**Cover sheet for submission of**

**work for assessment**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **UNIT DETAILS** | | | | | | | | | | |
| Unit name | |  | | | | | | Class day/time |  | Office use only |
| Unit code | |  | | | Assignment no. | |  | Due date |  |  |
| Name of lecturer/teacher | | | |  | | | | | |  |
| Tutor/marker’s name | | |  | | | | | | | Faculty or school date stamp |
| **STUDENT(S)** | | | | | | | | | | |
| Family Name(s) | | | | | | Given Name(s) | | | | Student ID Number(s) |
| (1) |  | | | | |  | | | |  |
| (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:   1. 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) |  | (4) |  | |
| (2) |  | (5) |  | |
| (3) |  | (6) |  | |

# http://www.swinburne.edu.my/images/2013/logo.jpgSwinburne 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)

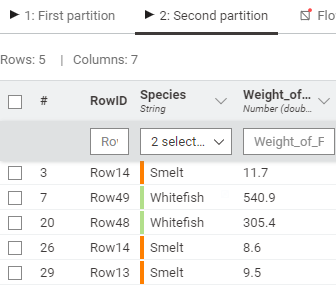
**Assessable Item:**

* 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]**

Ans: Check Assignment 1.1 for the KNIME workflow file.

* 1. How many tuples are included in the training set? **[2.5 marks]**

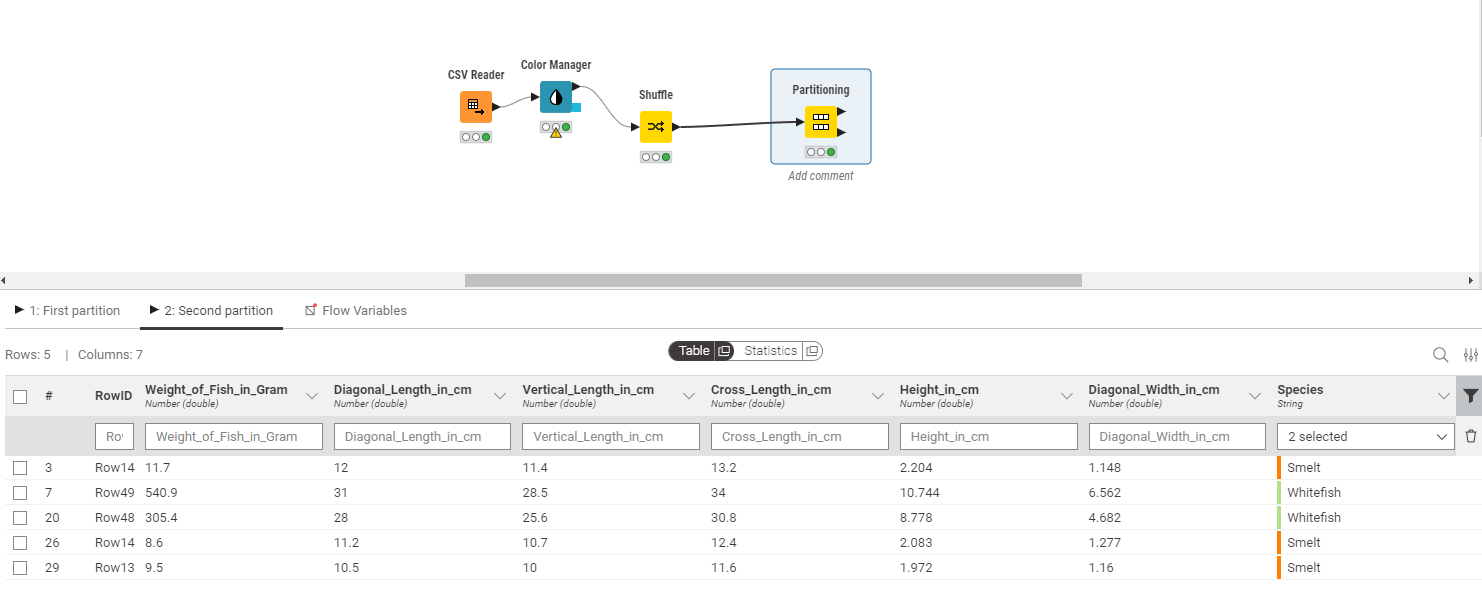
Ans: 120

* 1. How many species are included in the test set? **[2.5 marks]**

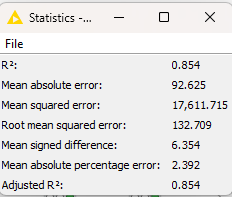
Ans: 7

* 1. Do species “Whitefish” and “Smelt” have the same number of tuples included in the test set?   
     **[2.5 marks]**

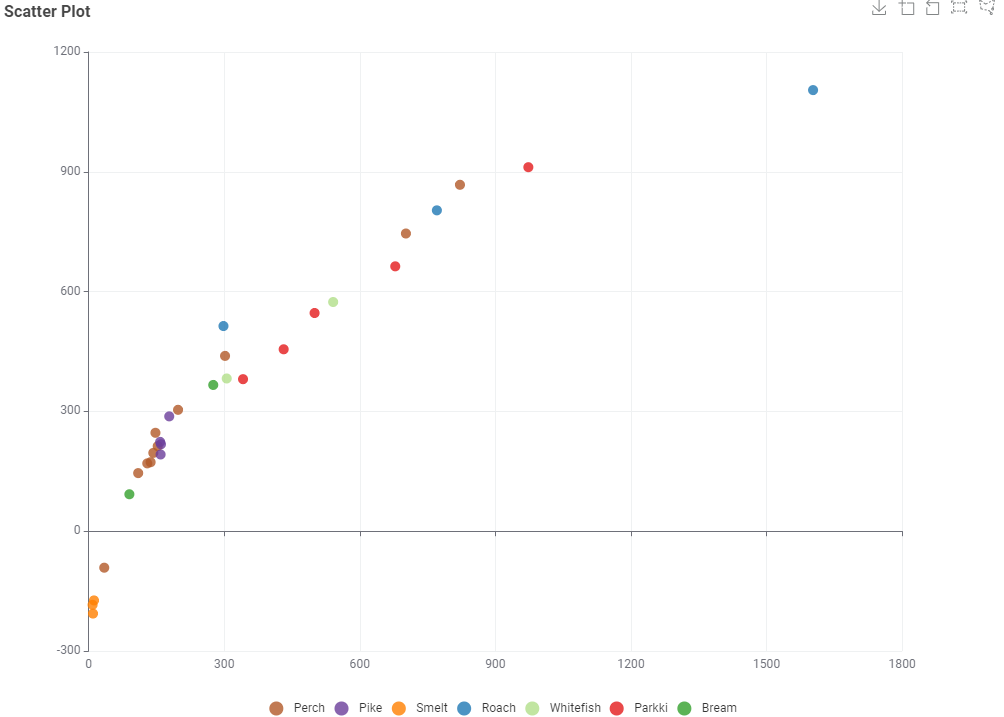
Ans: No, because Smelt has 3 records while Whitefish has 2 records in the test set



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

Ans: 0.854

1. 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]**



1. Which species has the heaviest predicted weight in your test result? **[5 marks]**

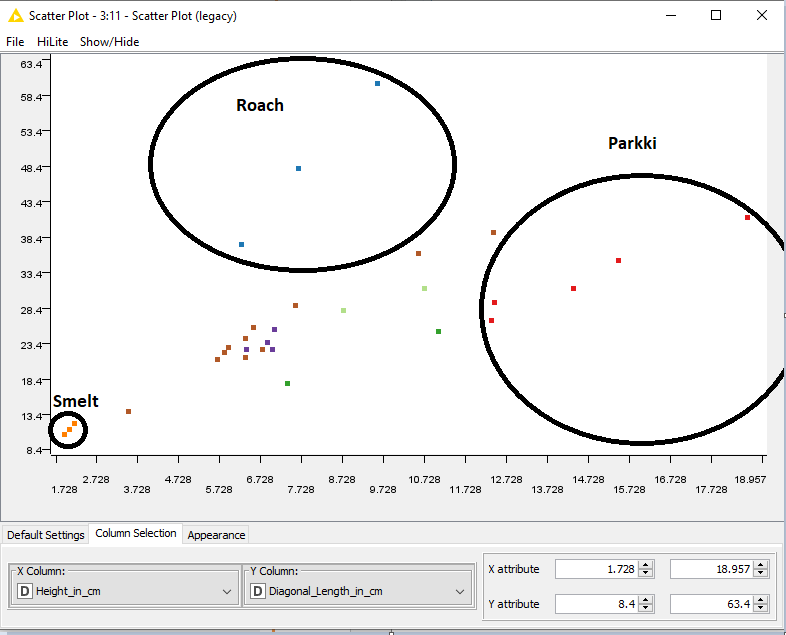
Ans: Roach has the heavies predicted weight in the test result

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

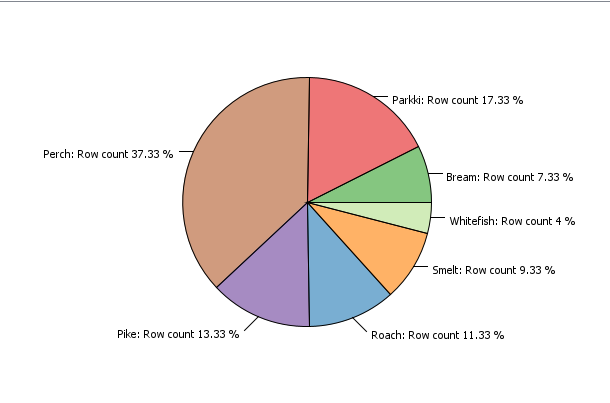
Ans: There are 3 infeasible prediction in the test result. Looking at the scatter plot, we can observe 3 points below x-axis (negative values) which are infeasible prediction for fish weight, as weight cannot be negative

1. 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: Smelt, Roach and Parkki can be easily separated from others based on Heigh and Diagonal\_length

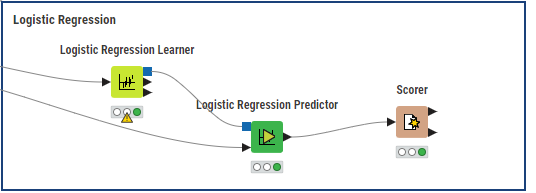


1. Draw a doughnut chart of the original input data with 0.55 as the doughnut hole ratio before splitting it into training and test sets. Use different colours for each species and show the percentage of data in the pie chart. **[5 marks]**



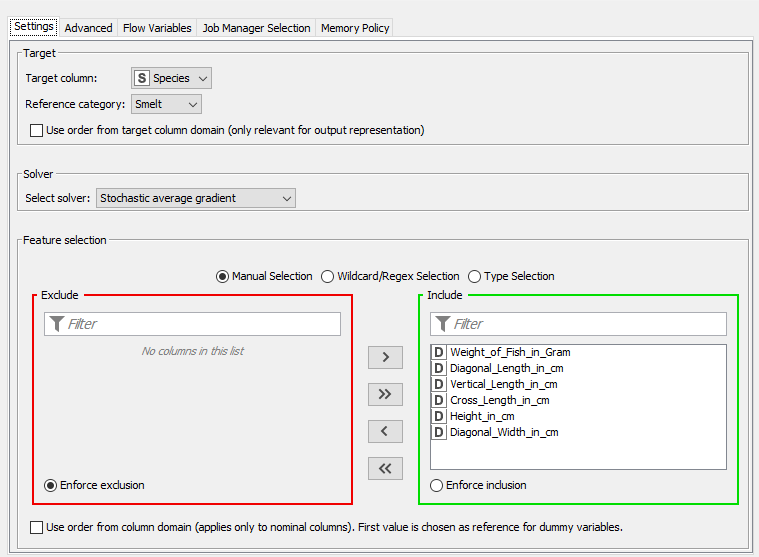
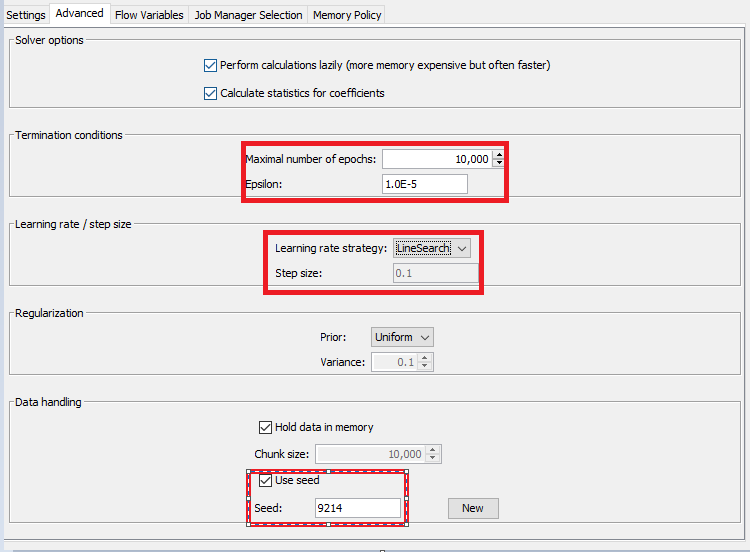
**We use Pie Chart (legacy) node in the “Data Visualization” section before partitioning.**

1. 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]**



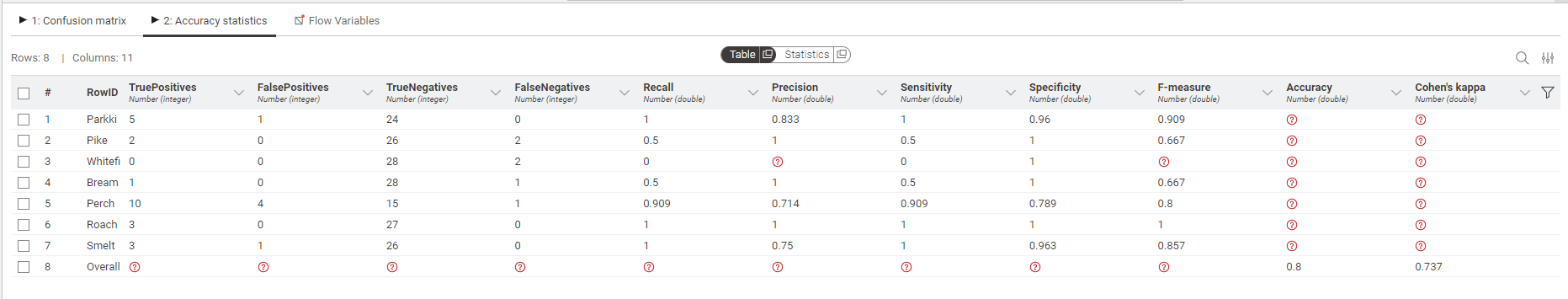
**We’ll split into 2 cases, because there is a warning on the Logistic Regression node**

**Case 1**: Proceed like normal. Here are workflow and the settings for logistic regression Learner node:

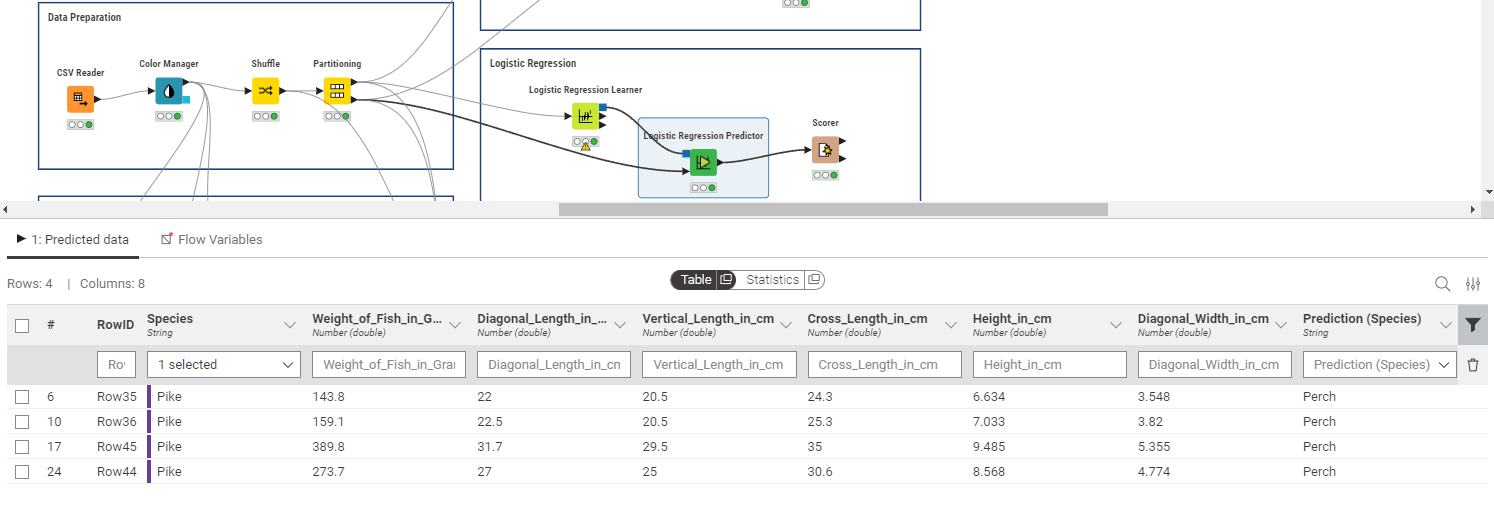


With all of the attributes, the "Smelt" species is utilised as the reference category for creating the Logistic Regression Model. Epsilon and the total number of epochs are limited to 10,000 and 0.00001, respectively. In the Learner Node for Logistic Regression, the seed value is set at 9214 and “LineSearch” is used as the learning rate strategy.

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

Ans: Pike. We can look at the scorer table, accuracy statistics

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

Ans: Perch. The species with no TP case is Pike, so we will go back to the Predictor to see the prediction result. As we can see after filtering “Pike”, we can see all of them are predicted Perch

1. What is the overall accuracy of the prediction result? **[5 marks]**

Ans: 0.7, which is 70%. Look at the Accuracy statistics above

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

Ans: Species with 100% correctly classified test results means they have recall of 100%. 🡪 They are Parkki, Roach, Smelt

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

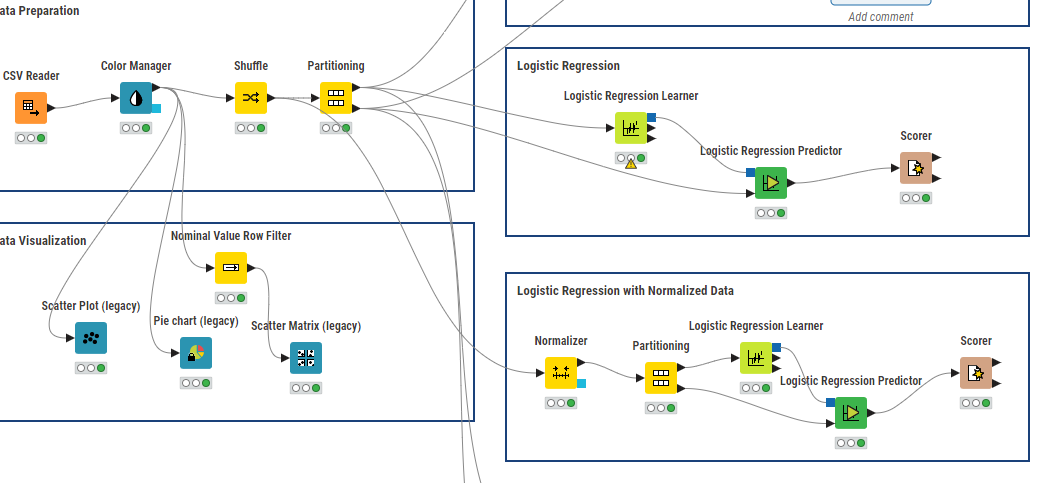
Ans: We also use recall to answer this question. No such species exists in statistic table above

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

Ans: In this case, we will use Precision to answer. Perch’s precision is 0.625 or 62.5%, so the percentage of the species is misplaced into others is 100% - 62.5% = 37.5% (Accuracy Statistics)

Case2: We add a “Normalizer” node before partitioning. We use normalizing method of Min Max Normalizing due to the following reason:

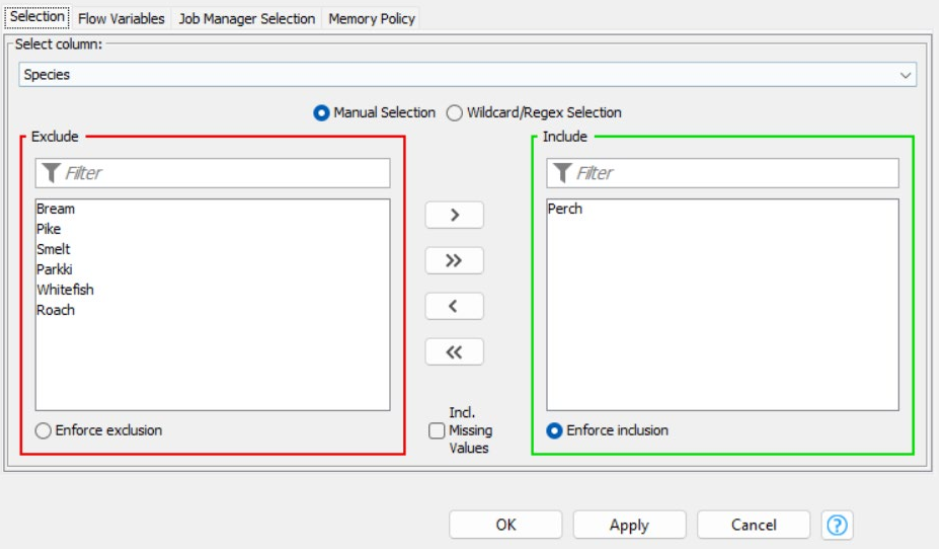
* The data do not have too many outliers.

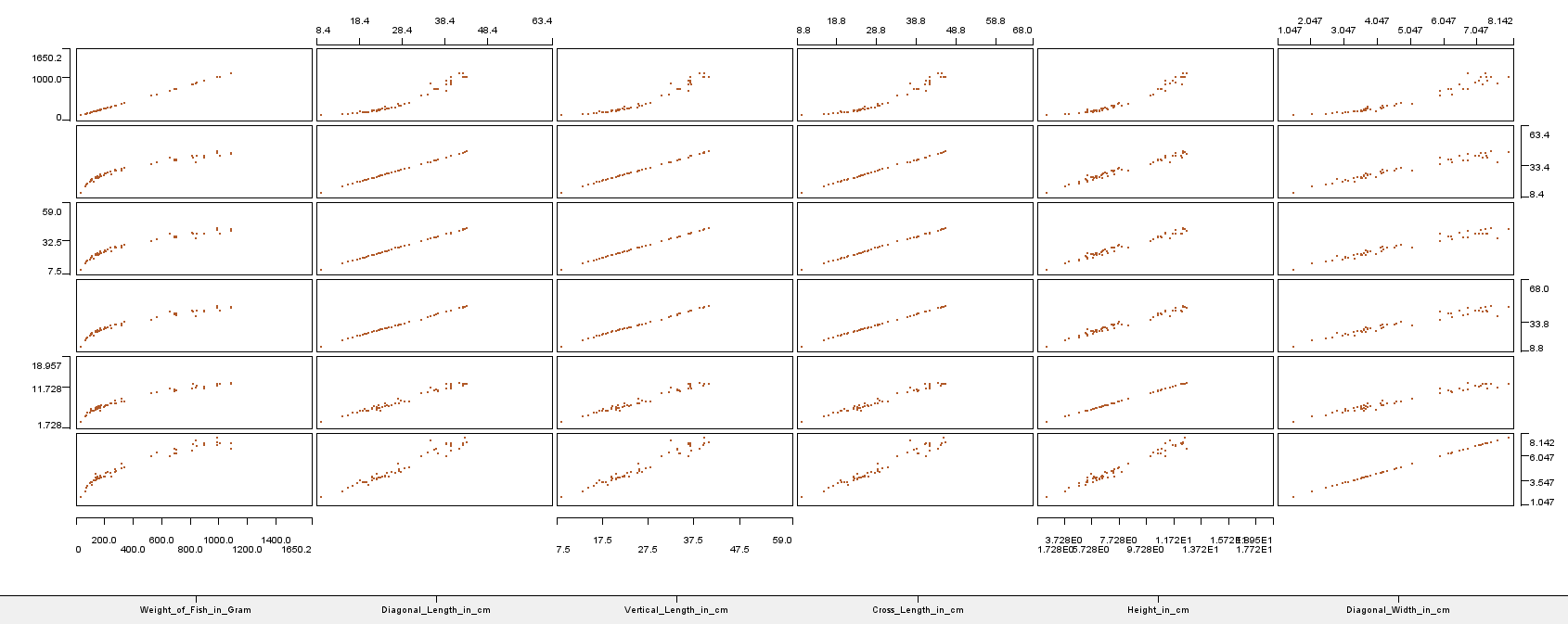
Here are the workflow and the settings: (the rest is the same as the previous logistic model, as well as how to find the answer)

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

Ans: Whitefish

1. 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]**

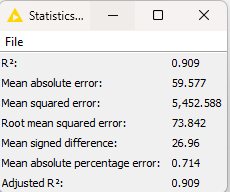
* Workflow of Improved Linear Regression model
* All 3 nominal value row filter is set to include “Peach” only

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

Ans: By observation trials, removing Diagonal\_Length and Height will yield the highest possible R2 I can find. To futher support this decision, in this matrix we can see that attributes Diagonal Length in cm and Heigh in cm has one of the most collinearity compares to others, so the two of them should be eliminated

1. List the 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]**

Ans: The new R^2 = 0.909

New model:  Old model 