**Hypothesis:**

**Different types of fish have different shapes and sizes then the species are predictable**

**Goal:**

* To outline the key ideas, strategies, and resources for working with data and creating predictive models.
* Use a logistic regression model to identify the species of fish, and a linear regression model to predict fish weight.
* Being able to explain the outputs

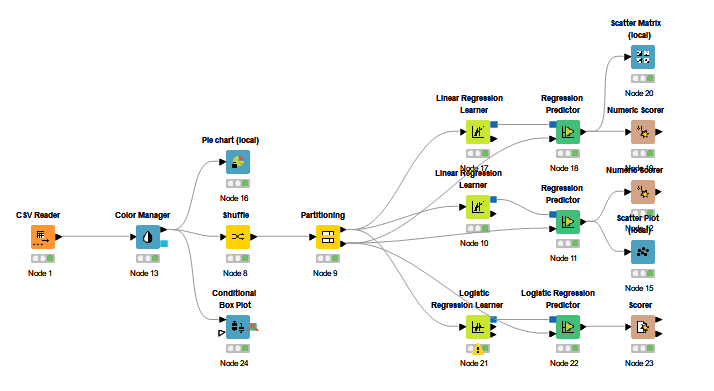
**Materials**

* KNIME Analytics Platform
* Fish\_Specises.csv

**Method**

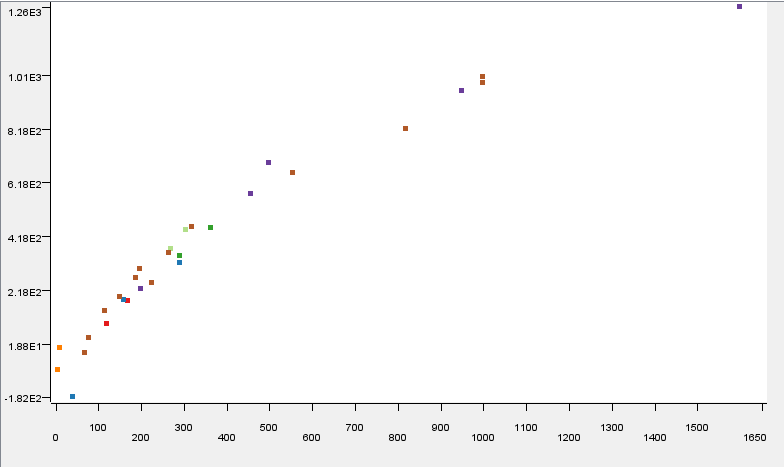
* Linear Regression Model in KNIME
* Logistic Regression Model in KNIME

**Procedure and Results:**



*\*The workflow..*

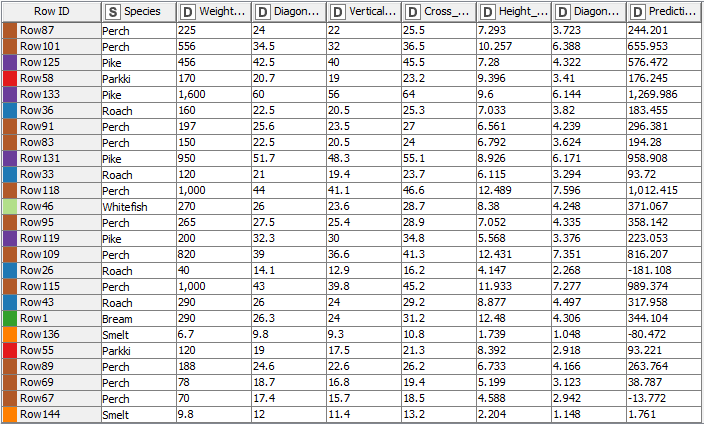
* The workflow includes 17 nodes which start with the CSV reader, where the Fish\_Specises.csv file is inserted. Subsequently, the Color Management node reads the file and assigns different colors to 7 species. Besides, the test set has 150 tuples.
* The amount of the Smelt tuples in the test set is not as that of the Whitefish tuples.
* After building a Linear Regression Model to predict the value of "Weight\_of\_the\_Fish\_in\_Gram", the R² value of the test result is 0.918 which can be seen in the "Numeric Scorer" node.



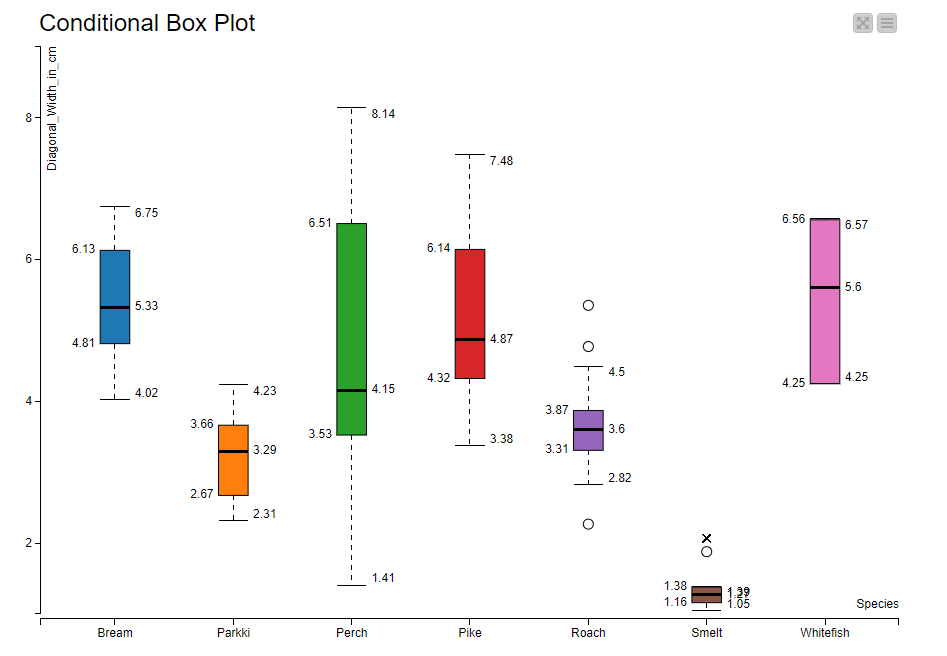
*\*The scatter plot result.*

*(The x-axis shows the "Weight of\_the\_Fish\_in\_Gram" while the y-axis represents the predicted value.)*

The screenshot indicates that the Pike is the heaviest species.

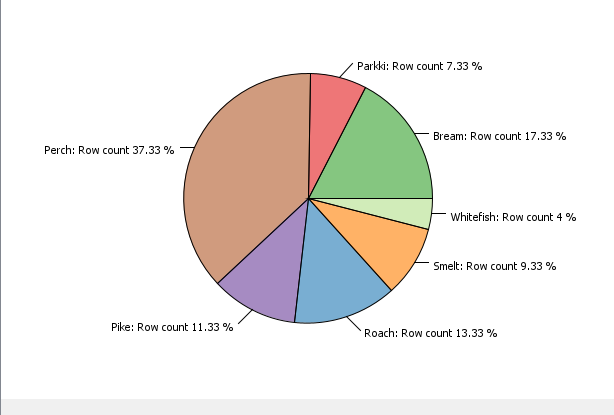


According to the data figure in the Regression Predictor above, There are three prediction outcomes that are infeasible in the test output (shown by the negative number in the prediction column) (row67, row26, row 136).



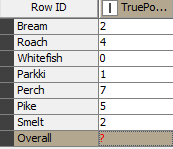
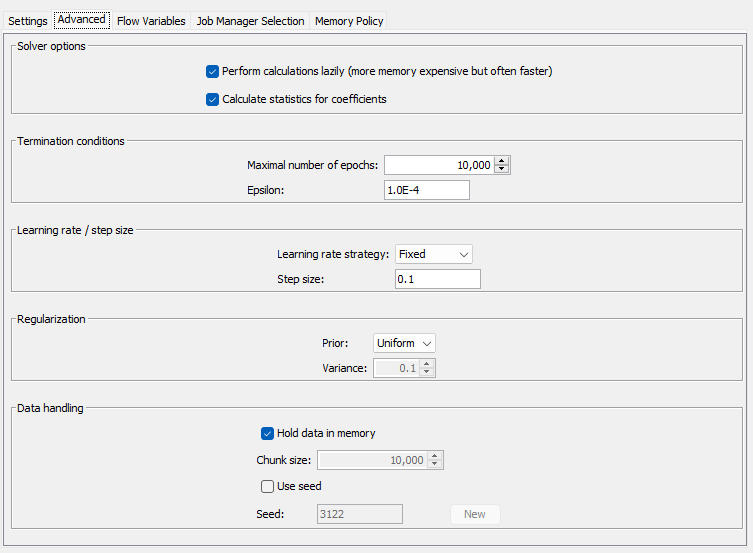
Applying the node "Conditional Box Plot", in term of the "Height\_in\_cm" and "Diagonal\_Width\_in cm" characteristics, it is apparent that the two species Bream(brown) and Perch(green) may be clearly distinguished from one another.

The visualization output is displayed above.

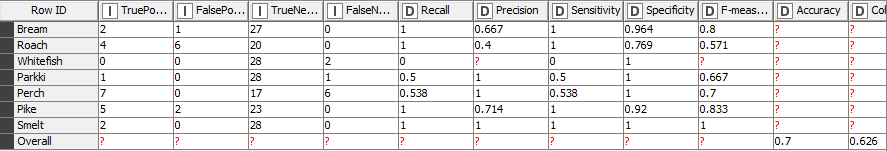


*\*The pie chart illustrates 7 species in 7 colors (unit: percentage)*

From the given case, the reference category is the Smelt and the maximum number of epochs and epsilon are 10000 and 0.0001, respectively. 3122 is used as a seed in the logistic regression model.



Whitefish type has no "TruePositive" case.

So the Whitefish might be misplaced. 

The output of the confusion matrix shows every value in the row and column of the Whitefish equals 0, except for the junction of the Whitefish and Roach type is 2. Thus, the Whitefish will be misplaced with the Roach species.

The overall accuracy number is 0.7 (70%), regarding the accuracy statistics table above.

The Bream, Roach, Pike, and Smelt are the four species that have 100% correctly classified test results. Additionally, Parkki is the species that has a 50%-possibility to be replaced by other species in the testing results.

It can be easily seen from the precision row that 28,6% of the Pike fish could be misplaced in comparison to others.

From the precision column, 28,6% (equals to 1 – 0.714) the Pike fish is misplaced into others.

