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	12/02/2023	
Name of lecturer	Dr. James Jackson					
Tutor/marker's name	Dr. James Jackson					Faculty or school date stamp

STUDENT(S)

Family Name	Given Name	Student ID Number
Luu	Tuan Hoang	104180391

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.

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 - *Semester 1, 2023*

I. Overview

This assignment focuses on:

- Defining the key concepts, procedures, and tools involved in data management and prediction model construction.
- Choosing and putting into practice models and features for a data science project.
- Partitioning the dataset and utilizing the linear and logistic regressions to build two predictive models in the KNIME analytics platform.
- Selecting the independent attributes, dividing the data into training and test sets, developing a practical predictive model, and interpreting the results.

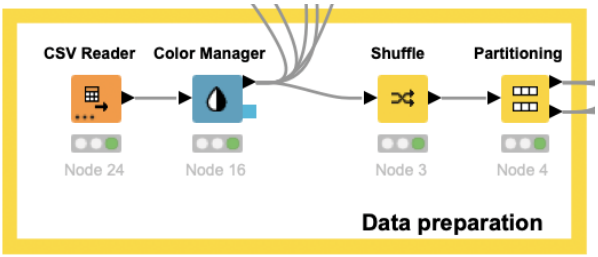
II. Abstract

The dataset includes 150 tuples representing 7 fish species in the market. The source data contains 6 attributes in total. This assignment has two objectives: the first is to build a linear regression model to predict the weight of the fish, such as the value in the "Weight_of_Fish_in_Gram" attribute, and the second is to build a logistic regression model to predict the species of fish.

Answers to questions in the assignment are included in the end of each assignment's task.

III. Data preparation (10%)

1. Data preparation for model training and test



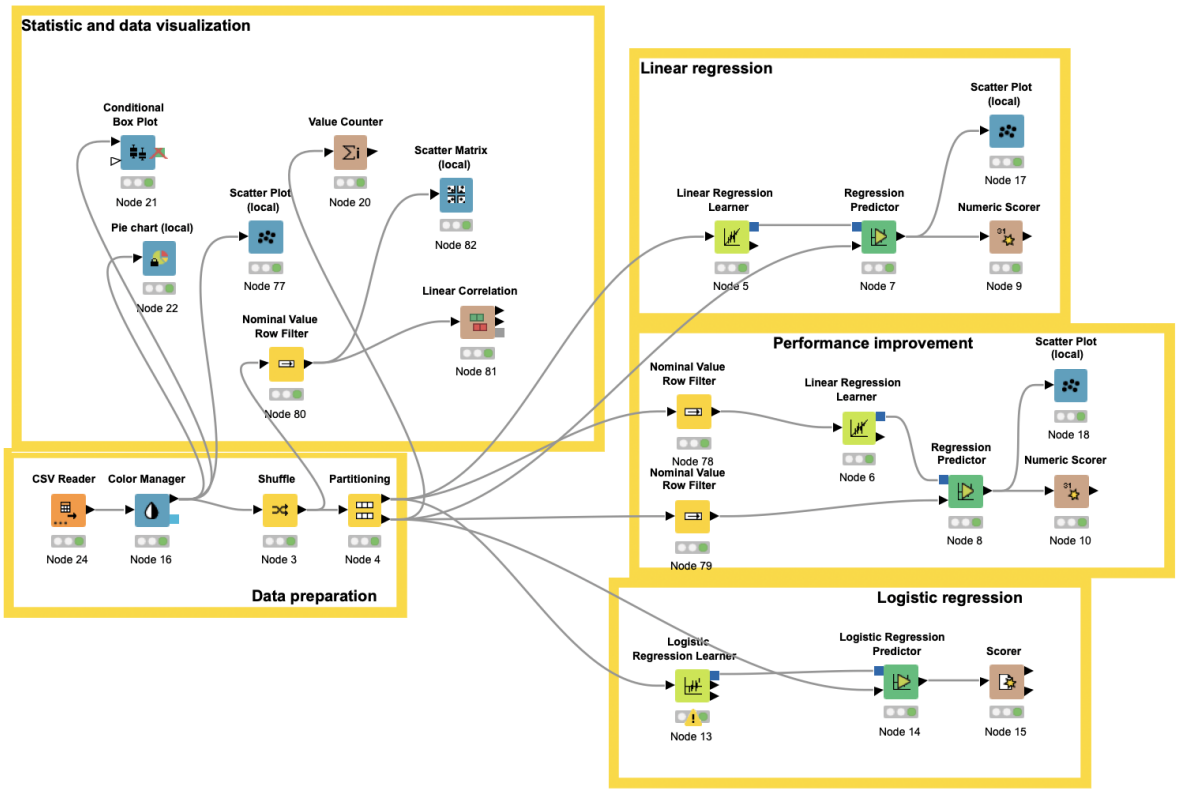
For data preparation, we use a total of 4 nodes for this task: CSV Reader, Color Manager, Shuffle and Partitioning.

Firstly, we use CSV Reader to input data from source file Fish_Species.csv which includes 150 tuples with header row. There are a total of 6 attributes in the source data: "Weight_of_Fish_in_Gram", "Diagonal_Length_in_cm", "Vertical_Length_in_cm", "Cross_Length_in_cm", "Height_in_cm", and "Diagonal_Width_in_cm". Secondly, a Color Manager is used to assign color to different species to highlight species and promote information recall for further data visualization. Thirdly, we shuffle data to prevent any possible bias during the training by using a Shuffle node with 3122 random seed for reproduction of random shuffling. Lastly, we use the Partition node to divide the data set

to 2 partitions: 80% for the training set and 20% for the test set and we divide it by drawing randomly with the same random seed used in the Shuffle node. As there are 150 tuples in the source data, we can easily calculate the number of tuples for each partition of the data. Training set consists of 120 tuples, while the number of tuples for the test set is 30.

2. Answer to the assignment questions

Q 1.1: **KNIME Workflow**



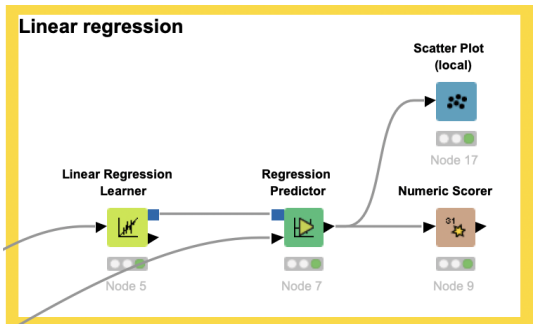
Q 1.2: There are a total of 120 tuples in the training set.

Q 1.3: There are 7 species included in the test set.

Q 1.4: “Whitefish” and “Smelt” have the same number of tuples (2) included in the test set.

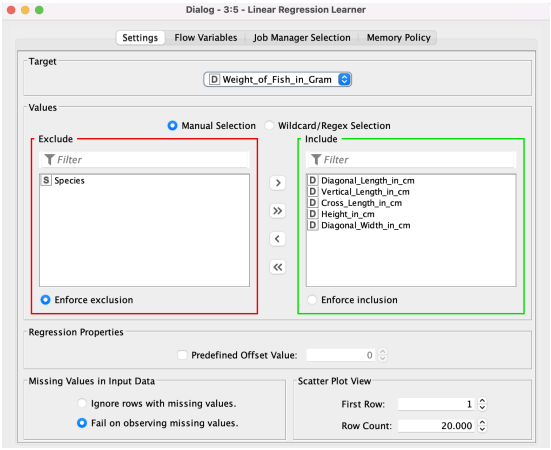
IV. Linear regression

1. Linear regression model building



To build a linear regression model, we use the Linear Regression Learner node for training the model and Regression Predictor node for data predicting. An addition of Numeric

Scorer node and Scatter Plot (local) node is used for evaluation of the test result after completion of training and testing the model. We use all possible attributes for training the model (Species is not considered as an attribute according to the description of the assignment). We split 80% of the source data for Learner while the Predictor gets the remaining 20%. The aim of the model is to predict "Weight_of_Fish_in_Gram" based on given attributes.



2. Predicted result evaluation

Row ID	Species	Weigh...	Diago...	Vertic...	Cross...	Height...	Diago...	Pre...
Row133	Pike	1,600	60	56	64	9.6	6.144	1,146.851
Row131	Pike	950	51.7	48.3	55.1	8.926	6.171	969.66
Row118	Perch	1,000	44	41.1	46.6	12.489	7.596	928.822
Row115	Perch	1,000	43	39.8	45.2	11.933	7.277	877.414
Row109	Perch	820	39	36.6	41.3	12.431	7.351	817.626
Row128	Pike	500	45	42	48	6.96	4.896	743.438
Row125	Pike	456	42.5	40	45.5	7.28	4.322	676.808
Row101	Perch	556	34.5	32	36.5	10.257	6.388	640.574
Row3	Bream	363	29	26.3	33.5	12.73	4.455	458.375
Row99	Perch	320	30	27.8	31.6	7.616	4.772	434.896
Row48	Whitefish	306	28	25.6	30.8	8.778	4.682	398.551
Row1	Bream	290	26.3	24	31.2	12.48	4.306	397.278
Row119	Pike	200	32.3	30	34.8	5.568	3.376	388.452
Row43	Roach	290	26	24	29.2	8.877	4.497	356.039
Row95	Perch	265	27.5	25.4	28.9	7.052	4.335	353.631
Row46	Whitefish	270	26	23.6	28.7	8.38	4.248	328.515
Row91	Perch	197	25.6	23.5	27	6.561	4.239	300.512
Row89	Perch	188	24.6	22.6	26.2	6.733	4.166	279.688
Row87	Perch	225	24	22	25.5	7.293	3.723	253.338
Row36	Roach	160	22.5	20.5	25.3	7.033	3.82	221.389
Row83	Perch	150	22.5	20.5	24	6.792	3.624	208.715
Row58	Parkki	170	20.7	19	23.2	9.396	3.41	202.007
Row33	Roach	120	21	19.4	23.7	6.115	3.294	159.478
Row74	Perch	115	21	19	22.5	5.918	3.308	148.569
Row55	Parkki	120	19	17.5	21.3	8.392	2.918	131.724
Row69	Perch	78	18.7	16.8	19.4	5.199	3.123	81.509
Row67	Perch	70	17.4	15.7	18.5	4.588	2.942	40.202
Row26	Roach	40	14.1	12.9	16.2	4.147	2.268	-59.355
Row144	Smelt	9.8	12	11.4	13.2	2.204	1.148	-172.479
Row136	Smelt	6.7	9.8	9.3	10.8	1.739	1.048	-229.868

From the predicted data table, we can clearly see that "Pike" has the highest predicted number of 1,146.851, while "Smelt" receives the lowest number of -229.868. There are 3 infeasible predictions in the predicted data as they are negative numbers ("Weight_of_Fish_in_Gram" should be in positive number).

We used a Numeric Scorer node which produces statistics between a numeric column's values (r_i) and predicted (p_i) values. It computes R^2, MSE, MAE, RMSE, MSD,

MAPE and adjusted R^2 for model evaluation. The R^2 value is 0.857, which means this model has a relatively high precision of its prediction. More accurate predictions are produced by models with higher R^2 values because these models have less errors.

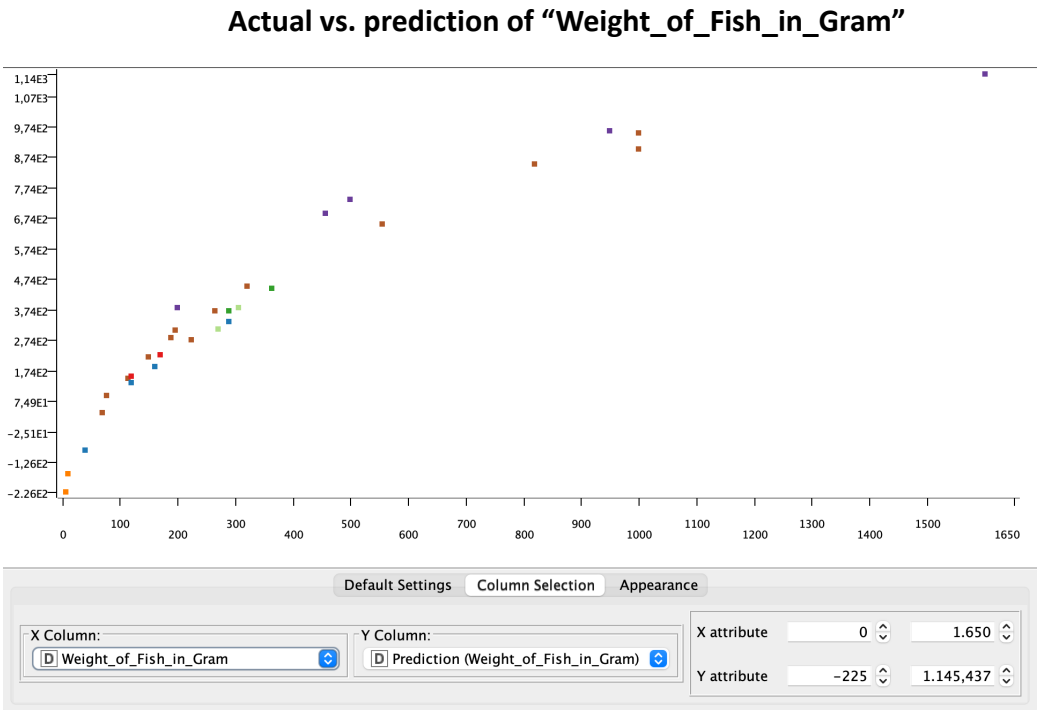
Statistic

R^2	0.857
MSE	101.02050625824289
MAE	18678.60341284359
RMSE	136.6696872493809
MSD	23.33801018808336
MAPE	2.118111058147091
Adjusted R^2	0.857

3. Answer to the assignment questions

Q 2.1. R^2 of the test result is 0.857.

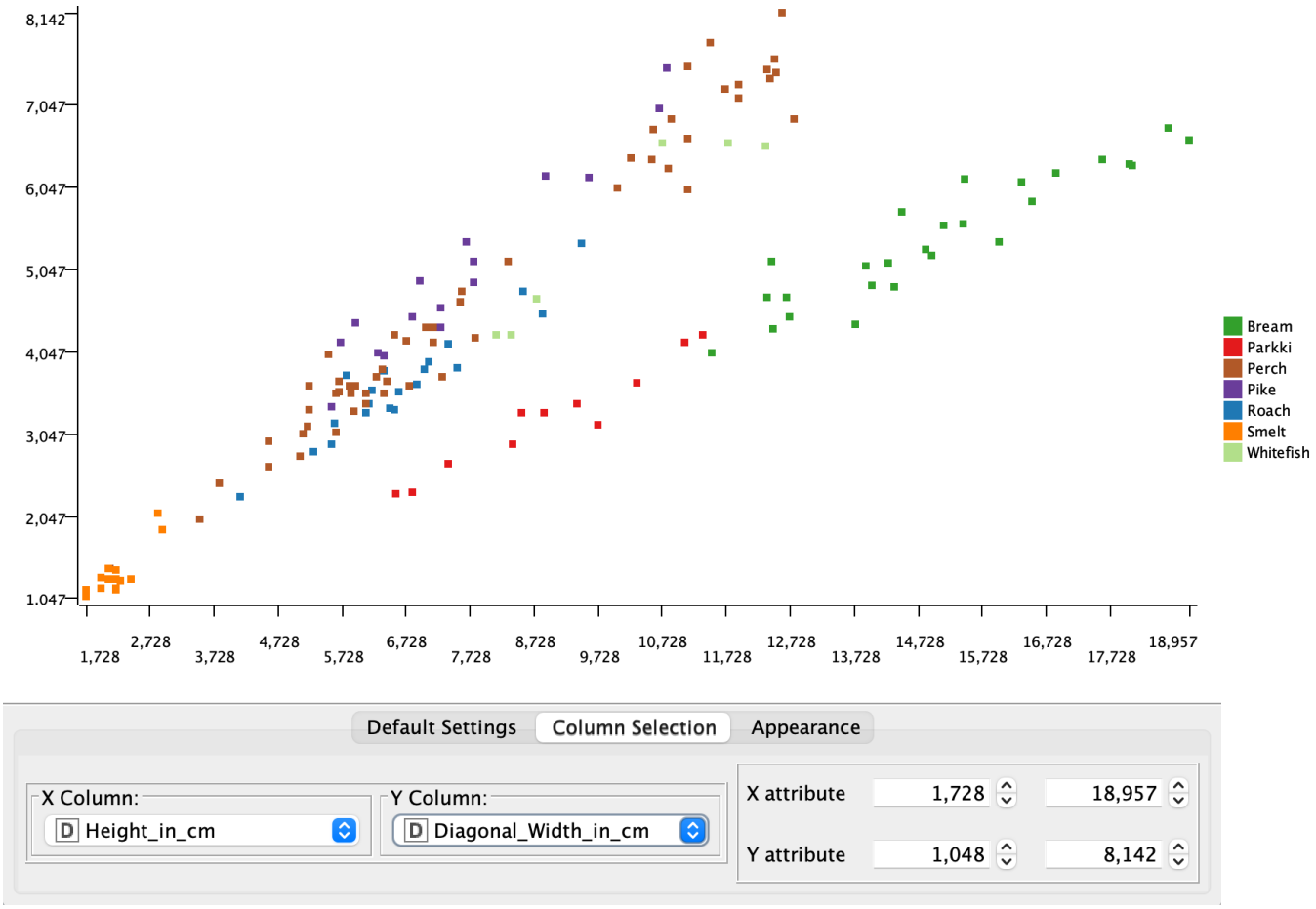
Q 2.2. Scatter plot results of the test output with "Weight_of_Fish_in_Gram" on the x-axis and the predicted value on the y-axis with assigned colors for each data point based on species name.



Q 2.3. "Pike" is the species that has the heaviest weight in the test result (1,146.851 gram).

Q 2.4. There are 3 infeasible predicted results in the test result, all of them are negative results.

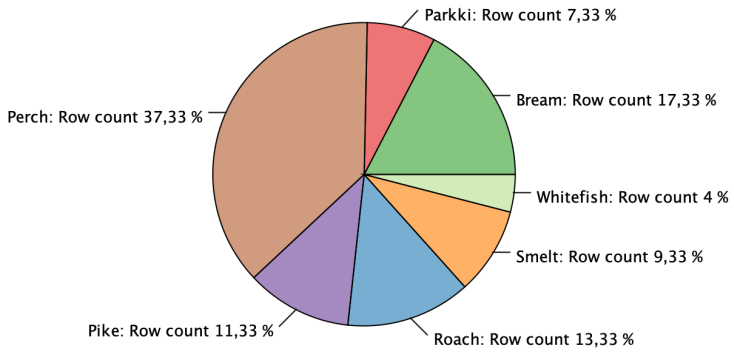
Q 2.5. **Scatterplot**



We can clearly distinguish "Bream" and "Smelt" from the rest of species just by looking at the scatter plot with attributes "Height_in_cm" and "Diagonal_Width_in_cm".

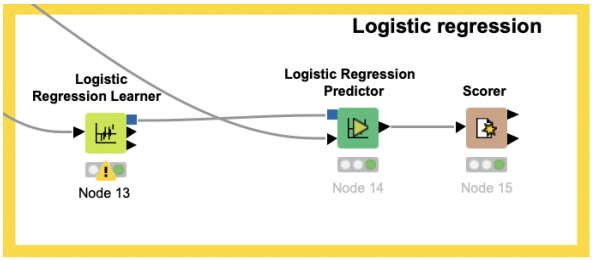
Q 2.6. An initial pie chart of the input data before it was divided into training and test sets. Each species has a different color and the pie chart also displays the percentage of the data.

Species distribution of original data



V. Logistic regression

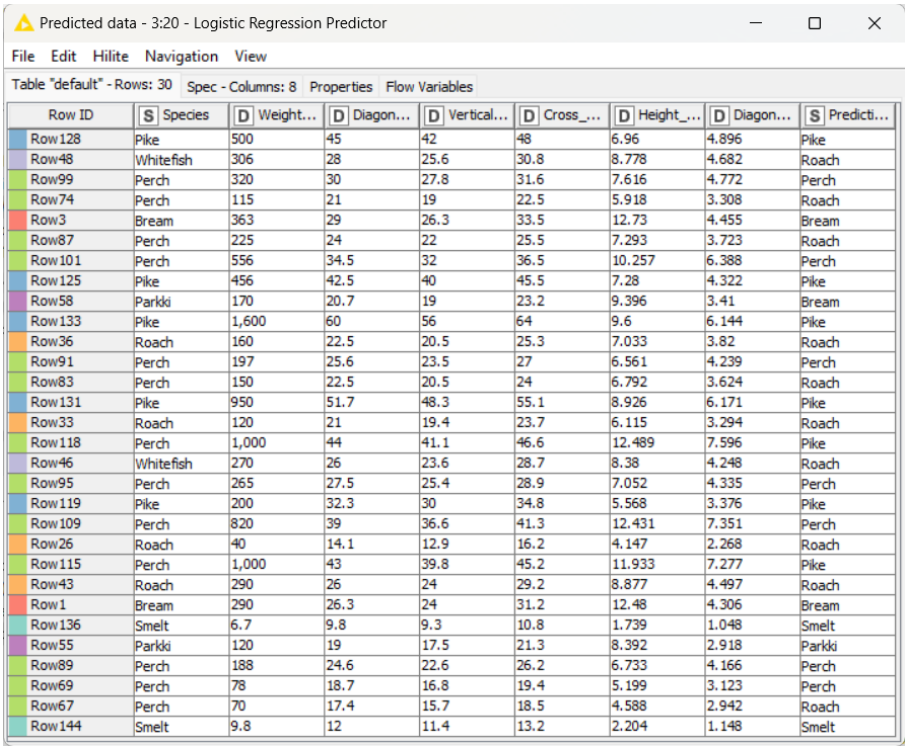
1. Logistic regression model building



Logistic regression model consists of a Logistic Regression Learner node and a Logistic Regression Predictor. Only Scorer node are used to evaluate the classification and show the confusion matrix. All attributes are used for model training, data splitting for Learner and Predictor is the same with Linear Regression model building in the previous part (80% for the training set and 20% for the test set). We use Smelt as a reference category and the termination points of epochs and epsilon are limited to 10.000 and 0.0001 respectively. A 3122 random seed is also used for data shuffling after each epoch completed during the model training which helps the training converge fast, and it also prevents any bias during the training as well as learning the order of the training of the model.

2. Predicted result evaluation

Overall accuracy of the prediction result is 70%, which is a decent accuracy for the prediction of the model.



Predicted data - 3:20 - Logistic Regression Predictor

File Edit Hilite Navigation View

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

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

3. Answer to the assignment questions

Q 3.1. "Whitefish" is the species that has no True Positive (TP) cases in the test result.

Q 3.2. For the species that has no TP case which is "Whitefish", "Roach" will be misplaced.

Q 3.3. Overall accuracy of the prediction result is 0.7 = 70%.

Q 3.4. Correctly classified test result means the sum of TP and TN, and correctly classified percentage is basically overall accuracy of the prediction. In this case, we use accuracy formula below to determine if a species has 100% correctly classified test result:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

The only species that has 100% correctly classified test results is "Smelt", which has 0 incorrectly classified test result (sum of FP and FN), turning the equation to $\frac{TP + TN}{TP + TN}$ which is always equal to 1 = 100%.

Q 3.5. To identify species name that has 50% chance of being misplaced into another species, which means 50% False Negative Rate, we will find species name that has the same Recall value based on this False Negative Rate formula below:

$$FNR = 1 - Recall \Leftrightarrow 50\% = 1 - Recall \Leftrightarrow Recall = 50\% = 0.5$$

The only species name that has Recall value equals to 0.5 is “Parkki”.

Q 3.6. The percentage that “Pike” is being misplaced into other species can be calculated using False Negative Rate formula below:

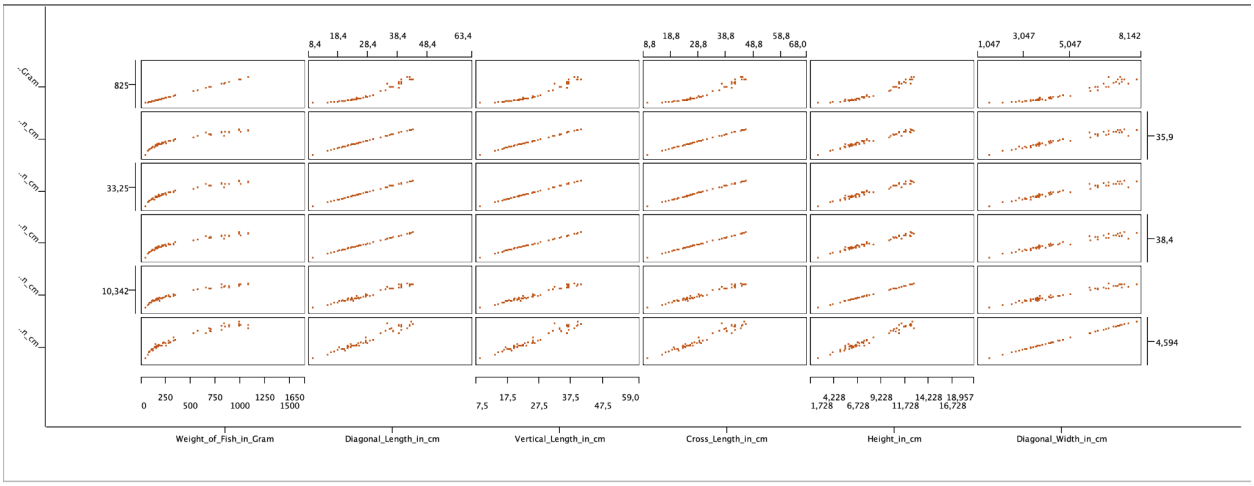
$$FNR = 1 - Recall = 1 - 1 = 0 = 0\%$$

VI. Performance Improvement

As mentioned before, more accurate predictions are produced by models with higher R^2 values because these models have less errors. The goal of the assignment is to improve the accuracy of the prediction result, which should result in a new model with higher R^2 value compared to the previous linear regression model.

Because this time we focus on specific species “Perch”, the Nominal Value Row Filter node is used to filter all tuples that include “Perch”. We placed those nodes into each of the Partition node’s output to ensure all tuples in the new training and test sets are fully the subset of the original training and test sets from the previous part of the assignment.

As we are reducing the dimension of the source data this time, input attributes must be chosen carefully to improve the model with better prediction accuracy. The number of input attributes are limited to 3, so 2 out of 5 input attributes must be eliminated. We use a Scatter Matrix (local) node to observe filtered data.



Q 4.1 Eliminated attributes with explanations:

- “Diagonal_Length_in_cm”: Very strong collinearity with number of attributes including “Vertical_Length_in_cm”, “Cross_Length_in_cm” and “Height_in_cm”,

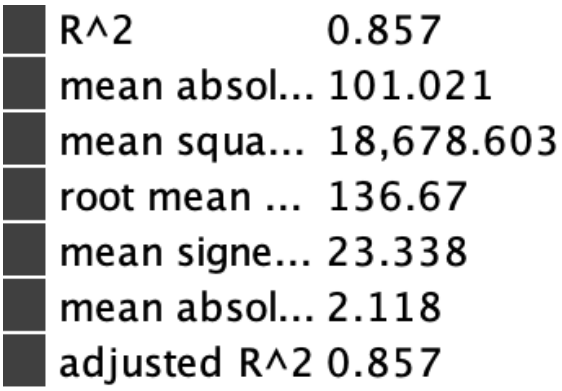
which decreases the statistical strength of the regression model by reducing the accuracy of the estimated coefficients.

- "Height_in_cm": This attribute also encounters the same issue as the mentioned attribute "Diagonal_Length_in_cm". It colinears with "Vertical_Length_in_cm", "Cross_Length_in_cm" and "Diagonal_Length_in_cm", creating multicollinearity which even worsen the accuracy of the regression model.

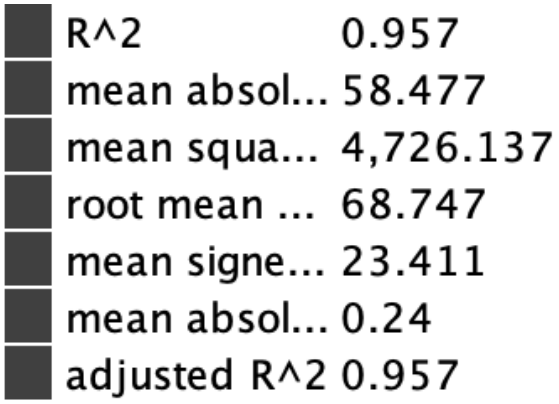
Based on given reasons, these attributes are eliminated, remaining attributes are "Vertical_Length_in_cm", "Cross_Length_in_cm" and "Diagonal_Width_in_cm".

Q 4.2

***R^2* comparison of 2 models**



First model



Second model

As was already established, a higher R^2 value indicates that the prediction findings are more accurate. By 0.1, the model's accuracy has increased. This outcome demonstrates that the model will be more accurate by removing some inappropriate attributes to minimize the dimension of the input data to train the model.

VII. Conclusion

The assignment's goals, which were to examine the data and create prediction models, have been met. By doing this assignment, I gained a lot of necessary skills of a data scientist such as preparing data for training, analyzing models, evaluating and improving existing models.