Cover sheet for submission of work for assessment

SWINBURNE UNIVERSITY OF TECHNOLOGY

UNIT DETAILS

Unit name	Data Science Principle			Class day/time	1pm Friday
Unit code	COS10022	Assignment no.	1	Due date	1/10/2023
Name of lecturer/teacher	Dang Vi Luan				
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)	Dang	Vi Luan	103802759
(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)	Nguyen Dinh Nhat Minh (peer reviewer)	(4)	
(2)	Dang Vi Luan (author)	(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/**

COS10022 Data Science Principles

Assignment 1 – Vi Luan Dang – 103802759

1. Introduction

Proficiency in conducting, utilizing, and elucidating predictive models and their outcomes is crucial for students aspiring to pursue a career in the IT and data science field. In this assignment, I aim to showcase my comprehension of fundamental data science concepts by constructing predictive models employing linear and regression models. Each section will encompass significant steps involved in this process, and within each section, I will address the assignment's provided questions.

2. Data Preparation

Data preparation plays a pivotal role in building predictive models, such as linear and logistic regression. The quality and reliability of the data directly impact the accuracy and effectiveness of our output, therefore, the very first step would be to prepare our data.

In order to process the assignment's input, we will utilize the "Fish_Species.csv" file by employing the "CSV Reader" tool. Subsequently, to enhance data visualization and facilitate attribute identification, we will incorporate the "Color Manager" tool, which assigns distinct colors to our key attribute, namely "species". Lastly, to ensure randomness and prevent any bias, we will shuffle the input data using the "Shuffle" tool, utilizing a seed value of "3122". The workflow for the process is as follows:

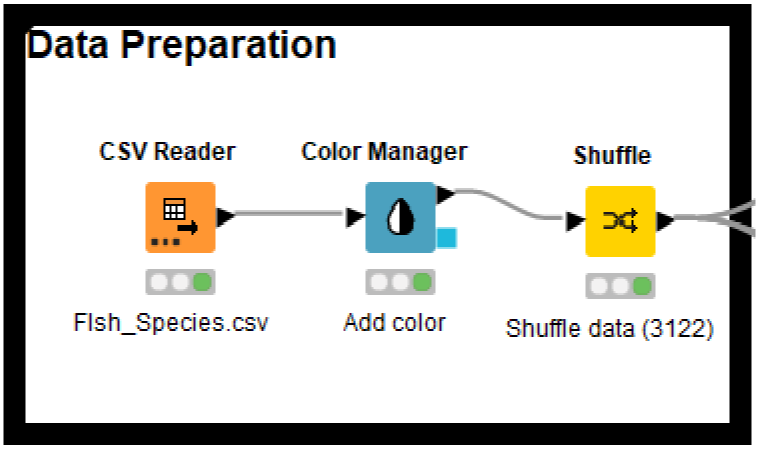


Figure 1: Data Preparation workflow

Answers to assignment's questions

1) Past a clear screenshot of the whole workflow of assignment 1 in the report. [2.5 marks]

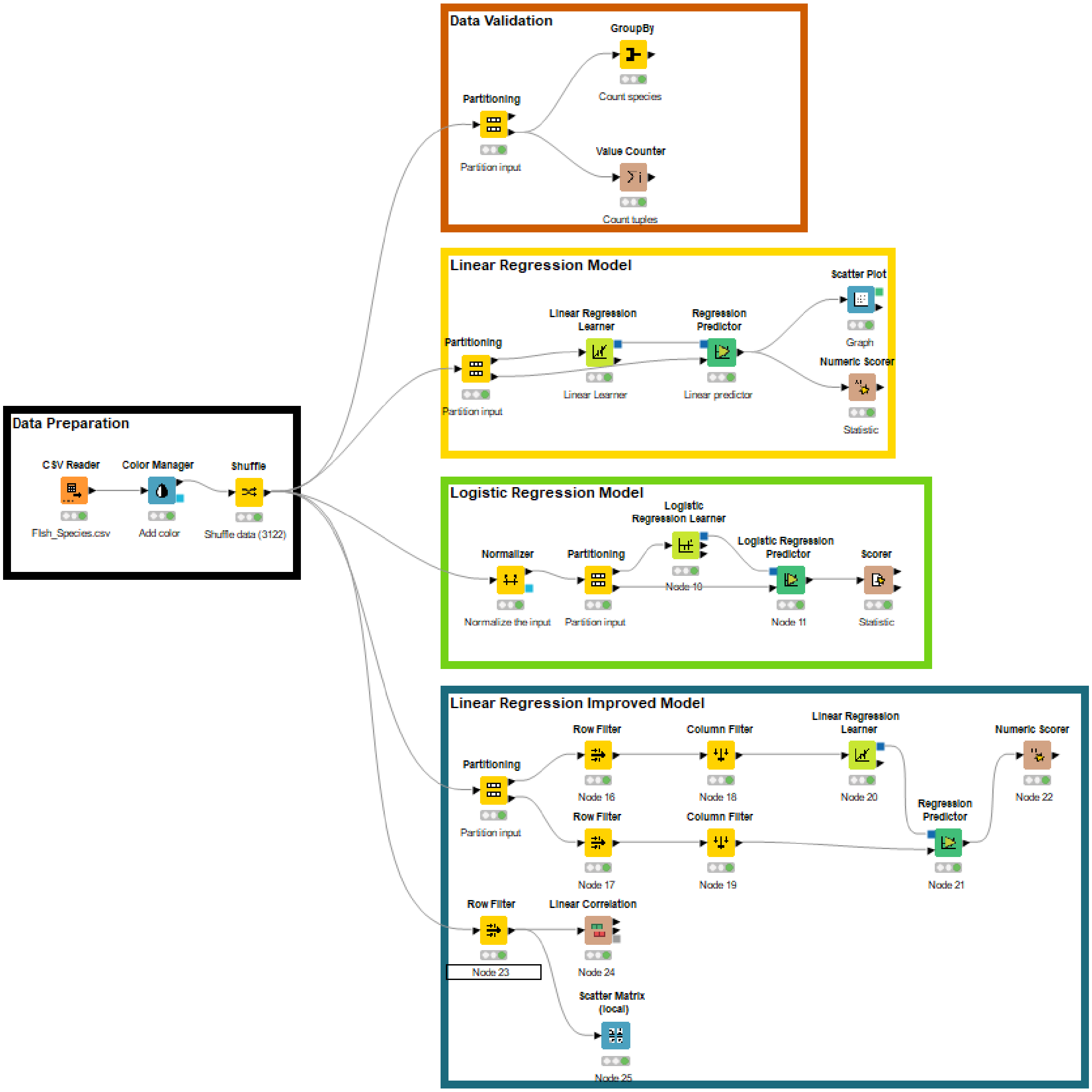


Figure 2: Overall workflow

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

The partition part will be done for each process; therefore, it is not included in the data preparation part. However, as we will use 80% of our tuples for training set, **the answer would be 120 tuples**.

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

Using "Group by" Tool in the data validation process, there are **7 species in the test set**.

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

Using the “Value Counter” Tool, we can confirm that **“Whitefish” and “Smelt” have the same number of tuples in the test set.**

3. Linear Regression model.

To construct the linear regression model, the initial step involves partitioning the dataset into training and test sets, utilizing an 80% - 20% split with a seed value of 3122. Subsequently, the training set is fed into the "Linear Regression Learner" tool, which employs the method of least squares to calculate the coefficients of the linear regression equation. Following this, the "Regression Predictor" tool utilizes the previously obtained coefficients to make predictions on the input feature values of the new data instances, specifically the test set. In order to visualize the model's output, the "Scatter Plot" tool is employed to generate a graph, while the "Numeric Scorer" tool provides statistical information pertaining to the output.

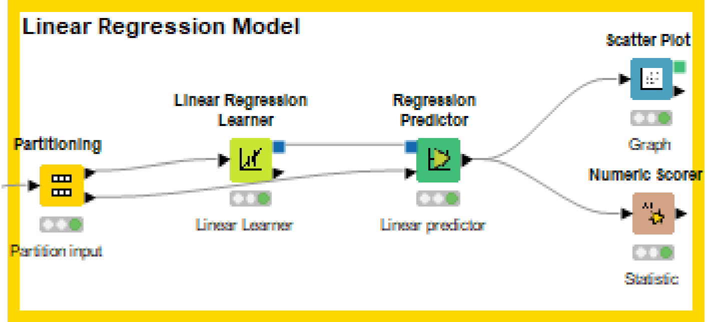
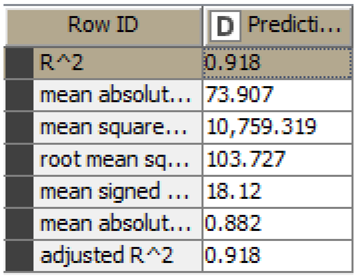


Figure 3: Linear Regression Model

The statical information produced by “Numeric Scorer” Tool is noteworthy:



Row ID	D Predicti...
R^2	0.918
mean absolut...	73.907
mean square...	10,759.319
root mean sq...	103.727
mean signed ...	18.12
mean absolut...	0.882
adjusted R^2	0.918

Figure 4: Linear Regression model statistics

Our model obtained an R-squared value of 0.918, indicates that approximately 91.8% of the variance in the dependent variables can be explained by the independent variables. This suggests a robust fit and substantial relationship between the variables. The Root Mean Squared Error (RMSE) of 103.727 reveals that, on average, the model’s predictions deviate by approximately 103.727 units from the actual values. This indicates a relatively low level of prediction errors and highlights the model’s accuracy in estimating the target variable. Additionally, the Mean Absolute Percentage Error (MAPE) of 0.882 indicates that, on average, the predictions deviate by approximately 88.2%. Overall, our model exhibits a high degree of explanatory power, low prediction error, and reasonable accuracy.

Answers to assignment’s questions

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

R^2 equals **0.918**

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.”

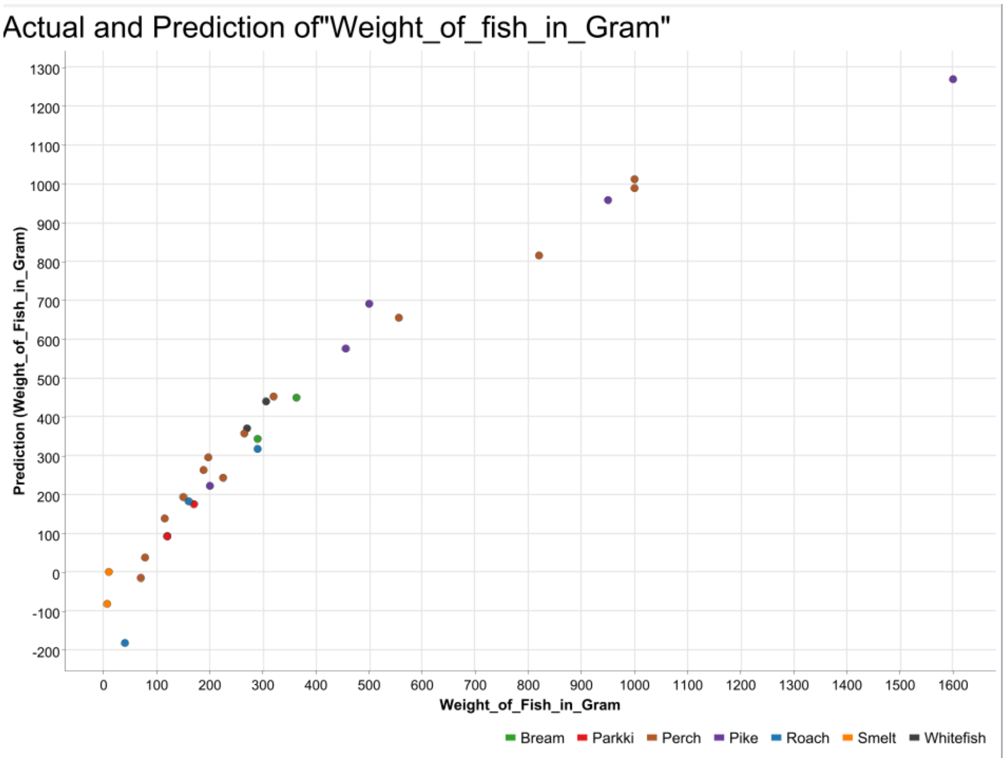
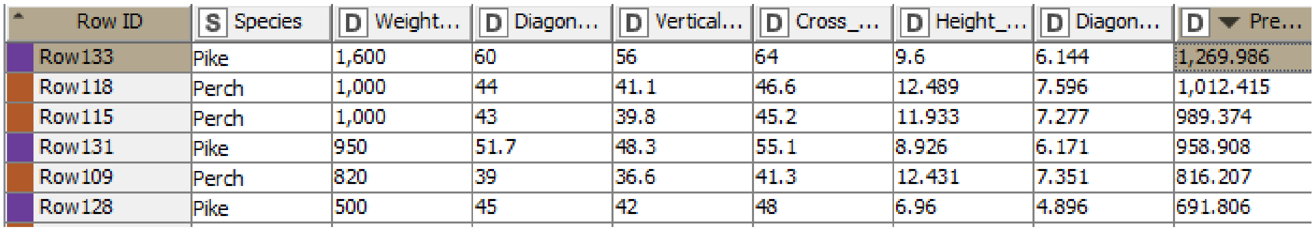


Figure 5: Prediction and actual data graph

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

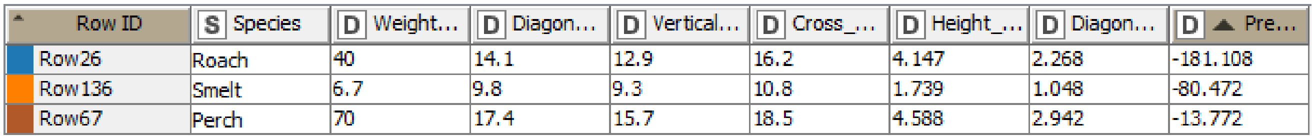


Row ID	S Species	D Weight…	D Diagon…	D Vertical…	D Cross_…	D Height_…	D Diagon…	D ▼ Pre…
🟪 Row133	Pike	1,600	60	56	64	9.6	6.144	1,269.986
🟫 Row118	Perch	1,000	44	41.1	46.6	12.489	7.596	1,012.415
🟫 Row115	Perch	1,000	43	39.8	45.2	11.933	7.277	989.374
🟪 Row131	Pike	950	51.7	48.3	55.1	8.926	6.171	958.908
🟫 Row109	Perch	820	39	36.6	41.3	12.431	7.351	816.207
🟪 Row128	Pike	500	45	42	48	6.96	4.896	691.806

Figure 6: Predicted weight.

Pike has the heaviest predicted weight in the test results.

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



Row ID	S Species	D Weight…	D Diagon…	D Vertical…	D Cross_…	D Height_…	D Diagon…	D ▲ Pre…
🟦 Row26	Roach	40	14.1	12.9	16.2	4.147	2.268	-181.108
🟧 Row136	Smelt	6.7	9.8	9.3	10.8	1.739	1.048	-80.472
🟫 Row67	Perch	70	17.4	15.7	18.5	4.588	2.942	-13.772

Figure 7: Infeasible values

There are 3 infeasible values, as weight cannot be a negative value.

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]

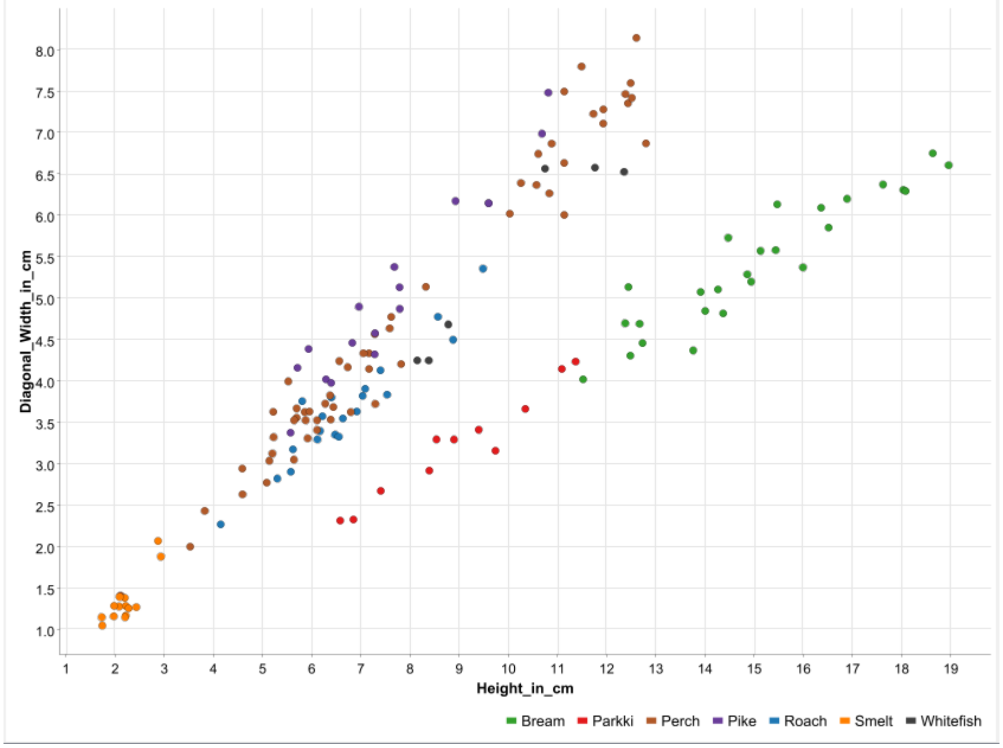


Figure 8: Distinguish species

We can easily separate “Bream” and “Smelt” from the rest of the species.

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]

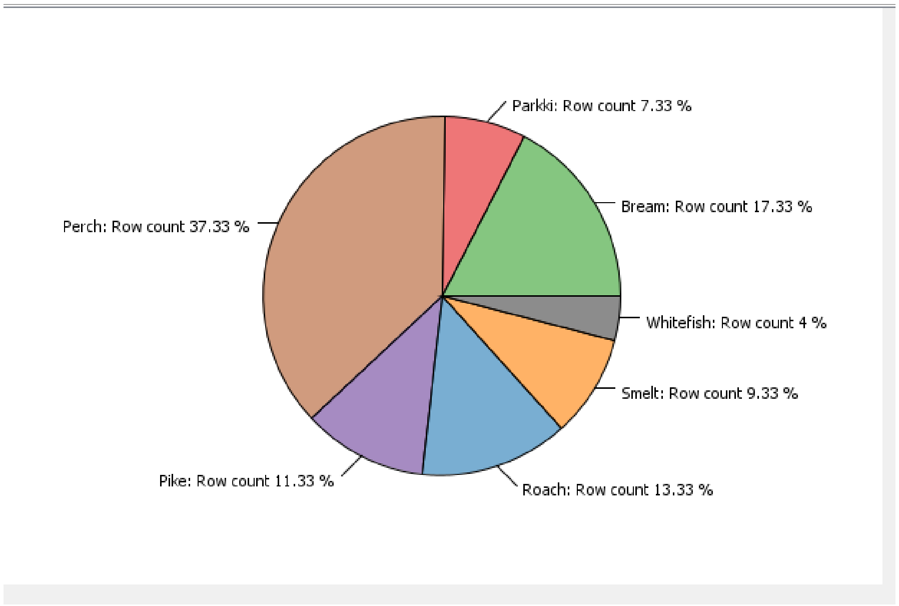


Figure 9: Pie chart for original data distribution

4. Logistic Regression model.

To build a logistic regression model, it is essential to normalize the data, scaling it from 0.1 to 1.0. This normalization step is crucial as logistic regression is sensitive to the scale of input features. When the features have varying scales, those with larger scales can dominate the model's learning process. Following normalization, the data is partitioned into training and test sets. We will configure "Smelt" as our reference category and the termination points of epochs and epsilon are limited to 10.000 and 0.0001 so as to control convergence behavior of the training process. The training set is then used as input for the "Logistic Regression Learner" tool, which estimates the regression coefficients that best fit the data. Finally, the trained model is applied using the "Logistic Regression Predictor" to make predictions on unseen data, specifically the test set.

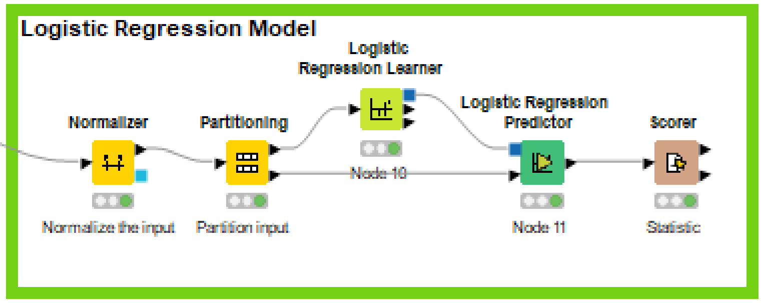
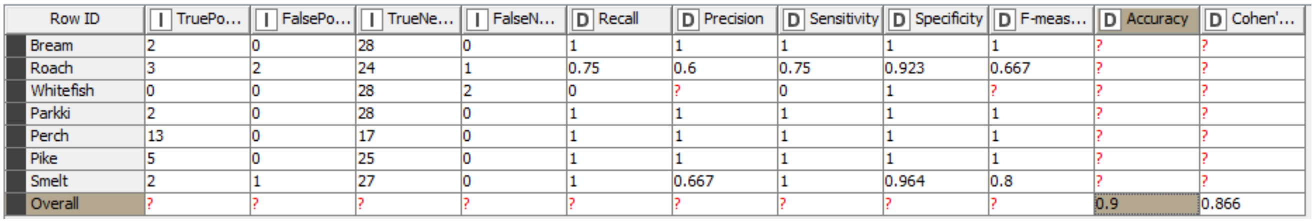


Figure 10: Logistic Regression model

After the configuration, we earned an overall accuracy of 90%, which is a decent accuracy for this predictive model.

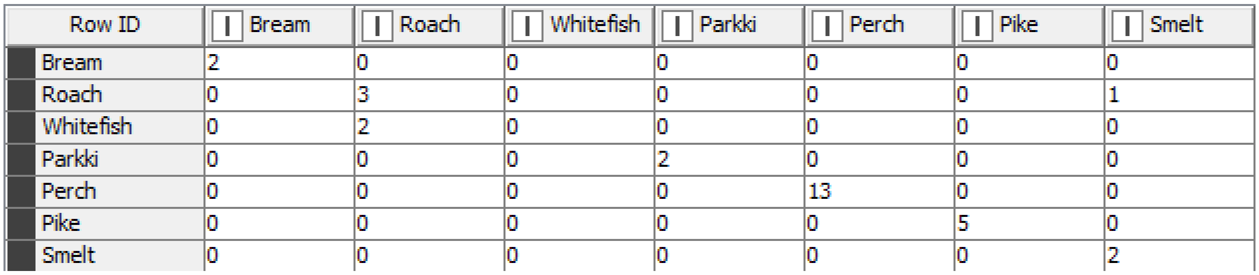


Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
Bream	2	0	28	0	1	1	1	1	1	?	?
Roach	3	2	24	1	0.75	0.6	0.75	0.923	0.667	?	?
Whitefish	0	0	28	2	0	?	0	1	?	?	?
Parkki	2	0	28	0	1	1	1	1	1	?	?
Perch	13	0	17	0	1	1	1	1	1	?	?
Pike	5	0	25	0	1	1	1	1	1	?	?
Smelt	2	1	27	0	1	0.667	1	0.964	0.8	?	?
Overall	?	?	?	?	?	?	?	?	?	0.9	0.866

Figure 11: Overall accuracy

Answers to assignment's questions

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



Row ID	I Bream	I Roach	I Whitefish	I Parkki	I Perch	I Pike	I Smelt
Bream	2	0	0	0	0	0	0
Roach	0	3	0	0	0	0	1
Whitefish	0	2	0	0	0	0	0
Parkki	0	0	0	2	0	0	0
Perch	0	0	0	0	13	0	0
Pike	0	0	0	0	0	5	0
Smelt	0	0	0	0	0	0	2

Figure 12: Confusion Matrix

"Whitefish" has no True Positive (TP) case.

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

For the species that has no TP case, "Roach" will be misplaced.

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

The overall accuracy of the prediction result is 90%

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

The accuracy of formula states that:

$$\boldsymbol{Accuracy} = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

So in order for Accuracy to be 100%, the cases of incorrectly classification has to be 0, which will make the formula equals $\frac{True\ Positives+True\ Negatives}{True\ Positives+True\ Negatives} = 1$.

There are 4 species that has no case of incorrectly classification in *Figure 11,* which are **"Bream", "Parkki", "Perch", "Pike".**

5) Which species has a 50% chance of being misplaced into another species in the test result? [5 marks]

To find the species that has a 50% chance of being misplaced, we will use the False Negative Rate formula:

$$FNR = \frac{FN}{FN + TP}$$

With the data provided in *Figure 11*, there is no species that will satisfy this condition.

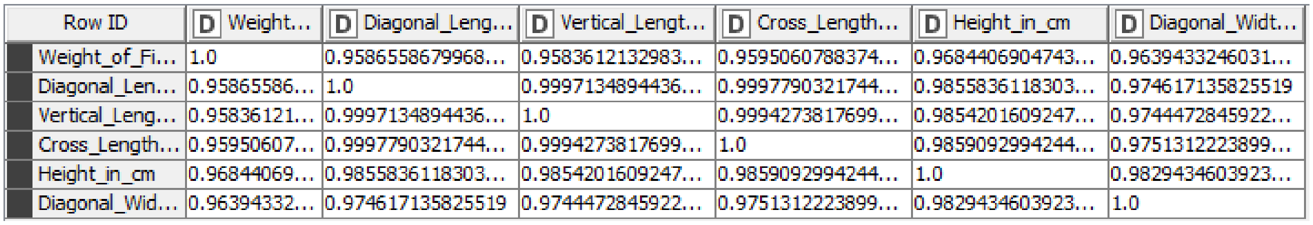
6) In the test result, what percentage of the species "Pike" is misplaced into others? [5 marks]

As "Pike" has the recall of 1 (according to *figure 11*), the percentage of "Pike" being misplaced into others is 0%.

5. Logistic Regression improved model.

In this section, our objective is to enhance our model in part 2 by specifically focusing on the species "Perch". As we reduce the dimensionality of the input variables, it becomes crucial to reassess the remaining attributes associated with the "Perch" species. This reassessment is necessary to mitigate any potential issues of collinearity among the remaining attributes.

We will first use "Linear Correlation" Tool to measure and quantify the strength and direction of the linear relationship between remaining attributes.

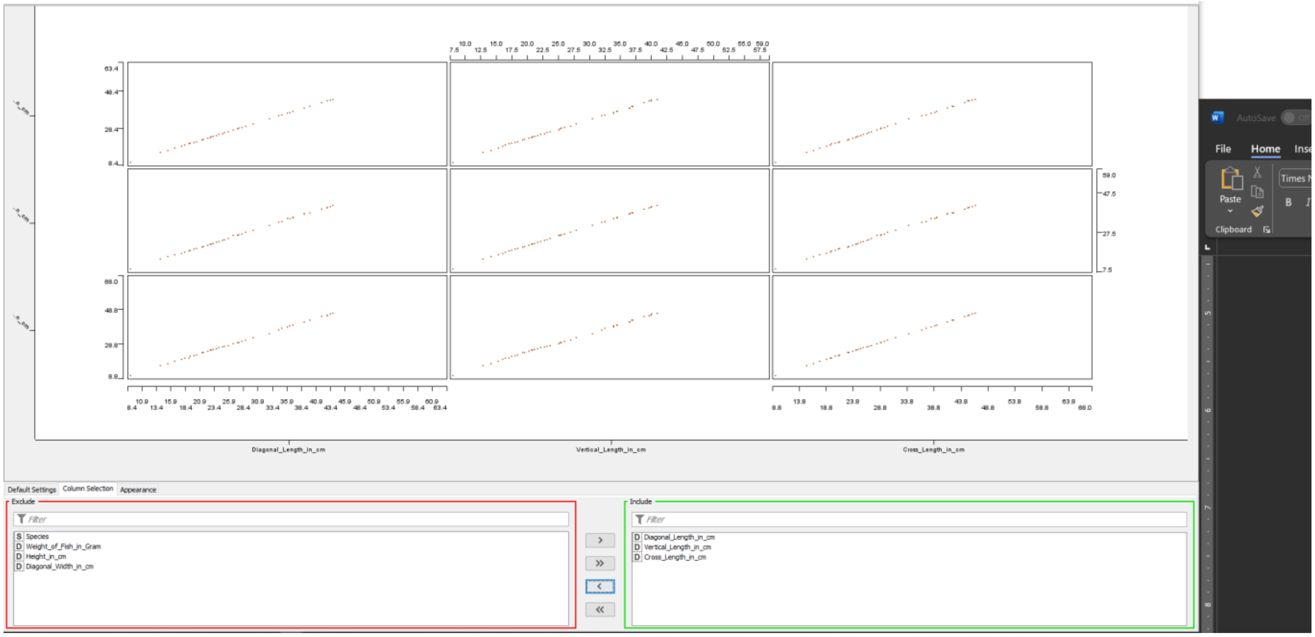


Row ID	D Weight...	D Diagonal_Leng...	D Vertical_Lengt...	D Cross_Length...	D Height_in_cm	D Diagonal_Widt...
Weight_of_Fi...	1.0	0.9586558679968...	0.9583612132983...	0.9595060788374...	0.9684406904743...	0.9639433246031...
Diagonal_Len...	0.95865586...	1.0	0.9997134894436...	0.9997790321744...	0.9855836118303...	0.974617135825519
Vertical_Leng...	0.95836121...	0.9997134894436...	1.0	0.9994273817699...	0.9854201609247...	0.9744472845922...
Cross_Length...	0.95950607...	0.9997790321744...	0.9994273817699...	1.0	0.9859092994244...	0.9751312223899...
Height_in_cm	0.96844069...	0.9855836118303...	0.9854201609247...	0.9859092994244...	1.0	0.9829434603923...
Diagonal_Wid...	0.96394332...	0.974617135825519	0.9744472845922...	0.9751312223899...	0.9829434603923...	1.0

Figure 13: Correlation matrix

As we can see from Figure 13, **Diagonal_Length_in_cm** and **Height_in_cm** possess high collinearity with other attributes.

This can be visualized using a Scatter Matrix, which displays the collinearity of these attributes with the others.



Exclude
- S Species
- D Weight_of_Fish_in_Gram
- D Height_in_cm
- D Diagonal_Width_in_cm

Include
- D Diagonal_Length_in_cm
- D Vertical_Length_in_cm
- D Cross_Length_in_cm

Figure 14: Scatter Matrix

Therefore, we need to exclude these attributes so that our model can be improved.

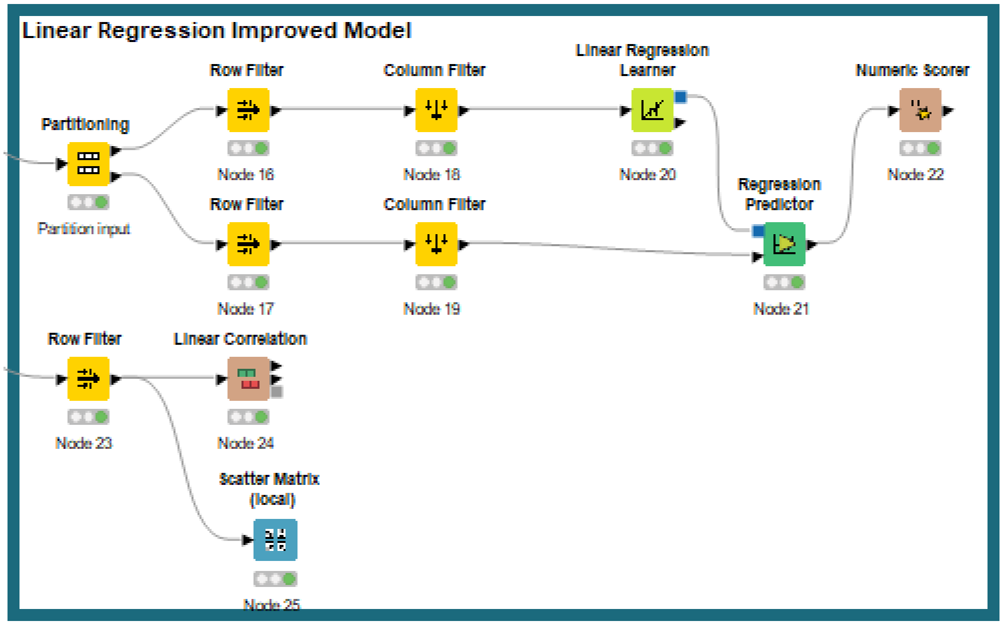


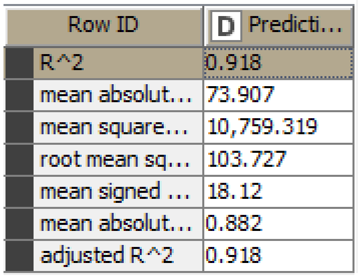
Figure 15: Improved Linear Regression Model

Answers to assignment's questions

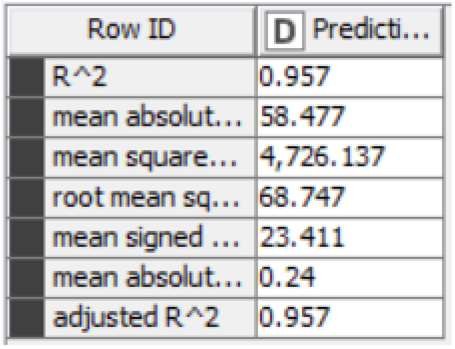
1) Give the reasons for each eliminated attribute and why they are not selected as the input. [5 marks]

The eliminated attributes are: **Diagonal_Length_in_cm** and **Height_in_cm** as explained above. The remaining attributes are: **Vertical_Length_in_cm, Cross_Length_in_cm, Diagonal_Width_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]



Row ID	D Predicti...
R^2	0.918
mean absolut...	73.907
mean square...	10,759.319
root mean sq...	103.727
mean signed ...	18.12
mean absolut...	0.882
adjusted R^2	0.918



Row ID	D Predicti...
R^2	0.957
mean absolut...	58.477
mean square...	4,726.137
root mean sq...	68.747
mean signed ...	23.411
mean absolut...	0.24
adjusted R^2	0.957

Figure 16: Improve Model

Our R^2 has improved from 0.918 to 0.957