Cover sheet for submission of work for assessment

SWINBURNE UNIVERSITY OF TECHNOLOGY

UNIT DETAILS

Unit name	Data Science Principles			Class day/time	Wednesday	Office use only
Unit code	COS10022	Assignment no.	1	Due date	12th February	
Name of lecturer/teacher	Dr Pham Kim Dung					
Tutor/marker's name	Dr Pham Kim Dung					Faculty or school date stamp

STUDENT(S)

	Family Name(s)	Given Name(s)	Student ID Number(s)
(1)	Luong	Trac Duc Anh	103488117
(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)	Trac Duc Anh Luong	(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, 12th February 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.

This report is a comprehensive view of assignment 1, with each step explained with detailed answers, screenshots, and explanations. The purpose of this report is to take a deeper look at model building, testing, and evaluation. The nodes used in this assignment are CSV Reader, Shuffle, Colour Manager, Scatter Matrix, Partitioning, Linear Regression Learner, Regression Predictor, Scatter Plot, Scorer, Numeric Scorer, Logistic Regression Learner, Linear Regression Learner, and Row Filter. The report on this assignment is written on the next page.

Question 1:

1.

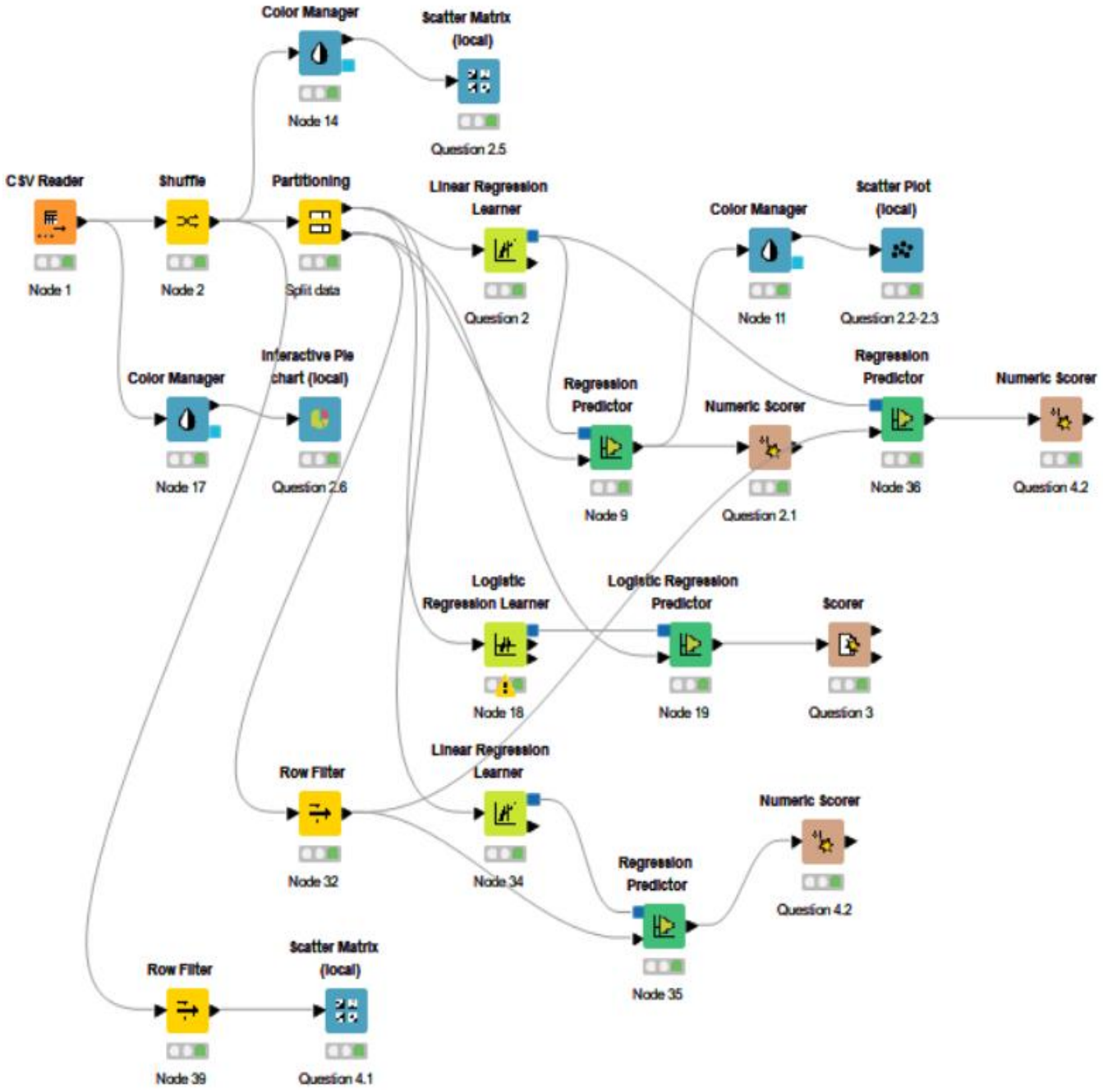

Figure 1: KNIME workflow

2. There are **120** tuples in the training set, as the requirement tells us to allocate 80% of the 150 original tuples for training. 150 * 80% = 120 tuples. For this step, we view the first partition in the Partitioning node.

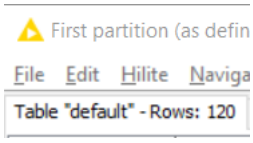

Figure 2: Number of tuples in the training set

3. There are **7** species in the test dataset, which is the same as the training dataset. For this step, we look at the possible values of Species in the second partition in the Partitioning node.

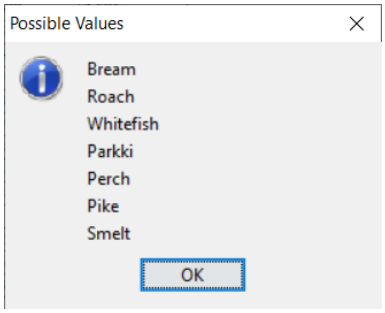

Figure 3: Number of species in the test set

4. Whitefish and Smelt both have 2 tuples in the test dataset, therefore they have the same number of tuples. For this step, we view the Species column in the Sort Ascending order.

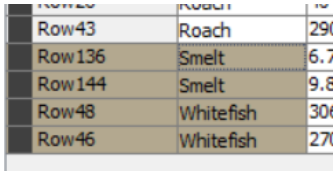

Figure 4: Number of Whitefish and Smelt

Question 2:

To predict a numerical value, we can use linear regression to analyse the value of a different variable. The related variable will be the one we want to predict while the unrelated variable will be used as the dataset to predict the value of the remaining variable.

1. In my test result, R^2 is equal to **0.857**. For this step, we look at the Statistics in the Numeric Scorer node.

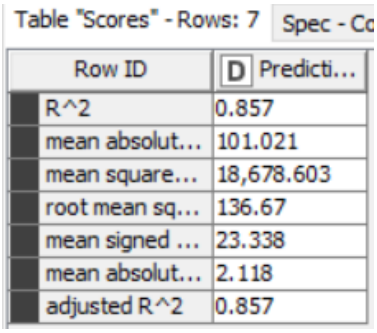

Figure 5: R^2 value is question 2

2. Here is my scatter plot table when the x-axis is Weight_of_Fish_in_Gram while the y-axis displays the prediction result. For this step we use the Scatter Plot node after the data output from the Regression Predictor has been labelled with the Colour Manager node.

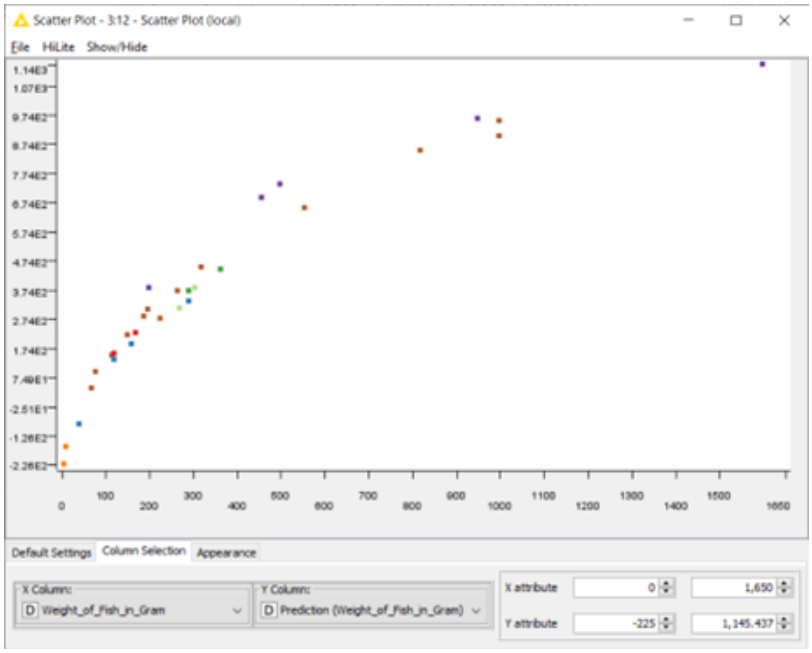

Figure 6: Scatter Plot

3. The species with the heaviest weight is Pike, with 1 result registered more than 1500 grams, which is the purple data point located in the top right corner of the chart.
4. There are 3 unrealistic results in my prediction test, which the result returned negative (Row 26, 136, 144). For this step, we view the Predicted data in the Regression Predictor node.

Row ID	Species	Weight...	Diagon...	Vertical...	Cross_...	Height_...	Diagon...	Prediction (Weight_of_Fish_in_Gram)
Row128	Pike	500	45	42	48	6.96	4.896	735.296
Row48	Whitefish	306	28	25.6	30.8	8.778	4.682	381.559
Row99	Perch	320	30	27.8	31.6	7.616	4.772	450.876
Row74	Perch	115	21	19	22.5	5.918	3.308	147.597
Row3	Bream	363	29	26.3	33.5	12.73	4.455	445.544
Row87	Perch	225	24	22	25.5	7.293	3.723	274.787
Row101	Perch	556	34.5	32	36.5	10.257	6.388	654.402
Row125	Pike	456	42.5	40	45.5	7.28	4.322	691.361
Row58	Parkki	170	20.7	19	23.2	9.396	3.41	225.478
Row133	Pike	1,600	60	56	64	9.6	6.144	1,145.437
Row36	Roach	160	22.5	20.5	25.3	7.033	3.82	188.575
Row91	Perch	197	25.6	23.5	27	6.561	4.239	305.623
Row83	Perch	150	22.5	20.5	24	6.792	3.624	219.246
Row131	Pike	950	51.7	48.3	55.1	8.926	6.171	962.627
Row33	Roach	120	21	19.4	23.7	6.115	3.294	133.942
Row118	Perch	1,000	44	41.1	46.6	12.489	7.596	953.389
Row46	Whitefish	270	26	23.6	28.7	8.38	4.248	311.231
Row95	Perch	265	27.5	25.4	28.9	7.052	4.335	371.092
Row119	Pike	200	32.3	30	34.8	5.568	3.376	382.436
Row109	Perch	820	39	36.6	41.3	12.431	7.351	851.957
Row26	Roach	40	14.1	12.9	16.2	4.147	2.268	-85.98
Row115	Perch	1,000	43	39.8	45.2	11.933	7.277	901.737
Row43	Roach	290	26	24	29.2	8.877	4.497	335.482
Row1	Bream	290	26.3	24	31.2	12.48	4.306	372.197
Row136	Smelt	6.7	9.8	9.3	10.8	1.739	1.048	-225
Row55	Parkki	120	19	17.5	21.3	8.392	2.918	157.251
Row89	Perch	188	24.6	22.6	26.2	6.733	4.166	281.576
Row69	Perch	78	18.7	16.8	19.4	5.199	3.123	94.241
Row67	Perch	70	17.4	15.7	18.5	4.588	2.942	35.653
Row144	Smelt	9.8	12	11.4	13.2	2.204	1.148	-163.974

Figure 7: Infeasible results

5. From observations from the source data before splitting them, the 2 species which can easily be differentiated from others were Bream and Smelt. Bream was labelled in green and Smelt was labelled in orange. Bream stood out as it has the highest numbers in the “Height_in_cm” attribute while Smelt stood out as it has the lowest numbers in the “Diagonal_Width_in_cm” attribute. Below is an illustration. This comes from a Scatter Matrix node after the data output from the Shuffle node has been labelled using the Color Manager Node.

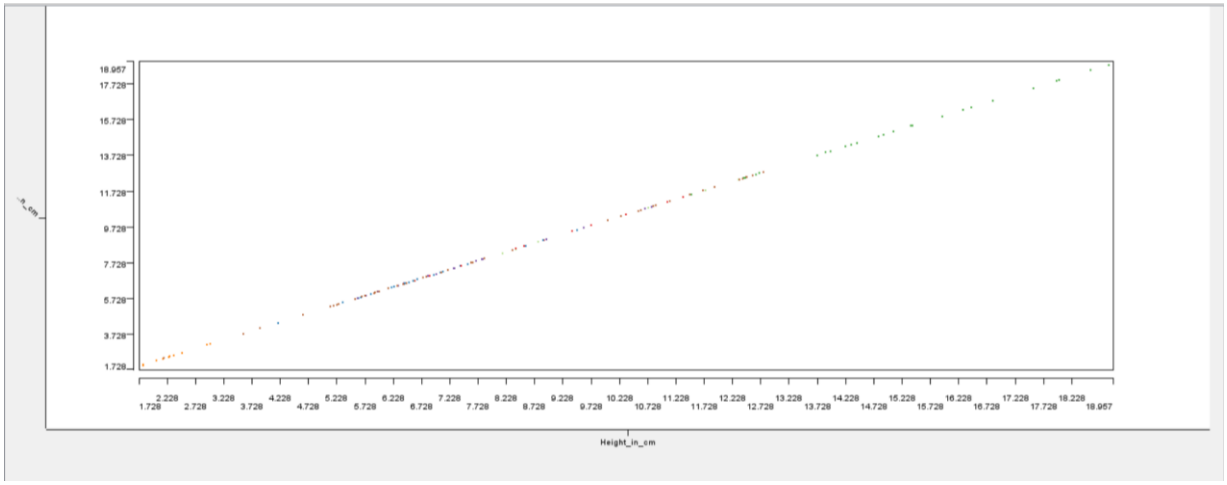

Figure 8: Height_in_cm

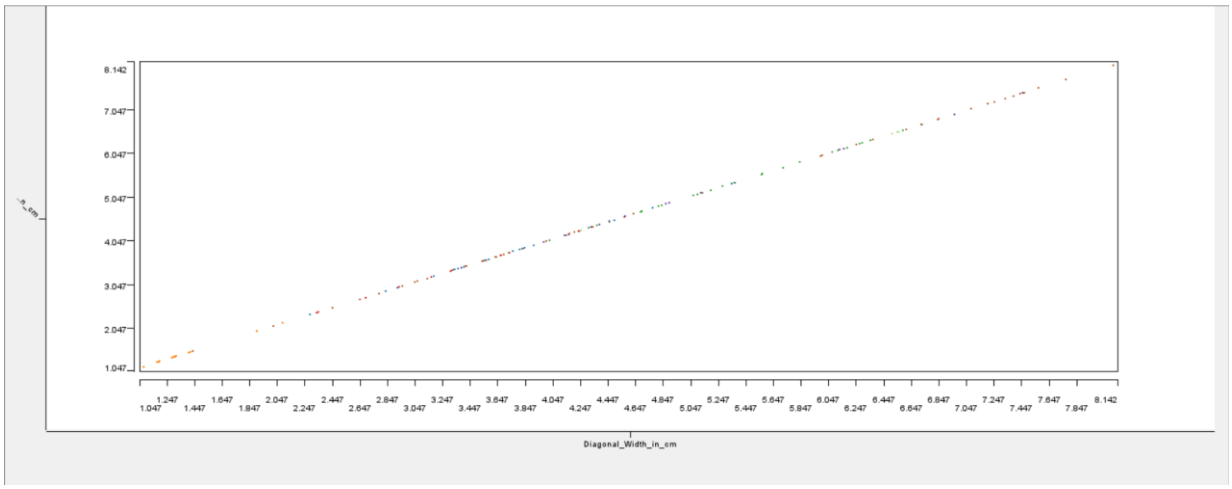

Figure 9: Diagonal_Width_in_cm

6. Below is a pie chart that describes the original data before it was put into the partitioning node and split into test and training sets. This used the Interactive Pie chart (local) node with data output from the CSV reader node.

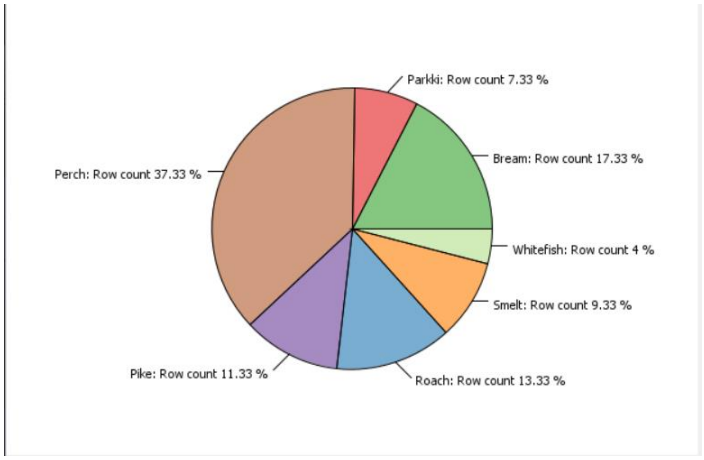

Figure 10: Pie chart

Question 3:

Dependent categorical values are predicted using logistic regression. The model will be built when the data we need to predict is categorical, with binary data like 0 or 1, true or false. The outcomes can be either of these values, both, but there is no value in the middle.

1. **Whitefish** as no True Positive (TP) case in the predicted result.

Row ID	TruePositives
Bream	2
Roach	4
Whitefish	0
Parkki	1
Perch	7
Pike	5
Smelt	2
Overall	?

Figure 11: True Positive cases

2. Since Whitefish has no True Positive (TP) case in the result, **Roach** will be misplaced into Whitefish.

Row ID	Bream	Roach
Bream	2	0
Roach	0	4
Whitefish	0	2
Parkki	1	0
Perch	0	4
Pike	0	0
Smelt	0	0

Figure 12: Misplaced TP cases

3. The overall accuracy of the results is **70%** or **0.7/1**.

Row ID	Accuracy
Overall	0.7

Figure 13: Accuracy of results

4. The species which have 100% correct classified results are **Bream, Roach, Pike, and Smelt.** The result is displayed as 1 in the Sensitivity column.

Row ID	Sensitivity
Bream	1
Roach	1
Whitefish	0
Parkki	0.5
Perch	0.538
Pike	1
Smelt	1
Overall	?

Figure 14: Correct classified species

5. The species with a 50% chance of being misplaced into a different type of fish is **Parkki**. 50% is corresponding to 0.5 in the sensitivity column

Row ID	Sensitivity
Bream	1
Roach	1
Whitefish	0
Parkki	0.5
Perch	0.538
Pike	1
Smelt	1
Overall	?

Figure 15: 50% of being misplaced.

6.

Row ID	TruePositives	FalsePositives	Specificity	Recall
Bream	2	1	0.964	1
Roach	4	6	0.769	1
Whitefish	0	0	1	0
Parkki	1	0	1	0.5
Perch	7	0	1	0.538
Pike	5	2	0.92	1
Smelt	2	0	1	1
Overall	?	?	?	?

Figure 16: Pike being misplaced into others

The percentage of Pike being misplaced into other fish types can be calculated in 2 ways:

- Answer = FalsePositives / [Total – (TruePositives / Recall)] * 100% = 2 / [30 – (5 / 1)] * 100% = 8%
- Answer = (1 – Specificity) * 100% = (1 – 0.92) * 100% = 8%

Question 4:

1. For the new linear regression model, we need to eliminate attributes so that there are only 3 left that would be used and increase the accuracy of the trained model. The eliminated attributes are "**Diagonal_Length_in_cm**" and "**Vertical_Length_in_cm**". This is because the data points observed when focusing on the species "Perch" were spread out across the table, therefore adding more noise and confusion to the model and lessening its accuracy. It can be seen that the lowest data point and the data point are separated by a big margin, with the rest of the points being scattered across the scale. On the other hand, the 3 attributes of "Cross_Length_in_cm", "Height_in_cm", and "Diagonal_Width_in_cm" is more compact with their data points, focusing on 1 or 2 hotspots only. From the figure below, we can see that graphs 4 and 5 both have 2 compact data regions, while graphs 1, 2, and 3 have more spread-out data points. We will pick the attribute from graph 1 (Diagonal_Length_in_cm) and graph 2 (Vertical_Length_in_cm) as the data is separated the most. For this, we use the Scatter Matrix after we filter the row from the Shuffle node so that only tuples with the attribute "Species" named "Perch" remains.

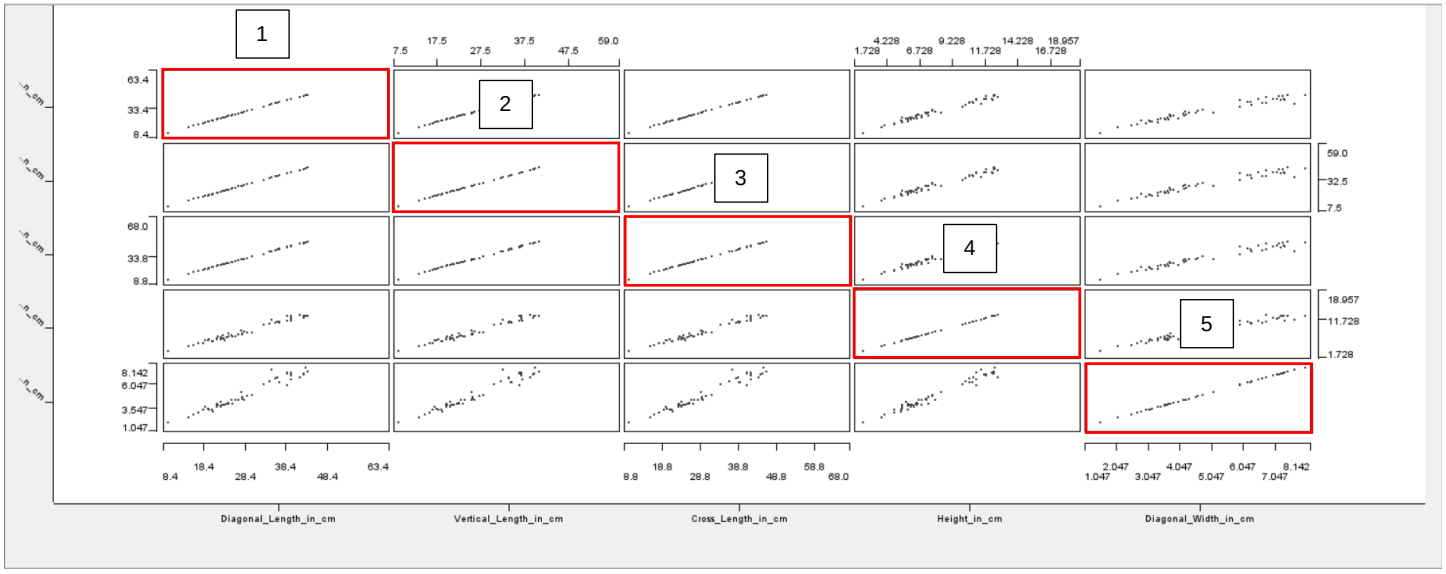

Figure 17: Graph for attributes removal decision

2. In question 2, R^2 when predicting all 7 species was 0.857, and when predicting only the species “Perch” was 0.943. The model that I have built in question 4 has R^2 = 0.947, which is an improved model when comparing both R^2 when predicting all species (better by 0.09) and “Perch” specifically (better by 0.004) in question 2. For this, we use the Numeric Scorer from their respective linear regression models.

Row ID	D Predicti...
R^2	0.947
mean absolut...	64.488
mean square...	5,752.543
root mean sq...	75.846
mean signed ...	23.966
mean absolut...	0.243
adjusted R^2	0.947

Figure 18: R^2 in question 4 model

Row ID	D Predicti...
R^2	0.943
mean absolut...	70.509
mean square...	6,250.14
root mean sq...	79.058
mean signed ...	42.937
mean absolut...	0.299
adjusted R^2	0.943

Figure 19: R^2 in question 2 model (Perch only)

Row ID	D Predicti...
R^2	0.857
mean absolut...	101.021
mean square...	18,678.603
root mean sq...	136.67
mean signed ...	23.338
mean absolut...	2.118
adjusted R^2	0.857

Figure 20: R^2 in question 2 model (all species)