

Cover sheet for submission of work for assessment



UNIT DETAILS

Unit name	Data Science Principles	Class day/time	1PM - Friday	Office use only	
Unit code	COS10022	Assignment no.	1	Due date	01/Oct/2023
Name of lecturer/teacher	Huy Truong				
Tutor/marker's name	Huy Truong			Faculty or school date stamp	

STUDENT(S)

Family Name(s)	Given Name(s)	Student ID Number(s)
Nguyen Dinh	Nhat Minh	103802490

DECLARATION AND STATEMENT OF AUTHORSHIP

1. I/we have not impersonated, or allowed myself/ourselves to be impersonated by any person for the purposes of this assessment.
2. 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.
3. 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.
4. I/we have not previously submitted this work for this or any other course/unit.
5. 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:

6. 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.

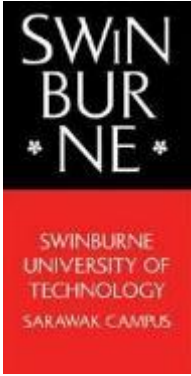
(1)

A handwritten signature in blue ink, appearing to read "Minh", written over a light blue grid background.

Further information relating to the penalties for plagiarism, which range from a formal caution to expulsion from the University is contained on the Current Students website at www.swin.edu.au/student/

Copies of this form can be downloaded from the Student Forms web page at www.swinburne.edu.au/studentforms/

PAGE 1 OF 1



Swinburne University of Technology Hawthorn Campus Dept. of Computer Science and Software Engineering

COS10022 Data Science Principles Assignment 1 - Semester 3, 2023

Assessment Title: Predictive Model Creation and Evaluation

Assessment Weighting: 20%

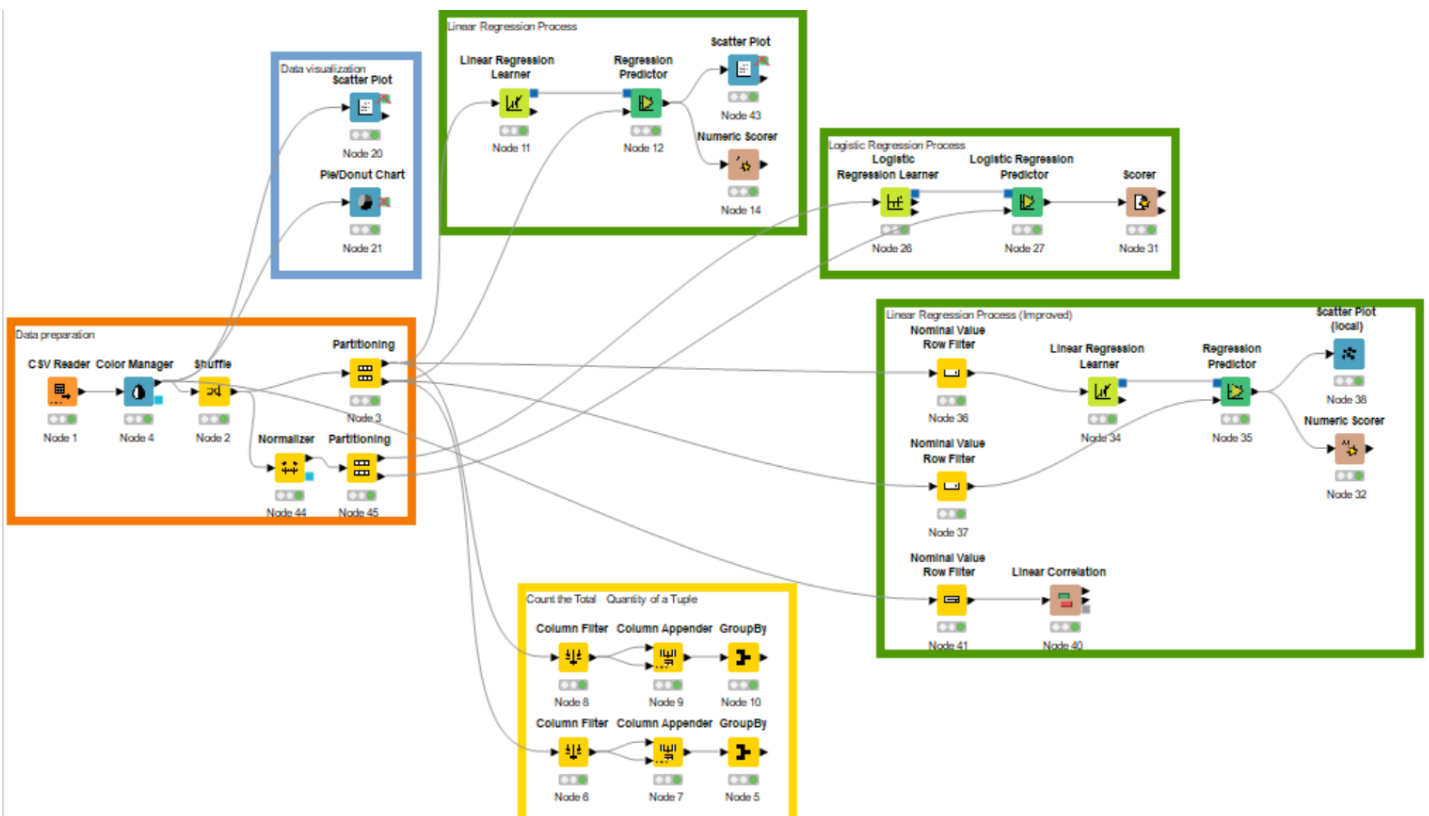
Due Date: Sunday, 1st October 2023 at 11.59 pm (GMT+7)

Assessable Item:

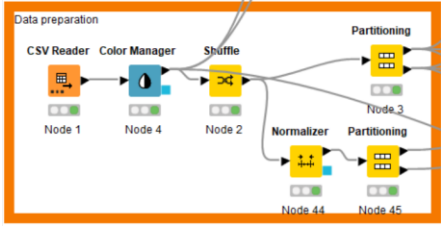
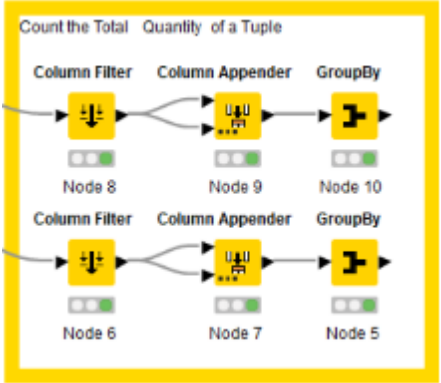
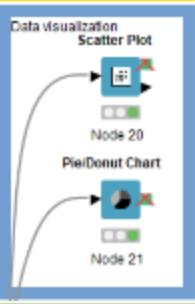
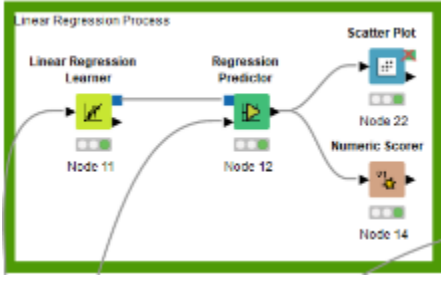
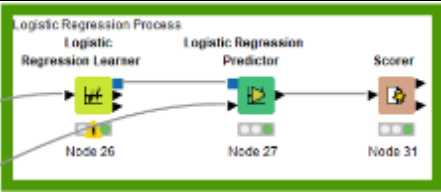
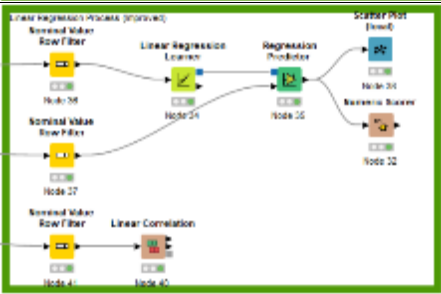
- One (1) piece of a written report no more than 10-page long with the signed Assignment Cover Sheet.
- A unit peer must review your submission before it can be marked.

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) Past a clear screenshot of the whole workflow of assignment 1 in the report. **[2.5 marks]**



The whole workflow contains 29 nodes divided into 6 different groups. The detail of each group is as follow:

Group	Designation	Functions	Explanation
	Data preparation group	<p>Read the raw CSV data and assign colours to different types of entries (i.e., fishes).</p> <p>Randomise the data set (with seed 3122). Separate data in the 80-20 ratio.</p> <p>Normalize data (for Logistic Regression) based on Min-Max normalization.</p>	<p>This group prepares the data to be used in the Linear Regression group, where 80% of the prepared data will be used as the training set, while the other 20% will be used for the test set.</p> <p>Prior to partitioning, the data is normalized to ensure the best result when calculating Linear Regression.</p>
	Counting Group	<p>Group and count the number of fish present in the training and testing dataset.</p>	<p>Counting the number of fishes in a particular species is obtained by appending a duplicated column of the species name and then counting the grouped number.</p>
	Visualizing group	<p>Visualized data using scatter plot and pie chart nodes.</p>	<p>This group visualizes the original data (i.e., data read from the CSV reader.). The visualized data is put in the scatter plot and the pie chart node to achieve the respective visualization.</p>
	Linear Regression Model 1	<p>Perform Linear Regression on the data.</p> <p>Output prediction, calculations of the Linear Regression Process</p> <p>Visualised scatter plot data.</p>	<p>This group is the initial linear regression model built to predict the value of the "Weight_of_Fish_in_Gram" attributes. Via Scatter plot, it visualised the original and the prediction data, allowing us to observe the LRM result better. Calculations (i.e., R2, means, etc.) made by the model are captured by the Numeric Scorer.</p>
	Logistic Regression Model	<p>Perform Logistic Regression on the data.</p> <p>Output categorization of each fish species</p>	<p>This model categorizes fish species and outputs the confusion matrix via the Scorer, which details the fish species' TP, FP, and FN cases.</p>
	Linear Regression Model 2	<p>Removes specified unwanted data and performs Linear Regression on the data.</p> <p>Output prediction, calculations of the Linear Regression Process, and visualized scatter plot data.</p> <p>Calculate collinearity between the attributes.</p>	<p>This model is the improved version of the initial model. It predicts the value of the "Weight_of_Fish_in_Gram" attributes of only one fish species (i.e., Perch). The improvement over the other model is made by having some data (high collinearity attributes) removed beforehand (via the nominal value row filter node) to increase the model's prediction accuracy.</p>

- 2) How many tuples are included in the training set? **[2.5 marks]**

The original data contains 150 tuples; using the partitioning tool, we received 80% of tuples in the training set, which amounted to 120.

Ans: 120 tuples.

- 3) How many species are included in the test set? **[2.5 marks]**

Row ID	S Species	I Species...
Row0	Bream	2
Row1	Parkki	2
Row2	Perch	13
Row3	Pike	5
Row4	Roach	4
Row5	Smelt	2
Row6	Whitefish	2

With the help of the groupby node, the number of species can be observed to be 7 in the test set.

Ans: 7 species.

- 4) Do species Whitefish and Smelt have the same number of tuples included in the test set? **[2.5 marks]**

Ans: The number of tuples from Whitefish & Smelt are similar, they both have 2 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]**

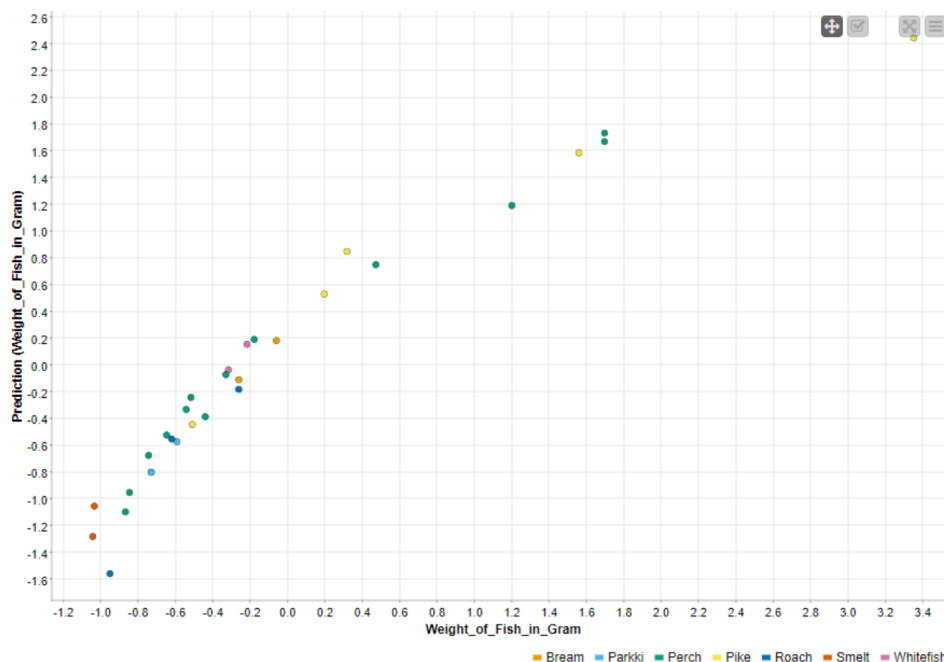
- 1) What is the R^2 value of your test result? **[5 marks]**

R^2 :	0.918
Mean absolute error:	0.204
Mean squared error:	0.082
Root mean squared error:	0.286
Mean signed difference:	0.05
Mean absolute percentage error:	0.588
Adjusted R^2 :	0.918

The coefficient calculation is inside the numeric scorer nodes; per inspecting the node, it is evident that the R^2 value is 0.918.

Ans: 0.918

- 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]

Row ID	S Species	D Weight...	D Diagon...	D Vertical...	D Cross_...	D Height_...	D Diagon...	D ▼ Pre...
Row133	Pike	1,600	60	56	64	9.6	6.144	1,269.986
Row118	Perch	1,000	44	41.1	46.6	12.489	7.596	1,012.415
Row115	Perch	1,000	43	39.8	45.2	11.933	7.277	989.374
Row131	Pike	950	51.7	48.3	55.1	8.926	6.171	958.908
Row109	Perch	820	39	36.6	41.3	12.431	7.351	816.207
Row128	Pike	500	45	42	48	6.96	4.896	691.806
Row101	Perch	556	34.5	32	36.5	10.257	6.388	655.953
Row125	Pike	456	42.5	40	45.5	7.28	4.322	576.472
Row99	Perch	320	30	27.8	31.6	7.616	4.772	453.347
Row3	Bream	363	29	26.3	33.5	12.73	4.455	450.338

From the data predicted via the Linear Regression Predictor, the Pike is the heaviest fish, with a weight reaching 1269.986 grams.

Ans: Pike

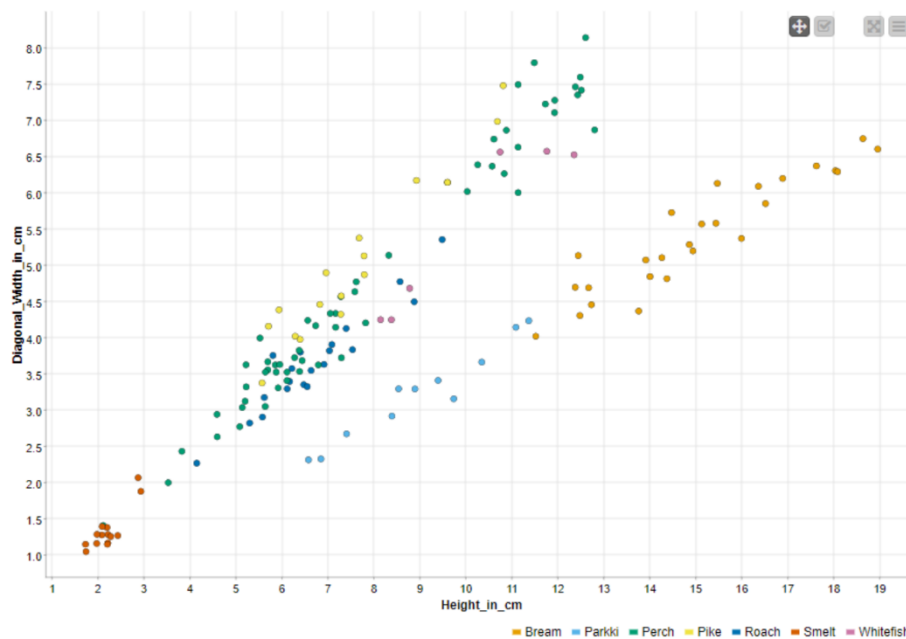
- 4) How many prediction results are infeasible in your test result? [5 marks]

Row67	Perch	70	17.4	15.7	18.5	4.588	2.942	-13.772
Row136	Smelt	6.7	9.8	9.3	10.8	1.739	1.048	-80.472
Row26	Roach	40	14.1	12.9	16.2	4.147	2.268	-181.108

3 infeasible records were found, as their weight are all in the negatives.

Ans: 3

- 5) Looking at your source data before splitting them, which two species can be easily separated from others if looking at the “Height_in_cm” and “Diagonal_Width_in_cm” attributes? Post your visualisation result on data observation in the report. [5 marks]

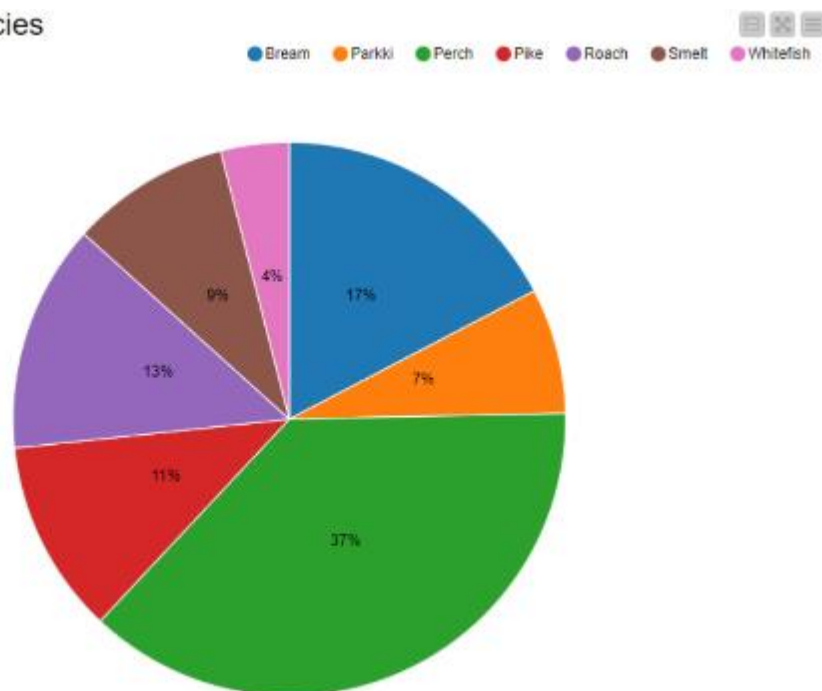


As inspected in the image, Beam and Smelt stand out most from the rest of the population. Smelt's values form a cluster at the bottom-left of the chart, while Beam's values spread out on the opposite end.

Ans: Beam and Smelt

- 6) Draw a pie chart of the original input data 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]**

Percentage of Species



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 **0.0001**, respectively. Use **3122** as the seed in the logistic regression node. Answer the following questions after completing the model training and test. **[40 marks in total]**

We can observe the full confusion matrix via the scorer node and draws some information from there:

ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall
Bream	2	0	28	0	1.0
Roach	3	1	23	3	0.5
Whitefish	0	2	28	0	?
Parkki	2	0	28	0	1.0
Perch	12	1	17	0	1.0
Pike	5	0	25	0	1.0
Smelt	2	0	27	1	0.6666666666666666

- 1) Which species has no "True Positive (TP)" case in the prediction result? **[5 marks]**

As seen in the summarised table, Whitefish has no TP cases.

Ans: Whitefish.

- 2) For the species with no TP case, which species will be misplaced? **[5 marks]**

Via the Logistic Regression Predictor node, we can observe what the Whitefish has been replaced with:

Row ID	Spe...	Predict...
Row48	Whitefish	Roach
Row46	Whitefish	Roach

Ans: Whitefish is misplaced with Roach.

- 3) What is the overall accuracy of the prediction result? **[5 marks]**

The overall accuracy can be calculated by manually counting the records, assuming there are few records in the data. However, for a quicker method, the accuracy can be inspected via the Scorer.

Correct classified: 26	Wrong classified: 4
Accuracy: 86.667%	Error: 13.333%
Cohen's kappa (κ): 0.824%	

Ans: The overall accuracy is **86.667%**

- 4) List all species names that have 100% correctly classified test results. **[15 marks]**
We use the following formula to calculate the accuracy of prediction of each species of fish.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} * 100$$

Based on the confusion matrix, the Beam, Parkki, and Pike have 0 FP and 0 FN. Hence, the accuracy rate is calculated, which amounts to a 100% accuracy rate.

Ans: Beam, Parkki, and Pike

- 5) Which species has a 50% chance of being misplaced into another species in the test result? **[5 marks]**
The chance of a species being misplaced into another one is calculated based on the False Negative rate:

$$FNR = 1 - Recall$$

We have the recall value for each species as follow:

Species	Recall
Whitefish	0.0
Smelt	0.(6)
Roach	0.5
Pike	1.0
Perch	1.0
Parkki	1.0
Beam	1.0

Using the FNR formula, we can calculate the recall rate and then compare the recall rate to all species to find the species that has a 50% chance of being misplaced:

$$50\% = 1 - Recall \Rightarrow Recall = 0.5$$

Roach is the species with the exactly 0.5 Recall's. Hence, the species with a 50% rate of being misplaced is Roach.

Ans: Roach

- 6) In the test result, what percentage of the species "Pike" is misplaced into others? **[5 marks]**
The treatment above will be used for this one:

$$FNR_{Pike} = 1 - Recall_{Pike} \Rightarrow FNR_{Pike} = 1 - 1 = 0$$

Ans: Pike has 0% chance of being misplaced into others.

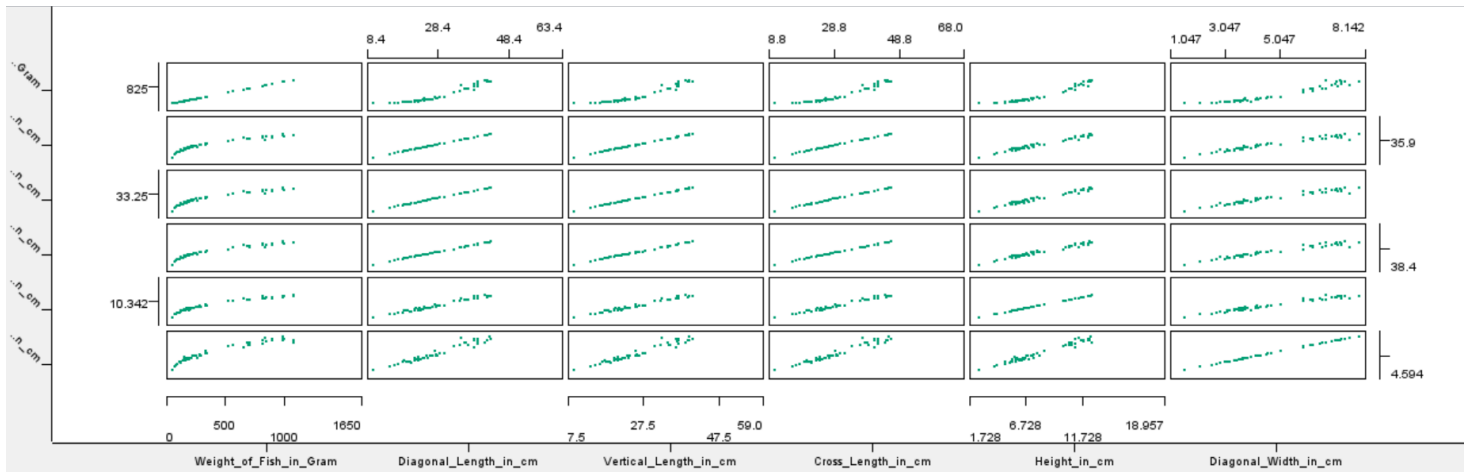
4. Build a new linear regression model different from the one built when answering question 2. This time let's focus on the species "Perch" only. You are limited to using three attributes in the input to predict the "Weight_of_Fish_in_Gram." Use a "Scatter Matrix (local)" node to observe your data and 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]**

- 1) Give the reasons for each eliminated attribute and why they are not selected as the input. **[5 marks]**

To determine which attribute should be removed, one of the most significant considerations is checking for the collinearity between the attributes; this can be achieved using the Linear Correlation (for the exact measurement) and the Scatter matrix (for a more abstract observation) node.

Row ID	S ▼ First column n...	S ▼ Second column...	D Correlation value
Row1	Weight_of_Fish_in_Gram	Vertical_Length_in_cm	0.958361213298317
Row3	Weight_of_Fish_in_Gram	Height_in_cm	0.9684406904743...
Row4	Weight_of_Fish_in_Gram	Diagonal_Width_in_cm	0.9639433246031...
Row0	Weight_of_Fish_in_Gram	Diagonal_Length_in_cm	0.9586558679968...
Row2	Weight_of_Fish_in_Gram	Cross_Length_in_cm	0.9595060788374...
Row10	Vertical_Length_in_cm	Height_in_cm	0.9854201609247...
Row11	Vertical_Length_in_cm	Diagonal_Width_in_cm	0.9744472845922...
Row9	Vertical_Length_in_cm	Cross_Length_in_cm	0.999427381769985
Row14	Height_in_cm	Diagonal_Width_in_cm	0.9829434603923...
Row5	Diagonal_Length_in_cm	Vertical_Length_in_cm	0.9997134894436...
Row7	Diagonal_Length_in_cm	Height_in_cm	0.9855836118303...
Row8	Diagonal_Length_in_cm	Diagonal_Width_in_cm	0.9746171358255...
Row6	Diagonal_Length_in_cm	Cross_Length_in_cm	0.9997790321744...
Row12	Cross_Length_in_cm	Height_in_cm	0.9859092994244...
Row13	Cross_Length_in_cm	Diagonal_Width_in_cm	0.9751312223899...

Row ID	S ▼ First column name	S ▼ Second colu...	D Correlation value
Row1	Weight_of_Fish_in_Gram	Vertical_Length_in_cm	0.958361213298317
Row5	Diagonal_Length_in_cm	Vertical_Length_in_cm	0.9997134894436...
Row3	Weight_of_Fish_in_Gram	Height_in_cm	0.9684406904743...
Row10	Vertical_Length_in_cm	Height_in_cm	0.9854201609247...
Row7	Diagonal_Length_in_cm	Height_in_cm	0.9855836118303...
Row12	Cross_Length_in_cm	Height_in_cm	0.9859092994244...
Row4	Weight_of_Fish_in_Gram	Diagonal_Width_in_cm	0.9639433246031...
Row11	Vertical_Length_in_cm	Diagonal_Width_in_cm	0.9744472845922...
Row14	Height_in_cm	Diagonal_Width_in_cm	0.9829434603923...
Row8	Diagonal_Length_in_cm	Diagonal_Width_in_cm	0.9746171358255...
Row13	Cross_Length_in_cm	Diagonal_Width_in_cm	0.9751312223899...
Row0	Weight_of_Fish_in_Gram	Diagonal_Length_in_cm	0.9586558679968...
Row2	Weight_of_Fish_in_Gram	Cross_Length_in_cm	0.9595060788374...
Row9	Vertical_Length_in_cm	Cross_Length_in_cm	0.999427381769985
Row6	Diagonal_Length_in_cm	Cross_Length_in_cm	0.9997790321744...



As inspected within the nodes, **Diagonal_Length_in_cm** and **Height_in_cm** are the attributes that ought to be eliminated as they present strong collinearity with each other as well as many other attributes. To get into more details:

- **Height_in_cm** collinear with: **Weight_of_Fish_in_Gram**, **Vertical_Length_in_cm**, **Diagonal_Length_in_cm**, **Cross_Length_in_cm**.
- **Diagonal_Length_in_cm** collinear with: **Weight_of_Fish_in_Gram**, **Height_in_cm**, **Vertical_Length_in_cm**, **Cross_Length_in_cm**.

The other items that were not opted for elimination fall into one or more of the categories listed below:

- Their correlation value is less significant than the two items chosen above.
- They don't collinear with many other attributes.
- Only two attributes should be removed.

This level of collinearity and quite possibly multicollinearity of **Diagonal_Length_in_cm** and **Height_in_cm** lowers the statistical significance of the original regression model as they reduce the accuracy of the estimated coefficients.

Ans:

The remaining attributes: **Vertical_Length_in_cm**, **Cross_Length_in_cm**, **Diagonal_Width_in_cm**.

The eliminated attributes: **Diagonal_Length_in** and **Height_in_cm**.

- 2) 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]**

Original Model

R^2 :	0.918
Mean absolute error:	73.907
Mean squared error:	10,759.319
Root mean squared error:	103.727
Mean signed difference:	18.12
Mean absolute percentage error:	0.882
Adjusted R^2 :	0.918

Improved Model

R^2 :	0.957
Mean absolute error:	58.477
Mean squared error:	4,726.137
Root mean squared error:	68.747
Mean signed difference:	23.411
Mean absolute percentage error:	0.24
Adjusted R^2 :	0.957

After completing the necessary steps to improve the model, the model has indeed been improved. The indicator for this is that the R^2 has been improved by 0.04 from the original 0.918 to 0.957, and the higher the R^2 , the higher the prediction strength, ultimately leading to an increase in model accuracy.

Ans: The R^2 value has increased by 0.04, which increased the prediction strength. The accuracy of the model has been improved.