

Cover sheet for submission of work for assessment

UNIT DETAILS

Unit name	Data Science Principles			Class day/time	Monday	Office use only
Unit code	COS10022	Assignment no.	1	Due date	Sunday, 01 June At 23:59 (VN Time)	
Name of lecturer/teacher	Mr. Hoang Anh Minh					
Tutor/marker's name	.Mr. Hoang Anh Minh					Faculty or school date stamp

STUDENT(S)

	Family Name(s)	Given Name(s)	Student ID Number(s)
(1)	Truong	Ngoc Gia Hieu	105565520
(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)



Swinburne University of Technology Hawthorn Campus
Dept. of Computing Technologies

COS10022 Data Science Principles
Assignment 1 - Semester 1, 2025

Assessment Title: Predictive Model Creation and Evaluation

Assessment Weighting: 20%

Due Date: Sunday, 01 June at 23:59 (VN Time)

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: There are 120 tuples in the training set.
 - 3) How many species are included in the test set? **[2.5 marks]**
Ans: According to **Figure 1**, there are 30 tuples in the test set.

▲ Second partition (remaining rows) - 3:4 - Partitioning

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

Figure 1

▲ Occurrences - 3:26 - Value Counter

File Edit Hilite Navigation View

Table "default" - Rows: 7 Spec - Column: 1 P

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

Figure 2

- 4) Do species “Whitefish” and “Smelt” have the same number of tuples included in the test set? **[2.5 marks]**
Ans: No, because Whitefish” has two tuples while “Smelt” has three tuples in the test set
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: The R^2 value of your test result is 0.873 in the test result (**Figure 3**)

Table "Scores" - Rows: 7		Spec - Column: 1	Properties	Flow Variables
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 3

- 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]

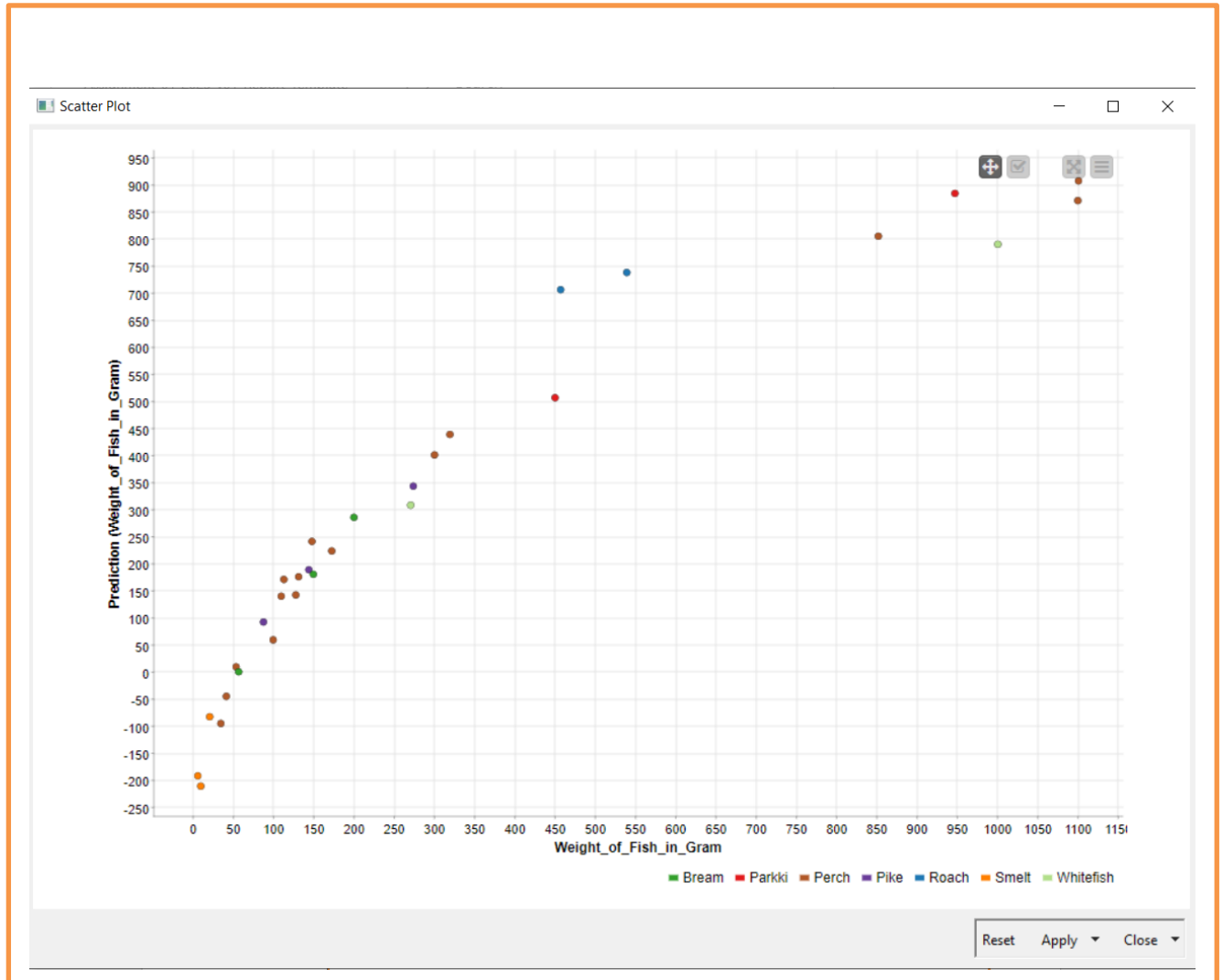


Figure 4

- 3) Which species has the heaviest predicted weight in your test result? [5 marks]
Ans: The heaviest predicted weight in my test result is “Perch”.
- 4) How many prediction results are infeasible in your test result? [5 marks]

Ans: According to **Figure 5**, there are 5 infeasible results in my test results because weight cannot be negative number.

▲ Input data and view selection - 3:10 - Scatter Plot

File Edit Hilite Navigation View

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 Predict...	B Selecte...
Row86	Perch	0.104	0.275	0.272	0.274	0.264	0.377	0.136	false
Row88	Perch	0.09	0.284	0.282	0.282	0.27	0.391	0.146	false
Row65	Perch	0.025	0.12	0.122	0.122	0.122	0.195	-0.027	false
Row25	Parkki	0.574	0.593	0.592	0.637	0.923	0.75	0.536	false
Row60	Bream	0.121	0.265	0.266	0.287	0.5	0.369	0.173	false
Row66	Perch	0.032	0.142	0.146	0.142	0.166	0.223	0.006	false
Row81	Perch	0.068	0.247	0.243	0.248	0.22	0.415	0.104	false
Row35	Pike	0.087	0.247	0.252	0.262	0.285	0.352	0.115	false
Row56	Bream	0.091	0.211	0.212	0.23	0.416	0.316	0.11	false
Row125	Roach	0.277	0.62	0.631	0.62	0.322	0.462	0.428	false
Row149	Smelt	0.012	0.12	0.122	0.125	0.07	0.117	-0.05	false
Row127	Roach	0.327	0.629	0.633	0.625	0.352	0.575	0.447	false
Row64	Perch	0.021	0.096	0.097	0.1	0.104	0.134	-0.057	false
Row93	Perch	0.182	0.344	0.344	0.336	0.383	0.576	0.243	false
Row99	Perch	0.193	0.393	0.394	0.385	0.342	0.525	0.266	false
Row46	Whitefish	0.164	0.32	0.313	0.336	0.386	0.451	0.187	false
Row79	Perch	0.079	0.247	0.243	0.248	0.254	0.349	0.107	false
Row68	Perch	0.06	0.175	0.169	0.176	0.203	0.321	0.036	false
Row51	Whitefish	0.606	0.575	0.579	0.586	0.617	0.772	0.479	false
Row29	Pike	0.053	0.207	0.208	0.226	0.226	0.3	0.056	false
Row110	Perch	0.516	0.575	0.571	0.566	0.592	0.854	0.488	false
Row116	Perch	0.667	0.629	0.633	0.62	0.626	0.898	0.55	false
Row44	Pike	0.166	0.338	0.34	0.368	0.397	0.525	0.208	false
Row114	Perch	0.666	0.611	0.612	0.605	0.643	0.82	0.528	false
Row52	Bream	0.034	0.115	0.117	0.13	0.297	0.18	0	false
Row137	Smelt	0.006	0.038	0.049	0.047	0.014	0.016	-0.128	false
Row75	Perch	0.077	0.229	0.223	0.231	0.23	0.369	0.086	false
Row73	Perch	0.066	0.229	0.223	0.231	0.23	0.353	0.085	false
Row141	Smelt	0.004	0.053	0.064	0.064	0.014	0.033	-0.116	false
Row6	Parkki	0.273	0.393	0.39	0.444	0.713	0.535	0.307	false

Figure 5

- 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: Based on **Figure 6**, species include "Bream", "Smelt", "Roach", and "Parkki" can be easily separated from others if looking at the Height_in_cm and "Diagonal_Length_in_cm" attributes.

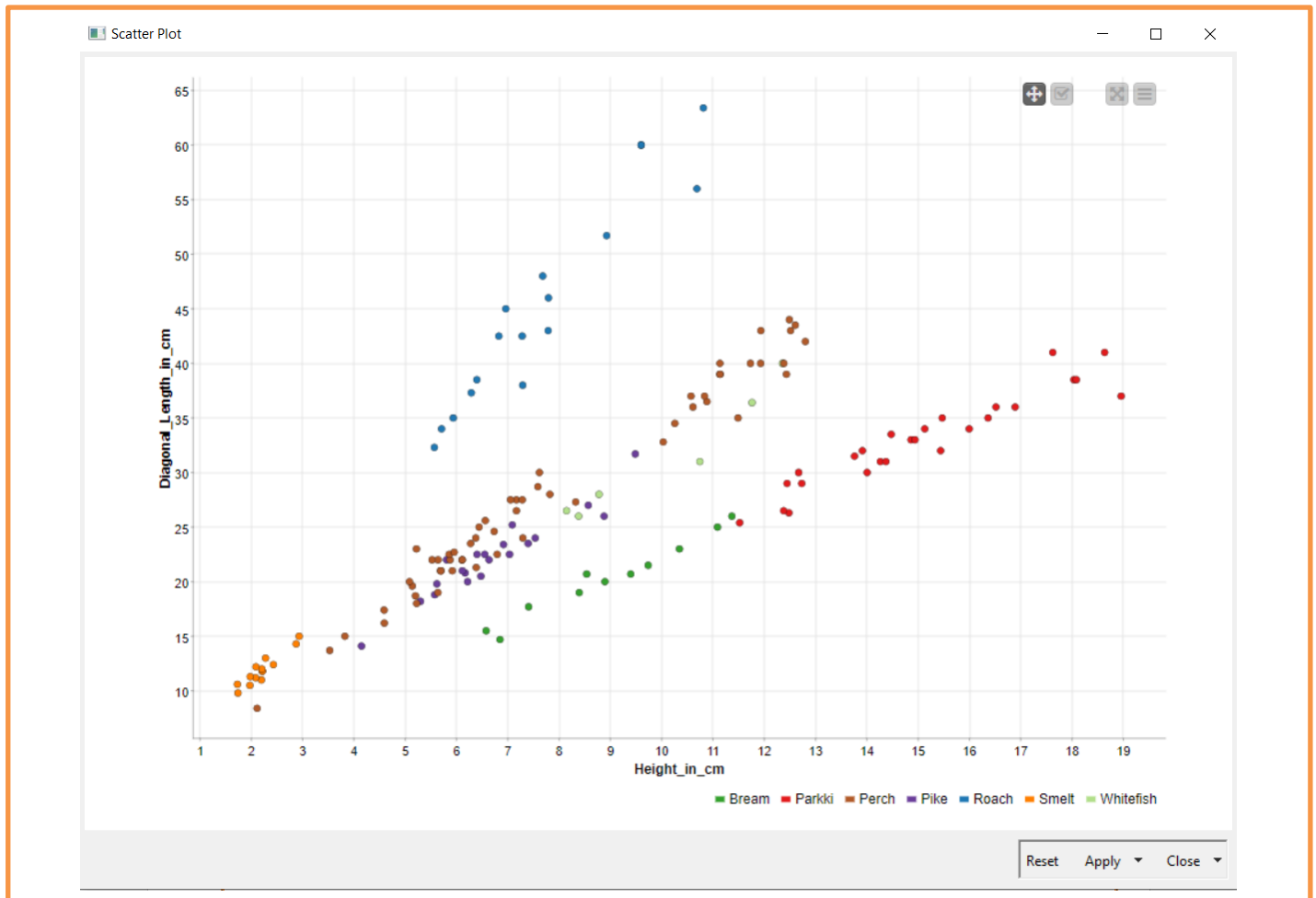


Figure 6

- 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]

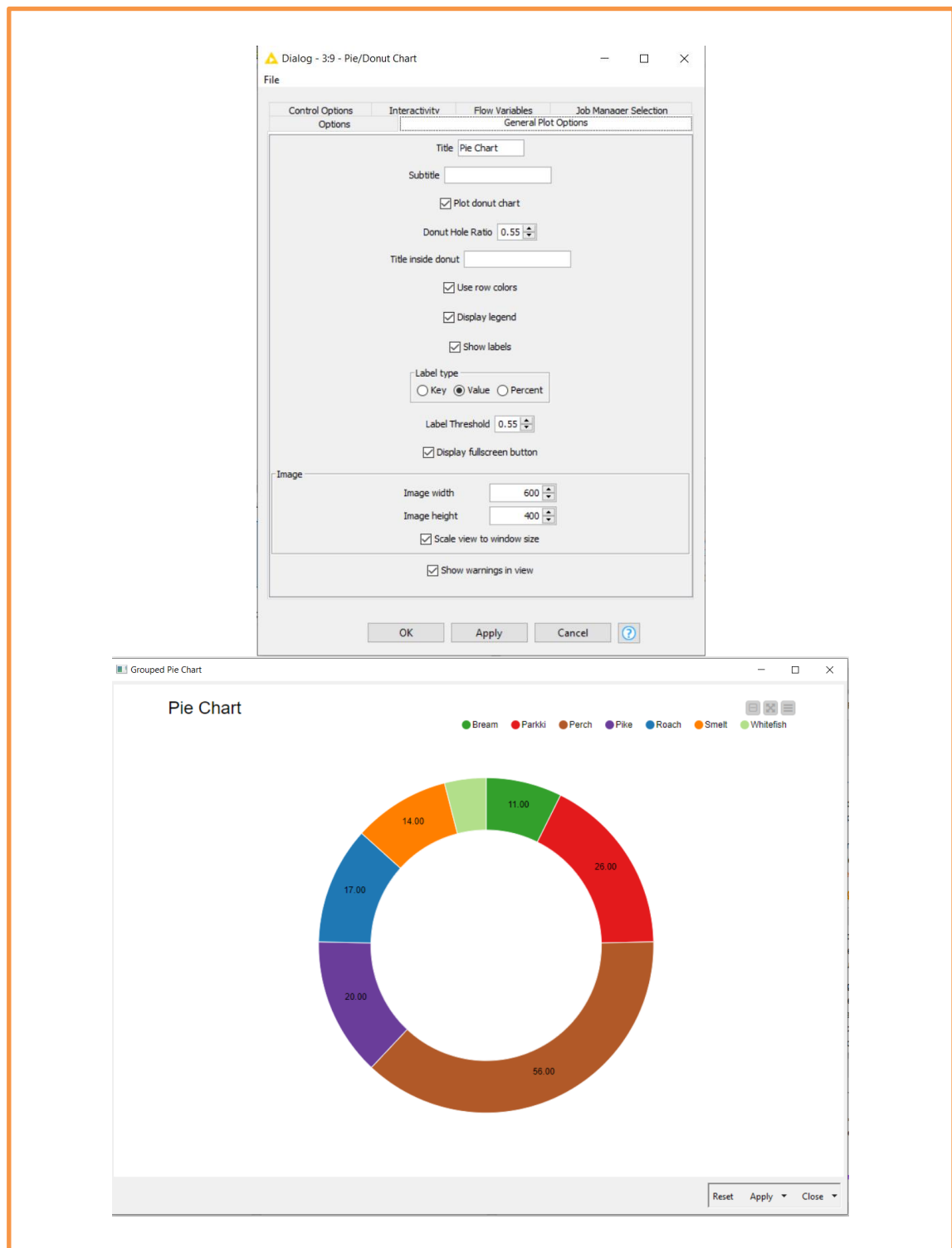


Figure 7

- 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]**

Dialog - 3:12 - Logistic Regression Learner

File

Settings Advanced Flow Variables Job Manager Selection Memory Policy

Target

Target column: Species

Reference category:

Dialog - 3:12 - Logistic Regression Learner

File

Settings Advanced Flow Variables Job Manager Selection Memory Policy

Solver options

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

☒ Calculate statistics for coefficients

Termination conditions

Maximal number of epochs:

Epsilon:

Learning rate / step size

Learning rate strategy:

Step size:

Regularization

Prior:

Variance:

Data handling

☒ Hold data in memory

Chunk size:

☒ Use seed

Seed:

Figure 8

- 1) Which species have/has no "True Positive (TP)" case in the prediction result? **[5 marks]**
 Ans: Based on the prediction result, the "Whitefish" has no "True Positive" case. (TP).
- 2) For the species with no TP case, which species will be misplaced? **[5 marks]**

Ans: For the “Whitefish” with no TP case, “Pike” and “Perch” will ne misplaced.

▲ Predicted data - 3:13 - Logistic Regression Predictor

File Edit Hilite Navigation View

Table "default" - Rows: 30 Spec - Columns: 8 Properties Flow Variables

Row ID	S Species	D Weight...	D Diagon...	D Vertical...	D Cross_...	D Height_...	D Diagon...	S Predicti...
Row86	Perch	0.104	0.275	0.272	0.274	0.264	0.377	Perch
Row88	Perch	0.09	0.284	0.282	0.282	0.27	0.391	Perch
Row65	Perch	0.025	0.12	0.122	0.122	0.122	0.195	Perch
Row25	Parkki	0.574	0.593	0.592	0.637	0.923	0.75	Parkki
Row60	Bream	0.121	0.265	0.266	0.287	0.5	0.369	Bream
Row66	Perch	0.032	0.142	0.146	0.142	0.166	0.223	Perch
Row81	Perch	0.068	0.247	0.243	0.248	0.22	0.415	Perch
Row35	Pike	0.087	0.247	0.252	0.262	0.285	0.352	Pike
Row56	Bream	0.091	0.211	0.212	0.23	0.416	0.316	Bream
Row125	Roach	0.277	0.62	0.631	0.62	0.322	0.462	Roach
Row149	Smelt	0.012	0.12	0.122	0.125	0.07	0.117	Smelt
Row127	Roach	0.327	0.629	0.633	0.625	0.352	0.575	Roach
Row64	Perch	0.021	0.096	0.097	0.1	0.104	0.134	Smelt
Row93	Perch	0.182	0.344	0.344	0.336	0.383	0.576	Perch
Row99	Perch	0.193	0.393	0.394	0.385	0.342	0.525	Perch
Row46	Whitefish	0.164	0.32	0.313	0.336	0.386	0.451	Pike
Row79	Perch	0.079	0.247	0.243	0.248	0.254	0.349	Perch
Row68	Perch	0.06	0.175	0.169	0.176	0.203	0.321	Perch
Row51	Whitefish	0.606	0.575	0.579	0.586	0.617	0.772	Perch
Row29	Pike	0.053	0.207	0.208	0.226	0.226	0.3	Pike
Row110	Perch	0.516	0.575	0.571	0.566	0.592	0.854	Perch
Row116	Perch	0.667	0.629	0.633	0.62	0.626	0.898	Perch
Row44	Pike	0.166	0.338	0.34	0.368	0.397	0.525	Pike
Row114	Perch	0.666	0.611	0.612	0.605	0.643	0.82	Perch
Row52	Bream	0.034	0.115	0.117	0.13	0.297	0.18	Bream
Row137	Smelt	0.006	0.038	0.049	0.047	0.014	0.016	Smelt
Row75	Perch	0.077	0.229	0.223	0.231	0.23	0.369	Perch
Row73	Perch	0.066	0.229	0.223	0.231	0.23	0.353	Perch
Row141	Smelt	0.004	0.053	0.064	0.064	0.014	0.033	Smelt
Row6	Parkki	0.273	0.393	0.39	0.444	0.713	0.535	Parkki

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

Ans: The overall accuracy of the prediction result is 0.9 (90%)

▲ Accuracy statistics - 3:14 - Scorer

File Edit Hilite Navigation View

Table "default" - Rows: 8 Spec - Columns: 11 Properties Flow Variables

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 9

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

Ans: Species with 100% correctly classified test results including “Pike”, “Smelt”, “Parkki”, “Bream”, and “Roach” because their TrueNegatives is 0, based on **Figure 9**.

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

Ans: Based on **Figure 9**, none of the species have a 33.33% chance of being misplaced into another species in the test result

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

Ans: In the test result as **Figure 9**, the percentage of species “Perch” misplaces into others is 6.7% because we have the formula FNR (False Negative Rate) = $1 - \text{Sensitivity/Recall} = 1 - 0.933 = 0.067$

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]**

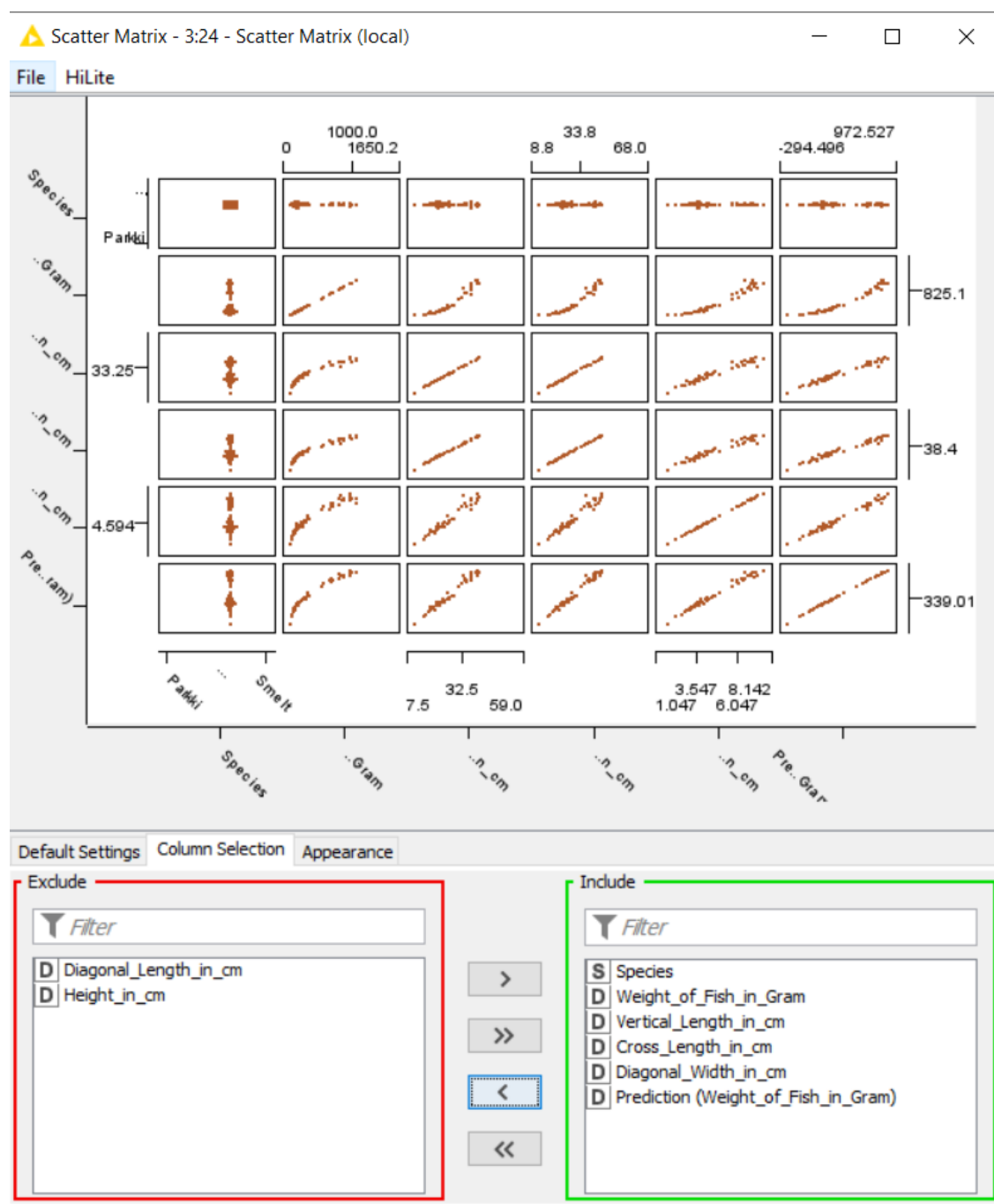


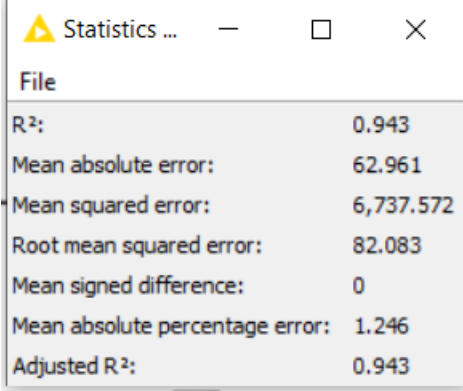
Figure 10

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

Ans: From my perspective, "Diagonal_Length_in_cm" and "Vertical_Length_in_cm" are not selected as the input because three remaining attributes including "Vertical_Length_in_cm", "Cross_Length_in_cm", and "Diagonal_Width_in_cm" demonstrate strong linear correlation while two exclude attributes do not create collinearity which reduces the regression model's statistical strength with each other.

- 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:



A screenshot of a software window titled 'Statistics ...'. The window contains a table of statistical metrics for a model. The metrics and their values are as follows:

File	
R ² :	0.943
Mean absolute error:	62.961
Mean squared error:	6,737.572
Root mean squared error:	82.083
Mean signed difference:	0
Mean absolute percentage error:	1.246
Adjusted R ² :	0.943

Figure 11

As the **Figure 11** mentioned above, a higher R^2 value showcases that the accuracy of the prediction results is improved. Additionally, the model accuracy is improved by 0.07, highlighting eliminates unsuitable attributes to reduce the dimension of the input data train the model. As a result, the model will be more accuracy.