Cover sheet for submission of work for assessment

UNIT DETAILS

Unit name	Data Science Principles		Class day/time	Saturday	Office use only
Unit code	COS10022	Assignment no. 1	Due date	29/09/2024	
Name of lecturer/teacher	Dr. Minh Hoang				
Tutor/marker's name	Dr. Minh Hoang				Faculty or school date stamp

STUDENT(S)

Family Name(s)	Given Name(s)	Student ID Number(s)
Huynh Trung	Chien	104848770

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.

HUYNH TRUNG CHIEN Digitally signed by HUYNH TRUNG CHIEN
Date: 2024.09.29 23:22:05 +07'00'

Swinburne University of Technology Hawthorn Campus

Dept. of Computing Technologies

COS10022 Data Science Principles

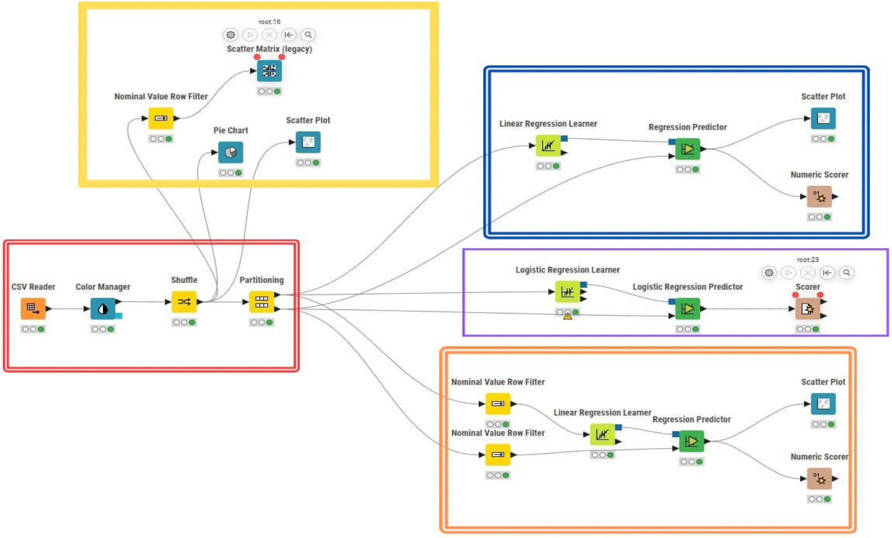
Assignment 1 - *Semester 1, 2024*

Assessment Title: Predictive Model Creation and Evaluation
Assessment Weighting: 20%
Due Date: Sunday, 24th 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]**



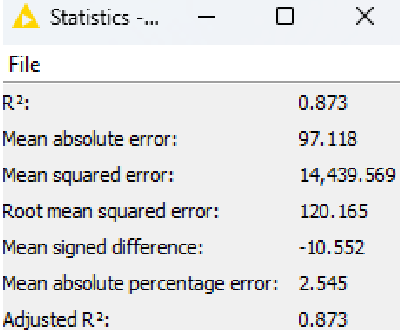
   2) How many tuples are included in the training set? **[2.5 marks]**
      - There are 120 tuples included in the training set.

   3) How many species are included in the test set? **[2.5 marks]**
      - The test set includes 7 species.

   4) Do species “Whitefish” and “Smelt” have the same number of tuples included in the test set?
      [2.5 marks]
      - “Whitefish” and “Smelt” do not have the same number of tuples – 2 for “Whitefish” and 3 for “Smelt”.

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]**

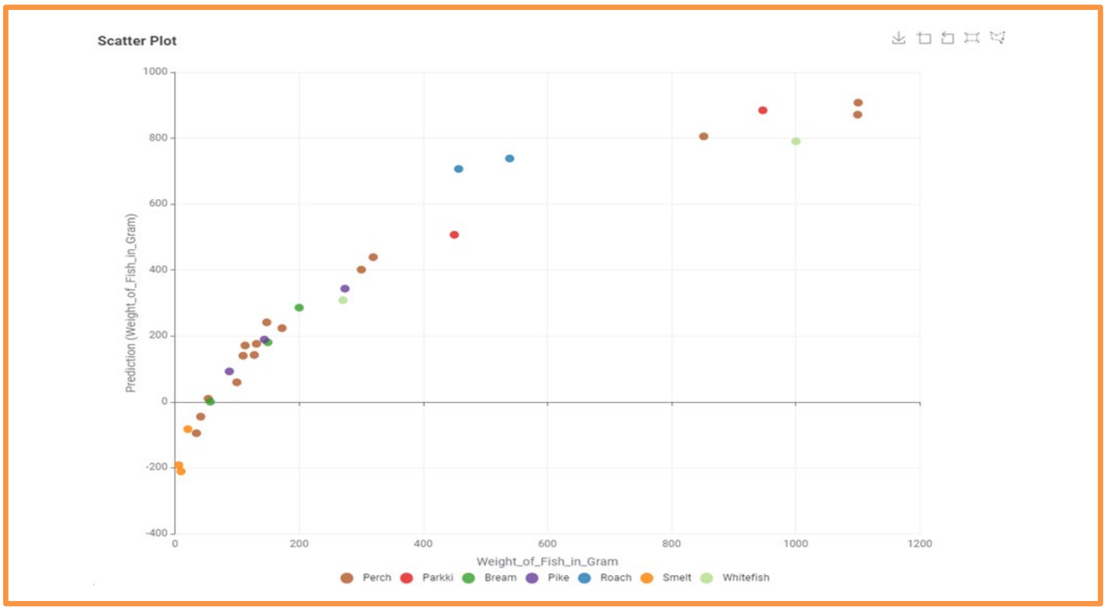
   1) What is the R^2 value of your test result? **[5 marks]**



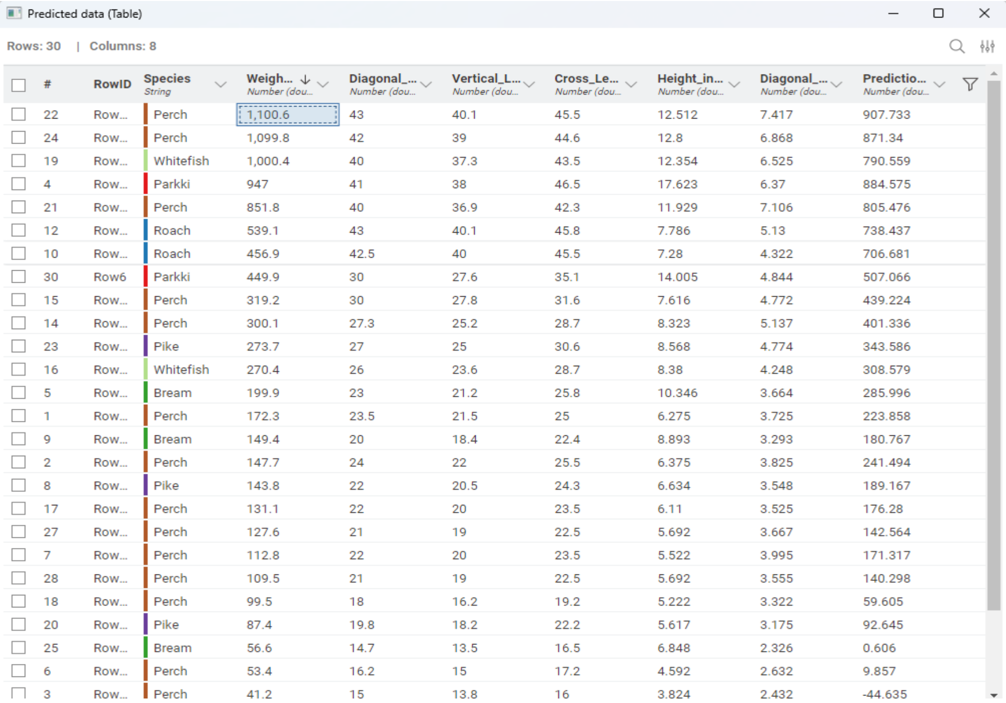
R²:	0.873
Mean absolute error:	97.118
Mean squared error:	14,439.569
Root mean squared error:	120.165
Mean signed difference:	-10.552
Mean absolute percentage error:	2.545
Adjusted R²:	0.873

- The value R2 of my test results is 0.873.

   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]**



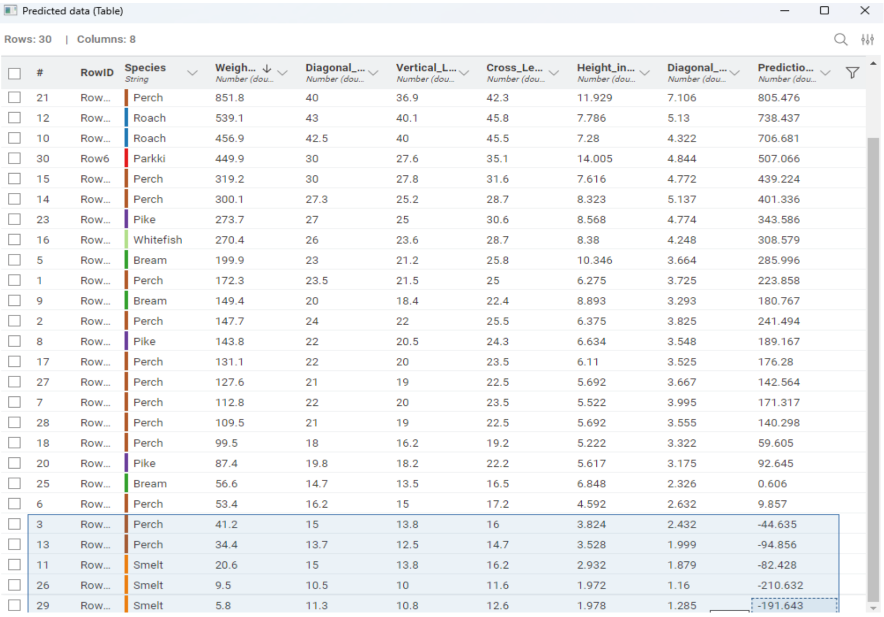
Predicted data (Table)

Rows: 30 | Columns: 8

□	#	RowID	Species String	Weigh... ↓ Number (dou...	Diagonal_... Number (dou...	Vertical_L... Number (dou...	Cross_Le... Number (dou...	Height_in... Number (dou...	Diagonal_... Number (dou...	Predictio... Number (dou...
□	22	Row...	Perch	1,100.6	43	40.1	45.5	12.512	7.417	907.733
□	24	Row...	Perch	1,099.8	42	39	44.6	12.8	6.868	871.34
□	19	Row...	Whitefish	1,000.4	40	37.3	43.5	12.354	6.525	790.559
□	4	Row...	Parkki	947	41	38	46.5	17.623	6.37	884.575
□	21	Row...	Perch	851.8	40	36.9	42.3	11.929	7.106	805.476
□	12	Row...	Roach	539.1	43	40.1	45.8	7.786	5.13	738.437
□	10	Row...	Roach	456.9	42.5	40	45.5	7.28	4.322	706.681
□	30	Row6	Parkki	449.9	30	27.6	35.1	14.005	4.844	507.066
□	15	Row...	Perch	319.2	30	27.8	31.6	7.616	4.772	439.224
□	14	Row...	Perch	300.1	27.3	25.2	28.7	8.323	5.137	401.336
□	23	Row...	Pike	273.7	27	25	30.6	8.568	4.774	343.586
□	16	Row...	Whitefish	270.4	26	23.6	28.7	8.38	4.248	308.579
□	5	Row...	Bream	199.9	23	21.2	25.8	10.346	3.664	285.996
□	1	Row...	Perch	172.3	23.5	21.5	25	6.275	3.725	223.858
□	9	Row...	Bream	149.4	20	18.4	22.4	8.893	3.293	180.767
□	2	Row...	Perch	147.7	24	22	25.5	6.375	3.825	241.494
□	8	Row...	Pike	143.8	22	20.5	24.3	6.634	3.548	189.167
□	17	Row...	Perch	131.1	22	20	23.5	6.11	3.525	176.28
□	27	Row...	Perch	127.6	21	19	22.5	5.692	3.667	142.564
□	7	Row...	Perch	112.8	22	20	23.5	5.522	3.995	171.317
□	28	Row...	Perch	109.5	21	19	22.5	5.692	3.555	140.298
□	18	Row...	Perch	99.5	18	16.2	19.2	5.222	3.322	59.605
□	20	Row...	Pike	87.4	19.8	18.2	22.2	5.617	3.175	92.645
□	25	Row...	Bream	56.6	14.7	13.5	16.5	6.848	2.326	0.606
□	6	Row...	Perch	53.4	16.2	15	17.2	4.592	2.632	9.857
□	3	Row...	Perch	41.2	15	13.8	16	3.824	2.432	-44.635

- Based on the predicted data table, “Perch” is the species that has the heaviest predicted weight in my test.

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



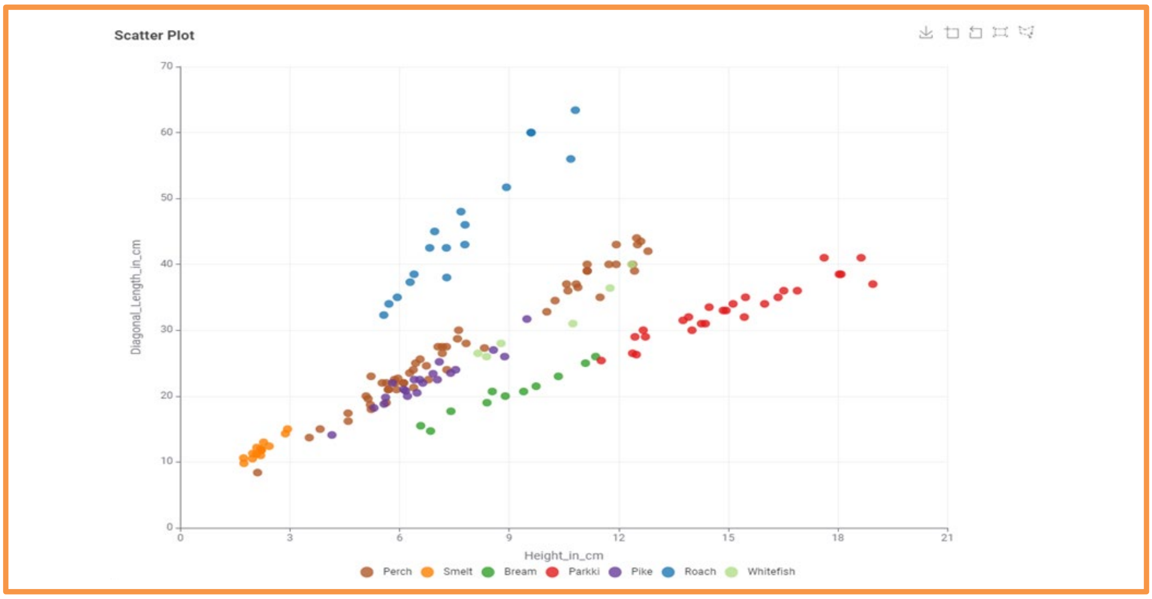
Predicted data (Table)

Rows: 30 | Columns: 8

□	#	RowID	Species String	Weigh... ↓ Number (dou...	Diagonal_... Number (dou...	Vertical_L... Number (dou...	Cross_Le... Number (dou...	Height_in... Number (dou...	Diagonal_... Number (dou...	Predictio... Number (dou...
□	21	Row...	Perch	851.8	40	36.9	42.3	11.929	7.106	805.476
□	12	Row...	Roach	539.1	43	40.1	45.8	7.786	5.13	738.437
□	10	Row...	Roach	456.9	42.5	40	45.5	7.28	4.322	706.681
□	30	Row6	Parkki	449.9	30	27.6	35.1	14.005	4.844	507.066
□	15	Row...	Perch	319.2	30	27.8	31.6	7.616	4.772	439.224
□	14	Row...	Perch	300.1	27.3	25.2	28.7	8.323	5.137	401.336
□	23	Row...	Pike	273.7	27	25	30.6	8.568	4.774	343.586
□	16	Row...	Whitefish	270.4	26	23.6	28.7	8.38	4.248	308.579
□	5	Row...	Bream	199.9	23	21.2	25.8	10.346	3.664	285.996
□	1	Row...	Perch	172.3	23.5	21.5	25	6.275	3.725	223.858
□	9	Row...	Bream	149.4	20	18.4	22.4	8.893	3.293	180.767
□	2	Row...	Perch	147.7	24	22	25.5	6.375	3.825	241.494
□	8	Row...	Pike	143.8	22	20.5	24.3	6.634	3.548	189.167
□	17	Row...	Perch	131.1	22	20	23.5	6.11	3.525	176.28
□	27	Row...	Perch	127.6	21	19	22.5	5.692	3.667	142.564
□	7	Row...	Perch	112.8	22	20	23.5	5.522	3.995	171.317
□	28	Row...	Perch	109.5	21	19	22.5	5.692	3.555	140.298
□	18	Row...	Perch	99.5	18	16.2	19.2	5.222	3.322	59.605
□	20	Row...	Pike	87.4	19.8	18.2	22.2	5.617	3.175	92.645
□	25	Row...	Bream	56.6	14.7	13.5	16.5	6.848	2.326	0.606
□	6	Row...	Perch	53.4	16.2	15	17.2	4.592	2.632	9.857
□	3	Row...	Perch	41.2	15	13.8	16	3.824	2.432	-44.635
□	13	Row...	Perch	34.4	13.7	12.5	14.7	3.528	1.999	-94.856
□	11	Row...	Smelt	20.6	15	13.8	16.2	2.932	1.879	-82.428
□	26	Row...	Smelt	9.5	10.5	10	11.6	1.972	1.16	-210.632
□	29	Row...	Smelt	5.8	11.3	10.8	12.6	1.978	1.285	-191.643

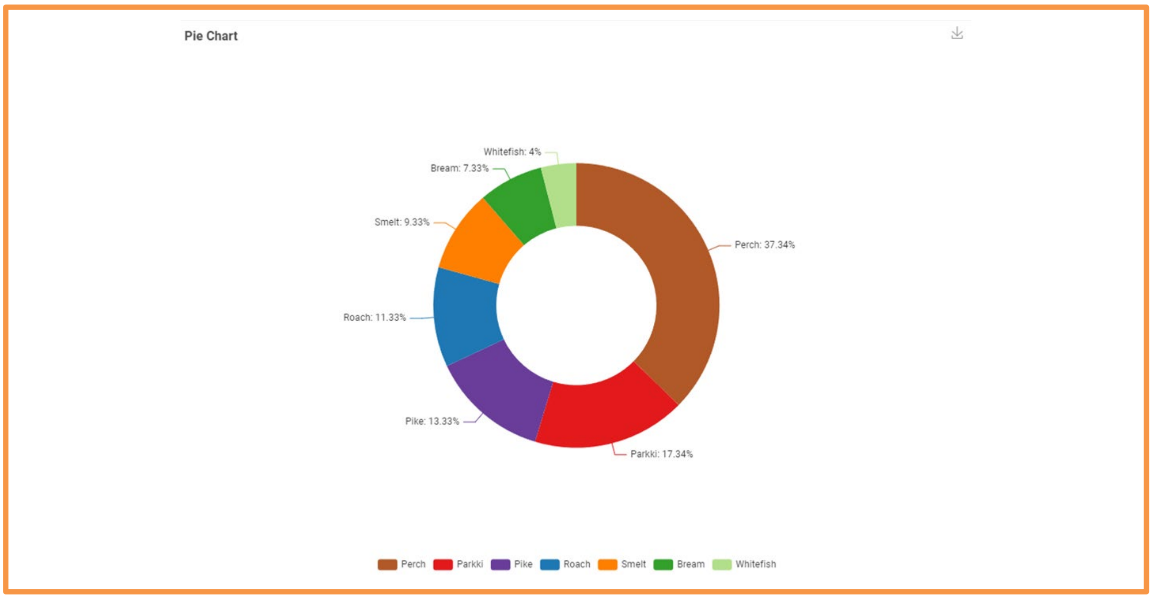
- Based on the predicted data table, there are 5 predictions which are infeasible in my test.

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]**



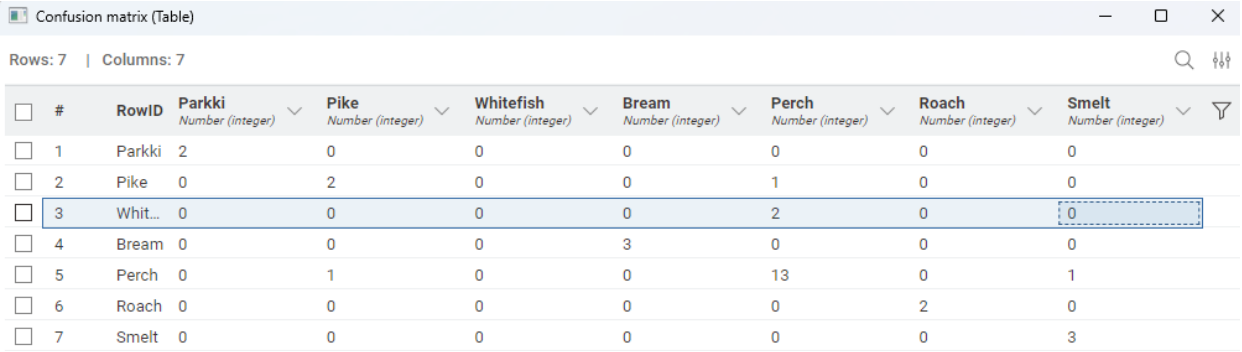
- When looking at my source data before splitting them, there are three species that can be easily separated from others if looking at the "Height_in_cm" and "Diagonal_Length_in_cm" attributes. Those three species are "Roach", "Parkki" and "Bream".

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]**

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



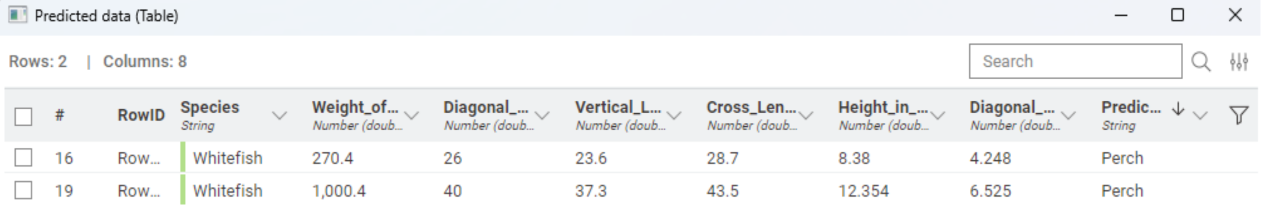
Confusion matrix (Table)

Rows: 7 | Columns: 7

□	#	RowID	Parkki *Number (integer)*	Pike *Number (integer)*	Whitefish *Number (integer)*	Bream *Number (integer)*	Perch *Number (integer)*	Roach *Number (integer)*	Smelt *Number (integer)*
□	1	Parkki	2	0	0	0	0	0	0
□	2	Pike	0	2	0	0	1	0	0
□	3	Whit...	0	0	0	0	2	0	0
□	4	Bream	0	0	0	3	0	0	0
□	5	Perch	0	1	0	0	13	0	1
□	6	Roach	0	0	0	0	0	2	0
□	7	Smelt	0	0	0	0	0	0	3

- By looking at the confusion matrix table, “Whitefish” is the only species that has no “True Positive (TP)” case in the prediction result.

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



Predicted data (Table)

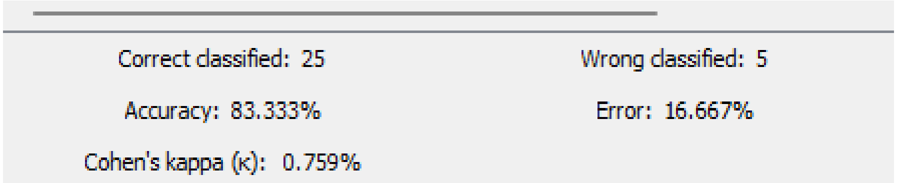
Rows: 2 | Columns: 8

Search

□	#	RowID	Species *String*	Weight_of... *Number (doub...*	Diagonal_... *Number (doub...*	Vertical_L... *Number (doub...*	Cross_Len... *Number (doub...*	Height_in_... *Number (doub...*	Diagonal_... *Number (doub...*	Predic... ↓ *String*
□	16	Row...	Whitefish	270.4	26	23.6	28.7	8.38	4.248	Perch
□	19	Row...	Whitefish	1,000.4	40	37.3	43.5	12.354	6.525	Perch

- For the “Whitefish” species, it has no TP case and it is misplaced into the “Perch” species.

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



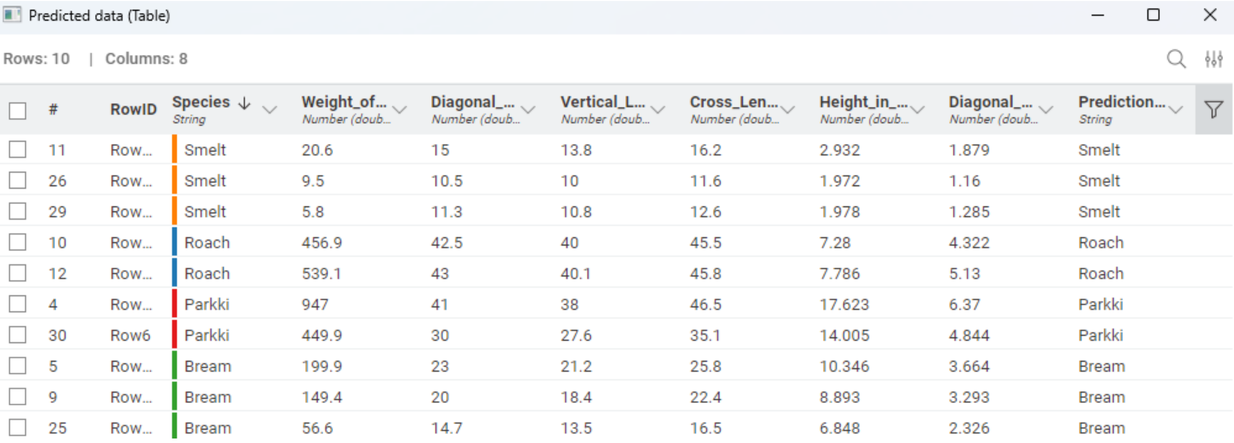
Correct classified: 25 Wrong classified: 5

Accuracy: 83.333% Error: 16.667%

Cohen's kappa (κ): 0.759%

- The overall accuracy of the prediction result is 83.33% (25 correct classified and 5 wrong classified).

4) List all species names with 100% correctly classified test results. **[15 marks]**



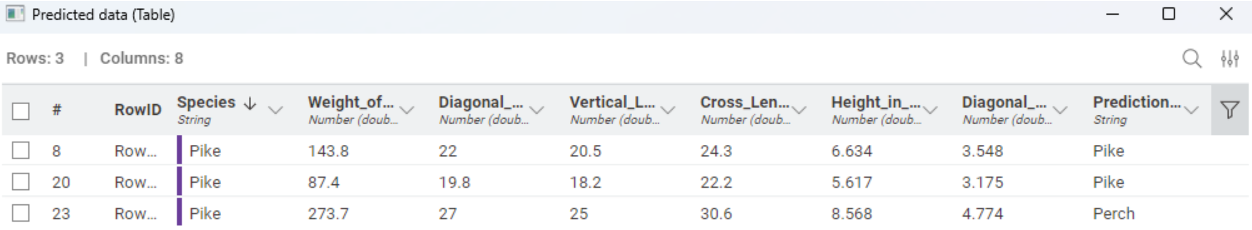
Predicted data (Table)

Rows: 10 | Columns: 8

□	#	RowID	Species ↓ *String*	Weight_of... *Number (doub...*	Diagonal_... *Number (doub...*	Vertical_L... *Number (doub...*	Cross_Len... *Number (doub...*	Height_in_... *Number (doub...*	Diagonal_... *Number (doub...*	Prediction... *String*
□	11	Row...	Smelt	20.6	15	13.8	16.2	2.932	1.879	Smelt
□	26	Row...	Smelt	9.5	10.5	10	11.6	1.972	1.16	Smelt
□	29	Row...	Smelt	5.8	11.3	10.8	12.6	1.978	1.285	Smelt
□	10	Row...	Roach	456.9	42.5	40	45.5	7.28	4.322	Roach
□	12	Row...	Roach	539.1	43	40.1	45.8	7.786	5.13	Roach
□	4	Row...	Parkki	947	41	38	46.5	17.623	6.37	Parkki
□	30	Row6	Parkki	449.9	30	27.6	35.1	14.005	4.844	Parkki
□	5	Row...	Bream	199.9	23	21.2	25.8	10.346	3.664	Bream
□	9	Row...	Bream	149.4	20	18.4	22.4	8.893	3.293	Bream
□	25	Row...	Bream	56.6	14.7	13.5	16.5	6.848	2.326	Bream

- There are 4 species with 100% correctly classified test results, which are “Roach”, “Bream”, “Parkki” and “Smelt”.

5) Which species has a 33.33% chance of being misplaced into another species in the test result? **[5 marks]**



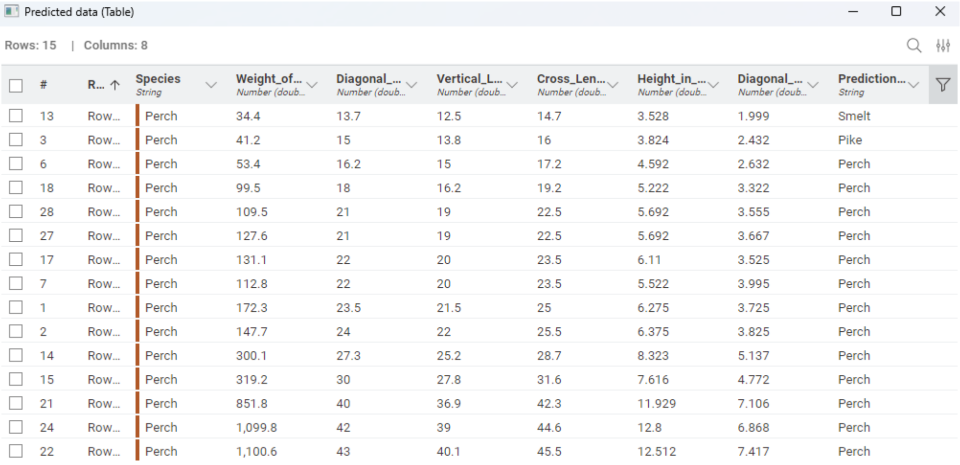
Predicted data (Table)

Rows: 3 | Columns: 8

□	#	RowID	Species String	Weight_of... Number (doub...	Diagonal_... Number (doub...	Vertical_L... Number (doub...	Cross_Len... Number (doub...	Height_in_... Number (doub...	Diagonal_... Number (doub...	Prediction... String
□	8	Row...	Pike	143.8	22	20.5	24.3	6.634	3.548	Pike
□	20	Row...	Pike	87.4	19.8	18.2	22.2	5.617	3.175	Pike
□	23	Row...	Pike	273.7	27	25	30.6	8.568	4.774	Perch

- “Pike” is the species that has a 33.33% of being misplaced into another species in the test result (1 out of 3).

6) In the test result, what percentage of the species “Perch” is misplaced into others? **[5 marks]**



Predicted data (Table)

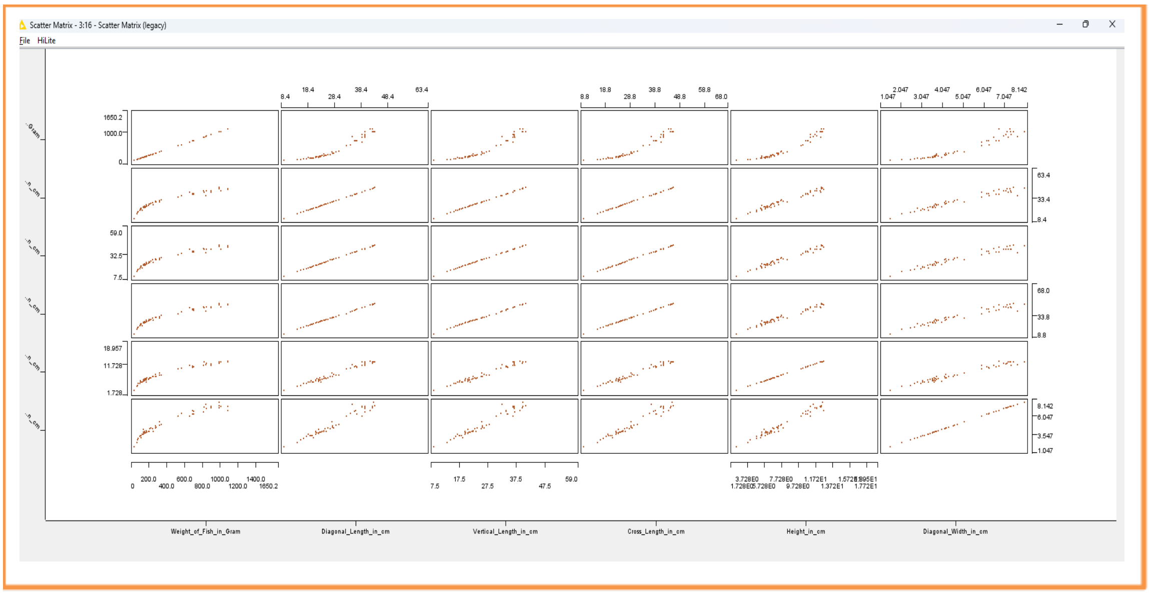
Rows: 15 | Columns: 8

□	#	R...	Species String	Weight_of... Number (doub...	Diagonal_... Number (doub...	Vertical_L... Number (doub...	Cross_Len... Number (doub...	Height_in_... Number (doub...	Diagonal_... Number (doub...	Prediction... String
□	13	Row...	Perch	34.4	13.7	12.5	14.7	3.528	1.999	Smelt
□	3	Row...	Perch	41.2	15	13.8	16	3.824	2.432	Pike
□	6	Row...	Perch	53.4	16.2	15	17.2	4.592	2.632	Perch
□	18	Row...	Perch	99.5	18	16.2	19.2	5.222	3.322	Perch
□	28	Row...	Perch	109.5	21	19	22.5	5.692	3.555	Perch
□	27	Row...	Perch	127.6	21	19	22.5	5.692	3.667	Perch
□	17	Row...	Perch	131.1	22	20	23.5	6.11	3.525	Perch
□	7	Row...	Perch	112.8	22	20	23.5	5.522	3.995	Perch
□	1	Row...	Perch	172.3	23.5	21.5	25	6.275	3.725	Perch
□	2	Row...	Perch	147.7	24	22	25.5	6.375	3.825	Perch
□	14	Row...	Perch	300.1	27.3	25.2	28.7	8.323	5.137	Perch
□	15	Row...	Perch	319.2	30	27.8	31.6	7.616	4.772	Perch
□	21	Row...	Perch	851.8	40	36.9	42.3	11.929	7.106	Perch
□	24	Row...	Perch	1,099.8	42	39	44.6	12.8	6.868	Perch
□	22	Row...	Perch	1,100.6	43	40.1	45.5	12.512	7.417	Perch

- In the test result, there is 13.33% of the species “Pearch” being misplaced into others (2 out of 15).

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]**

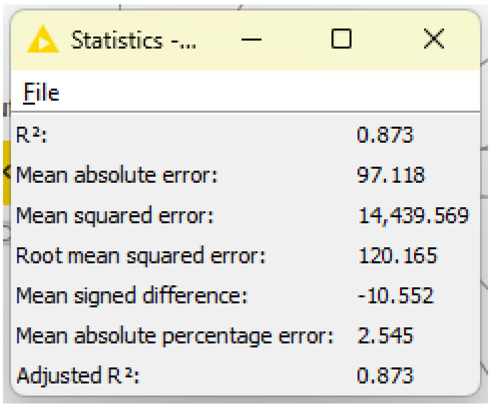


- In order to select the 3 attributes to train the linear regression model to predict the target "Weight_of_Fish_in_Gram", we need to consider 2 factors:
  + High Correlation with the "Weight_of_Fish_in_Gram": we need to choose the attributes that have a high correlation with the Target.
  + Avoid Collinearity: we also want to avoid collinearity, which is the high correlation between the attributes.

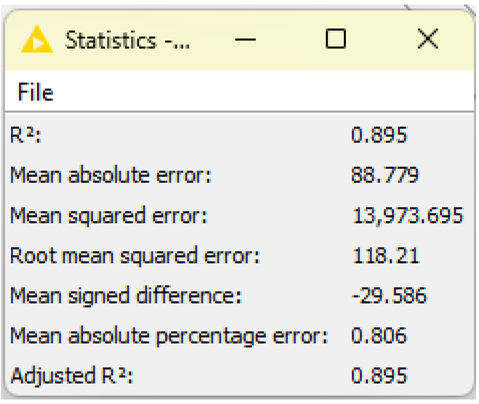
- From the output of the "Scatter Matrix (legacy)":
  + In terms of the "Diagonal_Width_in_cm", we can see that the correlation between it with the Target is slightly higher than the others, and this attribute does not have a high collinearity with the others as well.
  + For the "Height_in_cm", it also has a high correlation with the Target, and it has a moderate collinearity with "Diagonal_Width_in_cm" and the other length measures.
  + Considering the "Diagonal_Length_in_cm", "Vertical_Length_in_cm", and the "Cross_Length_in_cm", they have strong correlations with the Target, but they have a very high collinearity with each other.

⇨ In conclusion, by analyzing each attributes from the output of the "Scatter Matrix (legacy)", we can select the "Diagonal_Width_in_cm", "Height_in_cm" attributes for our model. The third attribute, we can choose any out of the three "Diagonal_Length_in_cm", "Vertical_Length_in_cm", "Cross_Length_in_cm" attributes because they have high collinearity, which means there is insignificant difference in the output of the model. Therefore, I am going to choose 3 attributes: **"Diagonal_Width_in_cm"**, **"Height_in_cm"**, **"Vertical_Length_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]**



```
R²:                             0.873
Mean absolute error:            97.118
Mean squared error:             14,439.569
Root mean squared error:        120.165
Mean signed difference:         -10.552
Mean absolute percentage error: 2.545
Adjusted R²:                    0.873
```



```
R²:                             0.895
Mean absolute error:            88.779
Mean squared error:             13,973.695
Root mean squared error:        118.21
Mean signed difference:         -29.586
Mean absolute percentage error: 0.806
Adjusted R²:                    0.895
```

- From the after statistics, we can see that the R^2 is larger and the means are lower, which means that the model has been improved. Moreover, we also avoid collinearity in our model so that the model will not be distorted.