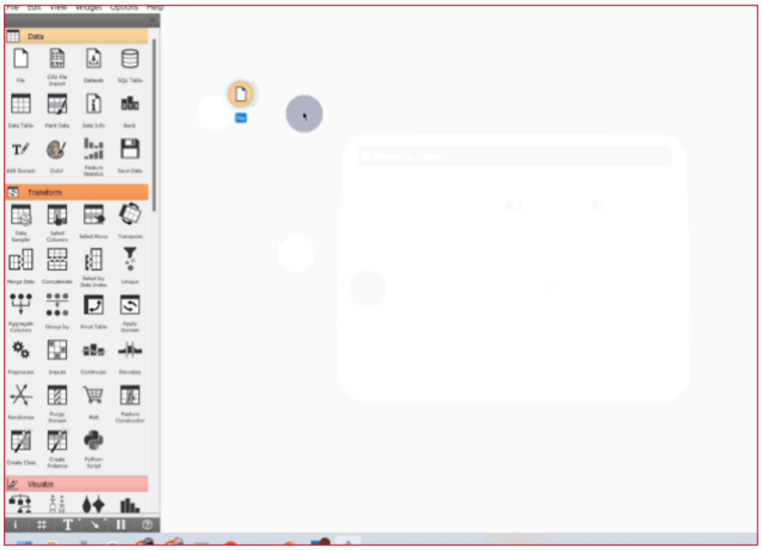
ORANGE DATA MINING

STEPS for AI model to predict the penguin species .

Step 1: Upload Dataset

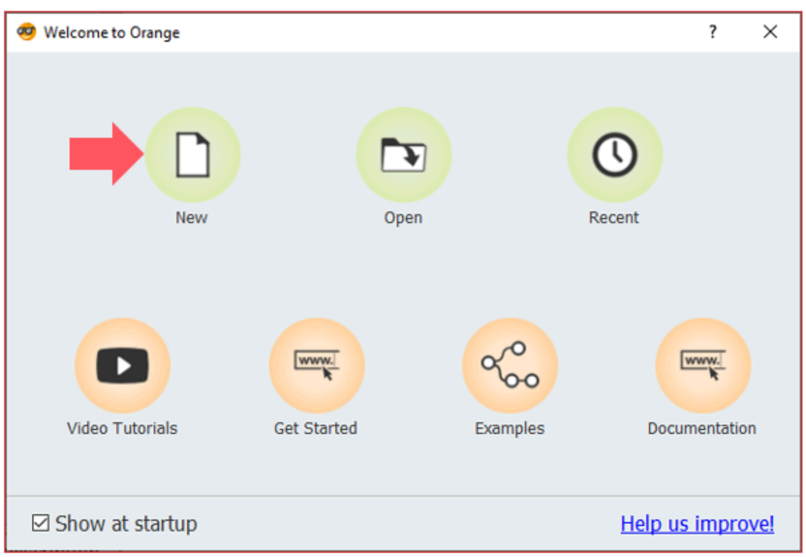
Data Acquisition



Step 1(a)

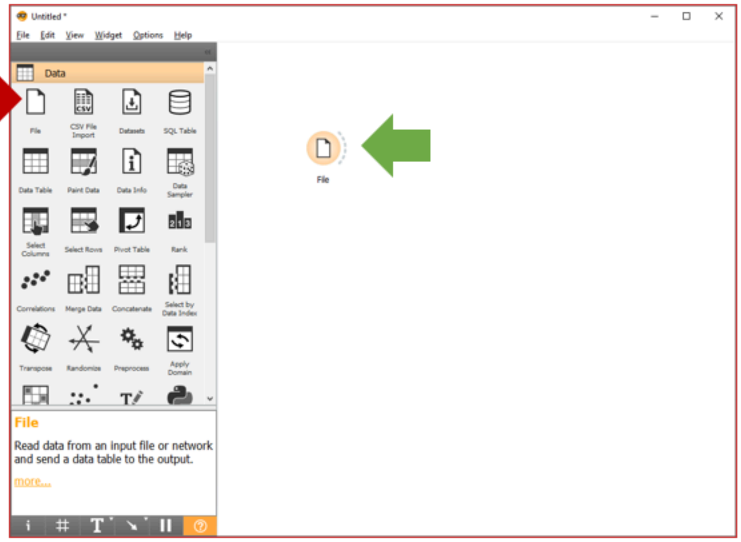
This is the Welcome page that you will see when you first launch Orange.

Click on New.

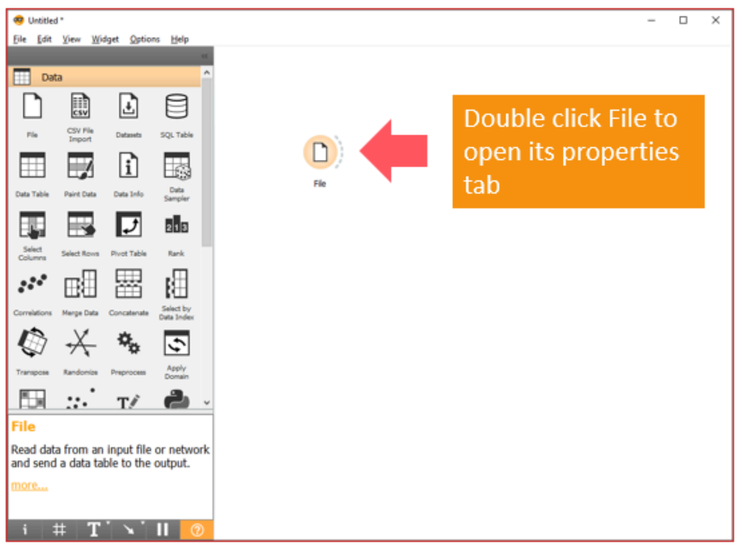


Step 1(b)

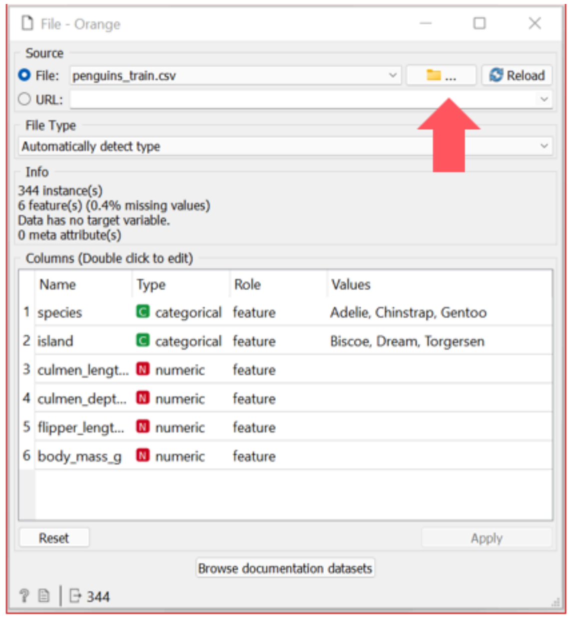
Insert the File widget onto the canvas



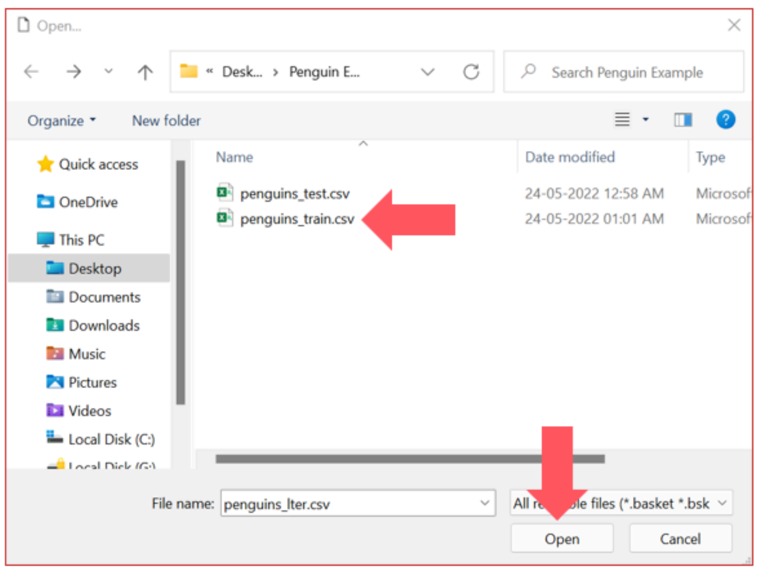
Step 1(c)



Step 1(d)

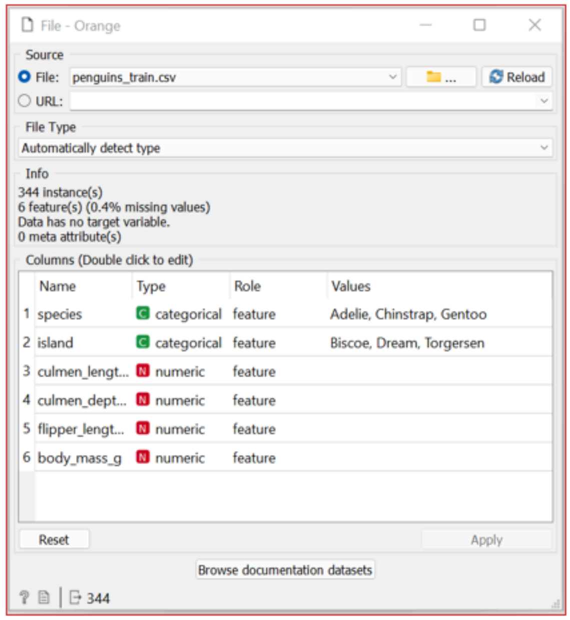


Step 1(e)



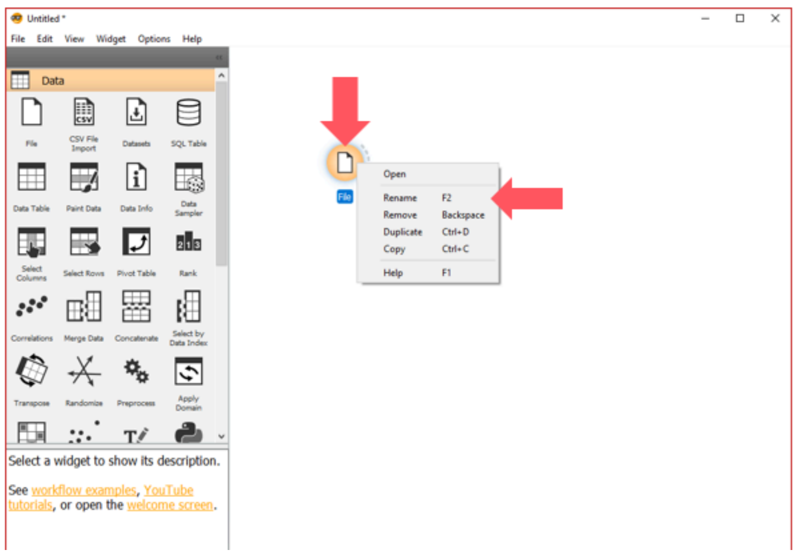
Step 1(f)

penguins_train.csv is uploaded

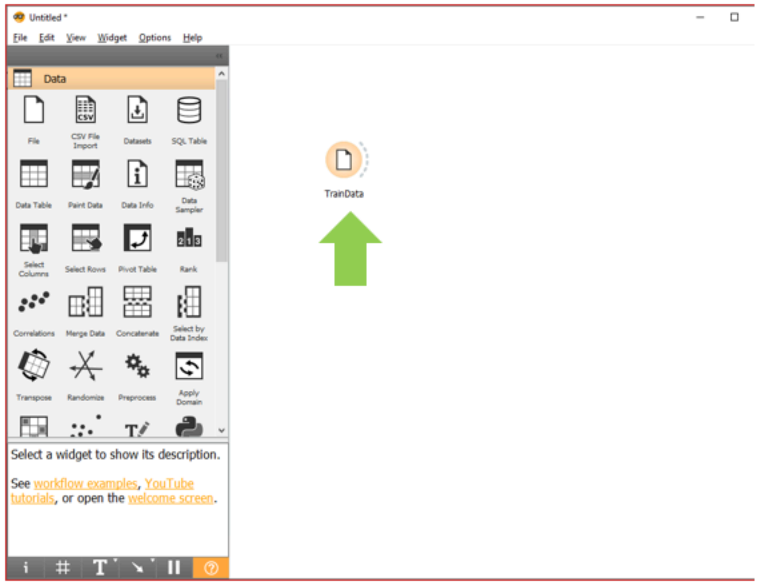


Step 1(g)

Right-click on the widget File and click on Rename.
We will rename it to 'Train Data' so that we do not confuse it with the testing data later

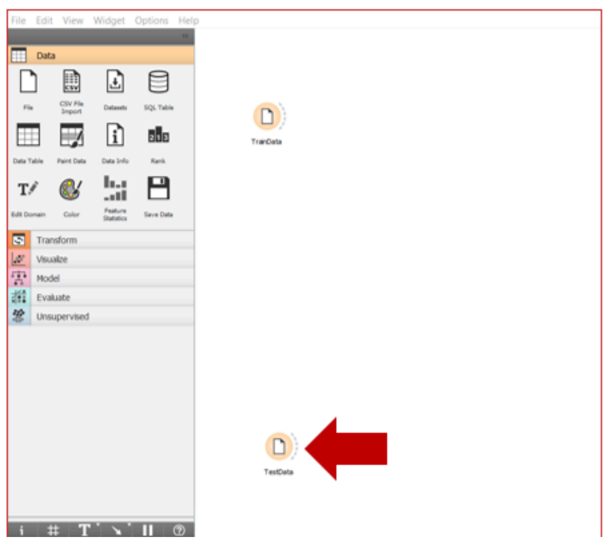


Step 1(h)



The file name for the uploaded dataset has changed

Step 1(i)

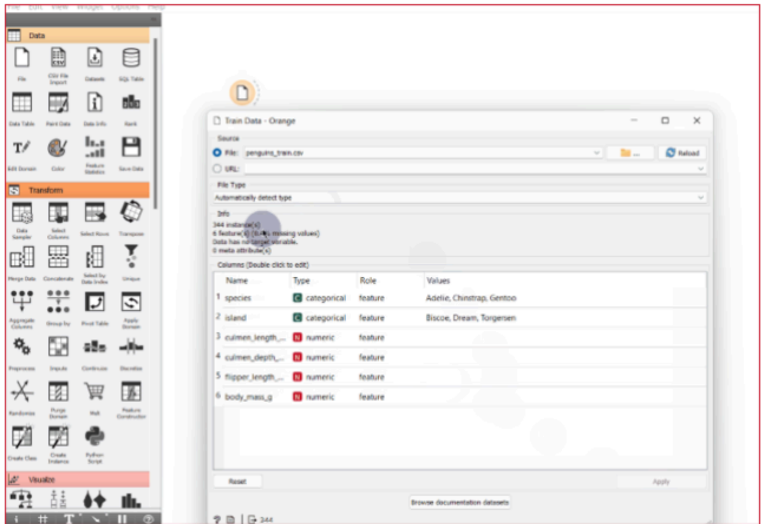


Repeat the same steps for 'Test Data' and upload testing dataset in the file widget

After Data Acquisition, what should we do next?

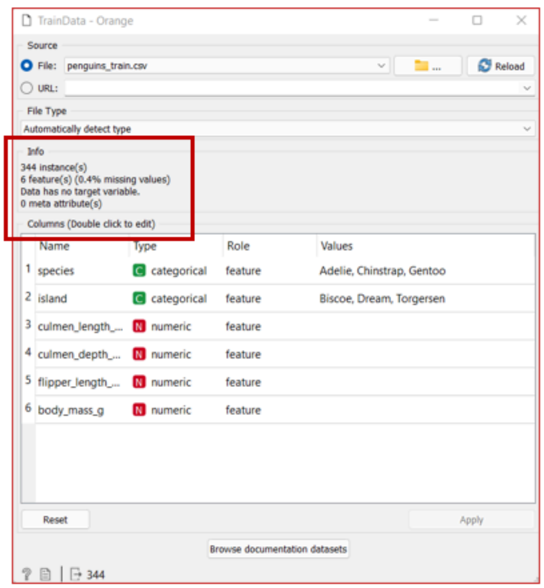
Data Exploration

Step 2: Clean Missing Data



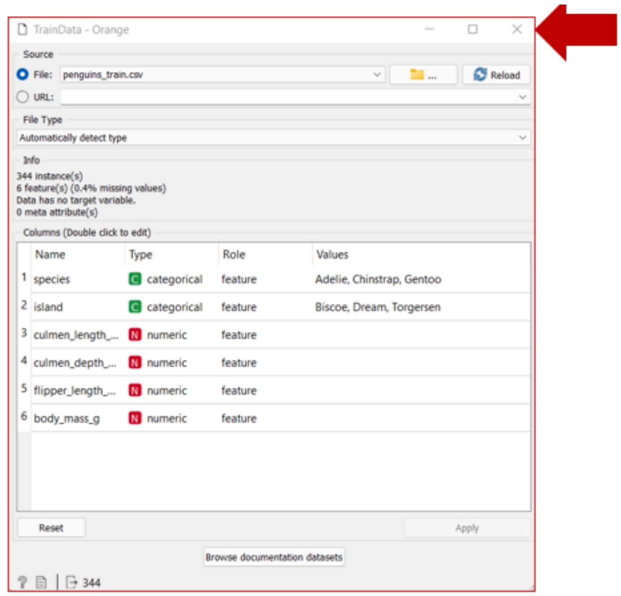
Step 2(a)

Check if there are any missing values
Notice that there are some missing values



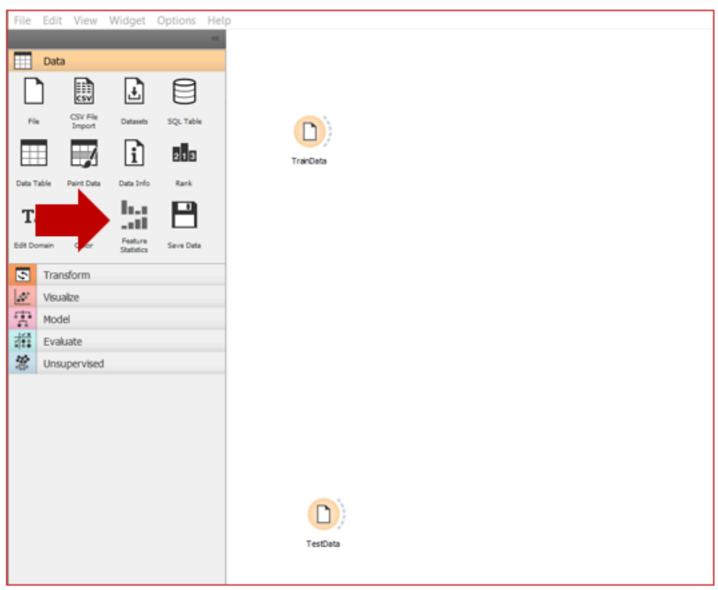
Step 2(b)

We will now look at another way to inspect on missing data.
Click X to close the pop up.



Step 2(c)

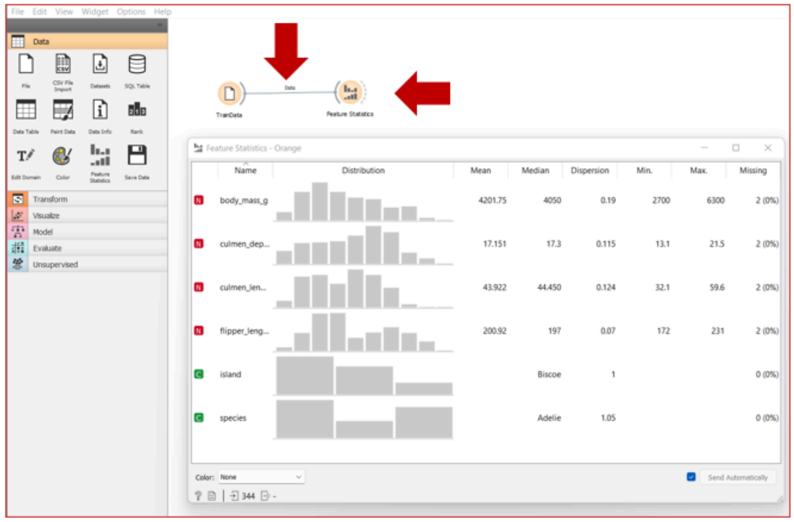
Insert the widget Feature Statistics onto the canvas



Step 2(d)

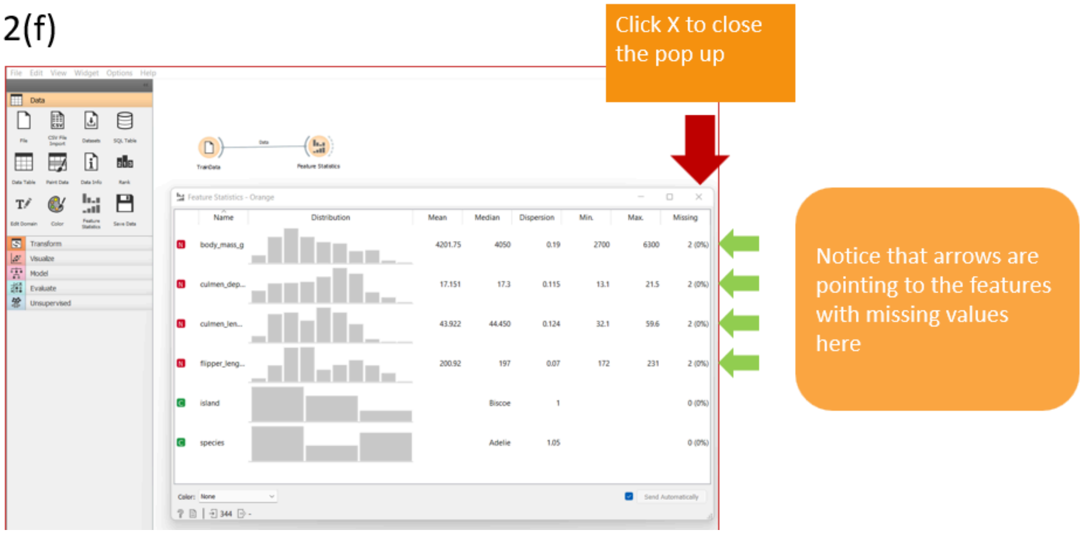
1. Connect widget 'Train Data' to widget Feature Statistics. We can do that by dragging the output from 'Train Data' to the input of Feature Statistics.

2. Double-click on Feature Statistics to see the results.

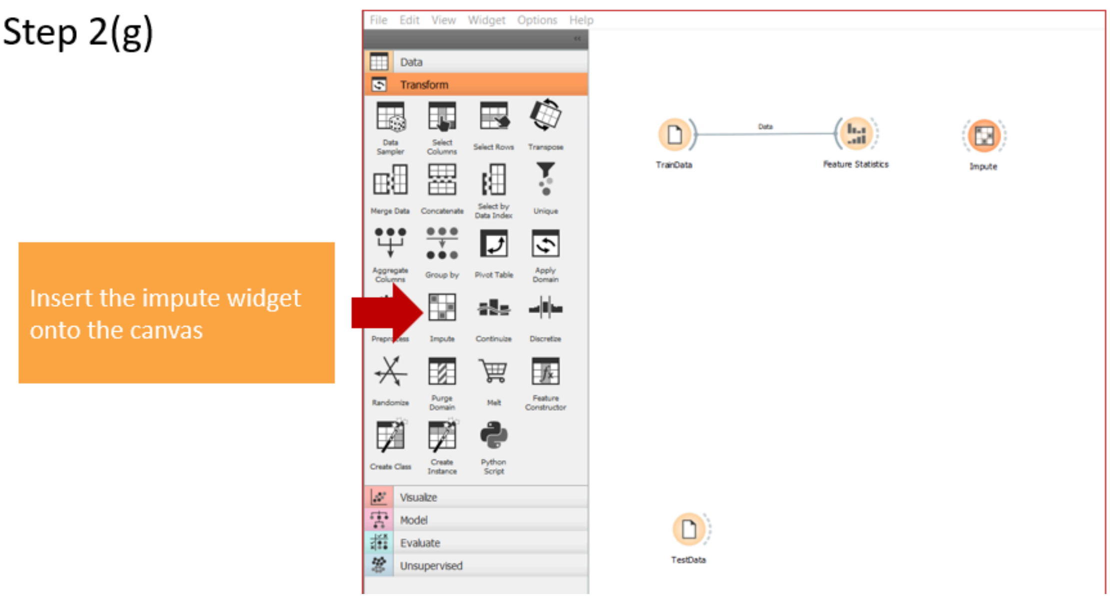


What is the mean value of culmen_length feature?

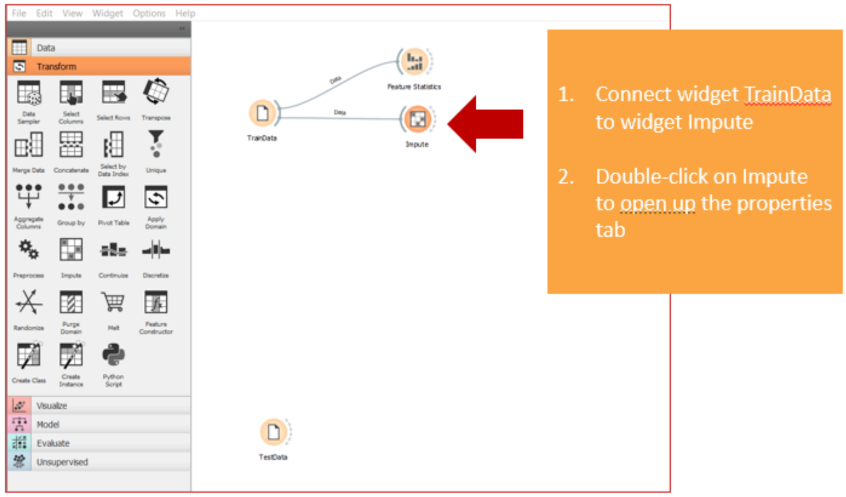
Step 2(f)



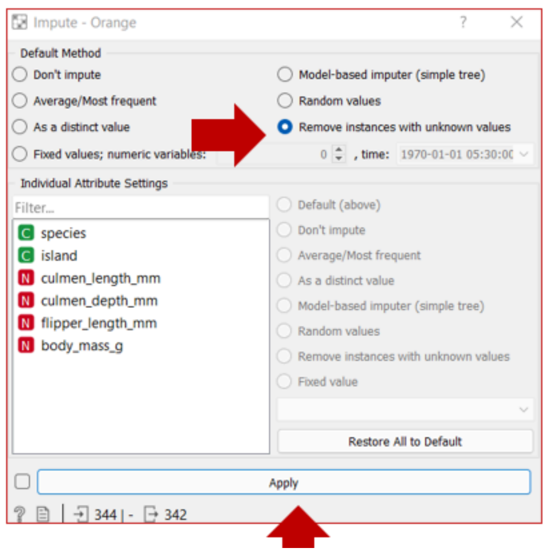
Step 2(g)



Step 2(h)



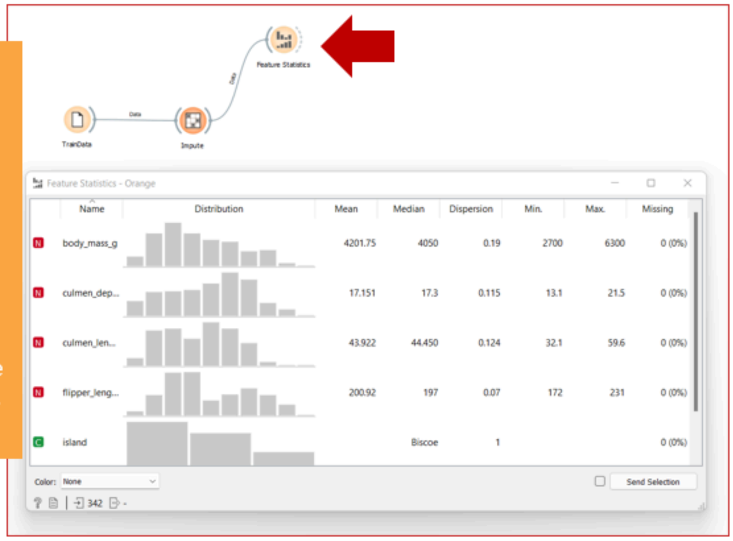
Step 2(i)



Step 2(j)

1. Connect the output of Impute to the input of the existing Feature Statistics (the previous connection between TrainData and Feature Statistics has been removed because only accepts one input)

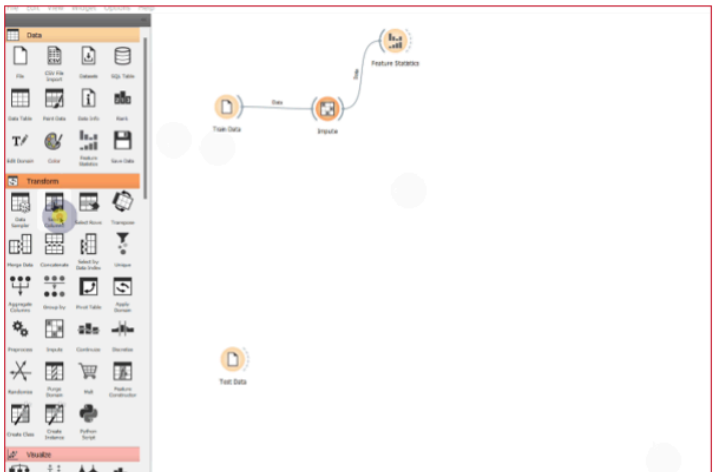
2. Double-click on the Feature Statistics to see the output.



Now that the data is clean and without any missing values, what next?

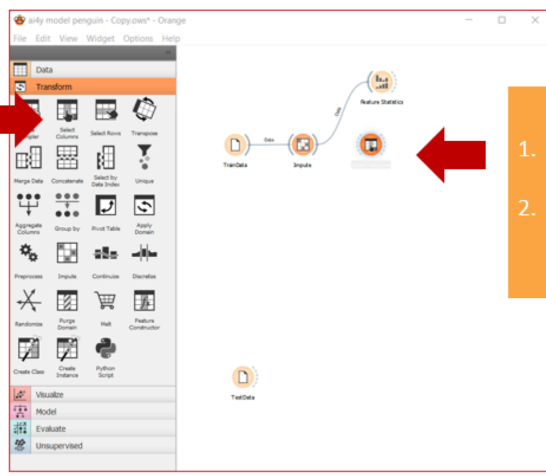
Step 3: Select Target Label

Data Exploration



Step 3(a)

Insert the Select columns widget onto the canvas



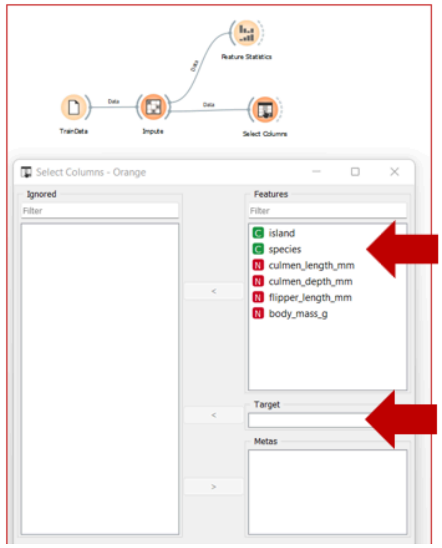
1. Connect with Impute widget
2. Double click on the widget

From TrainData, you would have noticed that the Feature Type for most of the columns is Numeric Feature. In supervised learning models, we have both the features and the labels. The labels are the output. Therefore, we need to define an output for our Palmer Penguin model. We will assign species as our label since that is what we want to identify.

Therefore, we will change the Feature Type for species, from Categorical Feature to Categorical Label. To do that, we will be using Select Columns.

Step 3(b)

A window displaying all the features will appear



Drag the ‘species’ feature to the ‘Target’ box

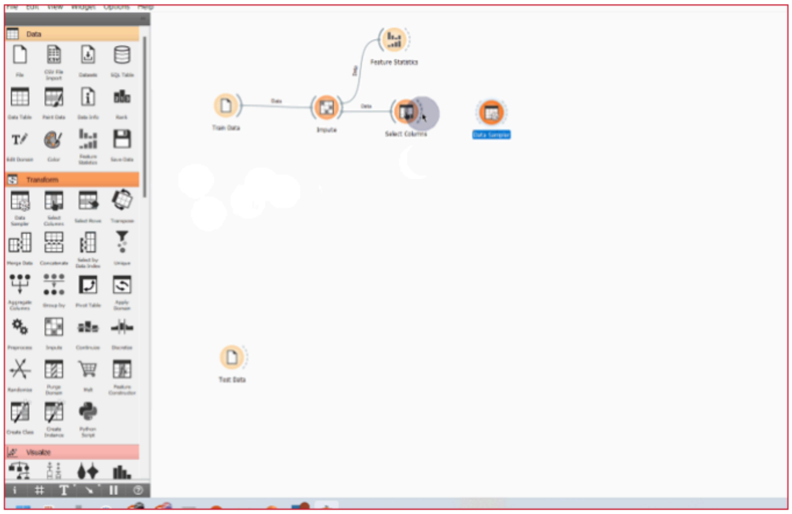
Step 3(b)

‘species’ is the Target label

After choosing a target label, we need to split the data

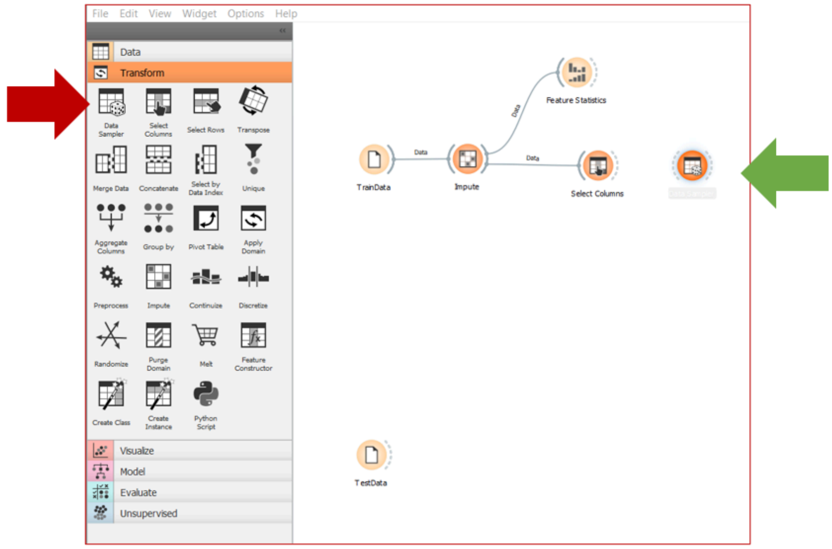
Step 4: Data Sampler

Data Exploration

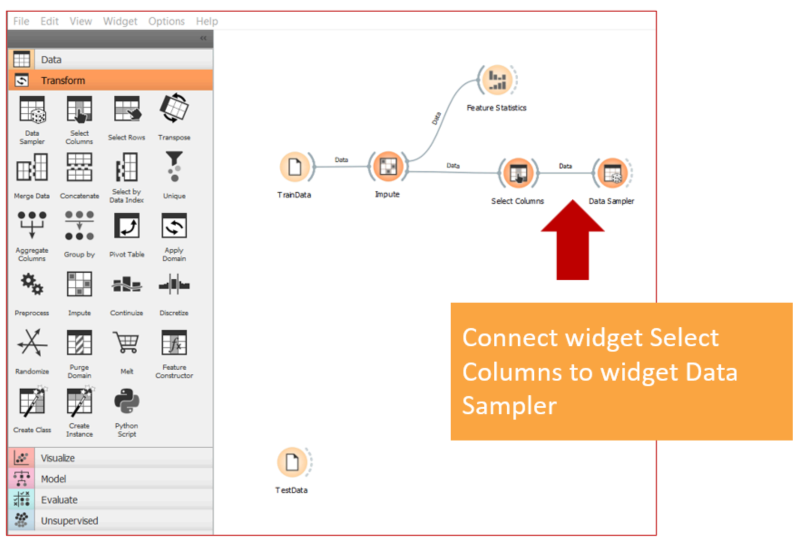


Step 4(a)

Insert the Data Sampler widget onto the canvas



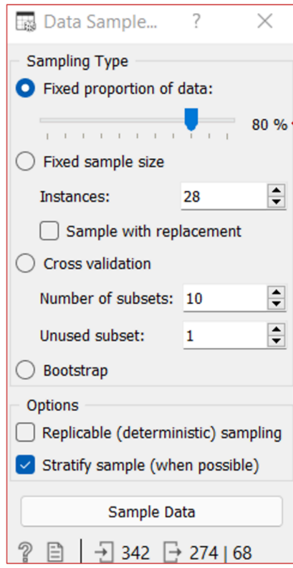
Step 4(b)



Connect widget Select Columns to widget Data Sampler. We can do that by dragging the output from Select Columns to the input of Data Sampler.

After the connection is made, double-click on Data Sampler to open the properties tab.

Step 4(c)

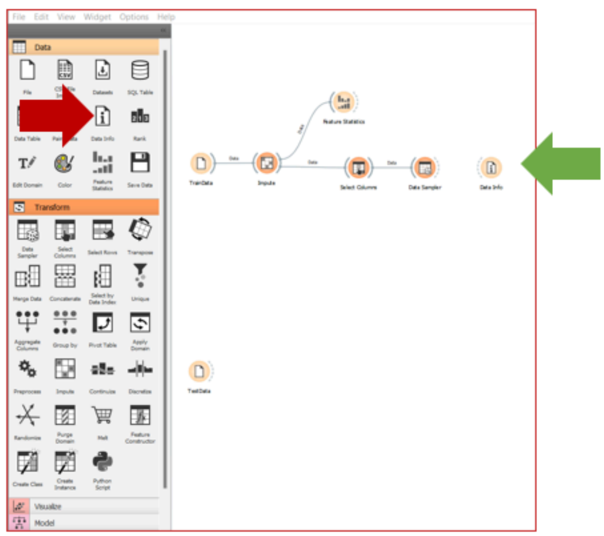


1. Set slider to 80%
2. Click on Sample Data to effect the changes
3. Click X to close the pop-up

How do we know if the data is actually split or not?

Step 4(d)

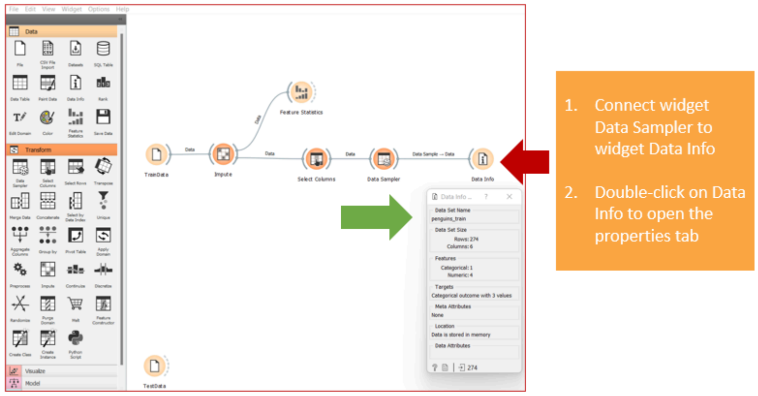
Insert the Data Info widget onto the canvas



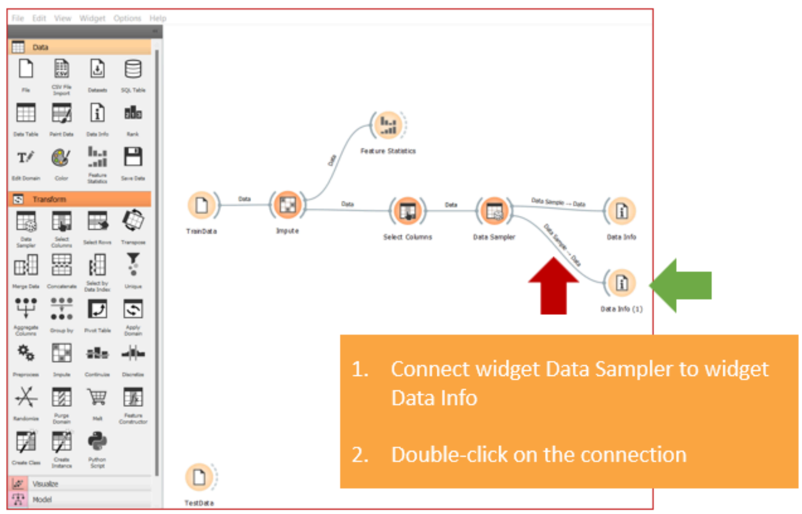
Let’s inspect on how the data is being split through Data Sampler.

We will be using Data Info.

Step 4(e)



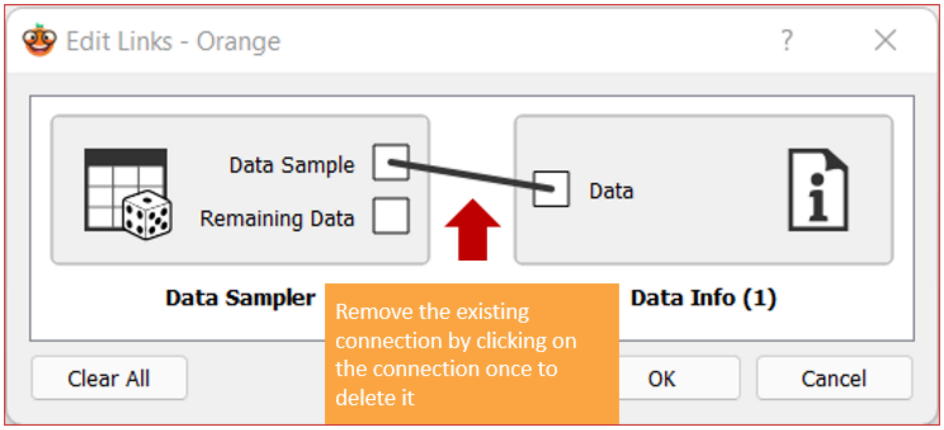
Step 4(f)



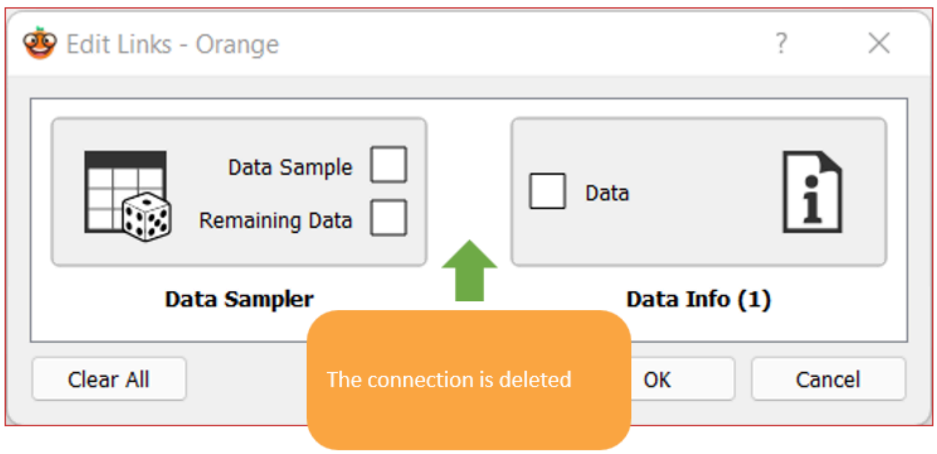
Connect widget Data Sampler to the second widget Data Info. We can do that by dragging the output from Data Sampler to the input of the second Data Info.

Take note of the connection name. We will change this. Double-click on the connection.

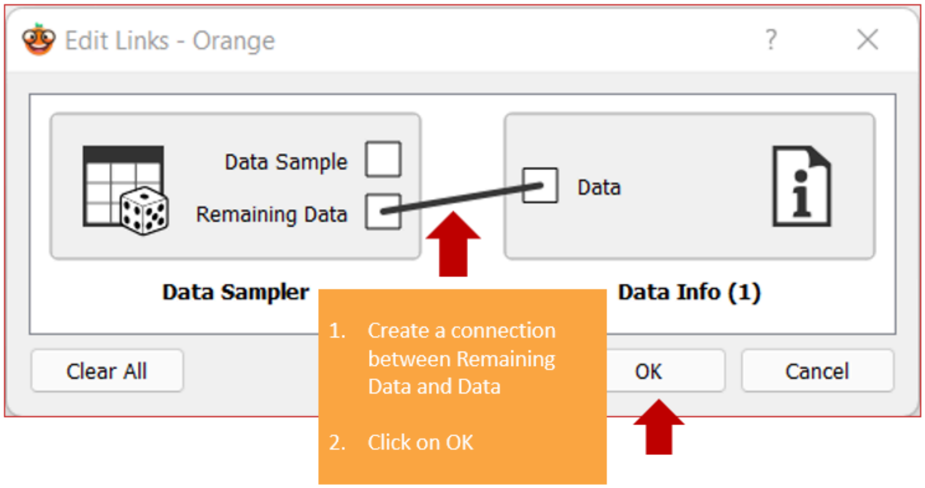
Step 4(g)



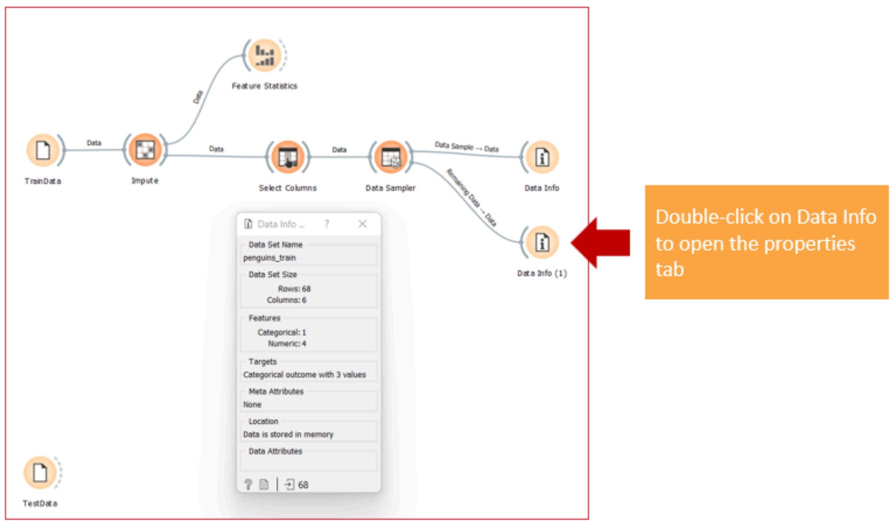
Step 4(g)



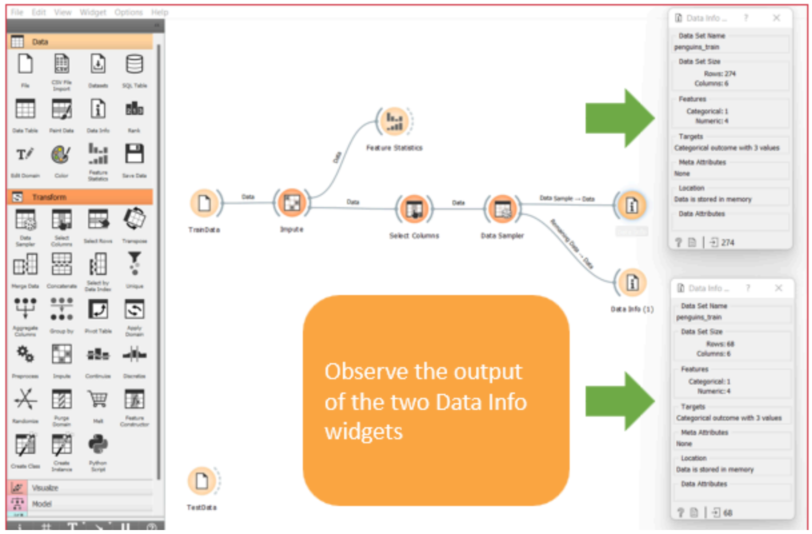
Step 4(h)



Step 4(i)



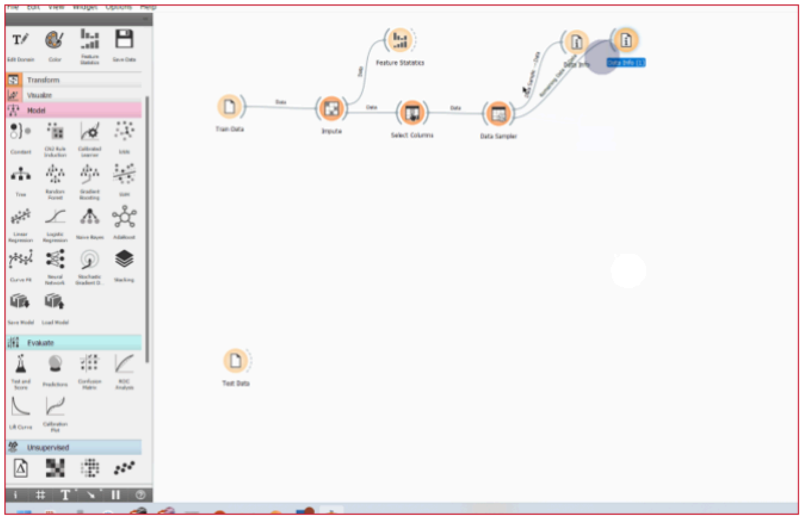
Step 4(j)



What do we do after having split the data?

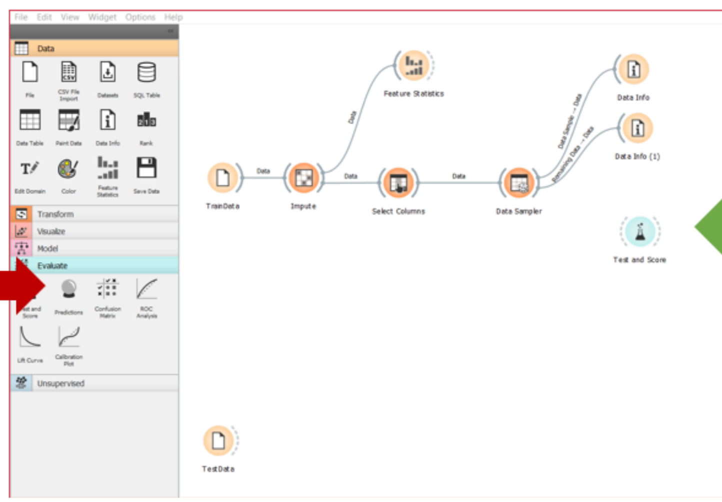
Modeling

Step 5: Train Model

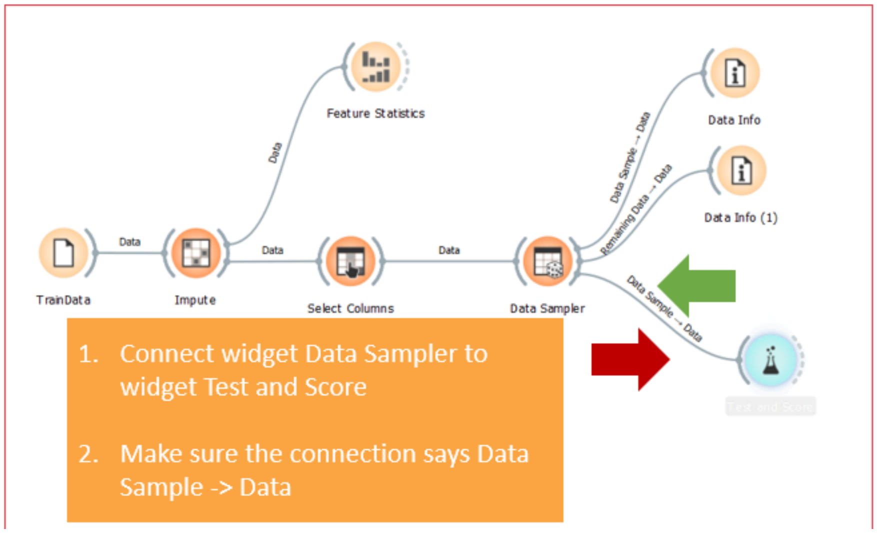


Step 5(a)

Insert the Test and Score widget onto the canvas

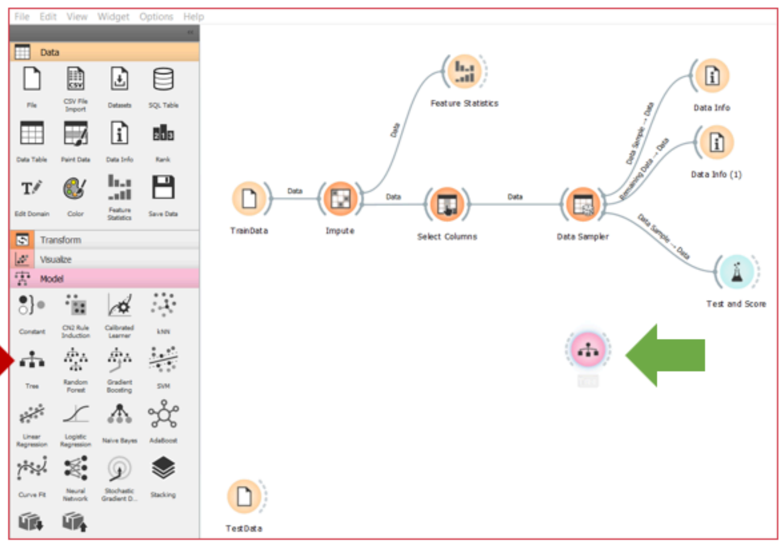


Step 5(b)



Step 5(c)

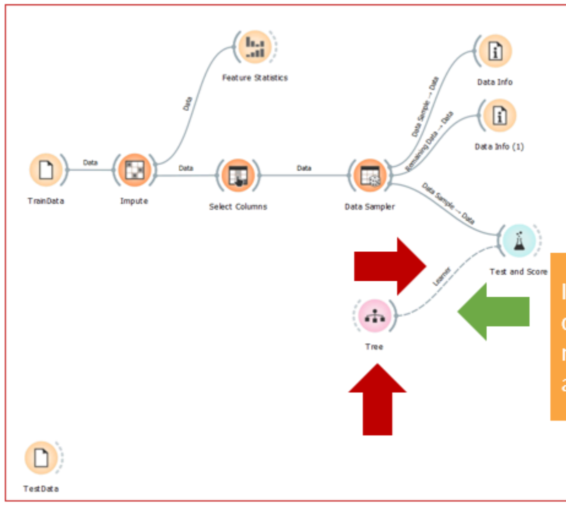
Insert the widget Tree into the canvas and put it to the left of widget Test and Score



Step 5(d)

1. Connect widget Tree to widget Test and Score

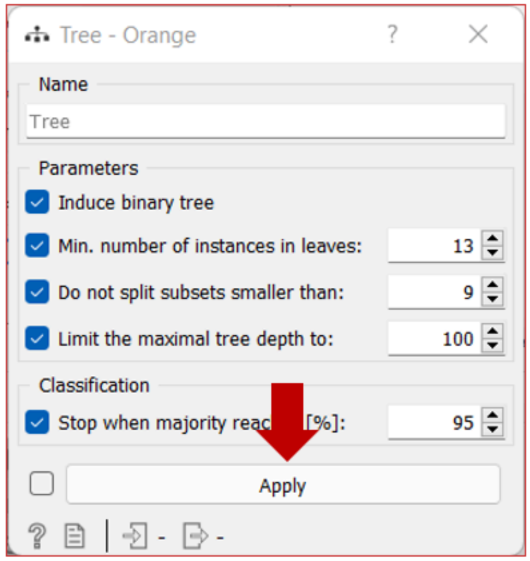
2. Double click on the model icon to open its properties



If the connection is dotted it means the model has not been applied yet

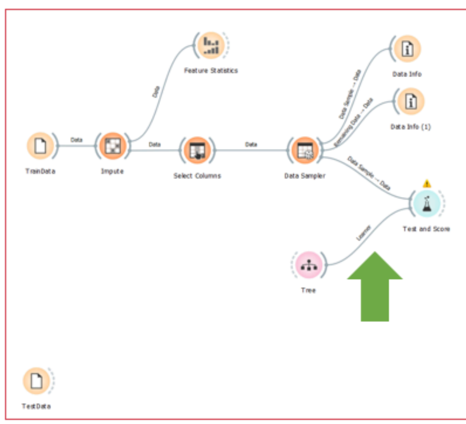
Step 5(e)

1. Click on Apply if automatically apply has not been checked
2. Click X to close the pop-up window



Step 5: Train Model

The model is ready

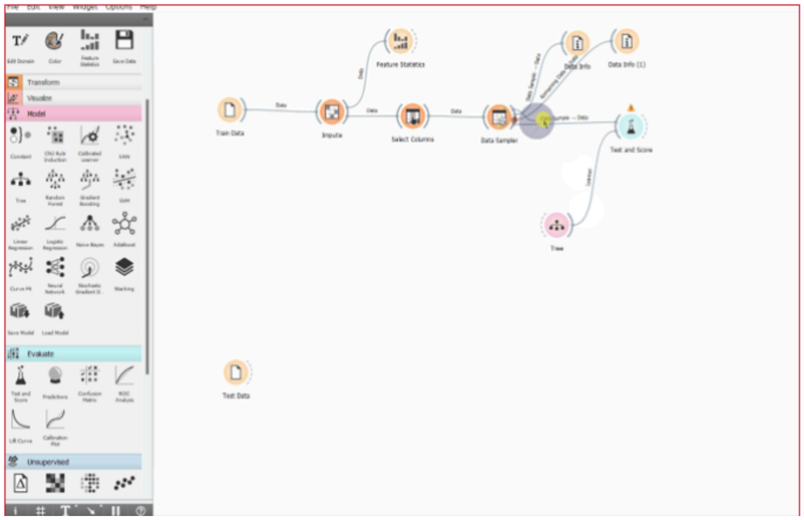


If there is a warning sign on the test and score widget, we will need to change its properties. Let's talk about it next

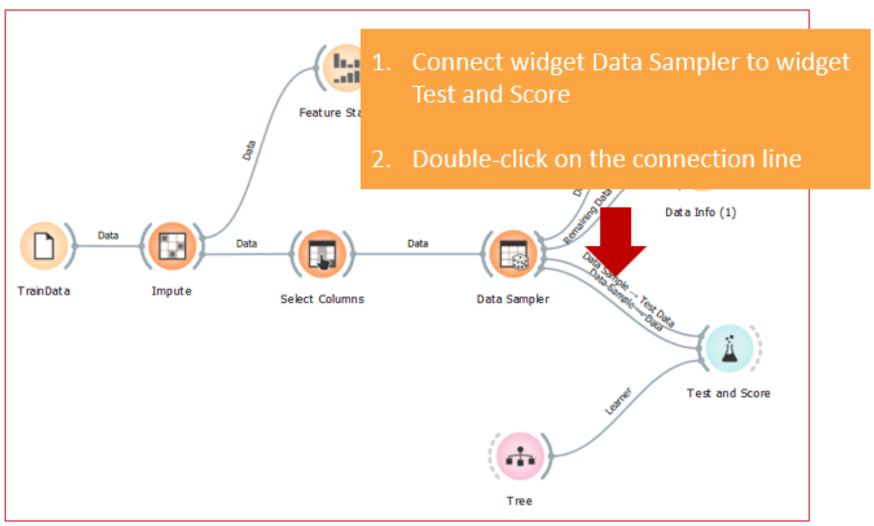
After creating a model, we need to test the model and check its accuracy

Evaluation

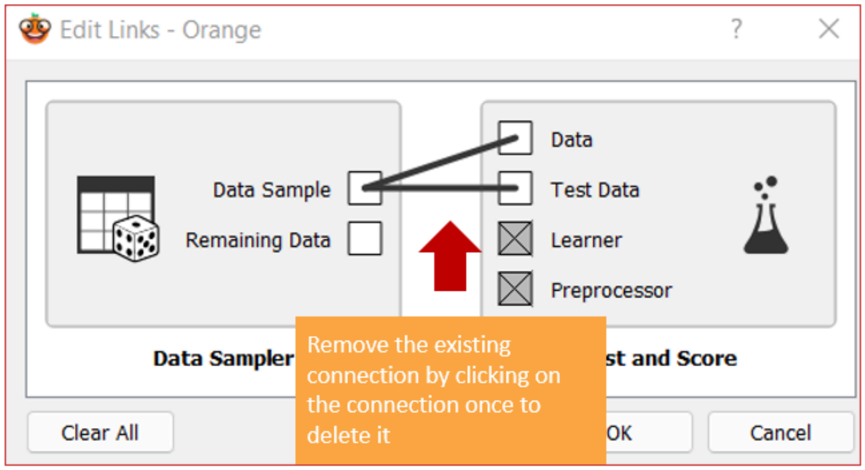
Step 6: Evaluate Model



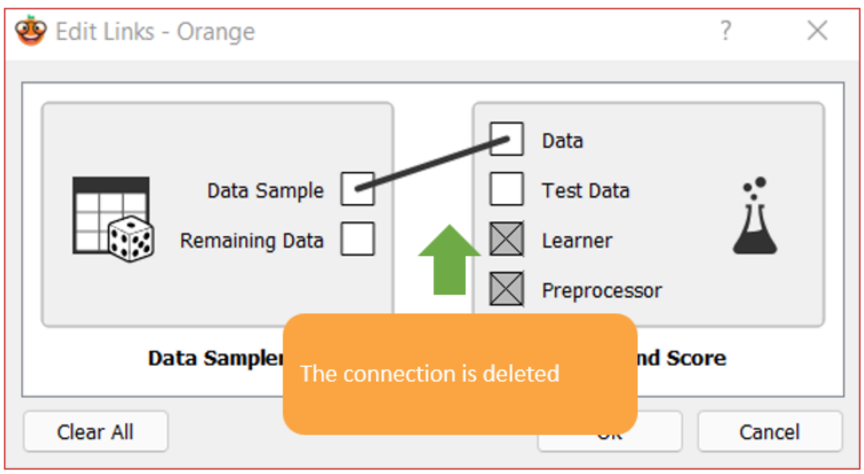
Step 6(a)



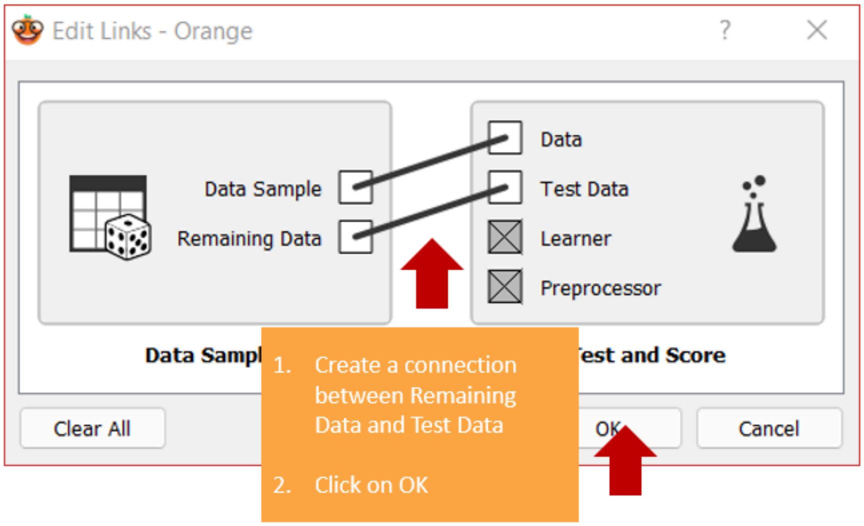
Step 6(b)



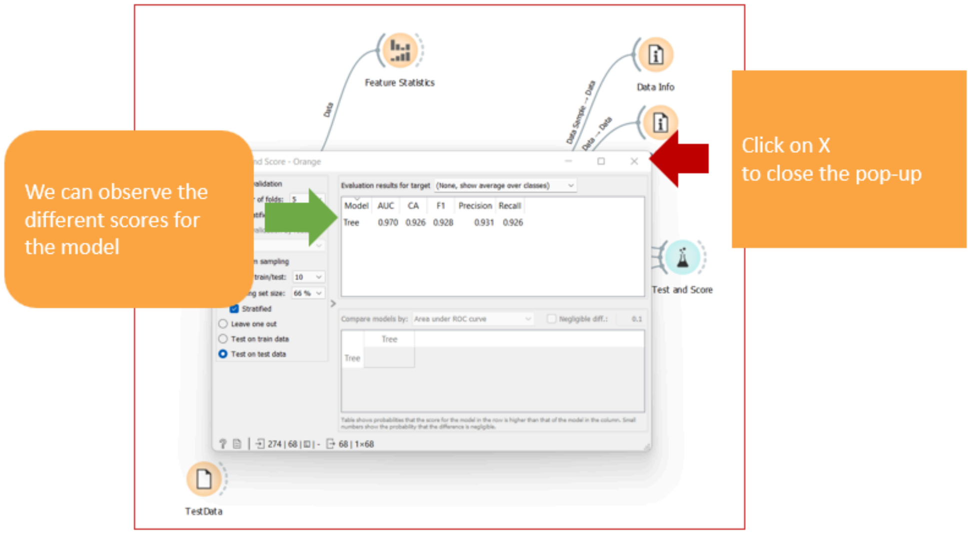
Step 6(b)

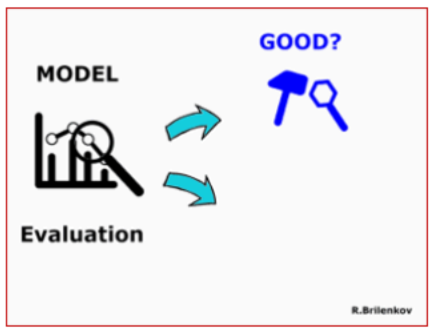


Step 6(c)


Edit Links - Orange
Data Sample
Remaining Data
Data
Test Data
Learner
Preprocessor
Data Sampl
est and Score
Clear All
OK
Cancel
1. Create a connection between Remaining Data and Test Data
2. Click on OK

Step 6(d)



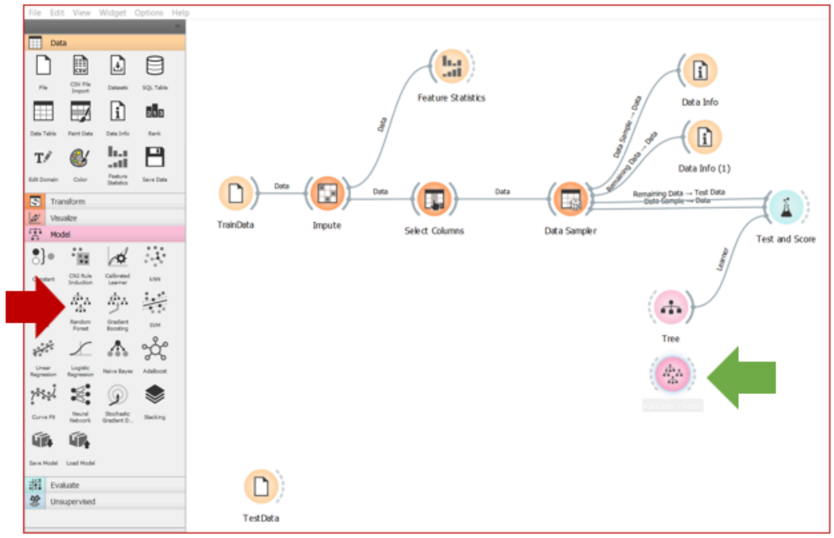


Through Evaluation we know
if a model is good or bad

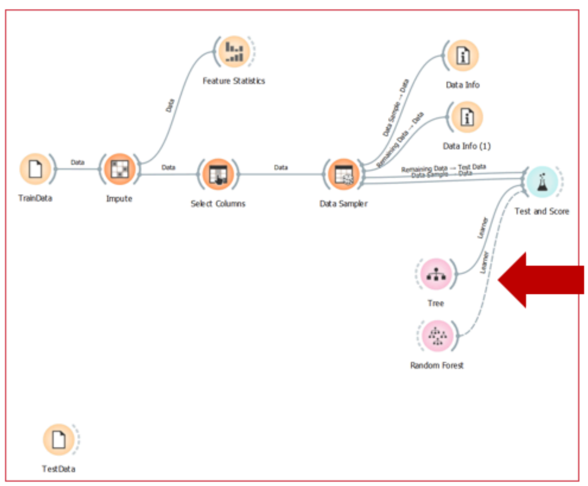
Let’s try a couple of other classification algorithms

Step 6(h)

Insert the widget Random Forest into the canvas



Step 6(i)

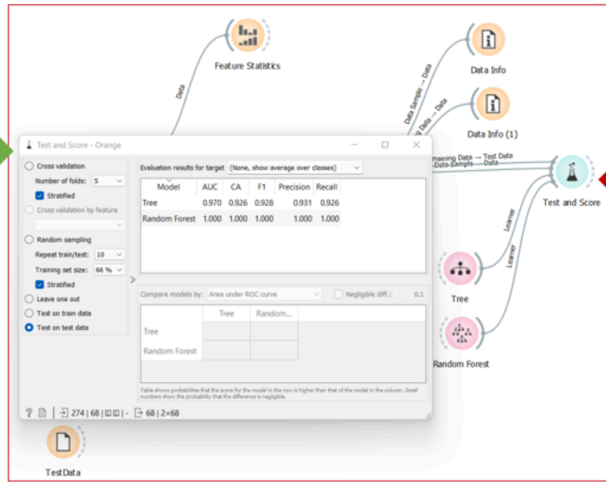


Connect the Random Forest widget to the Test and Score widget

Step 6(j)

This pop up will appear

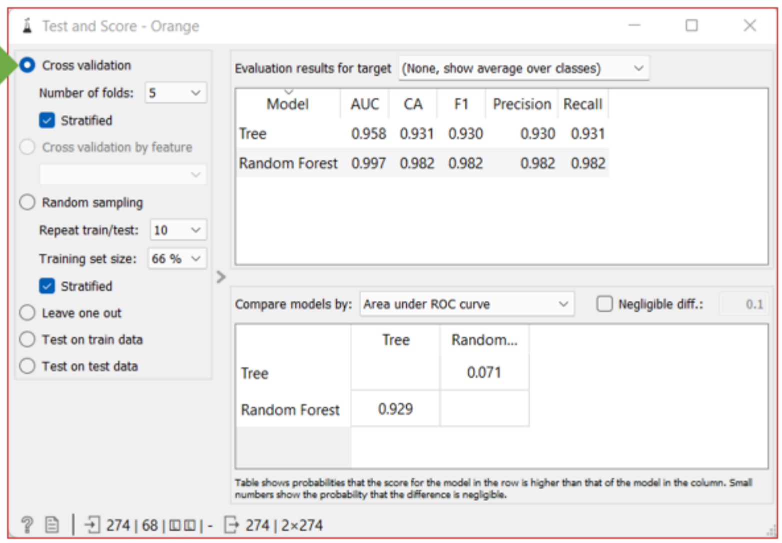
We can observe the individual scores for each algorithm here



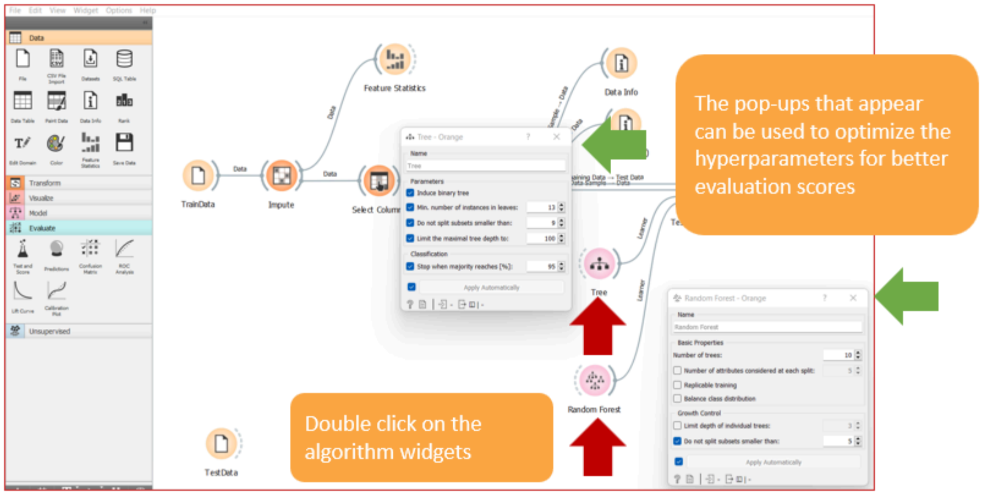
Double click on the Test and Score widget

Step 6(k)

We can choose the methods for evaluation in this pop up



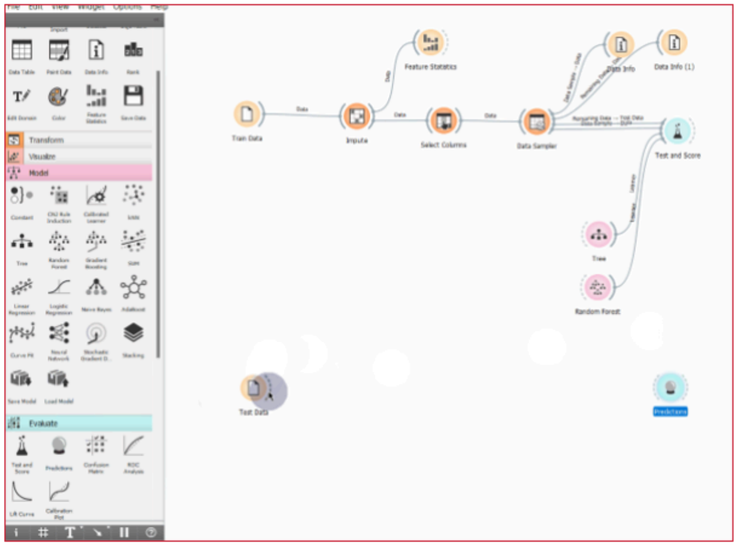
Step 6(l)



Now that we have found which model gives us the best results, we can use that one!

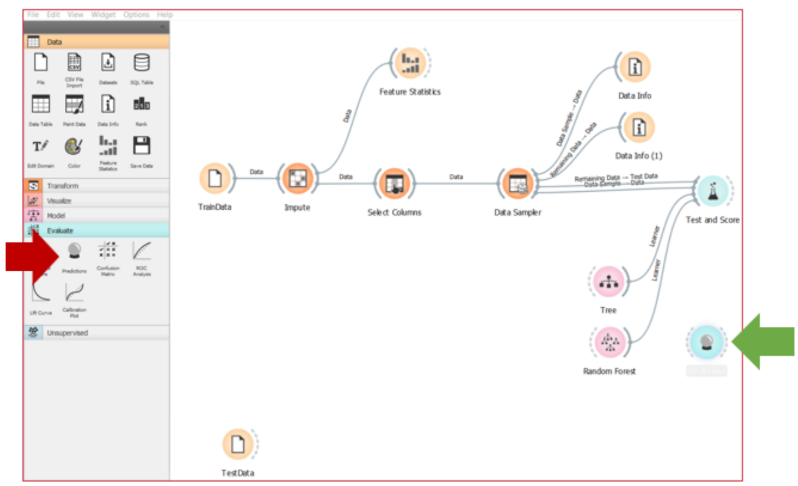
Predictions

Step 7: Predictions

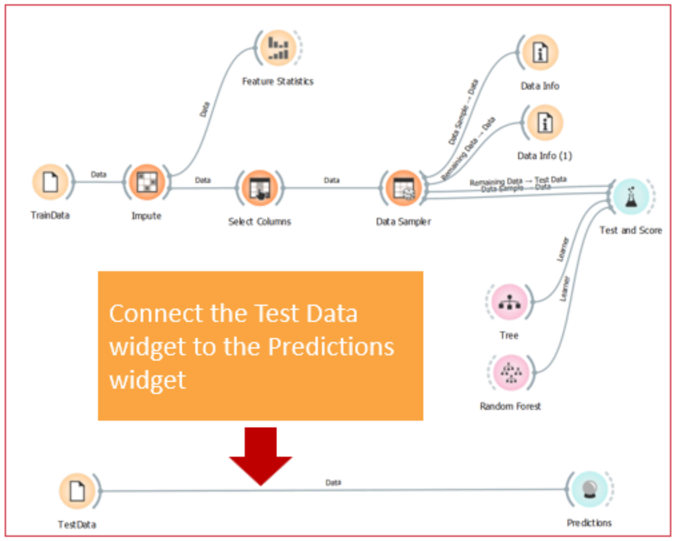


Step 7(a)

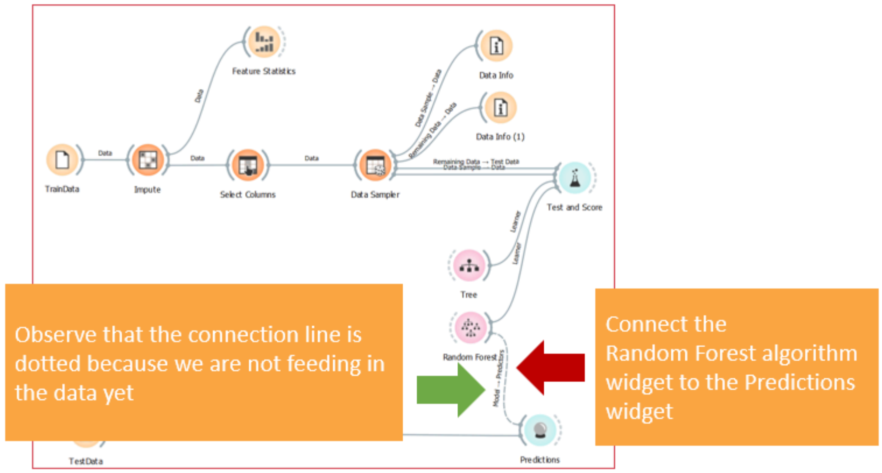
Insert the widget Predictions into the canvas



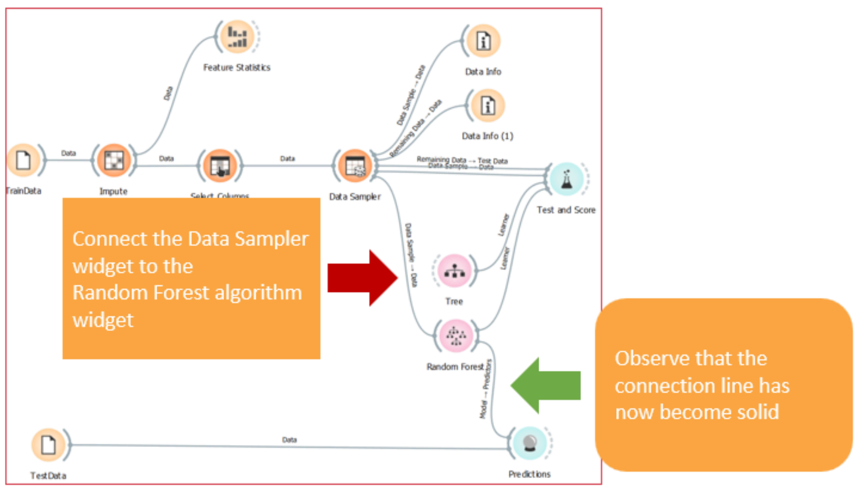
Step 7(b)


Feature Statistics
Data Info
Data Info (1)
TrainData
Impute
Select Columns
Data Sampler
Test and Score
Tree
Random Forest
Connect the Test Data widget to the Predictions widget
TestData
Predictions

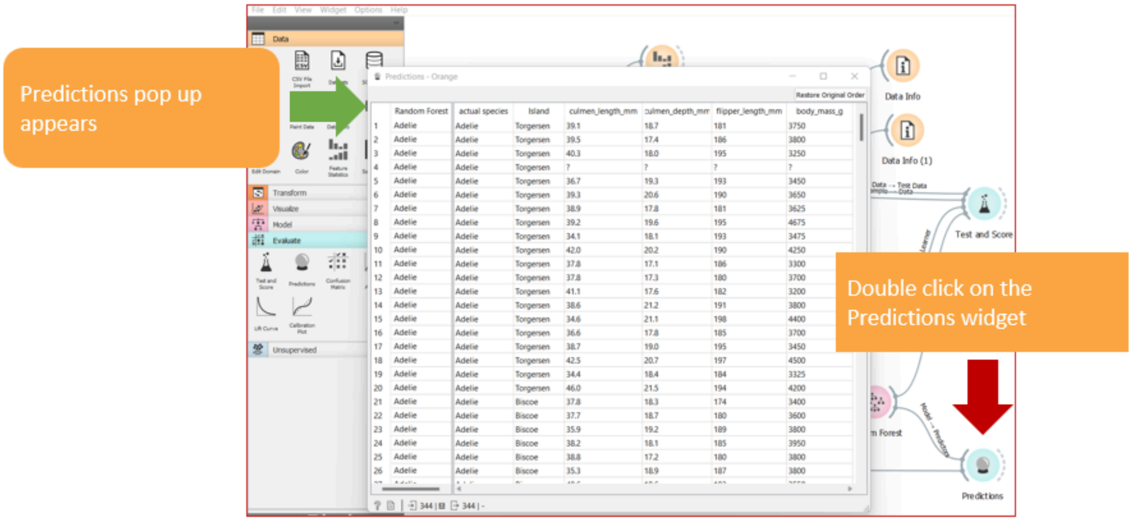
Step 7(b)

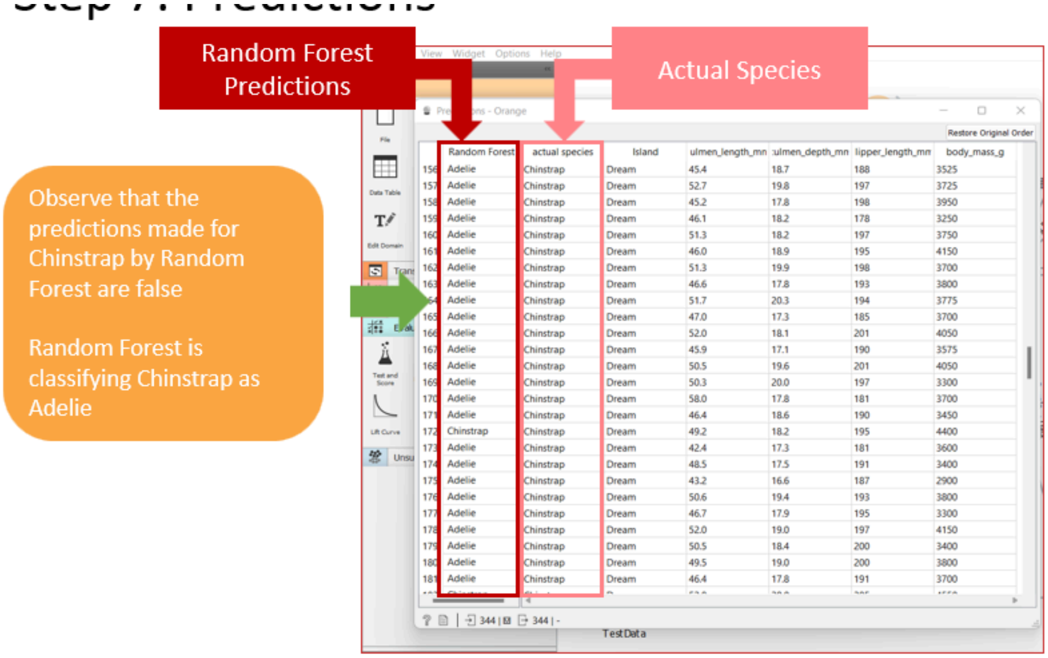


Step 7(c)



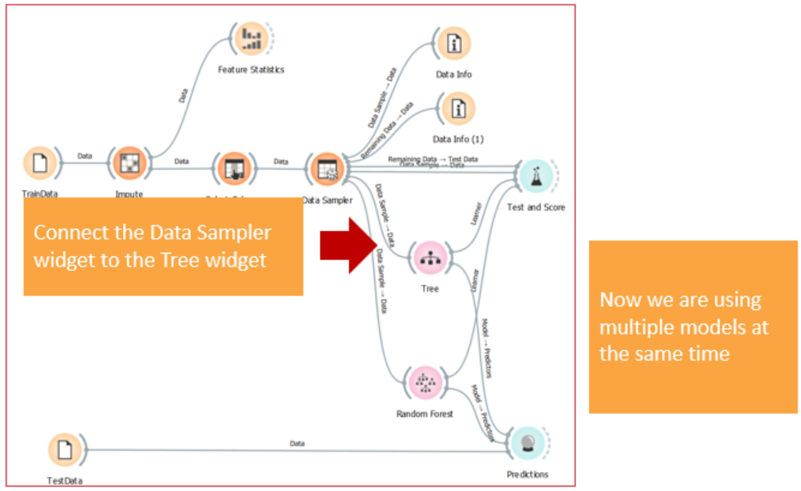
Step 7(d)





Since the Random Forest algorithm is not working well with one of the species, let's use another algorithm

Step 7(e)



Step 7: Predictions