	104974743
(2)			
(3)			
(4)			
(5)			
(6)			

## DECLARATION AND STATEMENT OF AUTHORSHIP


- I/we have not impersonated, or allowed myself/ourselves to be impersonated by any person for the purposes of this assessment.
- This assessment is my/our original work and no part of it has been copied from any other source except where due acknowledgement is made.
- No part of this assessment has been written for me/us by any other person except where such collaboration has been authorised by the lecturer/teacher concerned.
- I/we have not previously submitted this work for this or any other course/unit.
- I/we give permission for my/our assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.

I/we understand that:

- Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarised material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited.

### Student signature/s

I/we declare that I/we have read and understood the declaration and statement of authorship.

(1)		(4)	
(2)	Nguyen Minh Duy	(5)	
(3)		(6)	



**Swinburne University of Technology Hawthorn Campus**  
**Dept. of Computing Technologies**

**COS10022 Data Science Principles**  
**Assignment 1 - Semester 1, 2024**

**Assessment Title:** Predictive Model Creation and Evaluation

**Assessment Weighting:** 20%

**Due Date:** Sunday, 24<sup>th</sup> March 2024 at 11.59 pm (AEDT)

**Assessable Item:**

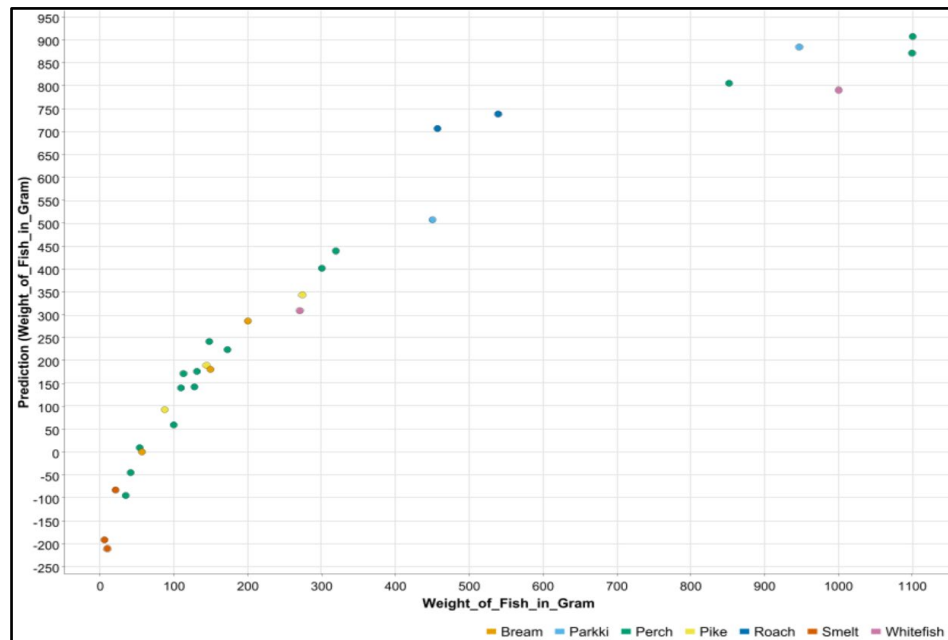
- One (1) piece of a written report no more than 10-page long with the signed Assignment Cover Sheet.
- The submitted report must be checked by Turnitin, and the similarity from **not the template part** should be less than 12%.

The submitted report should answer all questions listed in the assignment task section in sequence. You must include a digitally signed Assignment Cover Sheet with your submission.

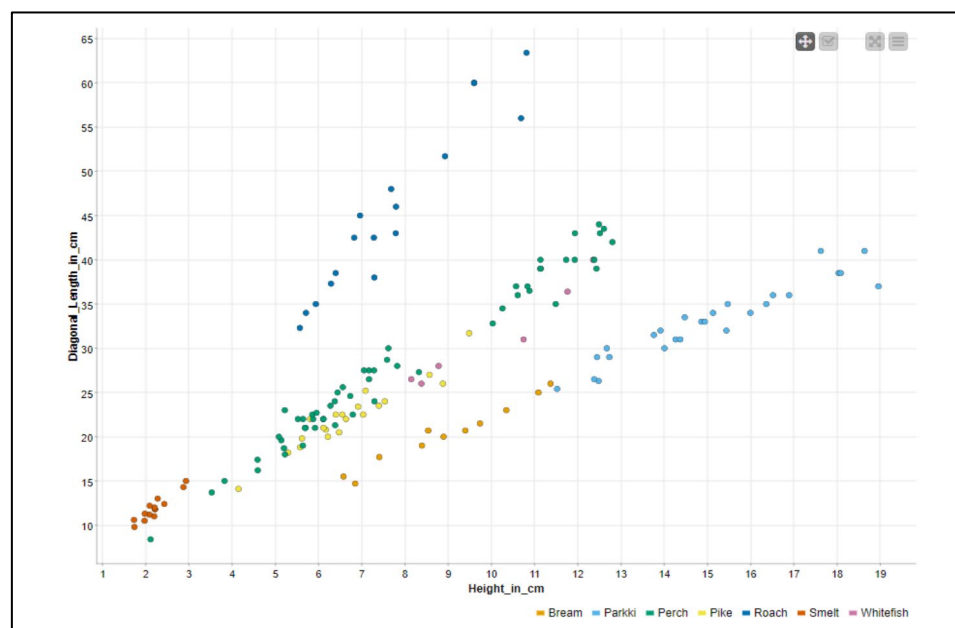
---

1. Follow the instructions above to split the source data into training and test sets. Answer the following questions after splitting the data. **[10 marks in total]**
  - 1) Submit the workflow of Assignment 1 via Assignment 1.1. **[2.5 marks]**  
*Ans:* [Check Assignment 1.1 for the KNIME workflow file.](#)
  - 2) How many tuples are included in the training set? **[2.5 marks]**  
*Ans:* [120 tuples \( 80% of 150 \)](#)
  - 3) How many species are included in the test set? **[2.5 marks]**  
*Ans:* [7](#)
  - 4) Do species "Whitefish" and "Smelt" have the same number of tuples included in the test set? **[2.5 marks]**  
*Ans:* [No, Whitefish has 2 tuples and Smelt has 3 tuples](#)
2. Build a Linear Regression Model using **all** available attributes to predict the value of the "Weight\_of\_Fish\_in\_Gram". Answer the following questions after completing the model training and test. **[40 marks in total]**
  - 1) What is the  $R^2$  value of your test result? **[5 marks]**  
*Ans:* [0.8732](#)

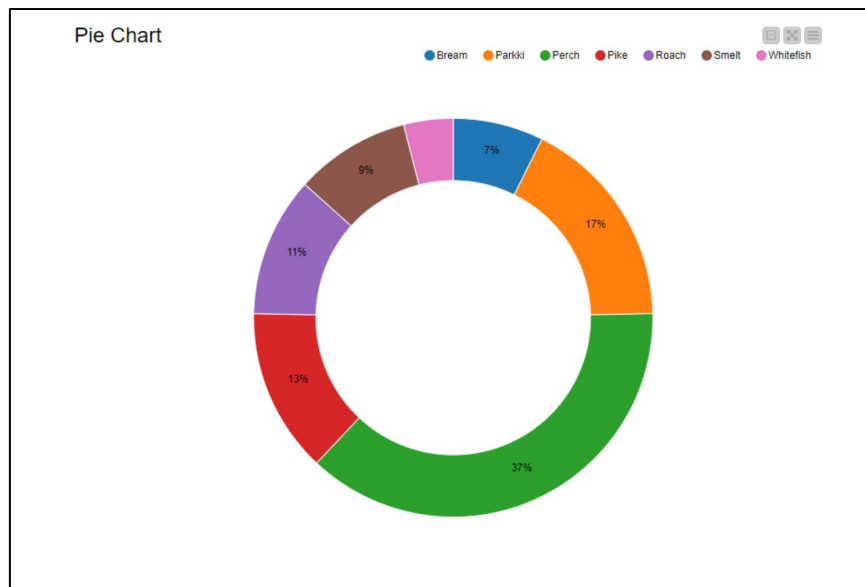
- 2) Give the screenshot of the scatter plot result of your test output using “Weight\_of\_Fish\_in\_Gram” on the x-axis and the prediction value on the y-axis. Assign different colours to the data points based on the “species.” [15 marks]



- 3) Which species has the heaviest predicted weight in your test result? [5 marks]  
 Ans: Perch
- 4) How many prediction results are infeasible in your test result? [5 marks]  
 Ans: 5 (negative results)
- 5) Looking at your source data before splitting them, which species can be easily separated from others if looking at the “Height\_in\_cm” and “Diagonal\_Length\_in\_cm” attributes? Post your visualisation result on data observation in the report. [5 marks]  
 Ans: Smelt, Roach, Parkki



- 6) Draw a doughnut chart of the original input data with 0.55 as the doughnut hole ratio before splitting it into training and test sets. Use different colours for each species and show the percentage of data in the pie chart. **[5 marks]**



3. Build a Logistic Regression Model with **all** attributes and use "Smelt" as the reference category. The maximal number of epochs and epsilon should be set to **10,000** and **0.00001**, respectively. Use "LineSearch" as the learning rate strategy. Use **9214** as the seed in the logistic regression node. Answer the following questions after completing the model training and test. **[40 marks in total]**
- Which species have/has no "True Positive (TP)" case in the prediction result? **[5 marks]**  
 Ans: Whitefish
  - For the species with no TP case, which species will be misplaced? **[5 marks]**  
 Ans: Perch
  - What is the overall accuracy of the prediction result? **[5 marks]**  
 Ans: 0.833
  - List all species names with 100% correctly classified test results. **[15 marks]**  
 Ans: Roach, Smelt, Bream, Parkki
  - Which species has a 33.33% chance of being misplaced into another species in the test result? **[5 marks]**  
 Ans: Pike
  - In the test result, what percentage of the species "Perch" is misplaced into others? **[5 marks]**  
 Ans: 13.333%
4. Build a new linear regression model different from the one built when answering question 2. This time, let's focus on the species "Perch" only. You are limited to using three attributes in the input to predict the "Weight\_of\_Fish\_in\_Gram." Use a "Scatter Matrix (local)" node to observe your data and decide the suitable attributes to be included. The linear regression model should be the same as the one used in question 2 except for the input attributes. Build, train, and test the model and then answer the questions below. **[10 marks in total]**
- Give the reasons for each eliminated attribute and why they are not selected as the input. **[5 marks]**  
 Ans:
    - "Diagonal\_length\_in\_cm": While this variable shows a positive linear relationship with "Weight\_of\_Fish\_in\_Gram," this correlation is primarily noticeable at the extremes of the scatter plot, indicating potential outliers. Including this variable could introduce instability into the model and lead to unreliable predictions.
    - "Vertical\_length\_in\_cm": The correlation between this variable and "Weight\_of\_Fish\_in\_Gram" is weaker compared to other attributes, and its inclusion is unlikely to significantly improve the model's predictive capability.
    - "Diagonal\_length\_in\_cm": The length-related variables show high multicollinearity, with correlation coefficients close to 1. Including multiple similar variables can cause multicollinearity issues, leading to unstable coefficient estimates and inaccurate predictions. Therefore, it is recommended to

exclude both "Vertical\_length\_in\_cm" and "Diagonal\_length\_in\_cm" to improve model stability.

List the  $R^2$  of your test result and compare it with the one in question 2. Reveal both  $R^2$  values obtained in question 2 and in question 4. If you can improve the model, you get the mark. **[5 marks]**

Ans:  $R^2$  of question 4: 0.9323 higher than  $R^2$  of question 2: 0.8732