Cover sheet for submission of work for assessment

SWINBURNE UNIVERSITY OF TECHNOLOGY

UNIT DETAILS

Unit name	Data Science Principles			Class day/time	Fridays	Office use only
Unit code	COS10022	Assignment no.	1	Due date	29/9/2024	
Name of lecturer/teacher	Mr. Minh					
Tutor/marker's name	Mr. Minh					Faculty or school date stamp

STUDENT(S)

Family Name(s)	Given Name(s)	Student ID Number(s)
Tran	Thien Thao Vy	104991221

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.

I. Introduction

The dataset, “Fish_Species_2024.csv”, comprises of 7 commonly found fish species in the market, featuring 6 attributes.

The assignment aims to give chances to practice on creating 2 kinds of regression model, which, firstly, is the linear regression model for predicting the weight of the fish. On the other hand, the logistic regression model helps predicting the species of the fish and a small enhancement is required for better working performance.

II. KNIME Workflow

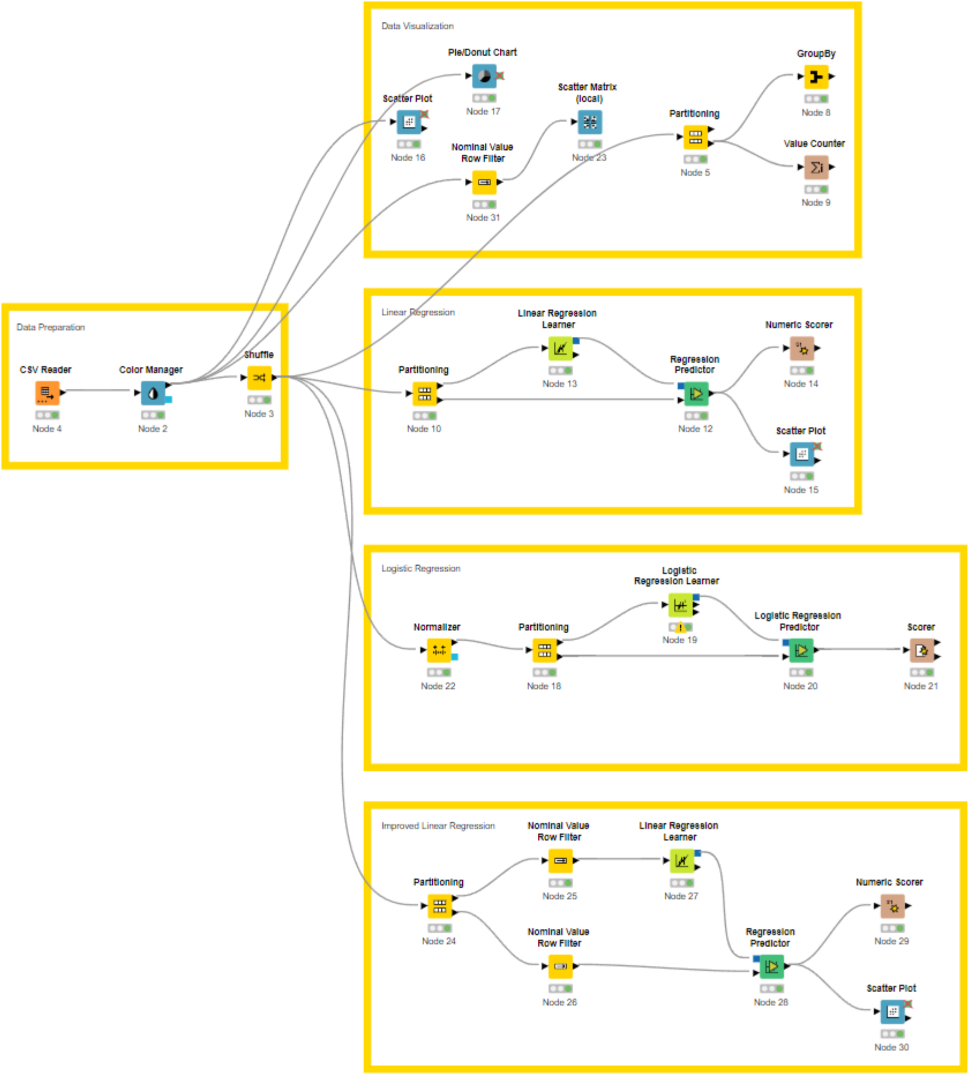


Figure 1: KNIME overall workflow

III. Assignment Tasks

1. Data Preparation

Data Preparation is an important threshold which keeps the workflow function, as well as its vital role in ensuring the accuracy and efficiency of the output data. Therefore, the first step is to ensure that the data is prepared properly. Before working with the data, or after importing data to “CSV reader”, it is in need to manage the data visualization through “Color Manager” node and shuffle the data in “Shuffle”. The “Partitioning” part is also important as we split the data into training and testing parts.

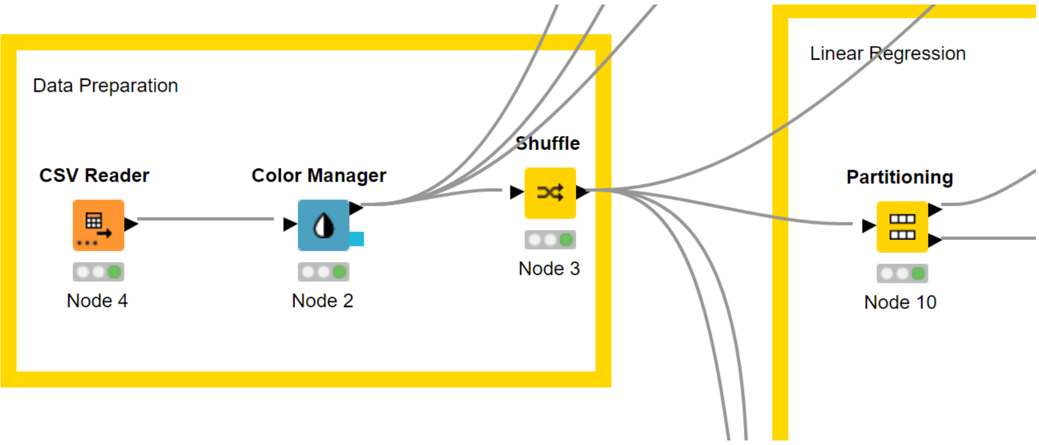


Figure 2: Data Preparation process

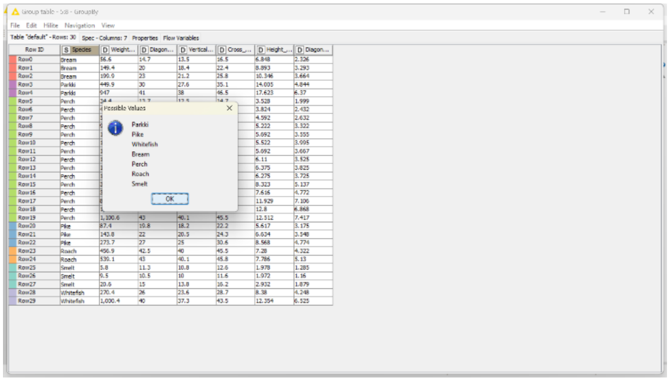
Answering the questions of the tasks:

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]**

   2) How many tuples are included in the training set? **[2.5 marks]**
   There are 150 tuples in total and 80% of them are included in the training set. Therefore, there are 120 tuples in the training set.

   3) How many species are included in the test set? **[2.5 marks]**
   There are 7 species are included in test set. As shown in the "GroupBy" node.


Group table - 5:8 - GroupBy
File Edit Hilite Navigation View
Table "default" - Rows: 30
Spec - Columns: 7
Properties
Flow Variables

Row ID	S Species	D Weight...	D Diagon...	D Vertical...	D Cross_...	D Height_...	D Diagon...
Row0	Bream	56.6	14.7	13.5	16.5	6.848	2.326
Row1	Bream	149.4	20	18.4	22.4	8.893	3.293
Row2	Bream	190.9	23	21.2	25.8	10.346	3.664
Row3	Parkki	449.9	30	27.6	35.1	14.005	4.844
Row4	Parkki	947	41	38	46.5	17.623	6.37
Row5	Perch	34.4	13.7	17.5	14.7	3.528	1.999
Row6	Perch	[illegible]	[illegible]	[illegible]	[illegible]	3.824	2.432
Row7	Perch	[illegible]	[illegible]	[illegible]	[illegible]	4.592	2.632
Row8	Perch	[illegible]	[illegible]	[illegible]	[illegible]	5.222	3.322
Row9	Perch	[illegible]	[illegible]	[illegible]	[illegible]	5.692	3.555
Row10	Perch	[illegible]	[illegible]	[illegible]	[illegible]	5.522	3.995
Row11	Perch	[illegible]	[illegible]	[illegible]	[illegible]	5.692	3.667
Row12	Perch	[illegible]	[illegible]	[illegible]	[illegible]	6.11	3.525
Row13	Perch	[illegible]	[illegible]	[illegible]	[illegible]	6.375	3.825
Row14	Perch	[illegible]	[illegible]	[illegible]	[illegible]	6.275	3.725
Row15	Perch	[illegible]	[illegible]	[illegible]	[illegible]	8.323	5.137
Row16	Perch	[illegible]	[illegible]	[illegible]	[illegible]	7.616	4.772
Row17	Perch	[illegible]	[illegible]	[illegible]	[illegible]	11.929	7.106
Row18	Perch	[illegible]	[illegible]	[illegible]	[illegible]	12.8	6.868
Row19	Perch	1,100.6	43	40.1	45.5	12.512	7.417
Row20	Pike	87.4	19.8	18.2	22.2	5.617	3.175
Row21	Pike	143.8	22	20.5	24.3	6.634	3.548
Row22	Pike	273.7	27	25	30.6	8.568	4.774
Row23	Roach	456.9	42.5	40	45.5	7.28	4.322
Row24	Roach	530.1	43	40.1	45.8	7.786	5.13
Row25	Smelt	5.8	11.3	10.8	12.6	1.978	1.285
Row26	Smelt	9.5	10.5	10	11.6	1.972	1.16
Row27	Smelt	20.6	15	13.8	16.2	2.932	1.879
Row28	Whitefish	270.4	26	23.6	28.7	8.38	4.248
Row29	Whitefish	1,000.4	40	37.3	43.5	12.354	6.525

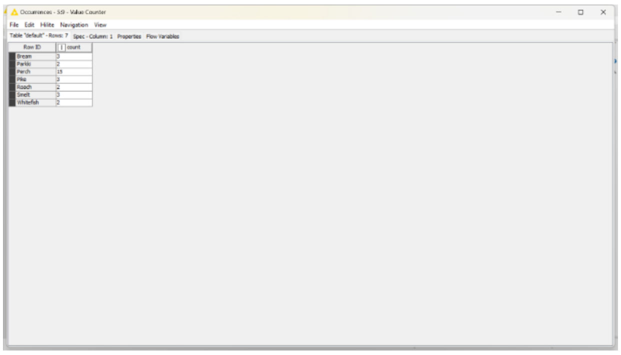
Possible Values

Parkki
Pike
Whitefish
Bream
Perch
Roach
Smelt

OK

Figure 3: GroupBy node

   4) Do species "Whitefish" and "Smelt" have the same number of tuples included in the test set? **[2.5 marks]**
   Using the "Value Counter" node to check the tuples of each type of fish, we can see that "Whitefish" and "Smelt" does not have the same number of tuples, in which "Whitefish" has 2 tuples while "Smelt" has 3 tuples.


Occurrences - 5:9 - Value Counter
File Edit Hilite Navigation View
Table "default" - Rows: 7
Spec - Column: 1
Properties
Flow Variables

Row ID	count
Bream	3
Parkki	2
Perch	15
Pike	3
Roach	2
Smelt	3
Whitefish	2

Figure 4: Value Counter node

2. Linear Regression Model

The initial step to build a proper linear regression model is to partition the data into training set and test. After that, the training set data will go through the "Linear regression learner" node, and in this node, model training begins, to fit the linear training into the data using the least squares method. The next process is testing the model by using "Regression Predictor" node, which has already obtained the coefficients from the Learner, the data that goes through this process is the one from test set. To visualize the output, we can employ "Scatter Plot" and "Numeric Scorer".

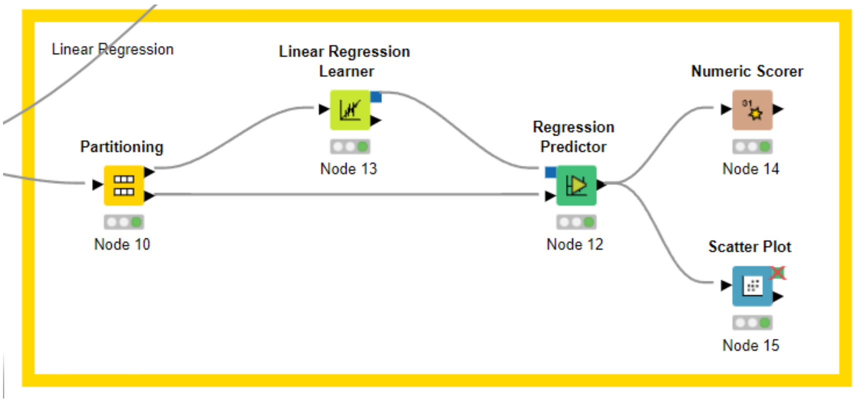
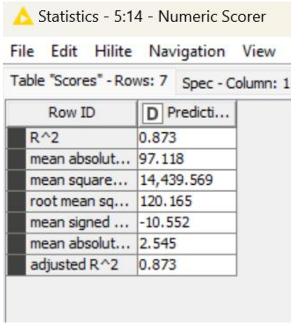


Figure 5: Linear Regression Workflow

Answering the questions of the tasks:

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]**

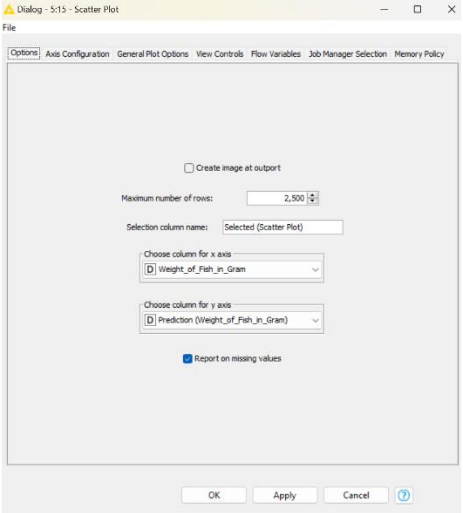
    1) What is the R^2 value of your test result? **[5 marks]**
    R^2 = 0.873.


Statistics - 5:14 - Numeric Scorer
File Edit Hilite Navigation View
Table "Scores" - Rows: 7
Spec - Column: 1

Row ID	D Predicti...
R^2	0.873
mean absolut...	97.118
mean square...	14,439.569
root mean sq...	120.165
mean signed ...	-10.552
mean absolut...	2.545
adjusted R^2	0.873

Figure 6: Numeric Scorer

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]**



□ Create image at outport

Maximum number of rows: 2,500

Selection column name: Selected (Scatter Plot)

Choose column for x axis: D Weight_of_Fish_in_Gram

Choose column for y axis: D Prediction (Weight_of_Fish_in_Gram)

☑ Report on missing values

Figure 7: Linear Regression Scatter Plot configuration

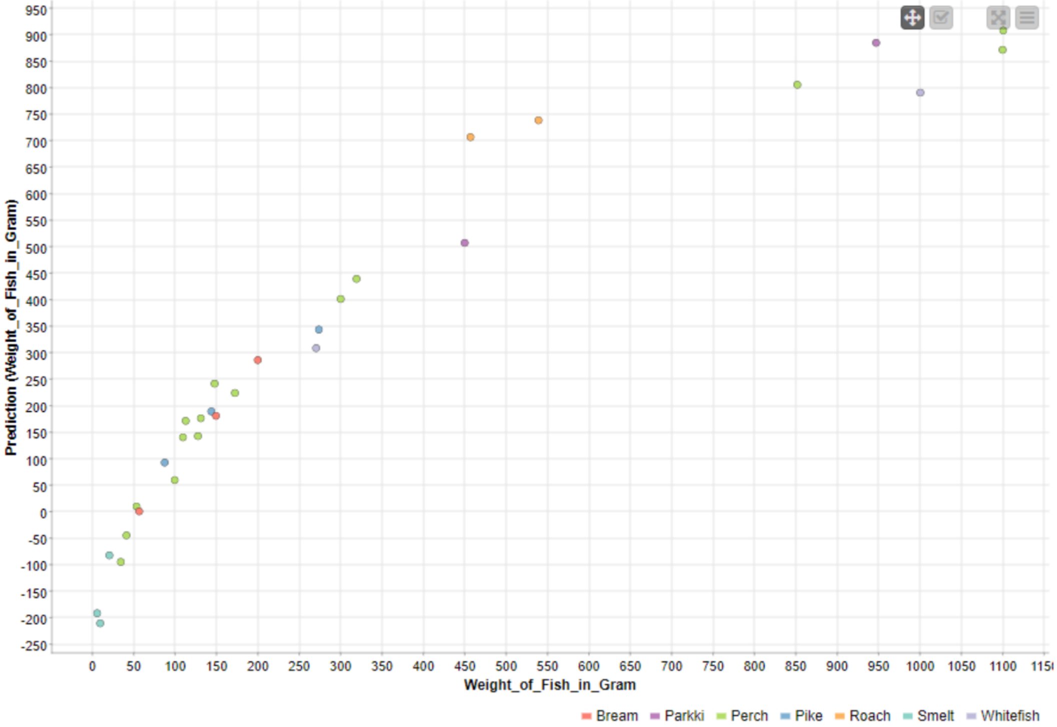


Figure 8: Scatter Plot (Actual vs. Prediction of weight of fish in gram)

3) Which species has the heaviest predicted weight in your test result? **[5 marks]**
The species has the heaviest predicted weight in my test result is “Perch”.

4) How many prediction results are infeasible in your test result? **[5 marks]**
There are 5 infeasible prediction results in my test result, because weight cannot be negative.

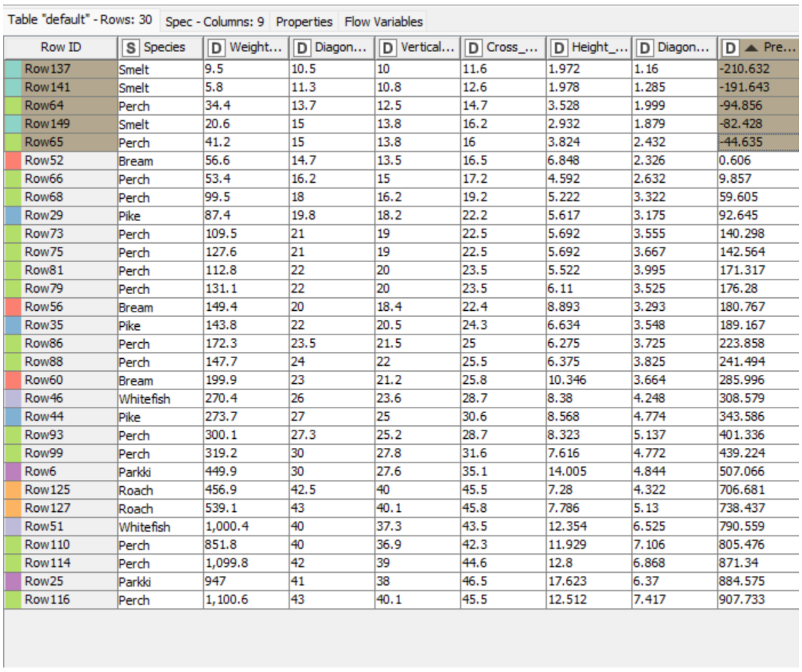


Table "default" - Rows: 30 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	S Species	D Weight...	D Diagon...	D Vertical...	D Cross_...	D Height_...	D Diagon...	D ▲ Pre...
Row137	Smelt	9.5	10.5	10	11.6	1.972	1.16	-210.632
Row141	Smelt	5.8	11.3	10.8	12.6	1.978	1.285	-191.643
Row64	Perch	34.4	13.7	12.5	14.7	3.528	1.999	-94.856
Row149	Smelt	20.6	15	13.8	16.2	2.932	1.879	-82.428
Row65	Perch	41.2	15	13.8	16	3.824	2.432	-44.635
Row52	Bream	56.6	14.7	13.5	16.5	6.848	2.326	0.606
Row66	Perch	53.4	16.2	15	17.2	4.592	2.632	9.857
Row68	Perch	99.5	18	16.2	19.2	5.222	3.322	59.605
Row29	Pike	87.4	19.8	18.2	22.2	5.617	3.175	92.645
Row73	Perch	109.5	21	19	22.5	5.692	3.555	140.298
Row75	Perch	127.6	21	19	22.5	5.692	3.667	142.564
Row81	Perch	112.8	22	20	23.5	5.522	3.995	171.317
Row79	Perch	131.1	22	20	23.5	6.11	3.525	176.28
Row56	Bream	149.4	20	18.4	22.4	8.893	3.293	180.767
Row35	Pike	143.8	22	20.5	24.3	6.634	3.548	189.167
Row86	Perch	172.3	23.5	21.5	25	6.275	3.725	223.858
Row88	Perch	147.7	24	22	25.5	6.375	3.825	241.494
Row60	Bream	199.9	23	21.2	25.8	10.346	3.664	285.996
Row46	Whitefish	270.4	26	23.6	28.7	8.38	4.248	308.579
Row44	Pike	273.7	27	25	30.6	8.568	4.774	343.586
Row93	Perch	300.1	27.3	25.2	28.7	8.323	5.137	401.336
Row99	Perch	319.2	30	27.8	31.6	7.616	4.772	439.224
Row6	Parkki	449.9	30	27.6	35.1	14.005	4.844	507.066
Row125	Roach	456.9	42.5	40	45.5	7.28	4.322	706.681
Row127	Roach	539.1	43	40.1	45.8	7.786	5.13	738.437
Row51	Whitefish	1,000.4	40	37.3	43.5	12.354	6.525	790.559
Row110	Perch	851.8	40	36.9	42.3	11.929	7.106	805.476
Row114	Perch	1,099.8	42	39	44.6	12.8	6.868	871.34
Row25	Parkki	947	41	38	46.5	17.623	6.37	884.575
Row116	Perch	1,100.6	43	40.1	45.5	12.512	7.417	907.733

Figure 9: Linear Regression Predictor

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]**
Species that can easily be separated from others are “Smelt”, “Roach”, “Bream”, and “Whitefish”.

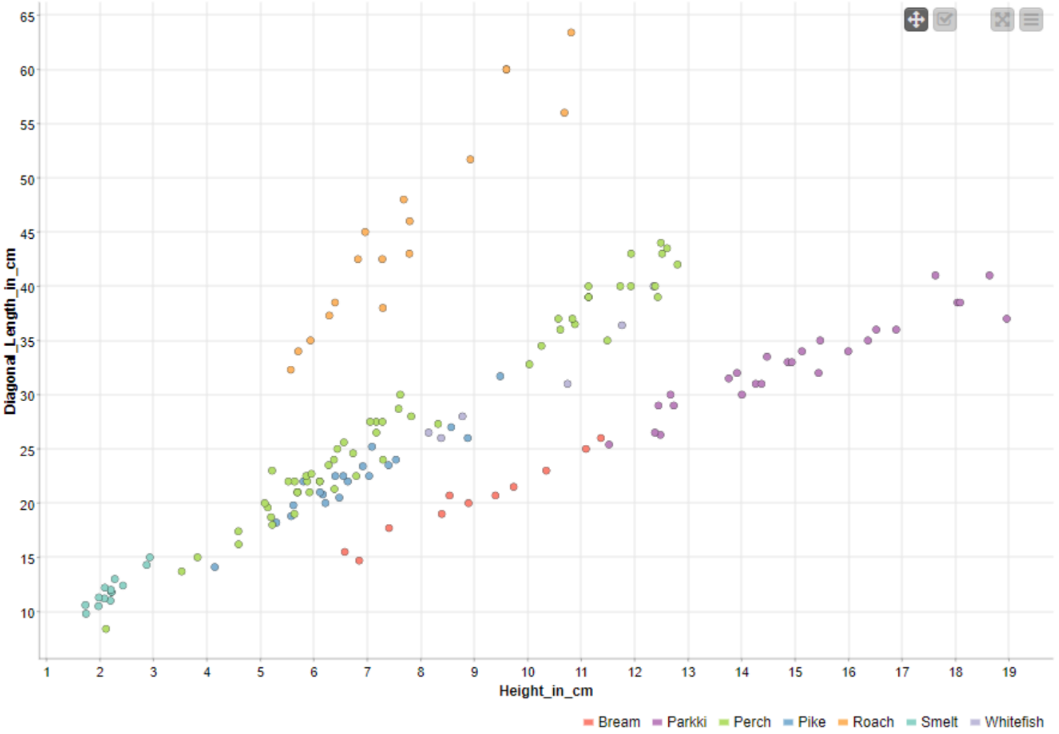
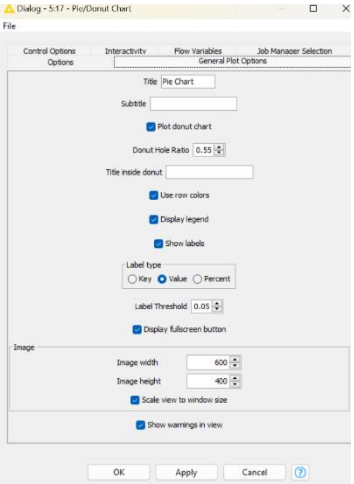


Figure 10: Scatter Plot

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]**



Title Pie Chart

Subtitle

☑ Plot donut chart

Donut Hole Ratio 0.55

Title inside donut

☑ Use row colors

☑ Display legend

☑ Show labels

Label type
○ Key ● Value ○ Percent

Label Threshold 0.05

☑ Display fullscreen button

Image
Image width 600
Image height 400
☑ Scale view to window size

☑ Show warnings in view

Figure 11: Pie Chart configuration

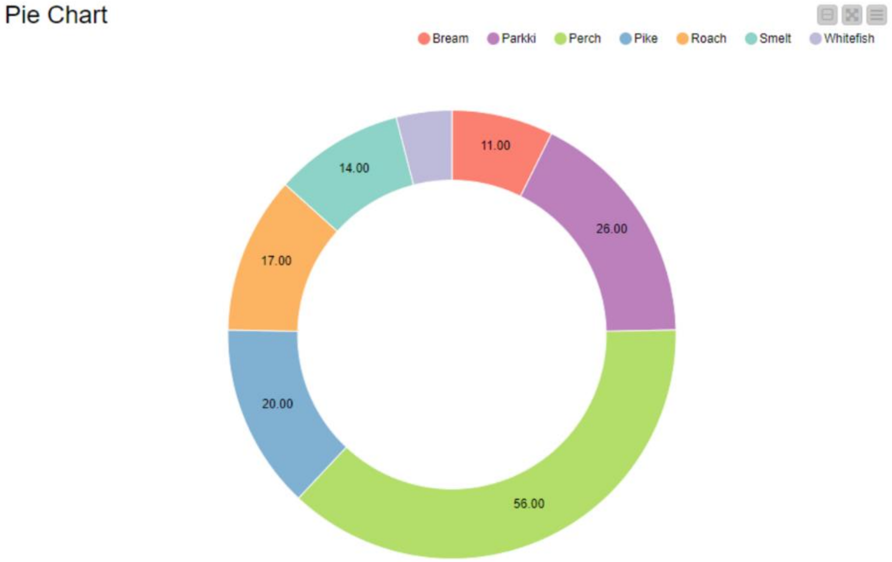
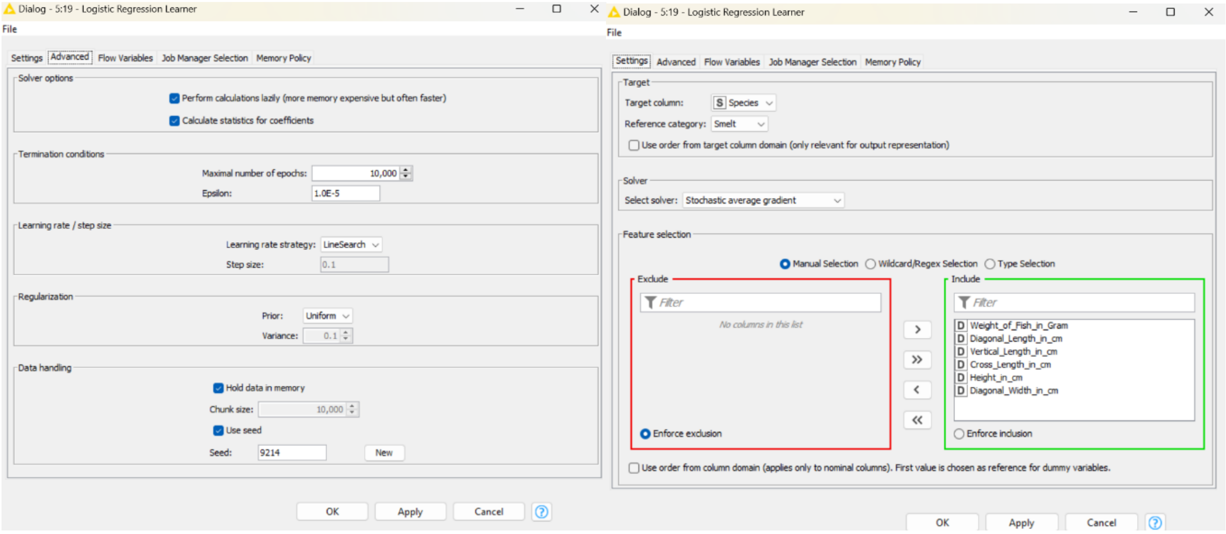


Figure 12: Pie Chart (distribution of the species)

3. Logistic Regression

The initial step to build a logistic regression model is to normalize the data, keep it in the range of 0.1 to 1.0, the reason is that it ensures all futures contribute to the model equally and enhances its accuracy. The next step is to partition the. After that, the training set data will go through the "Logistic regression learner" node, and in this node, configure "Smelt" as the reference species. Epochs and epsilons max point is set to 10000 and 0.00001 respectively. The next process is testing the model by using "Logistic Regression Predictor" node, which has already obtained the coefficients from the Learner, the data that goes through this process is the one from test set. To visualize the output, we can employ "Scorer".



Solver options

☑ Perform calculations lazily (more memory expensive but often faster)

☑ Calculate statistics for coefficients

Termination conditions

Maximal number of epochs: 10,000

Epsilon: 1.0E-5

Learning rate / step size

Learning rate strategy: LineSearch

Step size: 0.1

Regularization

Prior: Uniform

Variance: 0.1

Data handling

☑ Hold data in memory

Chunk size: 10,000

☑ Use seed

Seed: 9214 New

Target

Target column: S Species

Reference category: Smelt

□ Use order from target column domain (only relevant for output representation)

Solver

Select solver: Stochastic average gradient

Feature selection

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

Filter

No columns in this list

● Enforce exclusion

Include

Filter

D Weight_of_Fish_in_Gram
D Diagonal_Length_in_cm
D Vertical_Length_in_cm
D Cross_Length_in_cm
D Height_in_cm
D Diagonal_Width_in_cm

○ Enforce inclusion

□ Use order from column domain (applies only to nominal columns). First value is chosen as reference for dummy variables.

Figure 13: Logistic Regression Learner configuration

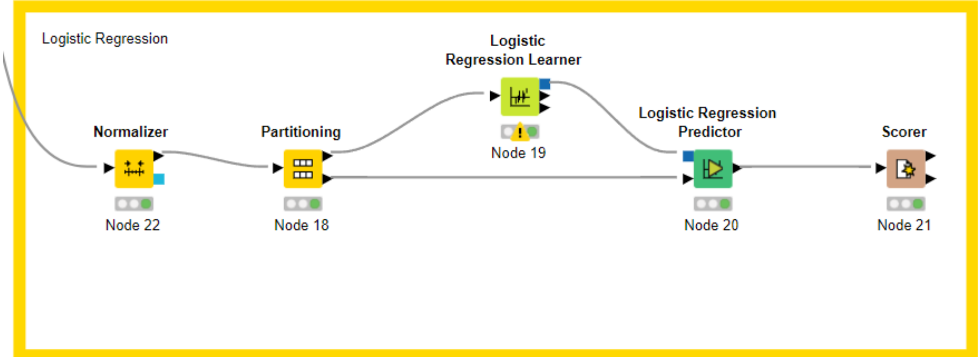
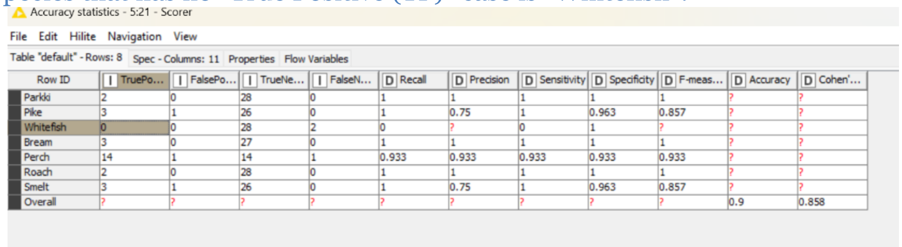


Figure 14: Logistic Regression Model

Answering the questions of the tasks:

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]**

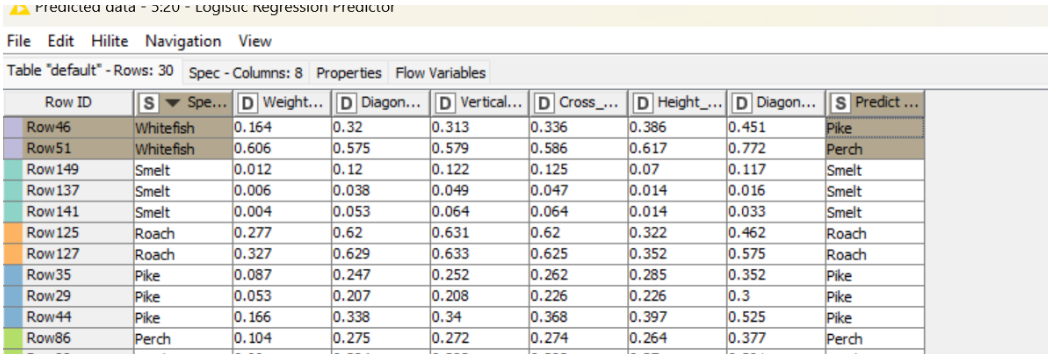
   1) Which species have/has no "True Positive (TP)" case in the prediction result? **[5 marks]**
   The species that has no "True Positive (TP)" case is "Whitefish".



Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
Parkki	2	0	28	0	1	1	1	1	1	?	?
Pike	3	1	26	0	1	0.75	1	0.963	0.857	?	?
Whitefish	0	0	28	2	0	?	0	1	?	?	?
Bream	3	0	27	0	1	1	1	1	1	?	?
Perch	14	1	14	1	0.933	0.933	0.933	0.933	0.933	?	?
Roach	2	0	28	0	1	1	1	1	1	?	?
Smelt	3	1	26	0	1	0.75	1	0.963	0.857	?	?
Overall	?	?	?	?	?	?	?	?	?	0.9	0.858

Figure 15: Accuracy statistics in Scorer

   2) For the species with no TP case, which species will be misplaced? **[5 marks]**
   For the species with no TP case which is "Whitefish", "Pike" and "Perch" will be misplaced.



Predicted data - 5:20 - Logistic Regression Predictor

File Edit Hilite Navigation View
Table "default" - Rows: 30 Spec - Columns: 8 Properties Flow Variables

Row ID	S Spe...	D Weight...	D Diagon...	D Vertical...	D Cross_...	D Height_...	D Diagon...	S Predict ...
Row46	Whitefish	0.164	0.32	0.313	0.336	0.386	0.451	Pike
Row51	Whitefish	0.606	0.575	0.579	0.586	0.617	0.772	Perch
Row149	Smelt	0.012	0.12	0.122	0.125	0.07	0.117	Smelt
Row137	Smelt	0.006	0.038	0.049	0.047	0.014	0.016	Smelt
Row141	Smelt	0.004	0.053	0.064	0.064	0.014	0.033	Smelt
Row125	Roach	0.277	0.62	0.631	0.62	0.322	0.462	Roach
Row127	Roach	0.327	0.629	0.633	0.625	0.352	0.575	Roach
Row35	Pike	0.087	0.247	0.252	0.262	0.285	0.352	Pike
Row29	Pike	0.053	0.207	0.208	0.226	0.226	0.3	Pike
Row44	Pike	0.166	0.338	0.34	0.368	0.397	0.525	Pike
Row86	Perch	0.104	0.275	0.272	0.274	0.264	0.377	Perch

Figure 16: Predicted data

3) What is the overall accuracy of the prediction result? **[5 marks]**
The overall accuracy of the prediction result is 0.9 or 90%. (*Figure 15*)

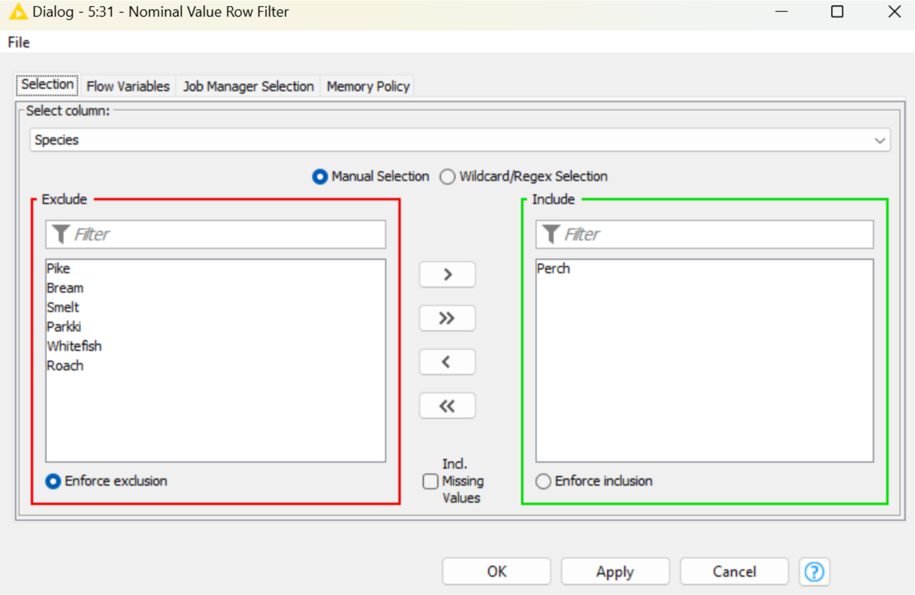
4) List all species names with 100% correctly classified test results. **[15 marks]**
The chance of species that has 100% of correctly classified test results means that its accuracy in classification is 100%, which moreover means that there are only correctly classified values (True Positives and True Negatives) and no incorrectly classified values (False Positives and False Negatives).
The formula of Accuracy rate goes: Accuracy rate = (TP + TN) / (TP + TN + FP + FN).
For its Accuracy to be 100% or (TP + TN) / (TP + TN + FP + FN) = 1, FP and FN (incorrectly classified values) should equal to 0, which makes the equation of (TP + TN) / (TP + TN) will always = 1.
Additionally, the Recall rate should be 1 so that FNR = 0.
With all the justification, the species name in this case should be: “Parkki”, “Bream”, “Roach”.

5) Which species has a 33.33% chance of being misplaced into another species in the test result? **[5 marks]**
To evaluate the chance of being misplaced into another species in the test result, we can use the False Negative Rate (FNR in short).
The fomula goes that FNR = FN / (FN + TP) = 33.33% or 1/3.
As can be seen from the *Figure 15*, there is no matching result.

6) In the test result, what percentage of the species “Perch” is misplaced into others? **[5 marks]**
To see the percentage of “Perch” is misplaced into others, we can also use the False Negative Rate (FNR).
FNR = 1 – Recall = 1 – 0.933 = 0.067% or 6.7% of “Perch” will be misplaced into other species.

4. Performance Improvement

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]**



Select column: Species

● Manual Selection ○ Wildcard/Regex Selection

Exclude: Pike, Bream, Smelt, Parkki, Whitefish, Roach

● Enforce exclusion

Include: Perch

○ Enforce inclusion

☐ Incl. Missing Values

Because this time we only focus on “Perch” so we can go to “Nominal Value Row Filter” and only includes “Perch”. The next step is to check on the “Scatter Matrix” for the performance of the attributes and how they contribute to the overall performance. As we use the “Linear Correlation” to check on the collinearity of the attributes, high collinearity suggests that the features are highly co-related to other variables and therefore, which worsen the performance.

Figure 17: Data Visualization

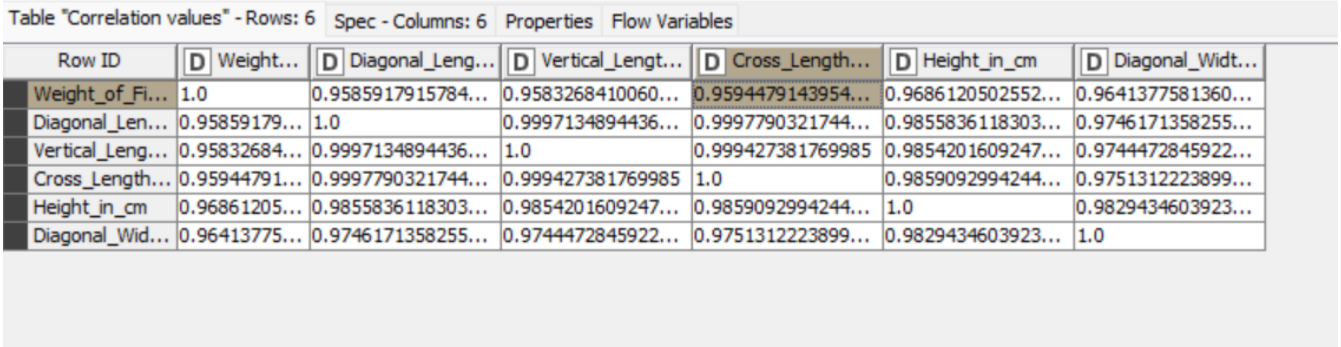
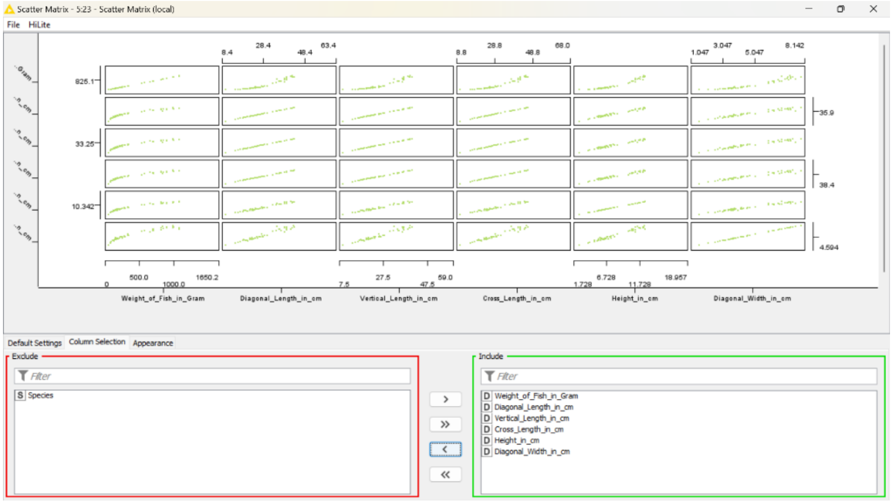


Table "Correlation values" - Rows: 6
Spec - Columns: 6
Properties
Flow Variables

Row ID	D Weight...	D Diagonal_Leng...	D Vertical_Lengt...	D Cross_Length...	D Height_in_cm	D Diagonal_Widt...
Weight_of_Fi...	1.0	0.9585917915784...	0.9583268410060...	0.9594479143954...	0.9686120502552...	0.9641377581360...
Diagonal_Len...	0.95859179...	1.0	0.9997134894436...	0.9997790321744...	0.9855836118303...	0.9746171358255...
Vertical_Leng...	0.95832684...	0.9997134894436...	1.0	0.999427381769985	0.9854201609247...	0.9744472845922...
Cross_Length...	0.95944791...	0.9997790321744...	0.999427381769985	1.0	0.9859092994244...	0.9751312223899...
Height_in_cm	0.96861205...	0.9855836118303...	0.9854201609247...	0.9859092994244...	1.0	0.9829434603923...
Diagonal_Wid...	0.96413775...	0.9746171358255...	0.9744472845922...	0.9751312223899...	0.9829434603923...	1.0

Figure 18: Correlation between attributes table



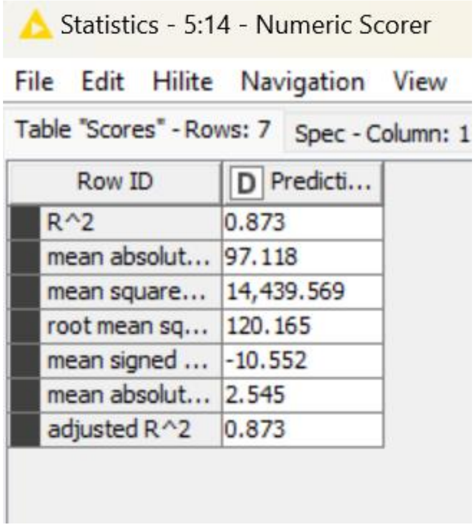
Exclude: S Species

Include: D Weight_of_Fish_in_Gram, D Diagonal_Length_in_cm, D Vertical_Length_in_cm, D Cross_Length_in_cm, D Height_in_cm, D Diagonal_Width_in_cm

Figure 19: Scatter Matrix

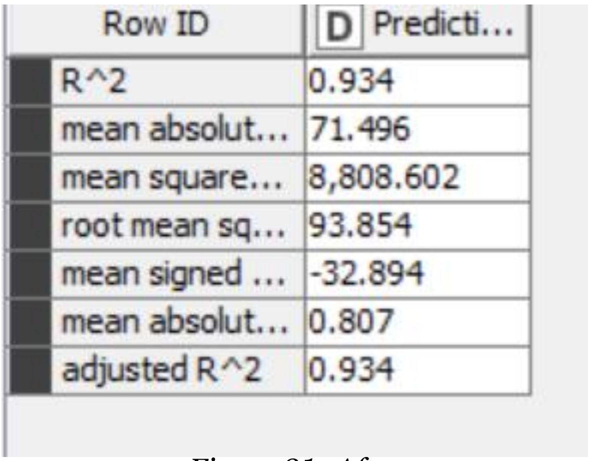
1) Give the reasons for each eliminated attribute and why they are not selected as the input. **[5 marks]**
With the given reason above, the eliminated attributes include “Diagonal_Width_in_cm” and “Vertical_Length_in_cm”.
“Diagonal_Width_in_cm”: High collinearity with “Weight_of_fish_in_gram”, “Heigh_in_cm” which will worsen the performance.
“Vertical_Length_in_cm”: High collinearity with “Diagonal_Length__in_cm”, “Cross_length_in_cm” which will worsen the performance.

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



Row ID	D Predicti...
R^2	0.873
mean absolut...	97.118
mean square...	14,439.569
root mean sq...	120.165
mean signed ...	-10.552
mean absolut...	2.545
adjusted R^2	0.873

Figure 20: Before



Row ID	D Predicti...
R^2	0.934
mean absolut...	71.496
mean square...	8,808.602
root mean sq...	93.854
mean signed ...	-32.894
mean absolut...	0.807
adjusted R^2	0.934

Figure 21: After

-