Cover sheet for submission of work for assessment

SWINBURNE UNIVERSITY OF TECHNOLOGY

UNIT DETAILS

Unit name	Data Science Principles			Class day/time	
Unit code	COS10022	Assignment no.	1	Due date	01/09/2023
Name of lecturer/teacher	Pham Thi Kim Dung				
Tutor/marker's name					

Office use only

Faculty or school date stamp

STUDENT(S)

	Family Name(s)	Given Name(s)	Student ID Number(s)
(1)	Le	Xuan Nhat	104169523
(2)			
(3)			
(4)			
(5)			
(6)			

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)		(4)	
(2)		(5)	
(3)		(6)	

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/**

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

COS10022 Data Science Principles

Assignment 1 - *Semester 1, 2023*

Assessment Title: Predictive Model Creation and Evaluation

***Assessment Weighting**: 20%*

Due Date: Saturday, 26th March 2023 at 11.59 pm (AEDT)

Assessable Item:

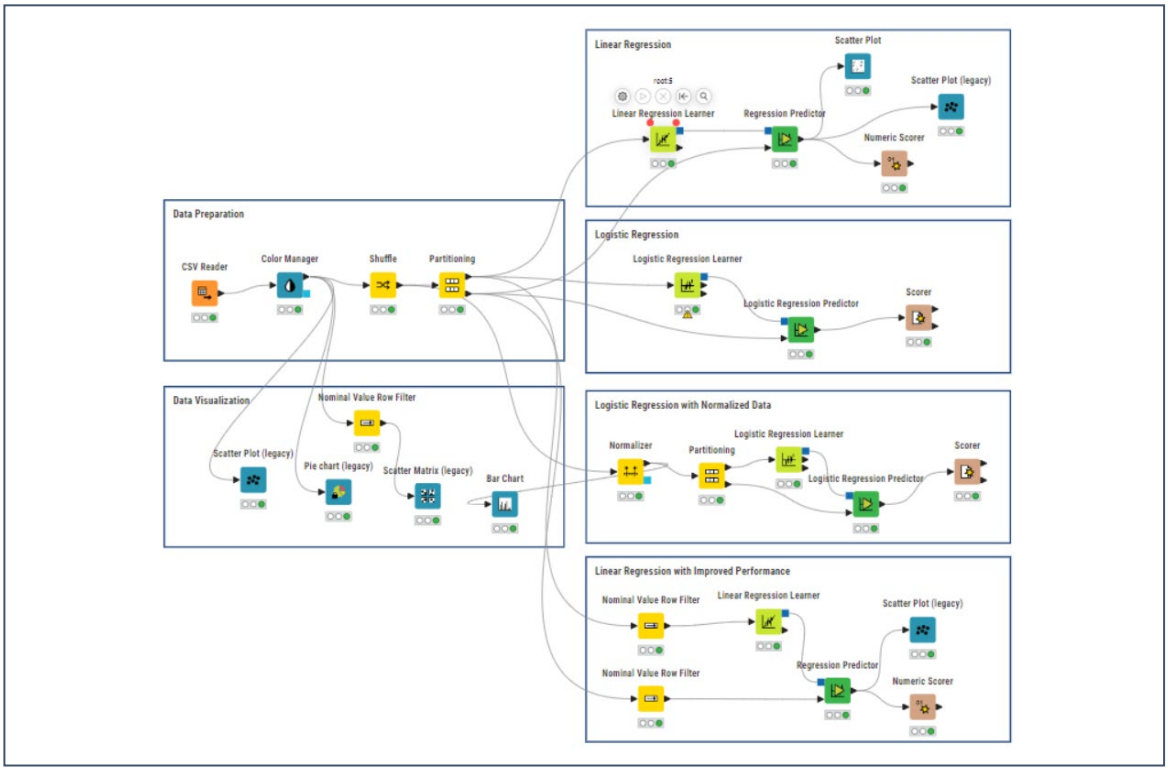
- One (1) piece of a written report no more than 10-page long with the signed Assignment Cover Sheet.
- A unit peer must review your submission before it can be marked.

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.

---

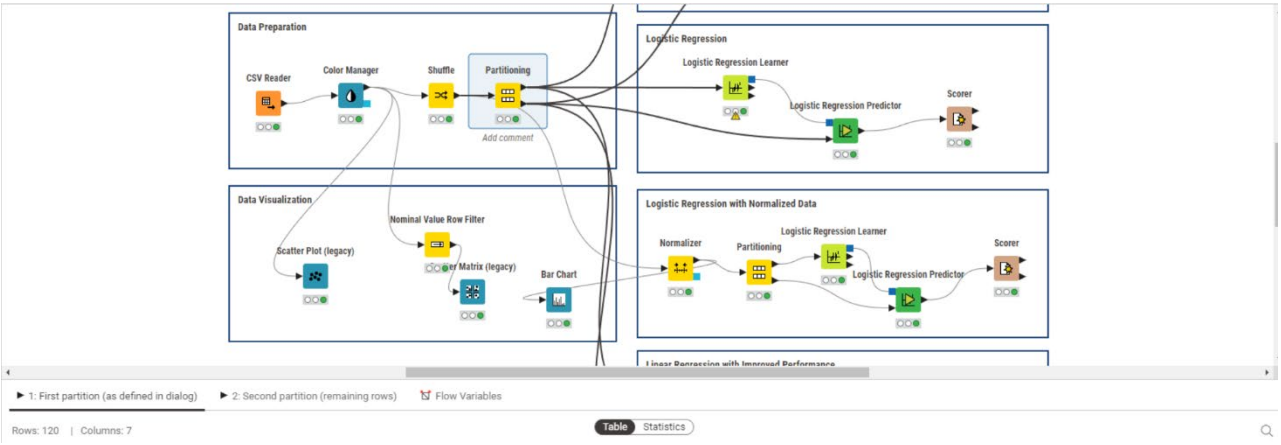
1. Follow the instructions above to split the source data into training and test sets. Answer the following



2. 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]**

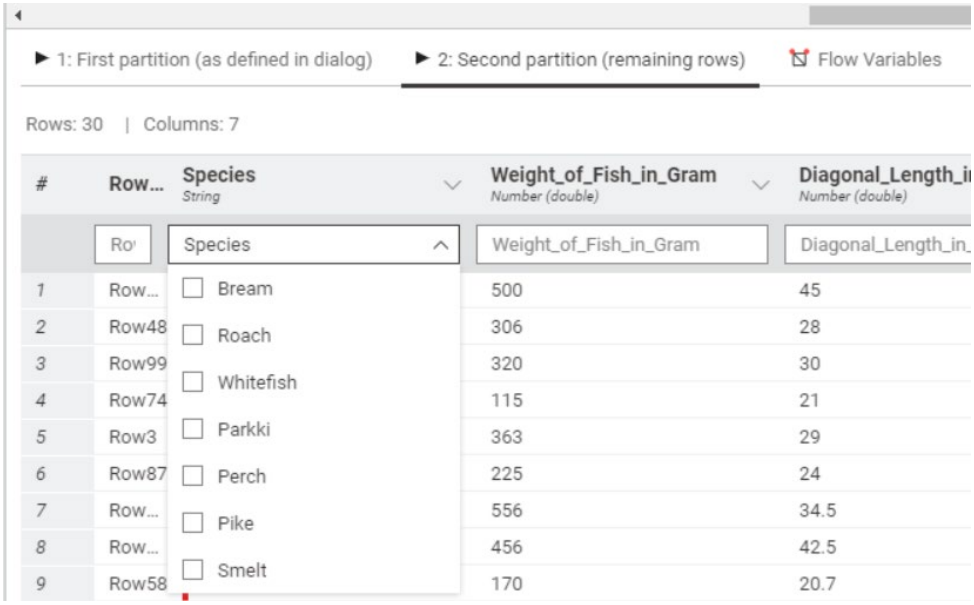
2) How many tuples are included in the training set? **[2.5 marks]**

Ans: 120 tuples. We can look at the first partition in “Partitioning” node and see the number of rows.



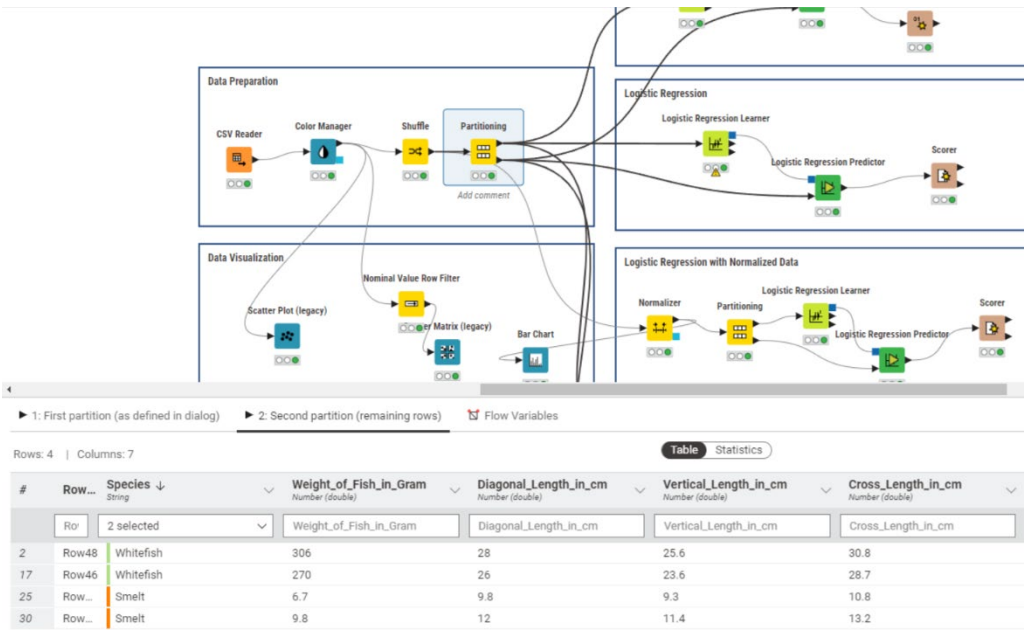
3) How many species are included in the test set? **[2.5 marks]**

Ans: 7 species total. We can look at the second partition, use the filter options to see the number of species.

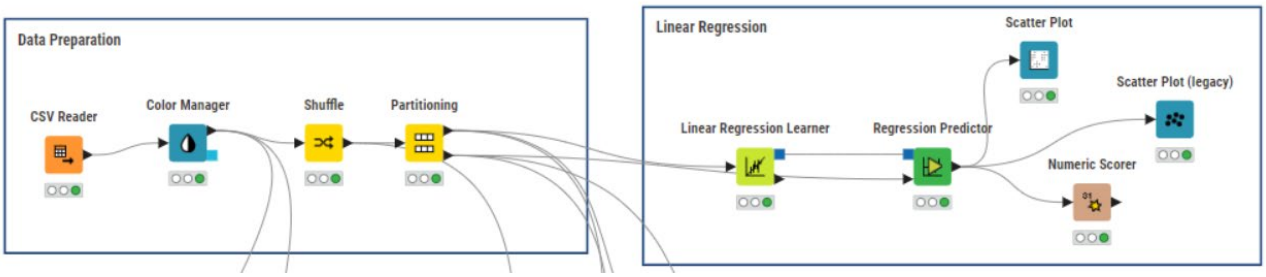


4) Do species “Whitefish” and “Smelt” have the same number of tuples included in the test set? **[2.5 marks]**

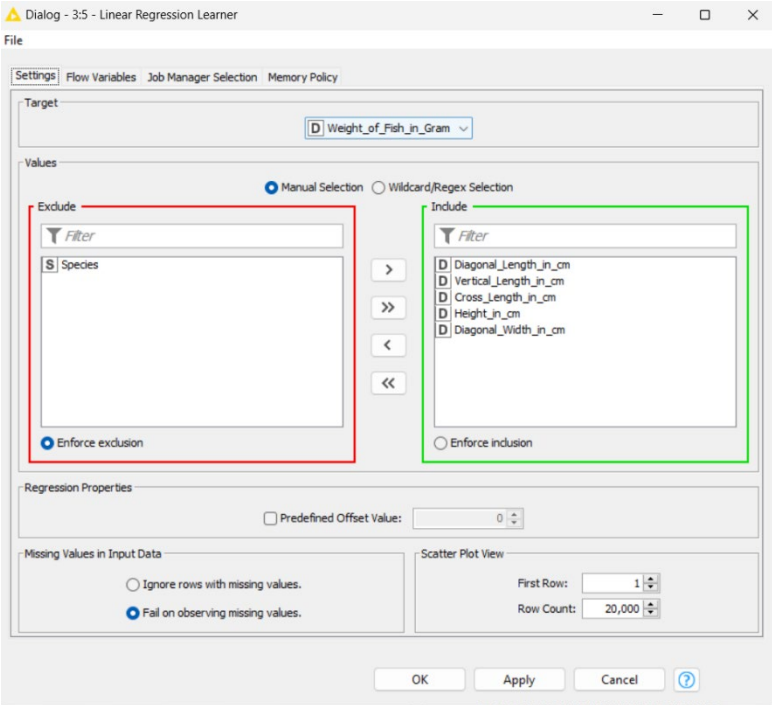
Ans: Yes. Both of them have 2 tuples. Again, we look at the second partition and use the filter for 2 species in the table.



3. 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]**



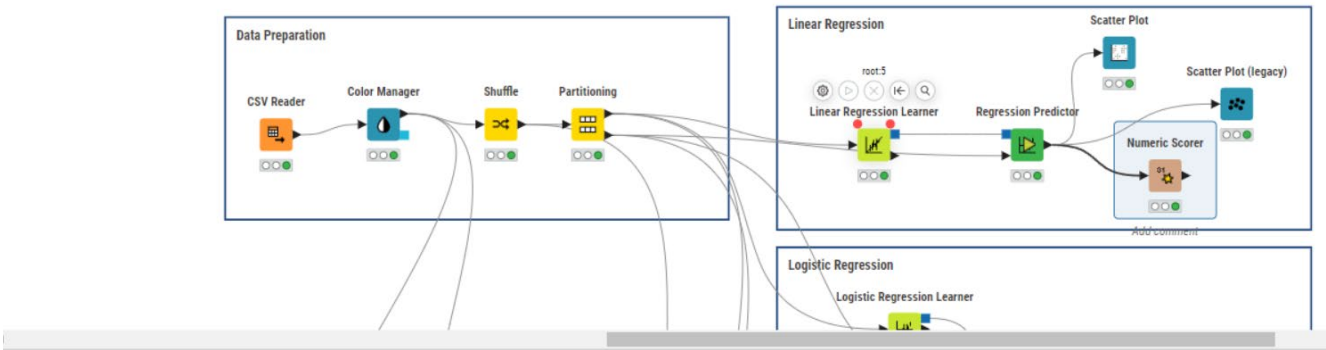
Here is the workflow of the Linear Regression model.



Here is the configuration for the Linear Regression Learner. We exclude “Species” because it is not a numerical attribute.

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

   Ans: 0.857. We use the Numeric Scorer node.

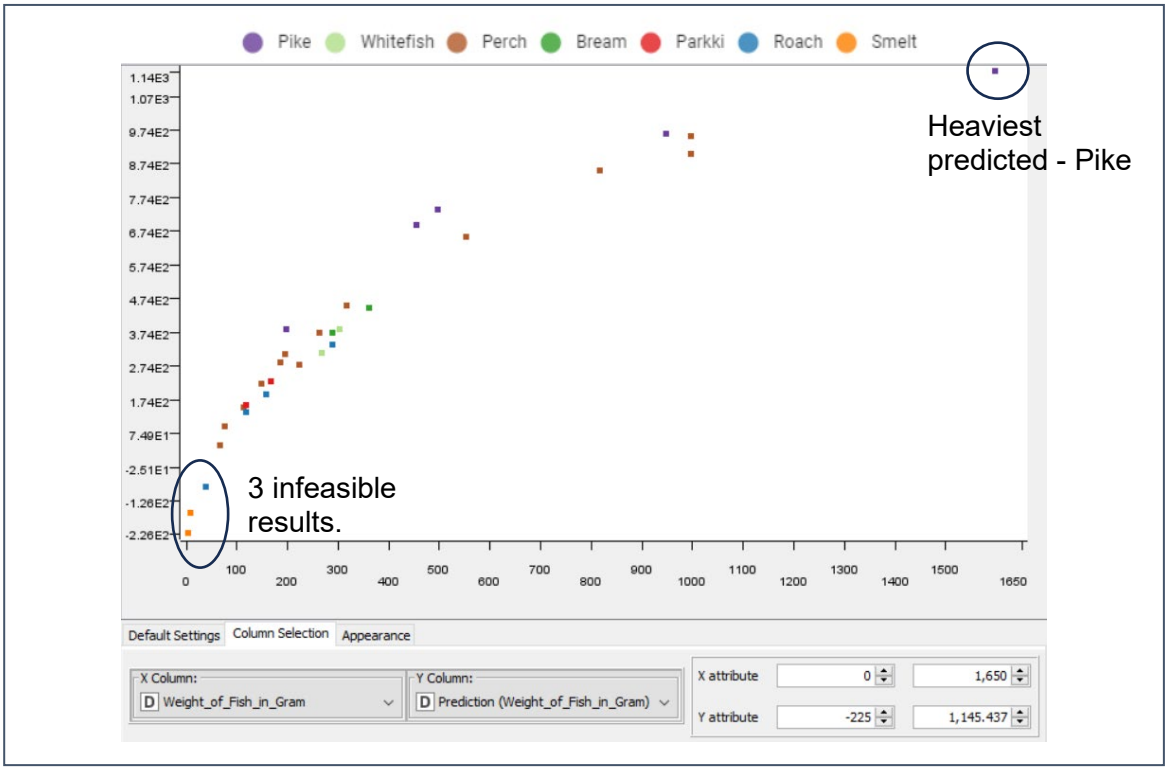


▶ 1: Statistics Flow Variables

Rows: 7 | Columns: 1 Table Statistics

#	Row...	Prediction (Weight_of_Fish_in_Gram) Number (double)
1	R^2	0.857
2	mea...	101.021
3	mea...	18,678.603
4	root ...	136.67
5	mea...	23.338
6	mea...	2.118
7	adju...	0.857

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]**



We already assign different colours in the Color Manager node.

3) Which species has the heaviest predicted weight in your test result? **[5 marks]**

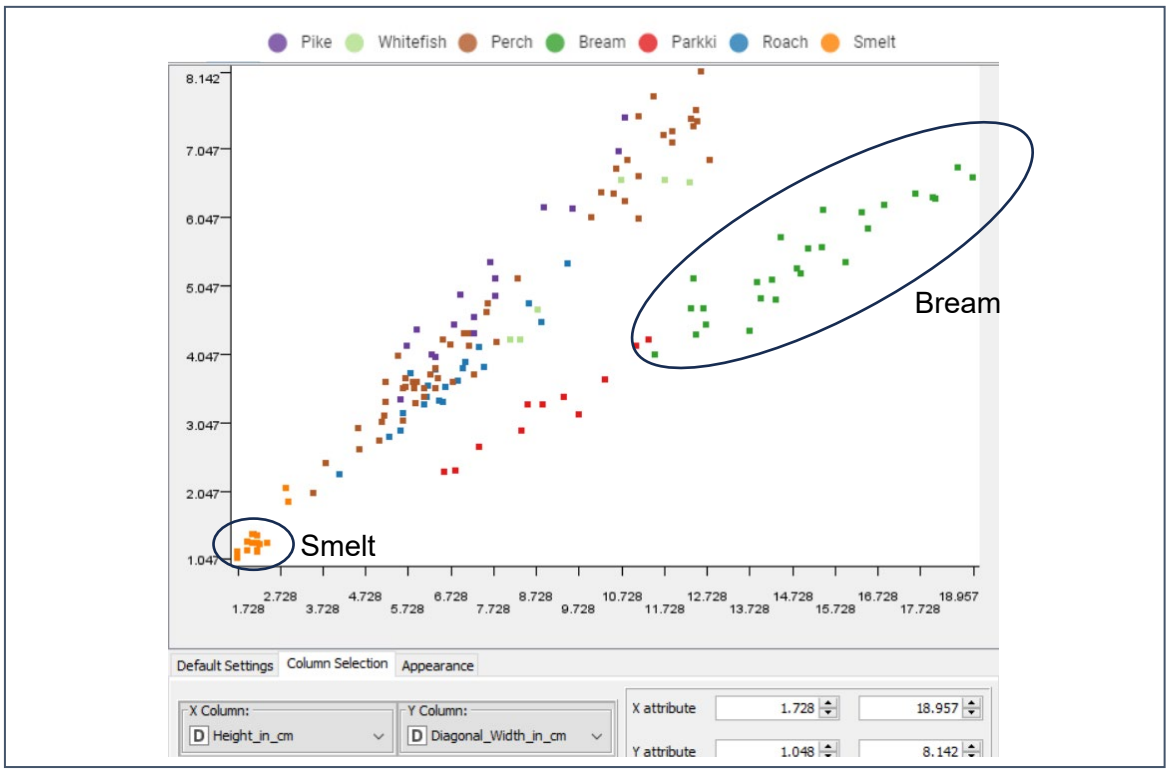
Ans: Pike

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

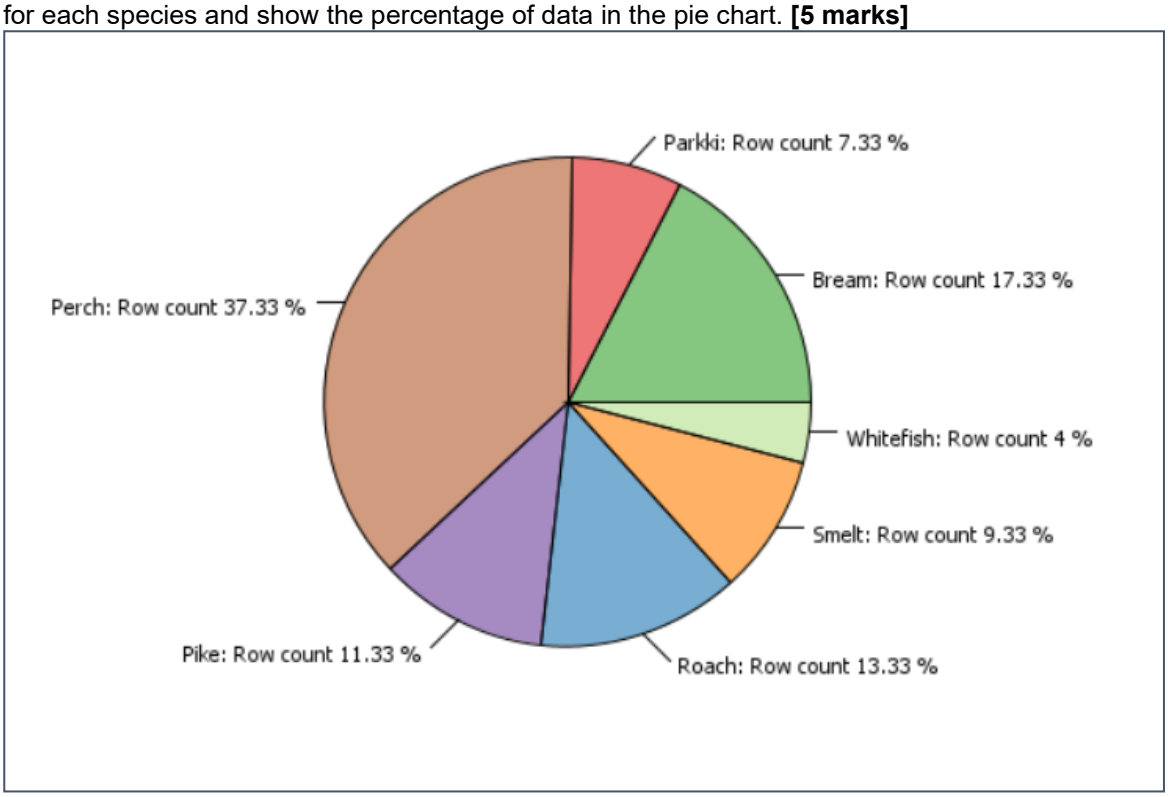
Ans: 3 are below 0, so they are infeasible.

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]**

Ans: Smelt and Bream. We use the Scatter Plot (legacy) node.



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]**

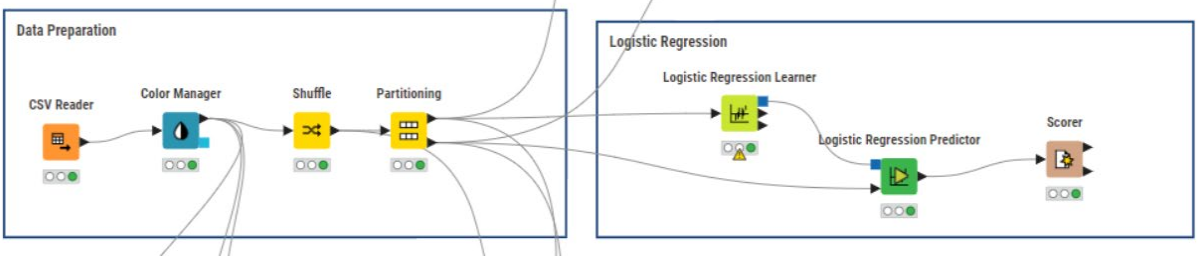


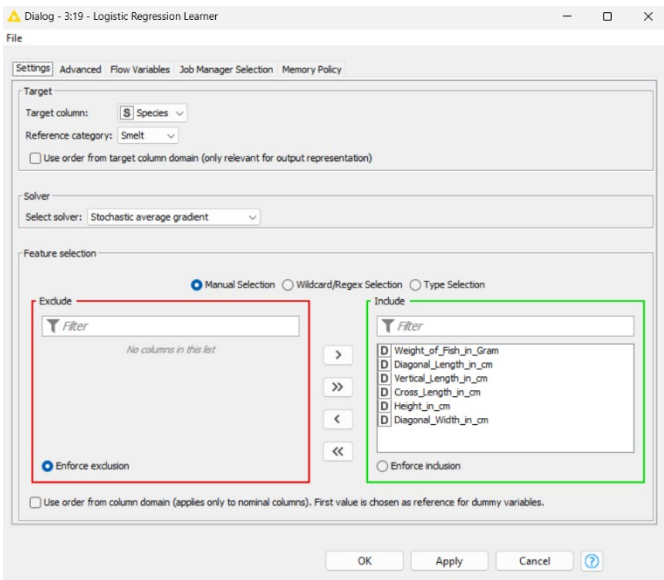
We use the Pie Chart (legacy) node in the "Data Visualization" section before partitioning.

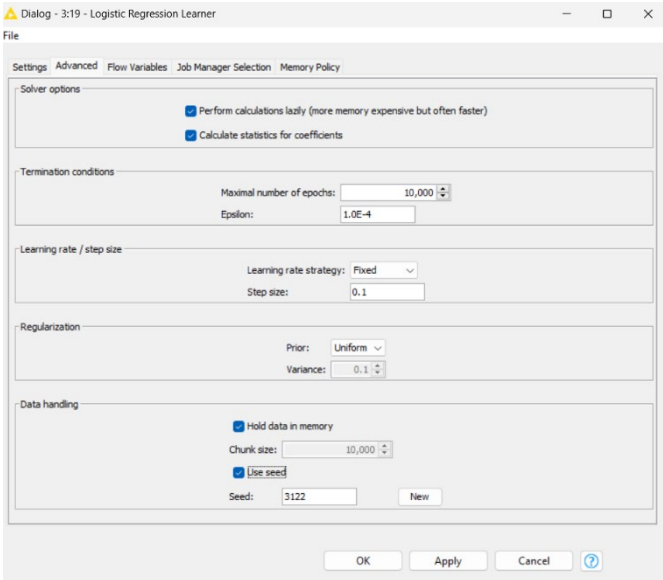
4. Build a Logistic Regression Model with **all** attributes and use "Smelt" as the reference category. The maximal number of epochs and epsilon should be set to **10,000** and **0.0001**, respectively. Use **3122** as the seed in the logistic regression node. Answer the following questions after completing the model training and test. **[40 marks in total]**

We will split into 2 cases, because there is a warning on the Logistic Regression Leaner node.

Case 1: Proceed like normal. Here are the workflow and the settings for the logistic regression Learner node:

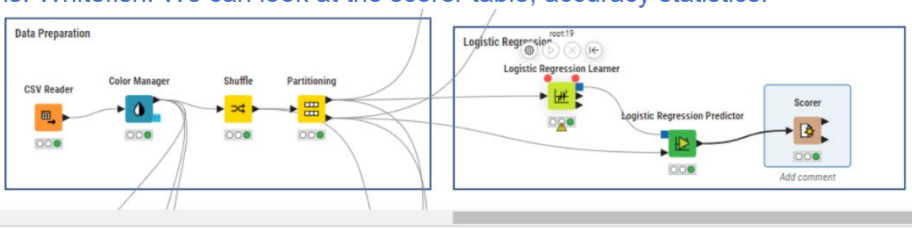






1) Which species has no "True Positive (TP)" case in the prediction result? **[5 marks]**

Ans: Whitefish. We can look at the scorer table, accuracy statistics:



1: Confusion matrix | 2: Accuracy statistics | Flow Variables

Rows: 8 | Columns: 11

Table | Statistics

#	RowID	TruePositives Number (integer)	FalsePositives Number (integer)	TrueNegatives Number (integer)	FalseNegatives Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)
1	Bream	2	1	27	0	1	0.667	1	0.964	0.8	?
2	Roach	4	6	20	0	1	0.4	1	0.769	0.571	?
3	Whitefish	0	0	28	2	0	?	0	1	?	?
4	Parkki	1	0	28	1	0.5	1	0.5	1	0.667	?
5	Perch	7	0	17	6	0.538	1	0.538	1	0.7	?
6	Pike	5	2	23	0	1	0.714	1	0.92	0.833	?
7	Smelt	2	0	28	0	1	1	1	1	1	?
8	Overall	?	?	?	?	?	?	?	?	?	0.7

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

Ans: Roach. The species with no TP case is Whitefish, so we will go back to the Predictor to see the prediction result. As we can see after filtering "Whitefish", we can see all of them are predicted as "Roach".

1: Predicted data | Flow Variables

Rows: 2 | Columns: 8

Table | Statistics

#	Row...	Species String	Weight_of_Fish_in_Gram Number (double)	Diagonal_Length_in_cm Number (double)	Vertical_Length_in_cm Number (double)	Cross_Length_in_cm Number (double)	Height_in_cm Number (double)	Diagonal_Width_in_cm Number (double)	Prediction (Species) String
	Ro	1 selected	Weight_of_Fish_in_Gram	Diagonal_Length_in_cm	Vertical_Length_in_cm	Cross_Length_in_cm	Height_in_cm	Diagonal_Width_in_cm	Prediction (Species)
2	Row48	Whitefish	306	28	25.6	30.8	8.778	4.682	Roach
17	Row46	Whitefish	270	26	23.6	28.7	8.38	4.248	Roach

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

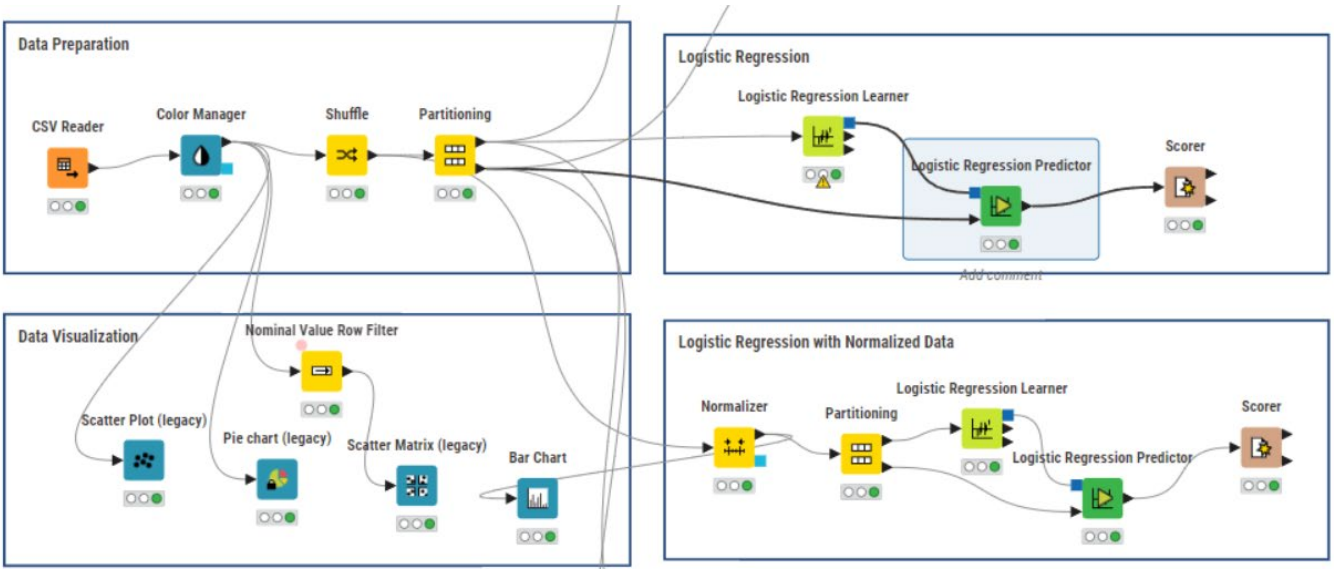
Ans: 0.7, which is 70%. Look at the Accuracy statistics above.

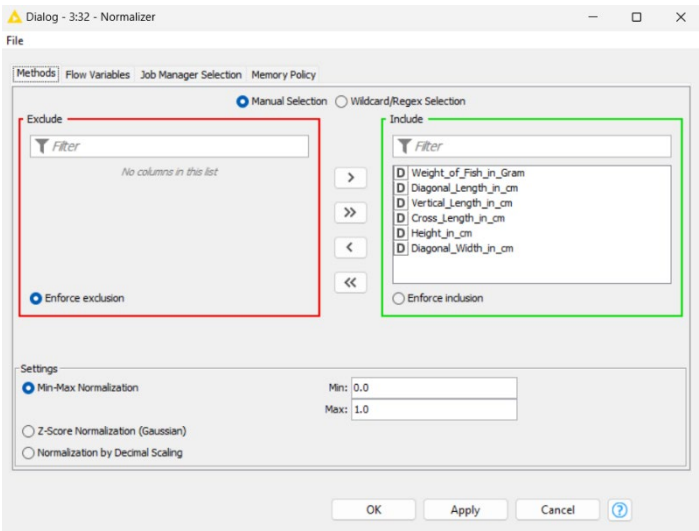
4) List all species names that have 100% correctly classified test results. **[15 marks]**
   Ans: Species with 100% correctly classified test results means they have recall of 100% or 1. They are: Bream, Roach, Pike, Smelt (Accuracy statistics)
5) Which species has a 50% chance of being misplaced into another species in the test result? **[5 marks]**
   Ans: We also use recall to answer this question. Only "Parkki" has 0.5 (or 50%) recall, so the answer is Parkki. (Accuracy Statistics)
6) In the test result, what percentage of the species "Pike" is misplaced into others? **[5 marks]**
   Ans: In this case, we will use Precision to answer. "Pike" 's precision is 0.0714 or 71.4%, so the percentage of the species is misplaced into others is 100% - 71.4% = 28.6% (Accuracy Statistics)

Case 2: We add a "Normalizer" node before partitioning. We use the normalizing method of Min-Max Normalization due to the following reason:

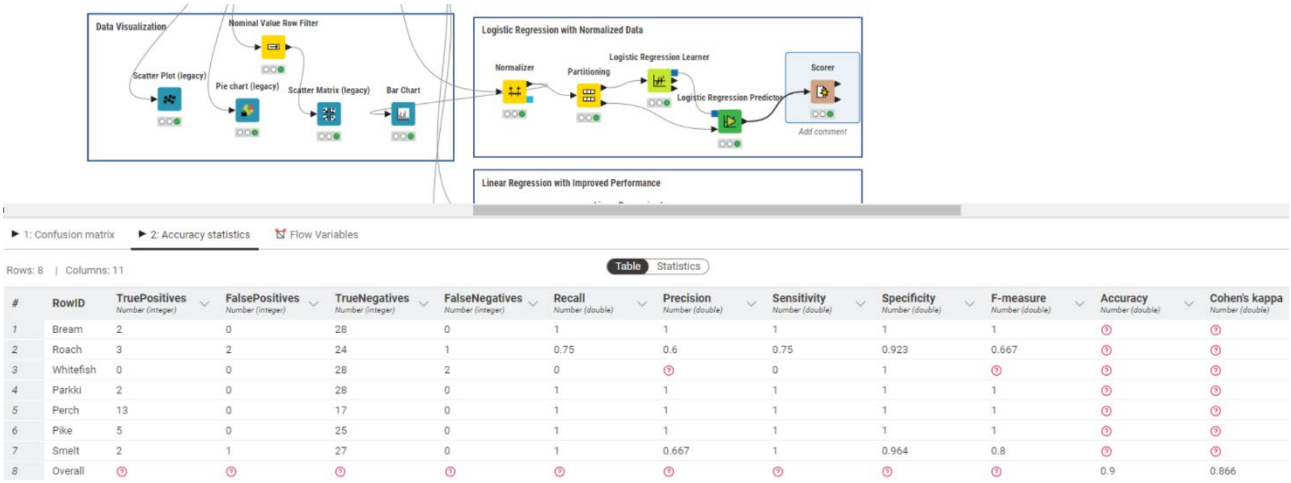
1. The data do not have too many outliers.
2. If we use z-score normalization, there will be negative numbers, thus make the model inaccurate (we can check by adding a bar chart after the normalizer)

Here are the workflow and the settings: (the rest is the same as the previous logistic model, as well as how to find the answers)



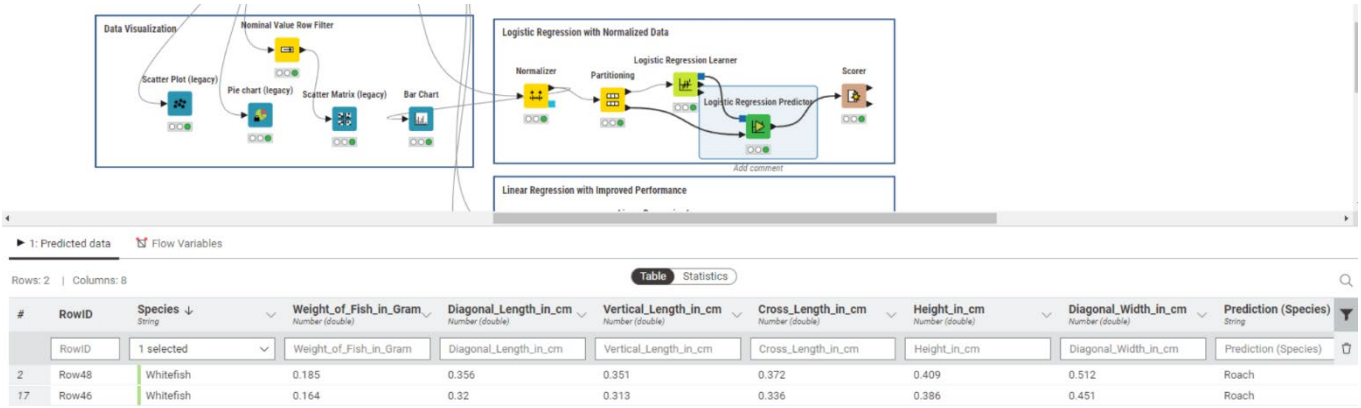


1) Which species has no “True Positive (TP)” case in the prediction result? **[5 marks]**
   Ans: Whitefish.



#	RowID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
1	Bream	2	0	28	0	1	1	1	1	1	?	?
2	Roach	3	2	24	1	0.75	0.6	0.75	0.923	0.667	?	?
3	Whitefish	0	0	28	2	0	?	0	1	?	?	?
4	Parkki	2	0	28	0	1	1	1	1	1	?	?
5	Perch	13	0	17	0	1	1	1	1	1	?	?
6	Pike	5	0	25	0	1	1	1	1	1	?	?
7	Smelt	2	1	27	0	1	0.667	1	0.964	0.8	?	?
8	Overall	?	?	?	?	?	?	?	?	?	0.9	0.866

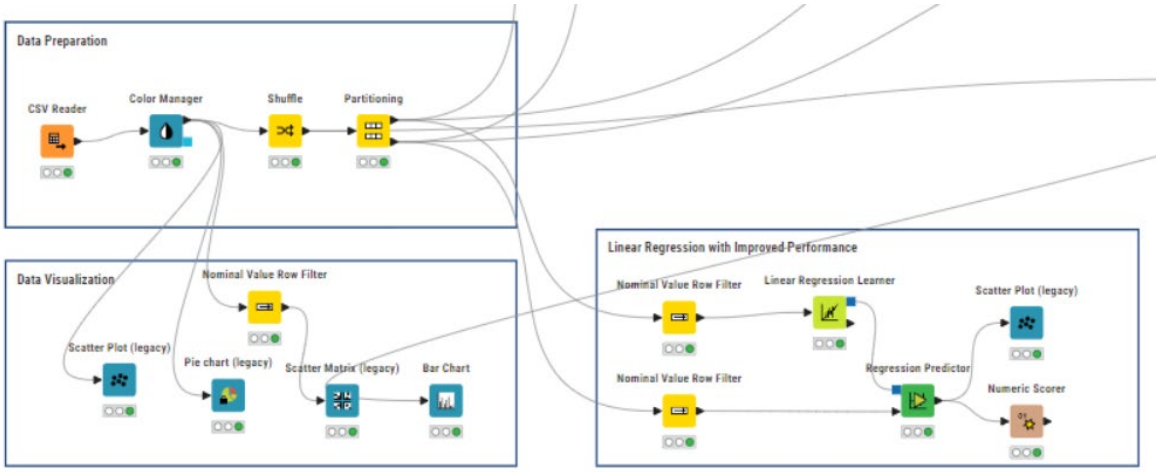
2) For the species with no TP case, which species will be misplaced? **[5 marks]**
   Ans: Roach.



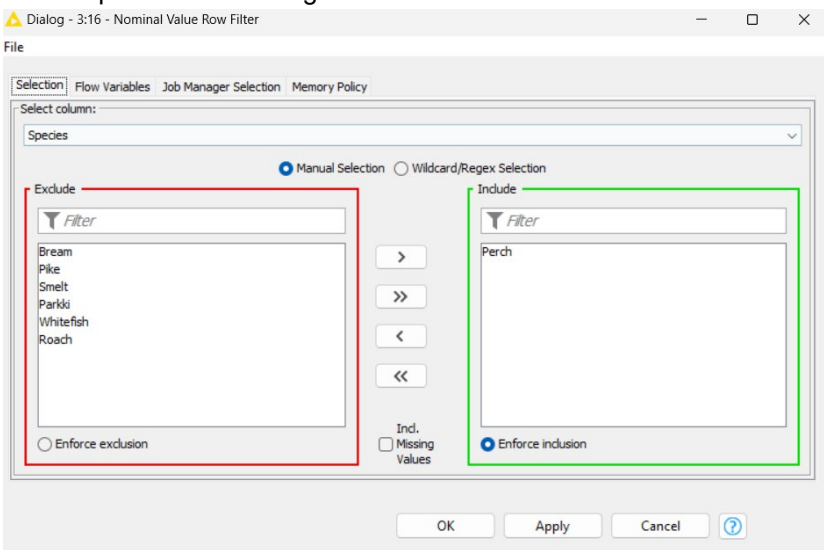
#	RowID	Species	Weight_of_Fish_in_Gram	Diagonal_Length_in_cm	Vertical_Length_in_cm	Cross_Length_in_cm	Height_in_cm	Diagonal_Width_in_cm	Prediction (Species)
	RowID	1 selected	Weight_of_Fish_in_Gram	Diagonal_Length_in_cm	Vertical_Length_in_cm	Cross_Length_in_cm	Height_in_cm	Diagonal_Width_in_cm	Prediction (Species)
2	Row48	Whitefish	0.185	0.356	0.351	0.372	0.409	0.512	Roach
17	Row46	Whitefish	0.164	0.32	0.313	0.336	0.386	0.451	Roach

3) What is the overall accuracy of the prediction result? **[5 marks]**
   Ans: 0.9, which is 90%
4) List all species names that have 100% correctly classified test results. **[15 marks]**
   Ans: Species with 100% correctly classified test results means they have recall of 100% or 1. They are: Bream, Parkki, Perch, Pike, Smelt
5) Which species has a 50% chance of being misplaced into another species in the test result? **[5 marks]**
   Ans: We also use recall to answer this question. None has 0.5 (or 50%) recall, so the answer is None.
6) In the test result, what percentage of the species “Pike” is misplaced into others? **[5 marks]**
   Ans: In this case, we will use Precision to answer. “Pike” ‘s precision is 1 or 100%, so the percentage of the species is misplaced into others is 100% - 100% = 0%

5. 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]**



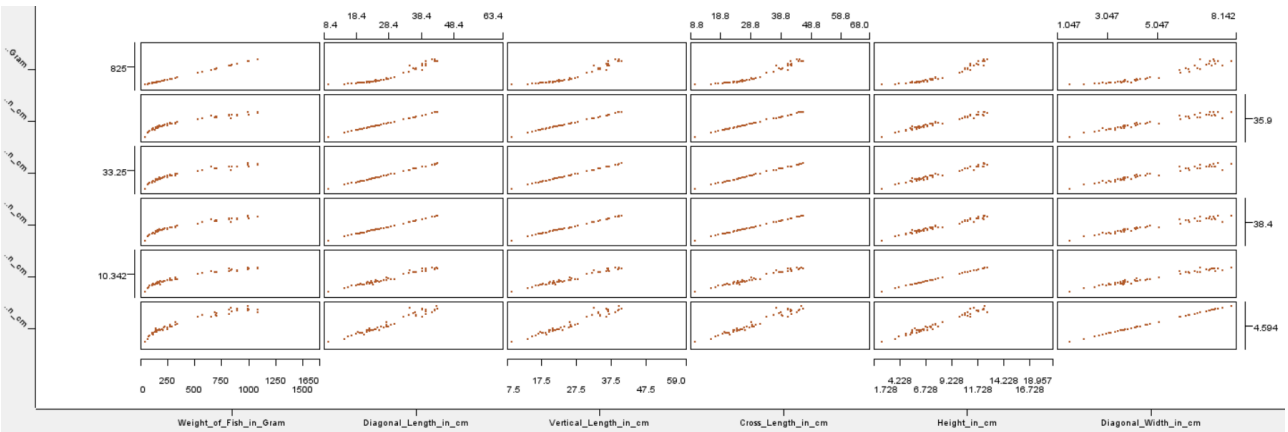
Here is the workflow of the Improved Linear Regression model.



All 3 Nominal Value row filter is set to include “Perch” only.

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

Ans: Here is the Scatter Matrix (local) node for the data we need:



By observation and trials, removing Diagonal_Length and Height will yield the highest possible R^2 I can find. To further support this decision, in this matrix, we can see that attributes Diagonal_Length_in_cm and Height_in_cm has one of the most collinearity values compares to others, so the two of them should be eliminated.

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]**
Ans: The new R^2 is 0.957.

New model:

#	RowID	Prediction (Weight_of_Fish_in_Gram) Number (double)
1	R^2	0.957
2	mean absolute error	58.477
3	mean squared error	4,726.137
4	root mean squared error	68.747
5	mean signed difference	23.411
6	mean absolute percentage error	0.24
7	adjusted R^2	0.957

Old model:

#	RowID	Prediction (Weight_of_Fish_in_Gram) Number (double)
1	R^2	0.857
2	mean absolute error	101.021
3	mean squared error	18,678.603
4	root mean squared error	136.67
5	mean signed difference	23.338
6	mean absolute percentage error	2.118
7	adjusted R^2	0.857