Cover sheet for submission of work for assessment

SWINBURNE UNIVERSITY OF TECHNOLOGY

UNIT DETAILS

Unit name	Data Science Principles			Class day/time	Wed, 8 – 12am	Office use only
Unit code	COS10022	Assignment no.	01	Due date	12/02/2023	
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	Given Name	Student ID Number
Hau	Linh Chi	104177160

DECLARATION AND STATEMENT OF AUTHORSHIP

1. For the sake of this evaluation, I have not impersonated anyone or let anyone else to impersonate me.
2. This evaluation is all original work from myself, with the exception of the places where proper credit has been given.
3. Except where such collaboration has been approved by the lecturer or instructor in question, no portion of this evaluation has been prepared for me by anyone else.
4. I have not previously submitted this work for this or any other course/unit.
5. I accept that my assessment response may be duplicated, shared, compared, stored, and used for benchmarking, plagiarism detection, or educational purposes.

I understand that:

6. Plagiarism is known as the act of presenting another person's work, idea, or creativity as your own. It is a sort of academic fraud that might get you kicked out of the university and is considered cheating. Written, graphic, and visual works, computer data, oral presentations, and other forms of presentation can all be used to source and deliver plagiarized content. When the source of the copied material is not properly cited, plagiarism occurs.

Student signature/s

I declare that I 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

PREDICTIVE MODEL CREATION & EVALUATION

I. ASSIGNMENT SUMMARY.

This assignment focuses on:

- Defining the key concepts, procedures, and tools involved in data management and prediction model construction.
- Working on choosing and implementing features and models for a data science project.
- In the KNIME analytical platform, the dataset is divided, and two models are built using linear and logistic regressions.
- Choosing the independent attributes, partitioning the data into training and test sets, developing an effective prediction model, and explaining the results.

II. INTRODUCTION.

The report is on the data about some commonly seen fish species in the market.

The dataset includes 150 tuples representing 7 fish species that are often purchased. The original data has 6 attributes in total.

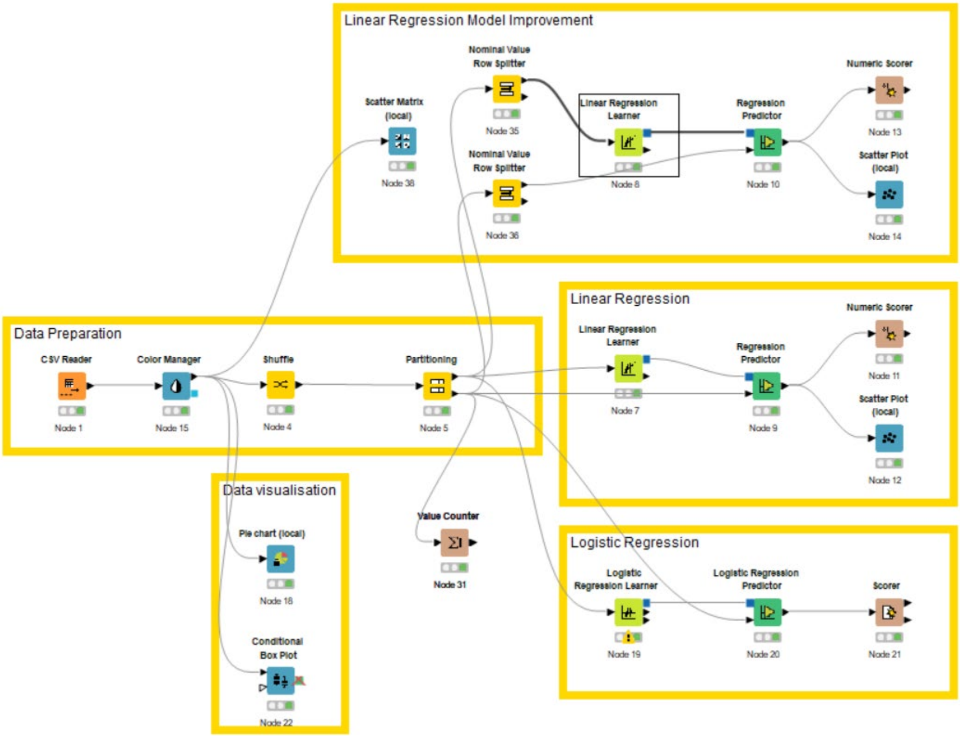
This assignment has two objectives: the first is to build a linear regression model to predict the weight of the fish, such as the value in the "Weight of Fish in Gram" attribute and the second is develop a logistic regression model to identify the fish species.

This report covers how I:

- Built the models (linear and logistic regression model) to predict the value and to classify fish species.
- Visualised the data.
- Chose and implemented features and models for this data project.
- Selected the independent attributes, divided the dataset into training and test sets, trained a usable predictive model, and explained the outputs.
- Improved the model.

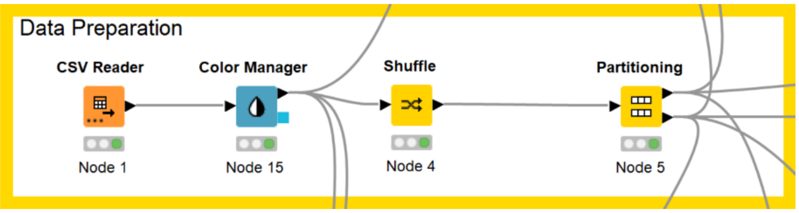
III. KNIME WORKFLOW

Question 1.1.



IV. DATA PREPARATION

1. Building models for prediction and classification.



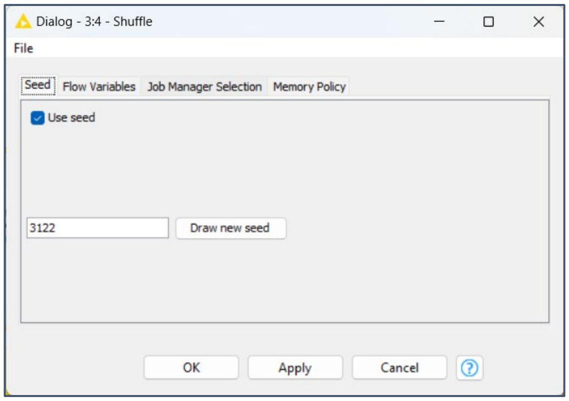
For data preparation, a total of 4 nodes are utilised: CSV Reader, Color Manager, Shuffle and Partitioning.

Data from the source file Fish Specises.csv, which contains 150 tuples with header rows and a total of 6 attributes (*"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")*, is imported to the CSV Reader Node.

To highlight various species and encourage information recall for additional data visualisation, a Color Manager Node is utilised.

2. Shuffling the dataset.

By using the Shuffle Node, the data is shuffled to eliminate any sampling bias in the algorithm. The dataset is being shuffled using the given seed of 3122 so that almost all of the output results match the right answers. The data set was shuffled before partitioning in order to guarantee that both the test and training data sets contain a wide range of data.

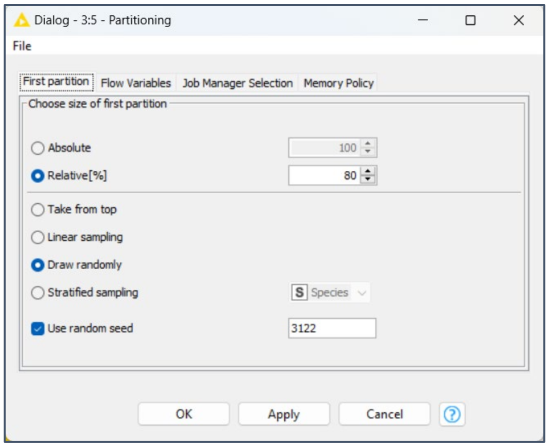


☑ Use seed

3122

3. Partitioning the dataset.

After shuffling the data, the Partitioning Node is used to divide the raw dataset in an 80:20 ratio, which means 80% of the raw data going to the training set and 20% going to the test set. The training and test data sets are separate from one another. The dataset is also being partitioned using the given seed of 3122 with the "draw randomly" method is applied. The training set is used to train the model in order to predict the output value from the test set.



Dialog - 3:5 - Partitioning

Choose size of first partition

○ Absolute 100

● Relative[%] 80

○ Take from top

○ Linear sampling

● Draw randomly

○ Stratified sampling S Species

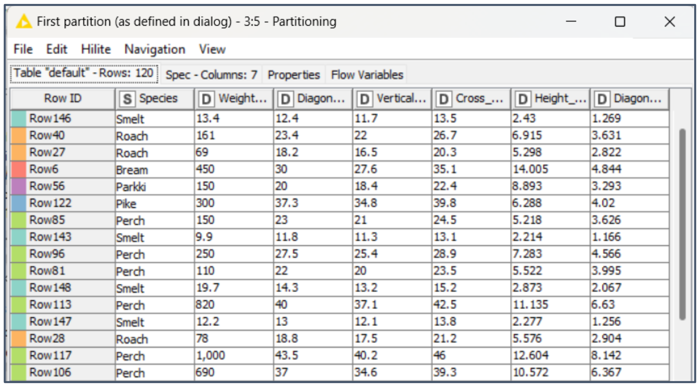
☑ Use random seed 3122

4. Test and training dataset.

Sample of the training dataset:

Question 1.2.

The training set contains 120 tuples (given that the source data contains 150 tuples, it is simple to calculate how many tuples are included in each data partition).

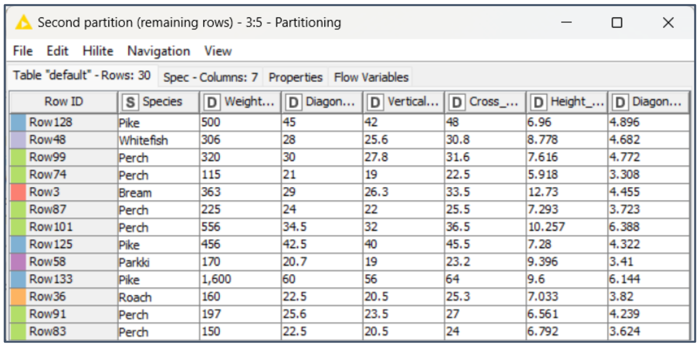


First partition (as defined in dialog) - 3:5 - Partitioning

Table "default" - Rows: 120 Spec - Columns: 7 Properties Flow Variables

Row ID	S Species	D Weight...	D Diagon...	D Vertical...	D Cross_...	D Height_...	D Diagon...
Row146	Smelt	13.4	12.4	11.7	13.5	2.43	1.269
Row40	Roach	161	23.4	22	26.7	6.915	3.631
Row27	Roach	69	18.2	16.5	20.3	5.298	2.822
Row6	Bream	450	30	27.6	35.1	14.005	4.844
Row56	Parkki	150	20	18.4	22.4	8.893	3.293
Row122	Pike	300	37.3	34.8	39.8	6.288	4.02
Row85	Perch	150	23	21	24.5	5.218	3.626
Row143	Smelt	9.9	11.8	11.3	13.1	2.214	1.166
Row96	Perch	250	27.5	25.4	28.9	7.283	4.566
Row81	Perch	110	22	20	23.5	5.522	3.995
Row148	Smelt	19.7	14.3	13.2	15.2	2.873	2.067
Row113	Perch	820	40	37.1	42.5	11.135	6.63
Row147	Smelt	12.2	13	12.1	13.8	2.277	1.256
Row28	Roach	78	18.8	17.5	21.2	5.576	2.904
Row117	Perch	1,000	43.5	40.2	46	12.604	8.142
Row106	Perch	690	37	34.6	39.3	10.572	6.367

Sample of the test dataset (input values):



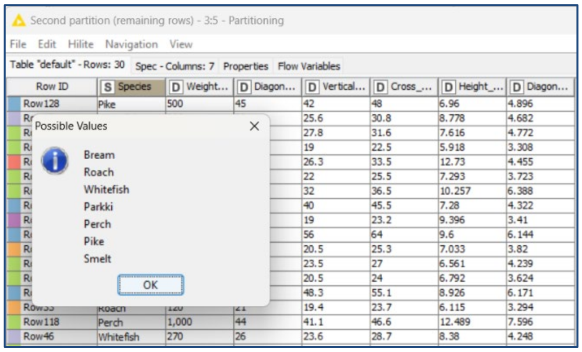
Second partition (remaining rows) - 3:5 - Partitioning

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...
Row128	Pike	500	45	42	48	6.96	4.896
Row48	Whitefish	306	28	25.6	30.8	8.778	4.682
Row99	Perch	320	30	27.8	31.6	7.616	4.772
Row74	Perch	115	21	19	22.5	5.918	3.308
Row3	Bream	363	29	26.3	33.5	12.73	4.455
Row87	Perch	225	24	22	25.5	7.293	3.723
Row101	Perch	556	34.5	32	36.5	10.257	6.388
Row125	Pike	456	42.5	40	45.5	7.28	4.322
Row58	Parkki	170	20.7	19	23.2	9.396	3.41
Row133	Pike	1,600	60	56	64	9.6	6.144
Row36	Roach	160	22.5	20.5	25.3	7.033	3.82
Row91	Perch	197	25.6	23.5	27	6.561	4.239
Row83	Perch	150	22.5	20.5	24	6.792	3.624

Question 1.3.

The test set contains all 7 species: *"Bream", "Roach", "Whitefish", "Parkki", "Perch", "Pike"* and *"Smelt".*



Second partition (remaining rows) - 3:5 - Partitioning

Table "default" - Rows: 30 Spec - Columns: 7 Properties Flow Variables

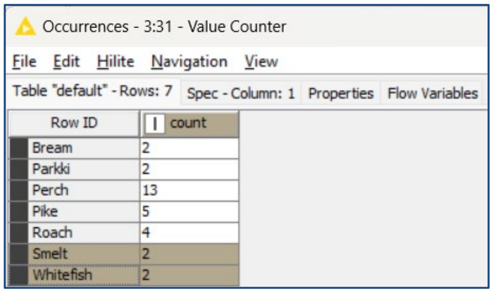
Row ID	S Species	D Weight...	D Diagon...	D Vertical...	D Cross_...	D Height_...	D Diagon...
Row128	Pike	500	45	42	48	6.96	4.896
R				25.6	30.8	8.778	4.682
R				27.8	31.6	7.616	4.772
R				19	22.5	5.918	3.308
R				26.3	33.5	12.73	4.455
R				22	25.5	7.293	3.723
R				32	36.5	10.257	6.388
R				40	45.5	7.28	4.322
R				19	23.2	9.396	3.41
R				56	64	9.6	6.144
R				20.5	25.3	7.033	3.82
R				23.5	27	6.561	4.239
R				20.5	24	6.792	3.624
R				48.3	55.1	8.926	6.171
Row33	Roach	120	21	19.4	23.7	6.115	3.294
Row118	Perch	1,000	44	41.1	46.6	12.489	7.596
Row46	Whitefish	270	26	23.6	28.7	8.38	4.248

Possible Values

Bream
Roach
Whitefish
Parkki
Perch
Pike
Smelt

Question 1.4.

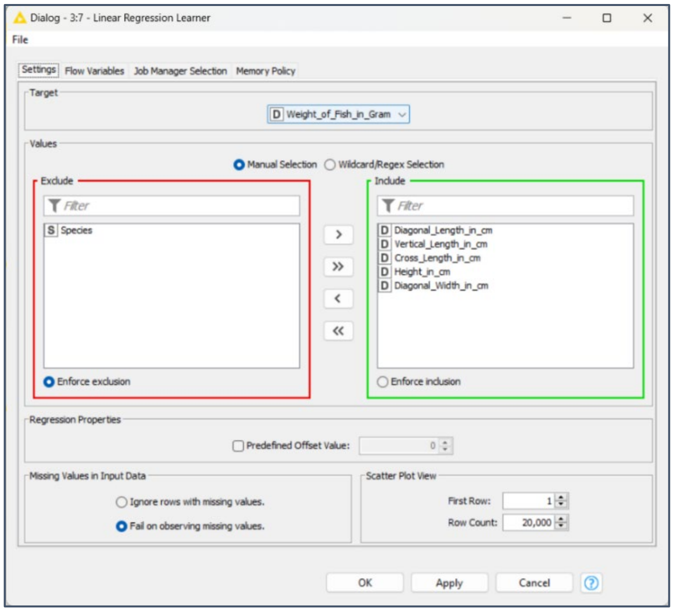
The number of species *"Whitefish"* and the number of species *"Smelt"* are the same, which is 2. They can be shown by using the Value Counter Node.


Occurrences - 3:31 - Value Counter
File Edit Hilite Navigation View
Table "default" - Rows: 7
Spec - Column: 1
Properties
Flow Variables

Row ID	I count
Bream	2
Parkki	2
Perch	13
Pike	5
Roach	4
Smelt	2
Whitefish	2

V. LINEAR REGRESSION

After having the training set and the test set divided, the model is trained using all available attributes, which are *"Diagonal_Length_in_cm"*, *"Vertical_Length_in_cm"*, *"Cross_Length_in_cm"*, *"Height_in_cm"* and *"Diagonal_Width_in_cm"* (*"Species"* does not count as an attribute) to predict the *"Weight_of_Fish_in_Gram"*.



Target: D Weight_of_Fish_in_Gram

Values: ● Manual Selection ○ Wildcard/Regex Selection

Exclude: S Species
● Enforce exclusion

Include:
D Diagonal_Length_in_cm
D Vertical_Length_in_cm
D Cross_Length_in_cm
D Height_in_cm
D Diagonal_Width_in_cm
○ Enforce inclusion

Regression Properties: ☐ Predefined Offset Value: 0

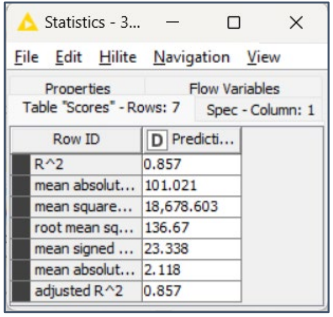
Missing Values in Input Data:
○ Ignore rows with missing values.
● Fail on observing missing values.

Scatter Plot View:
First Row: 1
Row Count: 20,000

Evaluating the results:

Question 2.1.

The test result R^2 value of a Linear Regression Model using all 6 available attributes is 0.857.

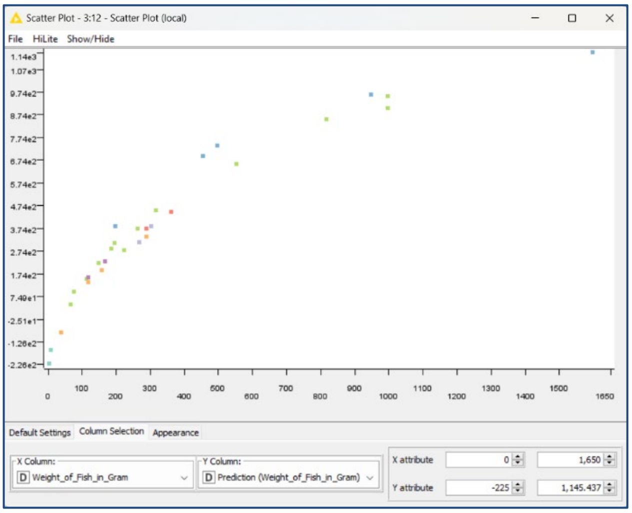


Row ID	D Predicti...
R^2	0.857
mean absolut...	101.021
mean square...	18,678.603
root mean sq...	136.67
mean signed ...	23.338
mean absolut...	2.118
adjusted R^2	0.857

Question 2.2.

The test output scatter plot with "Weight of Fish in Gram" on the x-axis and the prediction value on the y-axis. Different colours are given to the data points based on the *"Species"*.

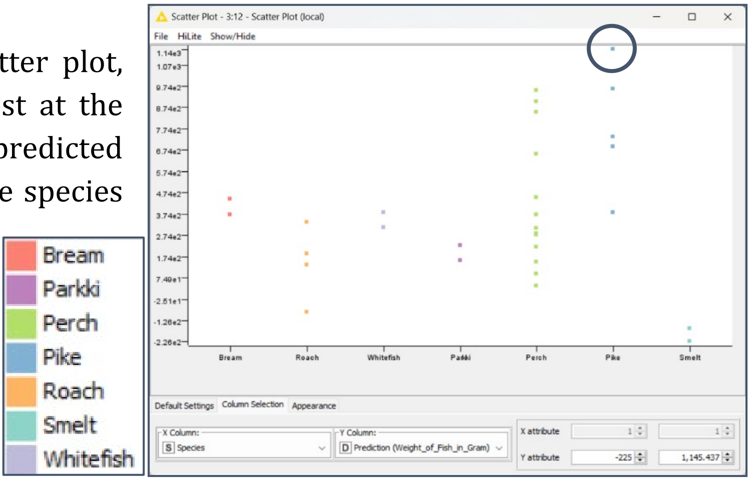
Bream
Parkki
Perch
Pike
Roach
Smelt
Whitefish



X Column: D Weight_of_Fish_in_Gram
Y Column: D Prediction (Weight_of_Fish_in_Gram)
X attribute: 0 – 1,650
Y attribute: -225 – 1,145.437

Question 2.3.

As can be observed from the scatter plot, the data point that appears highest at the top represents the heaviest predicted weight. So, the species "Pike" is the species with the heaviest projected weight, which is about 1.14e^3 gram.



X Column: Species

Y Column: Prediction (Weight_of_Fish_in_Gram)

X attribute 1 1

Y attribute -225 1,145.437

Question 2.4.

It can be clearly seen from the predicted data table that there are 3 prediction results being impracticable in the test result because they are negative numbers while the weight of a fish can never be a negative value (-225, -163.974 and -85.98).

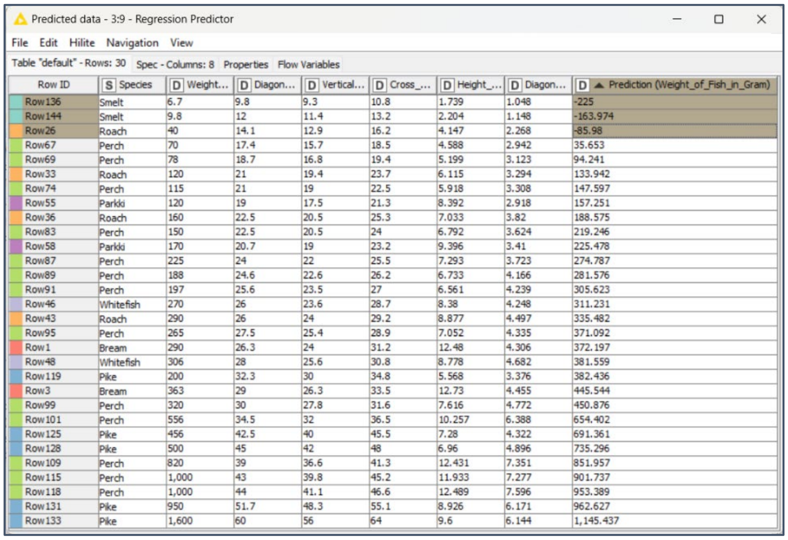

Predicted data - 3:9 - Regression Predictor
File Edit Hilite Navigation View
Spec - Columns: 8
Properties
Flow Variables

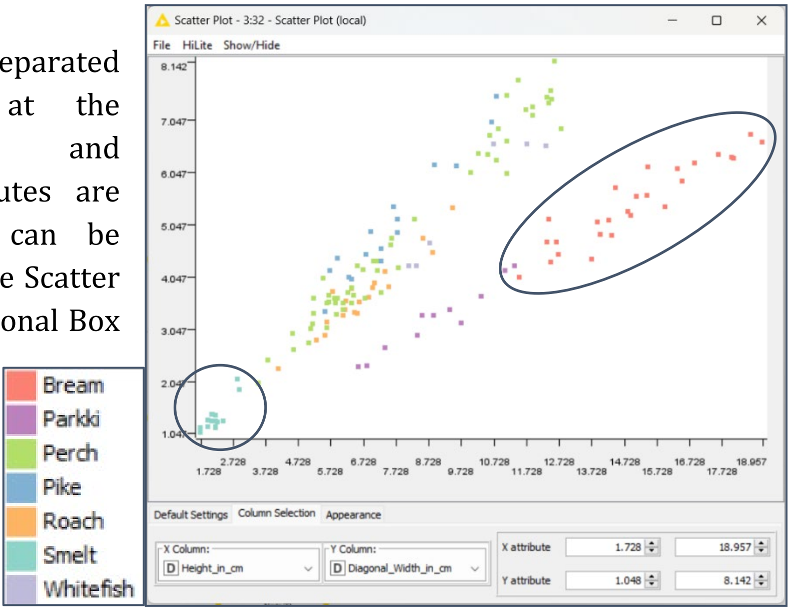
Table "default" - Rows: 30

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

VI. **DATA VISUALISATION** (original input data before being splitted).

Question 2.5.

Two species can be easily separated from others if looking at the *"Height_in_cm"* and *"Diagonal_Width_in_cm"* attributes are Bream and Smelt fish. It can be obviously observed by using the Scatter Plot (local) Node or the Conditional Box Plot Node with two selected attributes.

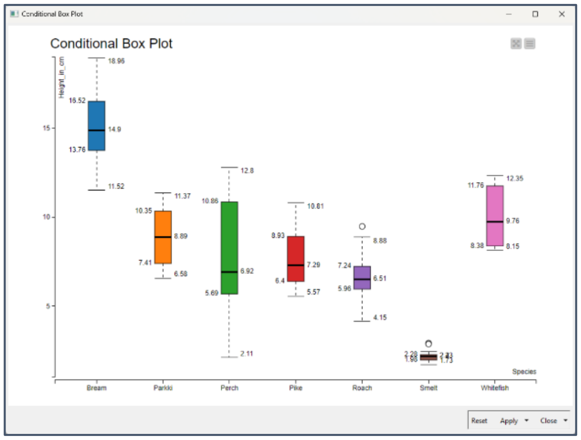


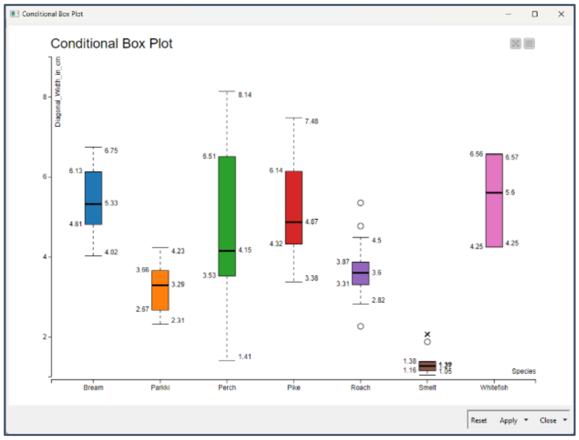
X Column: Height_in_cm

Y Column: Diagonal_Width_in_cm

X attribute 1.728 18.957

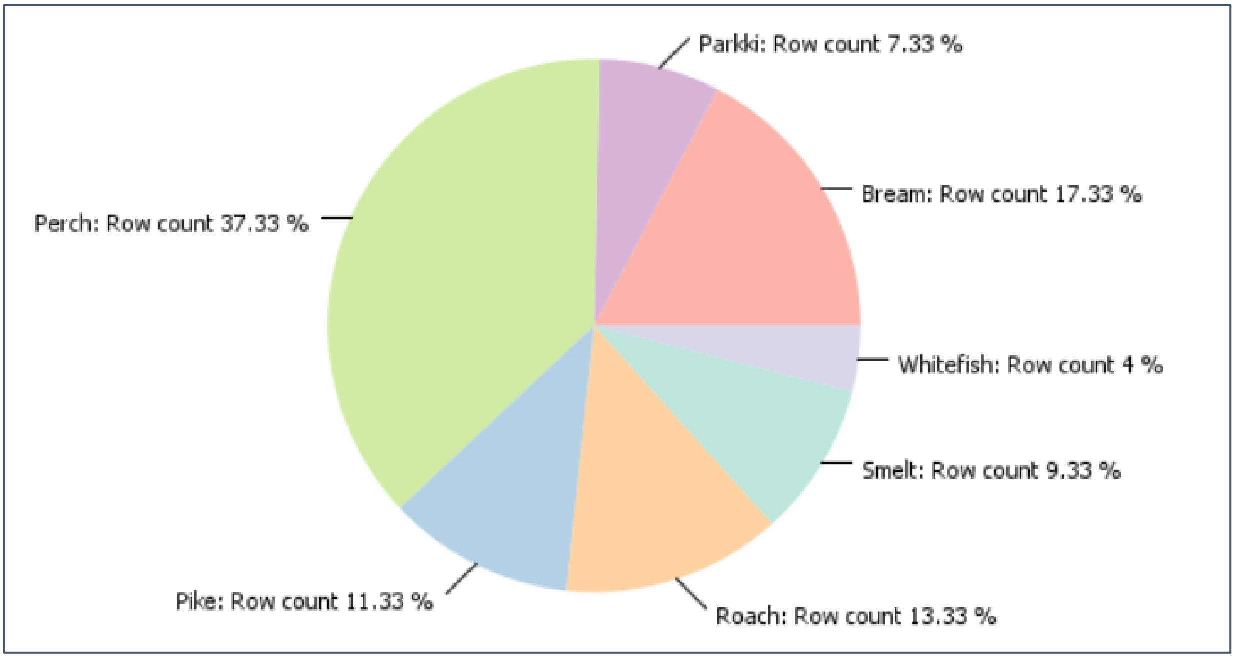
Y attribute 1.048 8.142





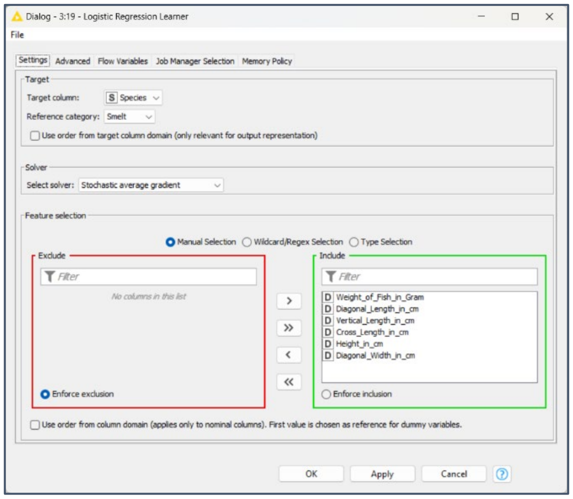
Question 2.6.

Pie chart (%) shows the distribution of species from the raw data with different colours are assigned for each of them:



VII. LOGISTIC REGRESSION

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 0.0001, respectively. In the Learner Node for Logistic Regression, the seed value is set at 3122.



Target

Target column: S Species

Reference category: Smelt

□ Use order from target column domain (only relevant for output representation)

Solver

Select solver: Stochastic average gradient

Feature selection

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

Filter

No columns in this list

● Enforce exclusion

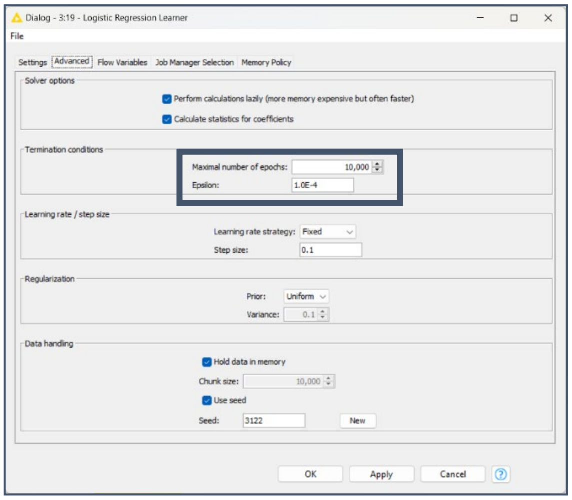
Include

Filter

D Weight_of_Fish_in_Gram
D Diagonal_Length_in_cm
D Vertical_Length_in_cm
D Cross_Length_in_cm
D Height_in_cm
D Diagonal_Width_in_cm

○ Enforce inclusion

□ Use order from column domain (applies only to nominal columns). First value is chosen as reference for dummy variables.



Solver options

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

☑ Calculate statistics for coefficients

Termination conditions

Maximal number of epochs: 10,000

Epsilon: 1.0E-4

Learning rate / step size

Learning rate strategy: Fixed

Step size: 0.1

Regularization

Prior: Uniform

Variance: 0.1

Data handling

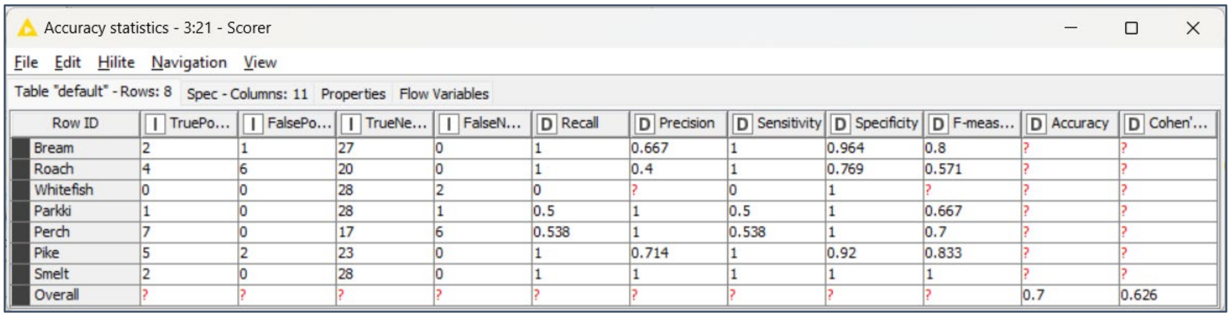
☑ Hold data in memory

Chunk size: 10,000

☑ Use seed

Seed: 3122

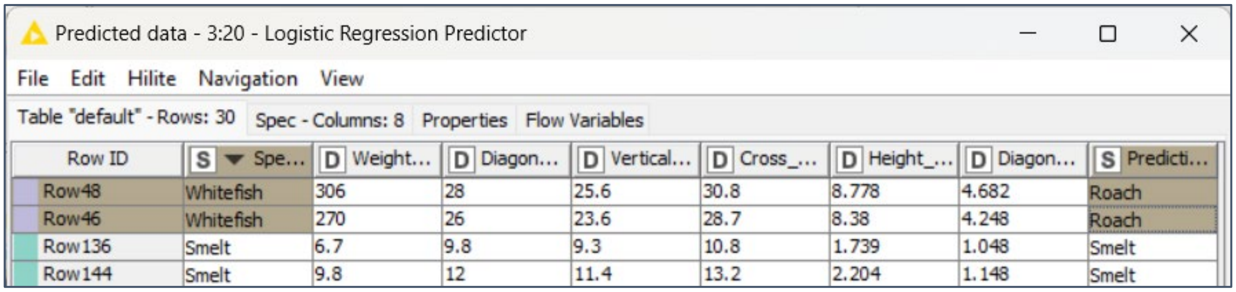
Evaluating the results:



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	1	27	0	1	0.667	1	0.964	0.8	?	?
Roach	4	6	20	0	1	0.4	1	0.769	0.571	?	?
Whitefish	0	0	28	2	0	?	0	1	?	?	?
Parkki	1	0	28	1	0.5	1	0.5	1	0.667	?	?
Perch	7	0	17	6	0.538	1	0.538	1	0.7	?	?
Pike	5	2	23	0	1	0.714	1	0.92	0.833	?	?
Smelt	2	0	28	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	0.7	0.626

Question 3.1.

"Whitefish" is the species that has no "True Positive" (TP) case in the prediction result.


Predicted data - 3:20 - Logistic Regression Predictor
File
Edit
Hilite
Navigation
View
Table "default" - Rows: 30
Spec - Columns: 8
Properties
Flow Variables

Row ID	S Spe...	D Weight...	D Diagon...	D Vertical...	D Cross_...	D Height_...	D Diagon...	S Predicti...
Row48	Whitefish	306	28	25.6	30.8	8.778	4.682	Roach
Row46	Whitefish	270	26	23.6	28.7	8.38	4.248	Roach
Row136	Smelt	6.7	9.8	9.3	10.8	1.739	1.048	Smelt
Row144	Smelt	9.8	12	11.4	13.2	2.204	1.148	Smelt

Question 3.2.

As it can be obviously observed from the predicted data table, the *"Whitefish"* species (with no TP case) is misplaced with *"Roach"* species.

Question 3.3.

The predicted results shows the overall accuracy of 0.7 (=70%).

Question 3.4.

A species having test results that are 100% correctly classified is one with an accuracy value of 1, which is the total of the True Positive and True Negative values over the total Positive and Negative values. The Accuracy value can be calculated using the below formula:

Accuracy = $\frac{\text{True Positive + True Negative}}{\text{True Positive + False Positive + True Negative + False Negative}}$

The species must not be misclassified in order for the Accuracy value to be 100%, hence the total of the False Positive and False Negative values must be 0.

Based on the statistics, species "Smelt" is the only species that is correctly classified.

Question 3.5.

The chance of being misplaced into another species means the possibility of a species A is identified incorrectly by the model to another species B but in actual, it is not and it belongs to species A. It is calculated using the False Negative Rate formula:

False Negative Rate = 1 – Sensitivity/Recall = $\frac{\text{False Negative}}{\text{True Positive + False Negative}}$

In order to get the possibility of being mistaken for a different species equal to 50%, the Sensitivity/Recall value have to be 0.5:

False Negative Rate = 50% = 1 – Sensitivity/Recall = $\frac{\text{False Negative}}{\text{True Positive + False Negative}}$ = 1 – 0.5 = $\frac{1}{1+1}$

Looking at the statistics, the only species that has a Sensitivity/Recall value equal to 0.5 is the species *"Parrki"*.

Question 3.6.

The percentage of the species *"Pike"* being misplaced into others means percentage of species *"Pike"* is identified incorrectly to another species. It is calculated using the False Negative Rate formula:

False Negative Rate = 1 – Sensitivity/Recall = $\frac{\text{False Negative}}{\text{True Positive + False Negative}}$ = 1 – 1 = $\frac{0}{0+5}$ = 0%

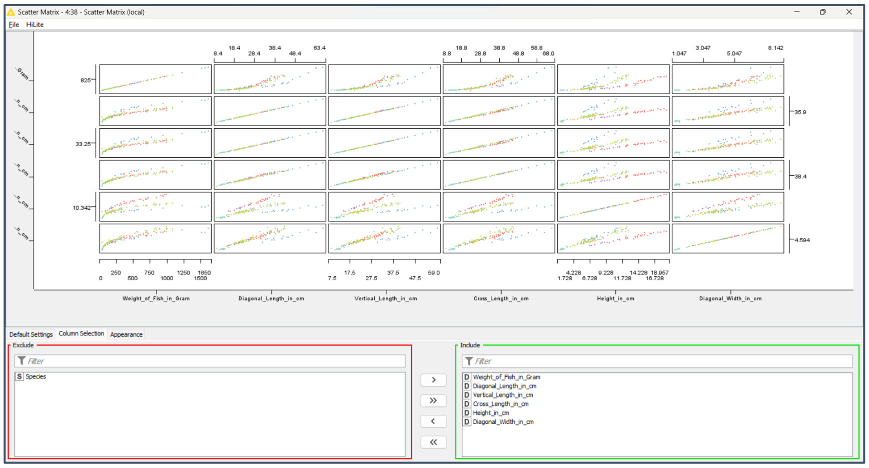
VIII. PERFORMANCE IMPROVEMENT

The model performance can be evaluated by looking at the R^2 value. The closer the R^2 number is to 1, the more variability of the response data around its mean is explained by the model, improving the accuracy of the prediction result. This value is a statistical measure of how close the data are to the fitted regression line.

As we have to focus on a single species of fish – *"Perch"*, two Nominal Value Row Splitter Node are used after partitioning the shuffled input data to ensure that all tuples in the new training and test sets of the split data are fully the subset of the original training and test sets.

In order to increase the R^2 number, the dimension of the model has to be reduced by removing 2 attributes and selecting only 3 most appropriate attributes. A Scatter Matrix (local) Node is utilised to observe and choose the right attributes from the original raw data.

Question 4.1.



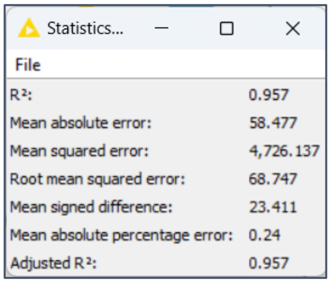
Default Settings | Column Selection | Appearance

Exclude
Filter
S Species

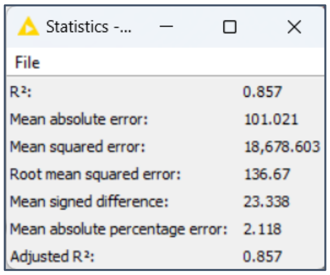
Include
Filter
D Weight_of_Fish_in_Gram
D Diagonal_Length_in_cm
D Vertical_Length_in_cm
D Cross_Length_in_cm
D Height_in_cm
D Diagonal_Width_in_cm

As it can be noticed from the scatter matrix, there are 2 unsuitable attribute that should be eliminated are *"Diagonal_Length_in_cm"* and *"Height_in_cm".* They are removed because the 3 remaining attributes, which are *"Vertical_Length_in_cm", "Cross_Length_in_cm"* and *"Diagonal_Width_in_cm"*, shows the strongest linear correlation with the *"Weight_of_Fish_in_Gram"* attribute and do not create the collinearity (reduces the regression model's statistical strength) with each other.

Question 4.2.



```
Statistics...
File
R²:                              0.957
Mean absolute error:             58.477
Mean squared error:              4,726.137
Root mean squared error:         68.747
Mean signed difference:          23.411
Mean absolute percentage error:  0.24
Adjusted R²:                     0.957
```



```
Statistics -...
File
R²:                              0.857
Mean absolute error:             101.021
Mean squared error:              18,678.603
Root mean squared error:         136.67
Mean signed difference:          23.338
Mean absolute percentage error:  2.118
Adjusted R²:                     0.857
```

As mentioned above, a higher R^2 value means that the accuracy of the prediction results is improved. The model accuracy is improved by 0.1. This result shows that by eliminating some unsuitable attributes to reduce the dimension of the input data to train the model, the model will be more accurate.

IX. CONCLUSION

The assignment's goals, which were to examine the data and create prediction and classification models, have been met.

To predict the fish's weight and classify fish species, a linear regression and a logistic regression model was created, respectively. Finally, a more accurate linear regression model was created by eliminating 2 unsuitable attribute and utilising only 3 attributes as the model's input with only one species is focused on.