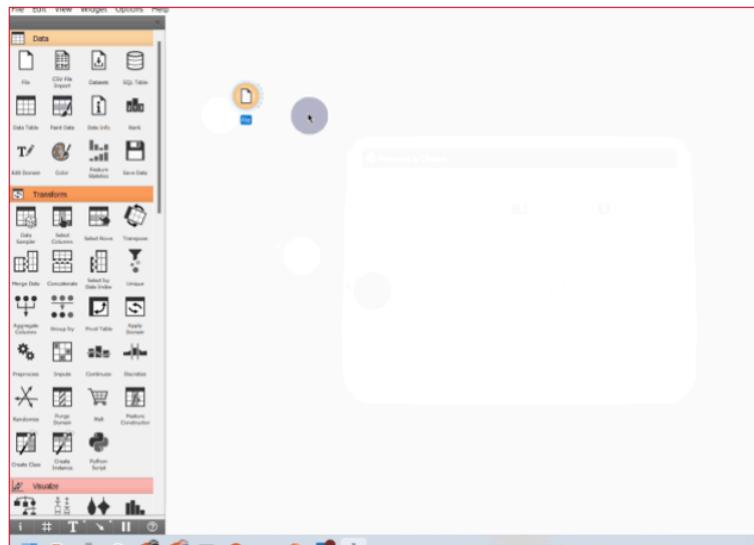


ORANGE DATA MINING

STEPS for AI model to predict the penguin species .

Step 1: Upload Dataset

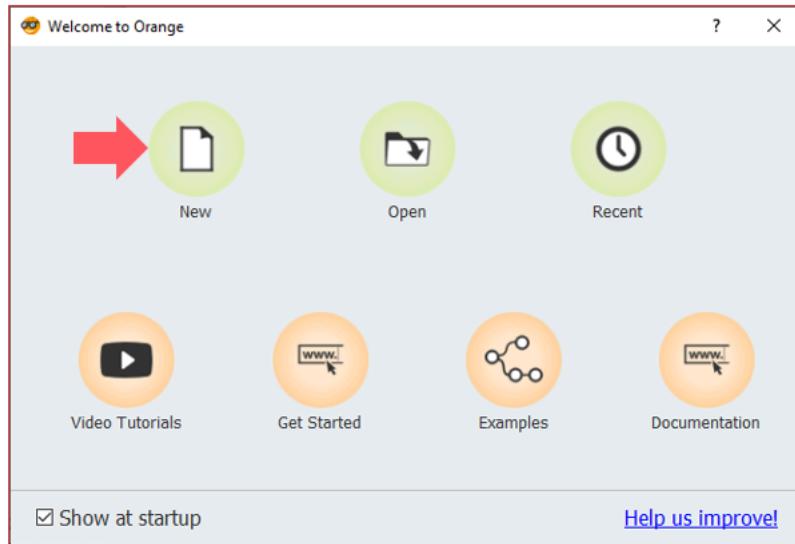


Data
Acquisition

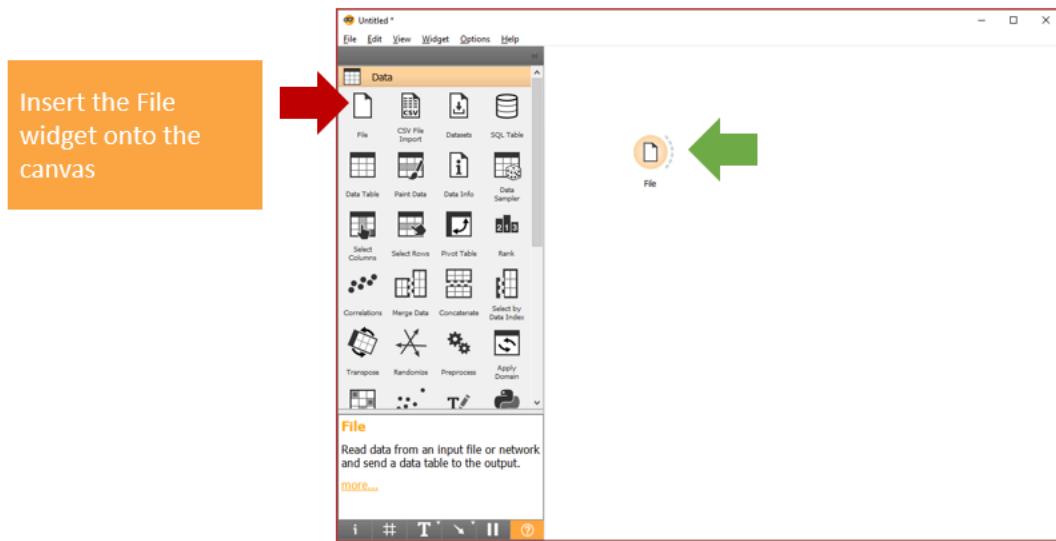
Step 1(a)

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

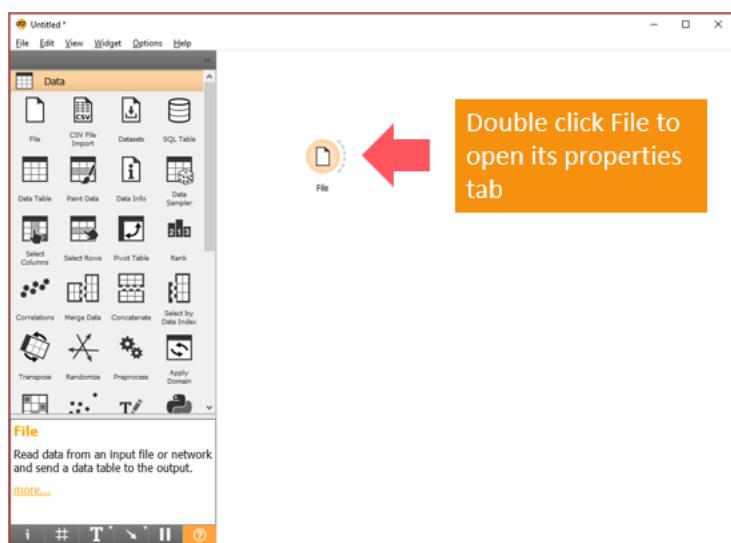
Click on New.



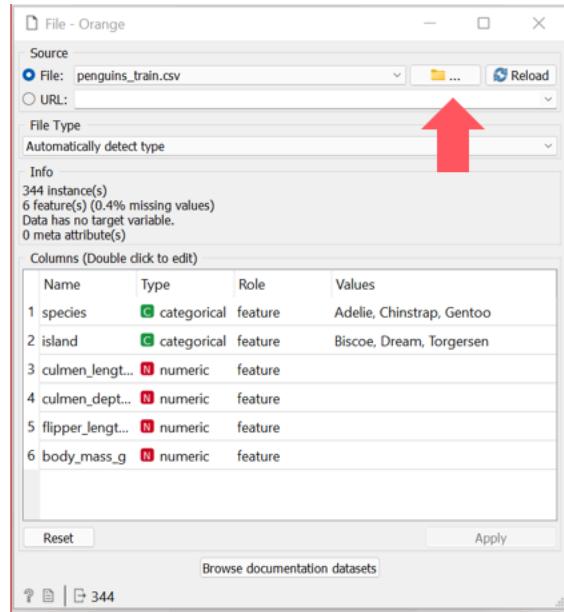
Step 1(b)



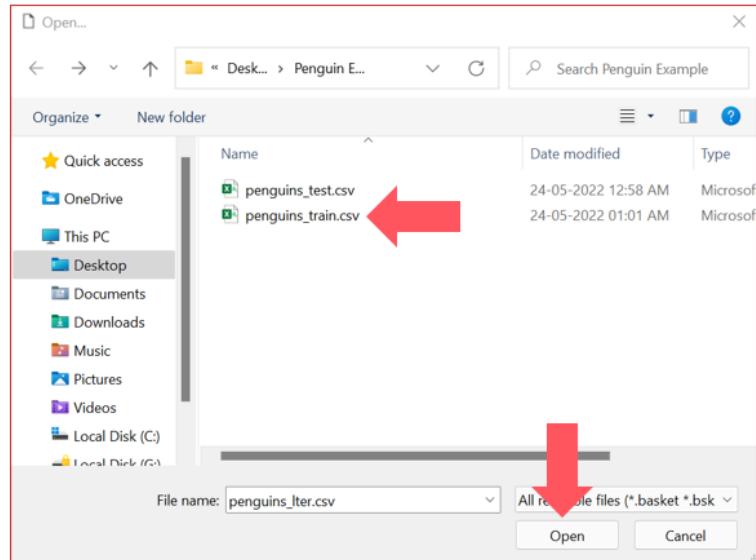
Step 1(c)



Step 1(d)



Step 1(e)



Step 1(f)

penguins_train.csv is uploaded



File - Orange

Source

File: penguins_train.csv

URL:

File Type

Automatically detect type

Info

344 instance(s)
6 feature(s) (0.4% missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

Name	Type	Role	Values
1 species	C categorical	feature	Adelie, Chinstrap, Gentoo
2 island	C categorical	feature	Biscoe, Dream, Torgersen
3 culmen_length_mm	N numeric	feature	
4 culmen_depth_mm	N numeric	feature	
5 flipper_length_mm	N numeric	feature	
6 body_mass_g	N numeric	feature	

Reset Apply

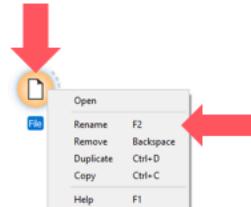
Browse documentation datasets

344

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



Untitled*

File Edit View Widget Options Help

Data

File CSV File Import Datasets SQL Table

Data Table Paint Data Data Info Data Sampler

Select Columns Select Rows Pivot Table Rank

Correlations Merge Data Concatenate Select by Data Index

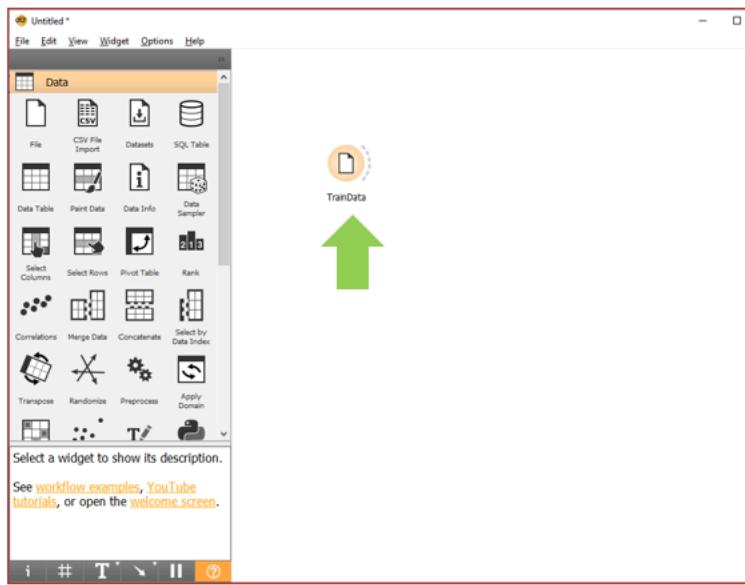
Transpose Randomize Preprocess Apply Domain

Select a widget to show its description.

See [workflow examples](#), [YouTube tutorials](#), or open the [welcome screen](#).

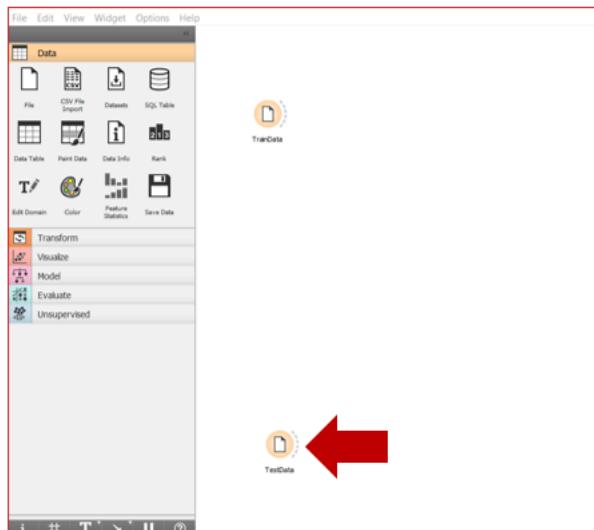
Step 1(h)

The file name for the uploaded dataset has changed



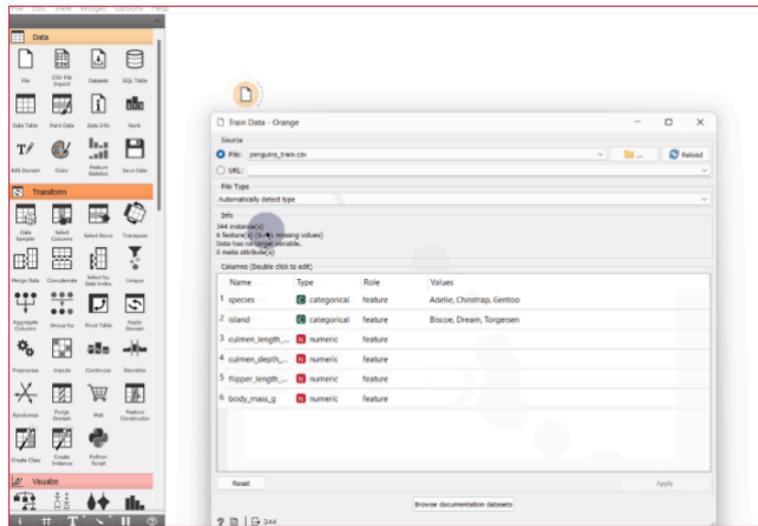
Step 1(i)

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



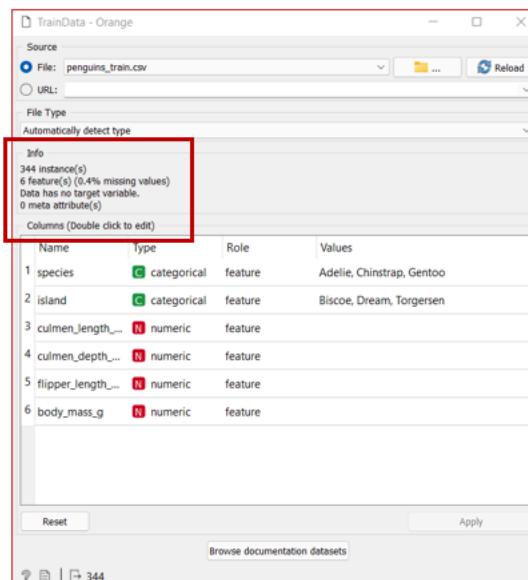
After Data Acquisition, what should we do next?

Step 2: Clean Missing Data



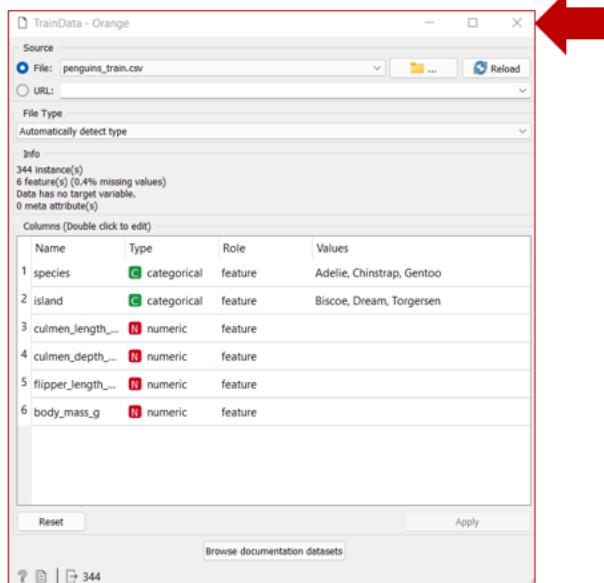
Step 2(a)

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



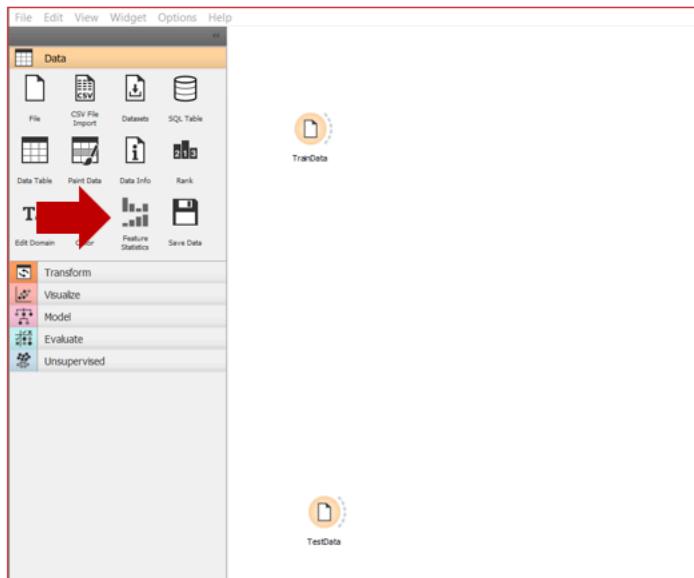
Step 2(b)

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



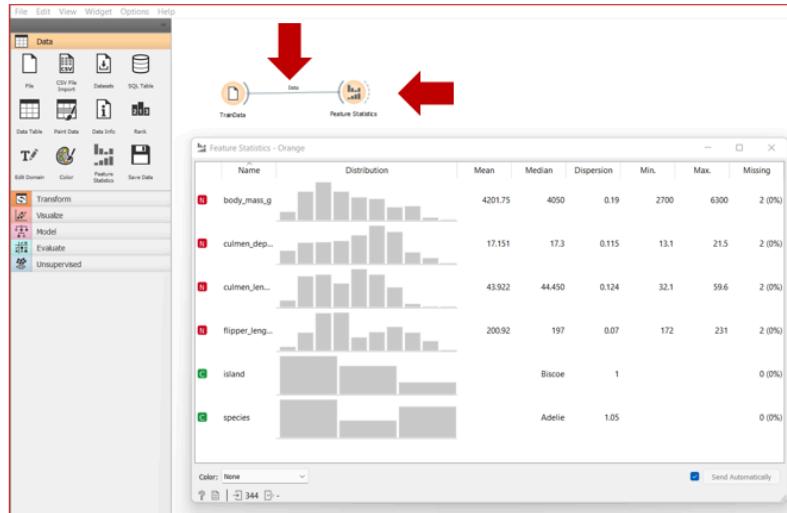
Step 2(c)

Insert the widget Feature Statistics onto the canvas



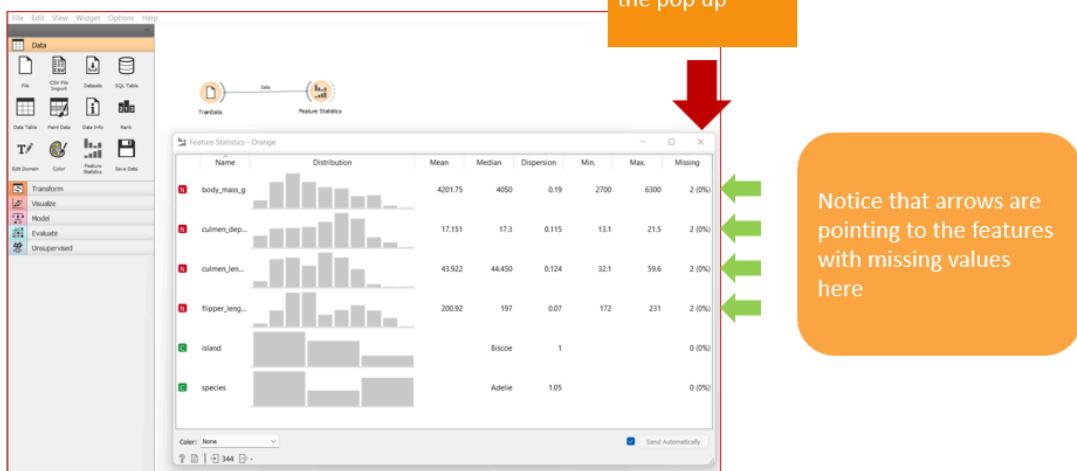
Step 2(d)

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.

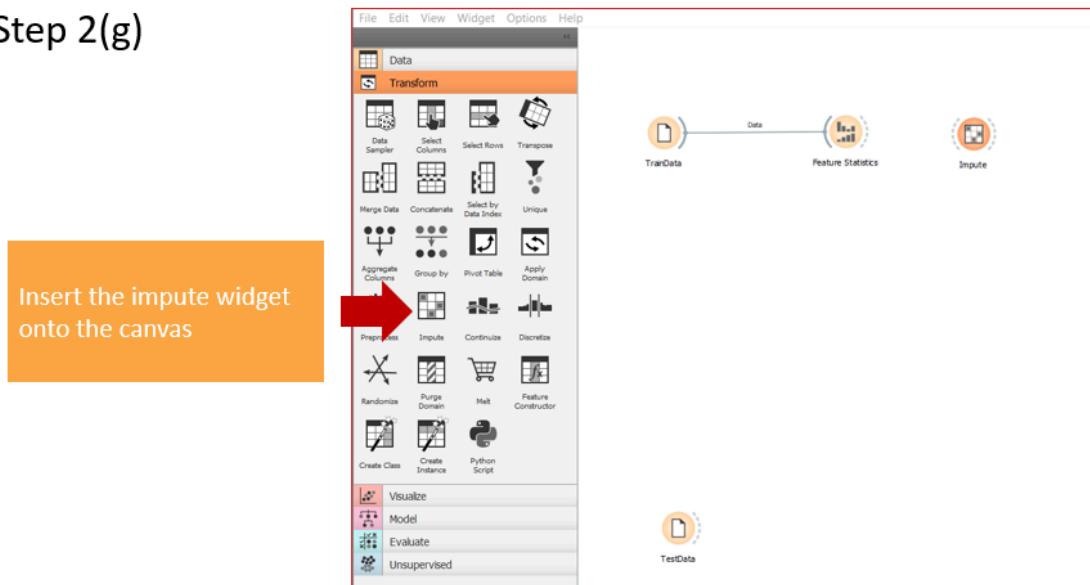


What is the mean value of culmen_length feature?

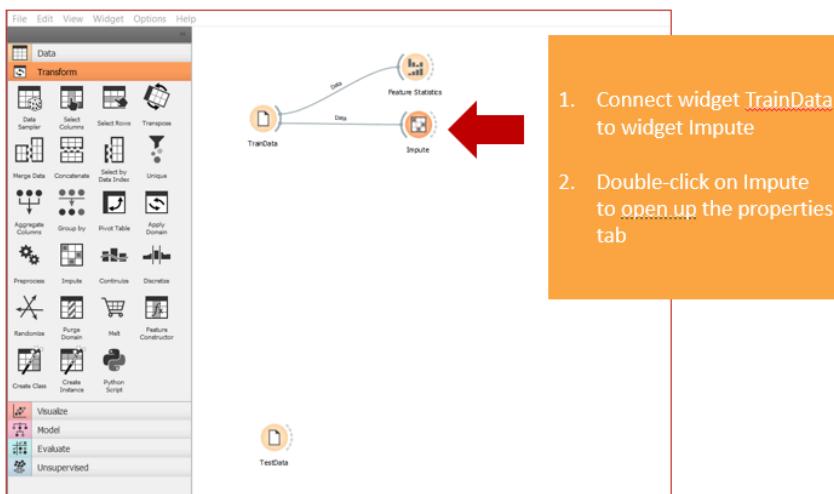
Step 2(f)



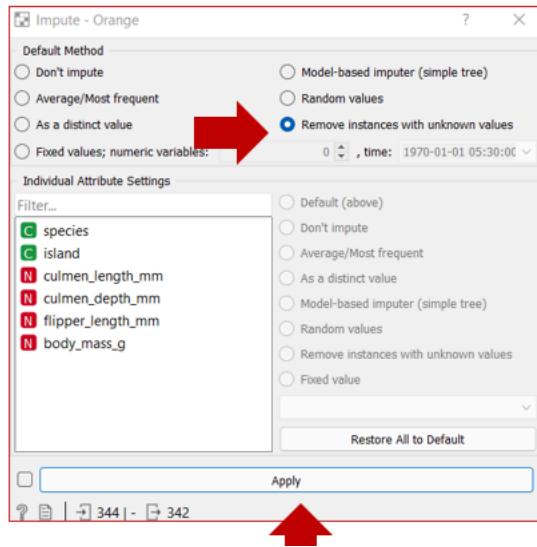
Step 2(g)



Step 2(h)

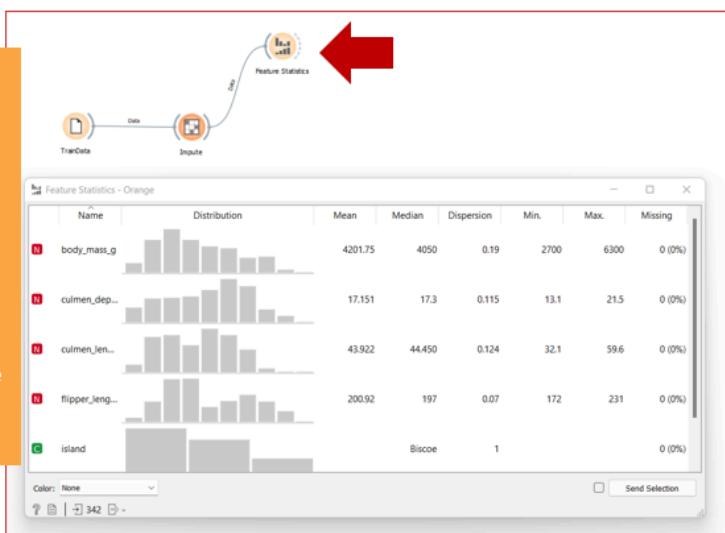


Step 2(i)



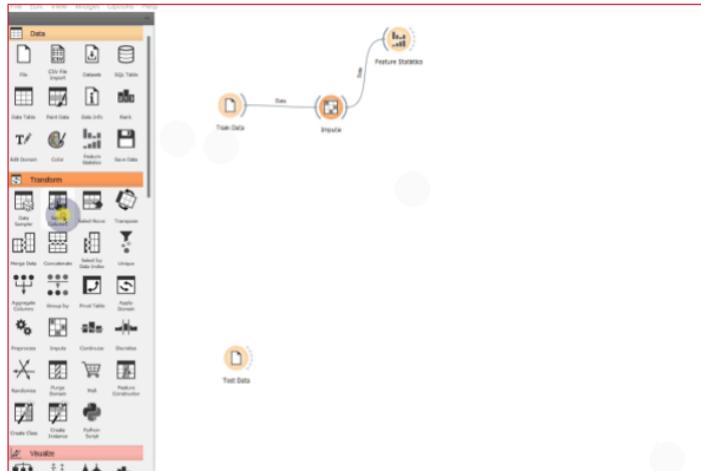
Step 2(j)

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?

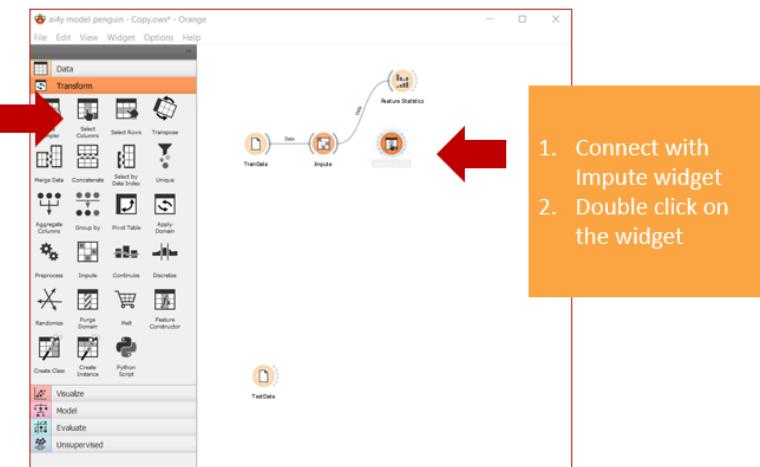
Step 3: Select Target Label



Step 3(a)

Insert the Select columns widget onto the canvas

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

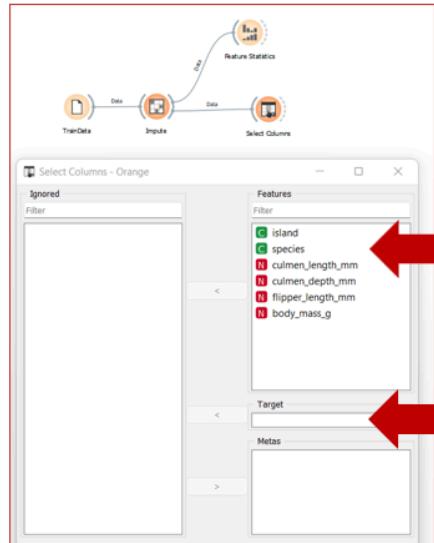


From TrainData, you would have noticed that the Feature Type for most of the columns is Numeric Feature. In supervised learning models, we have both the features and the labels. The labels are the output. Therefore, we need to define an output for our Palmer Penguin model. We will assign species as our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

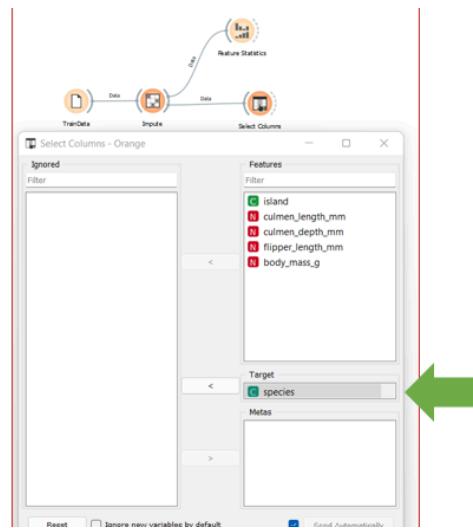
Step 3(b)

A window displaying all the features will appear



Drag the 'species' feature to the 'Target' box

Step 3(b)

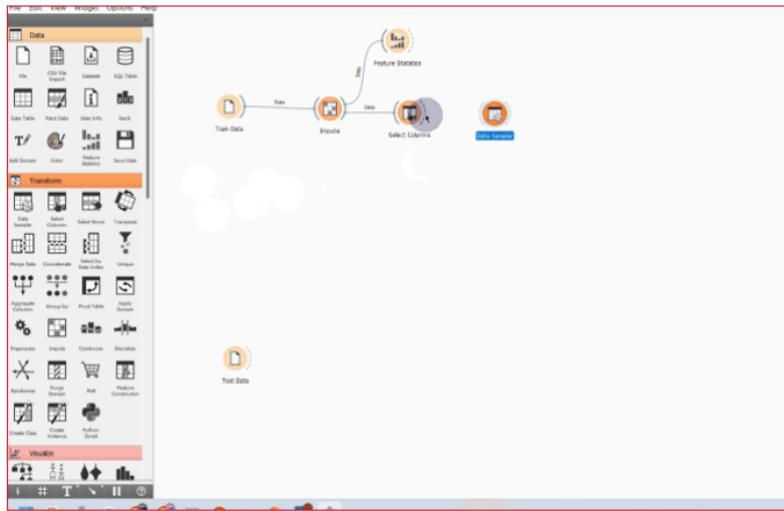


'species' is the Target label

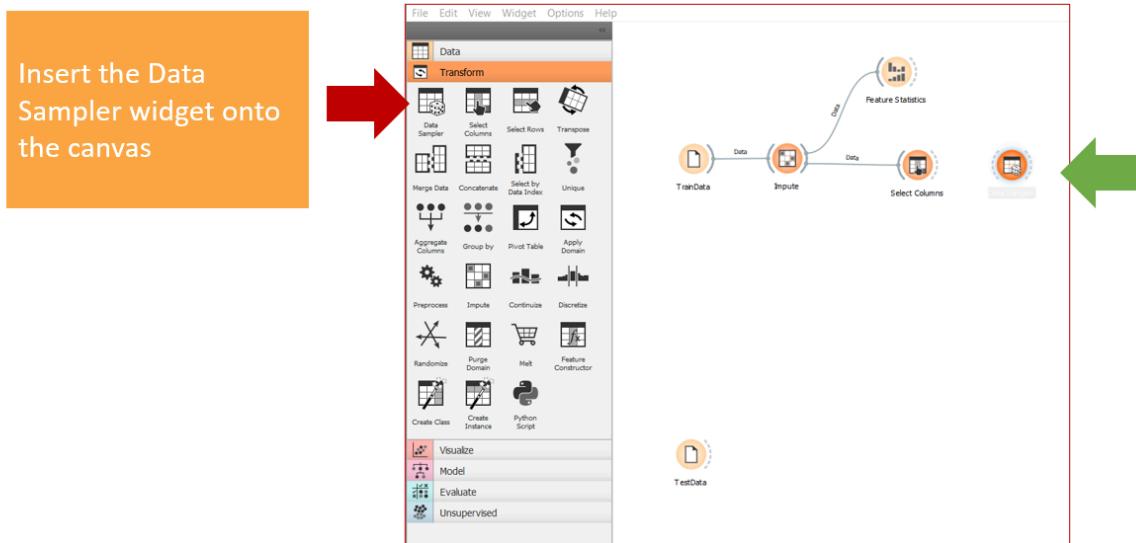
After choosing a target label, we need to split the data

Step 4: Data Sampler

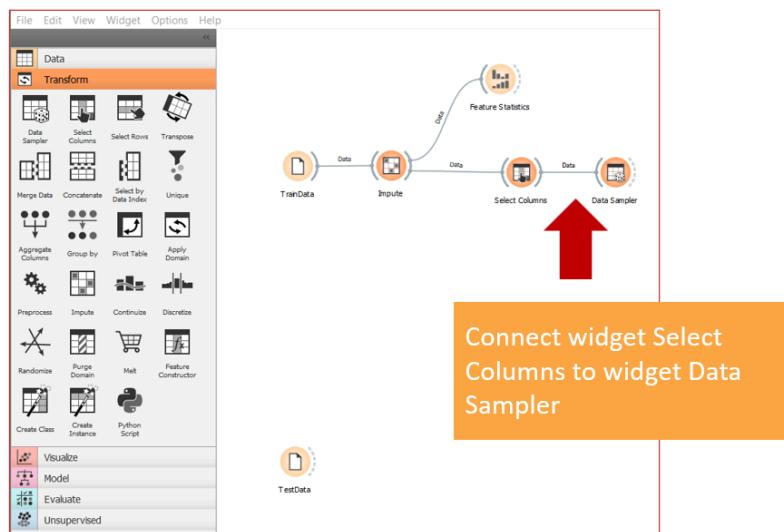
Data
Exploration



Step 4(a)



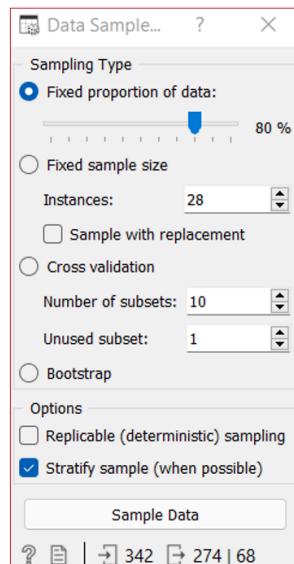
Step 4(b)



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)

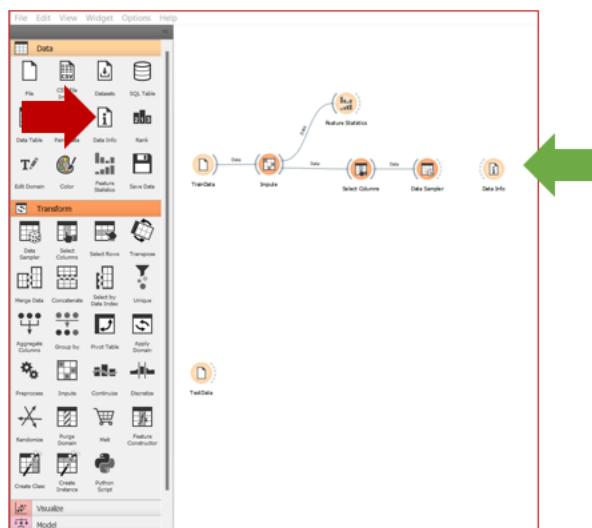


1. Set slider to 80%
2. Click on Sample Data to effect the changes
3. Click X to close the pop-up

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

Step 4(d)

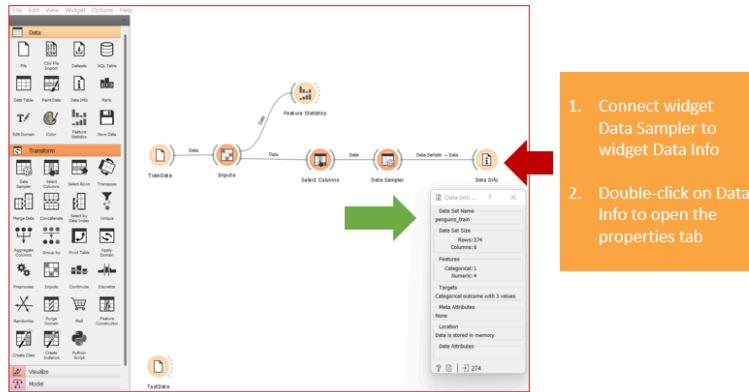
Insert the Data Info
widget onto the
canvas



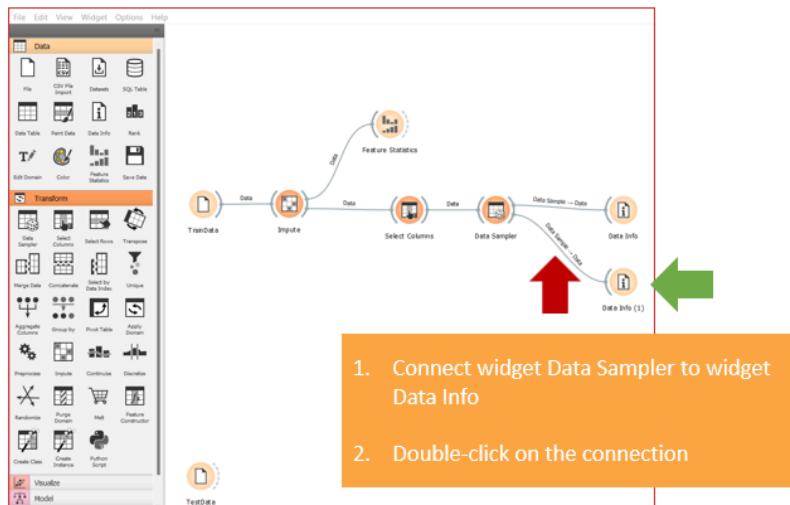
Let's inspect on how the data is being split through Data Sampler.

We will be using Data Info.

Step 4(e)



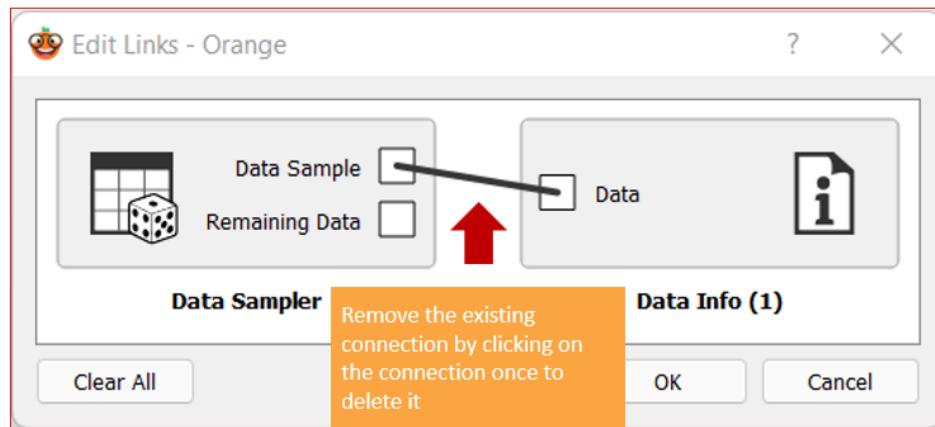
Step 4(f)



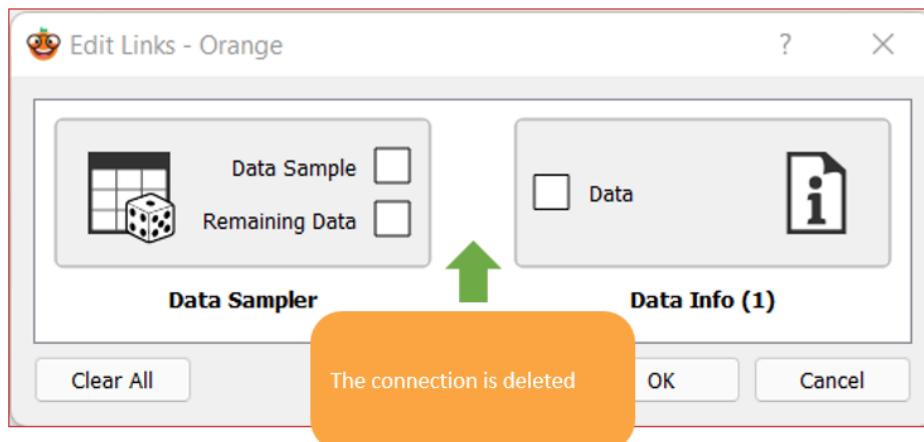
Connect widget Data Sampler to the second widget Data Info. We can do that by dragging the output from Data Sampler to the input of the second Data Info.

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

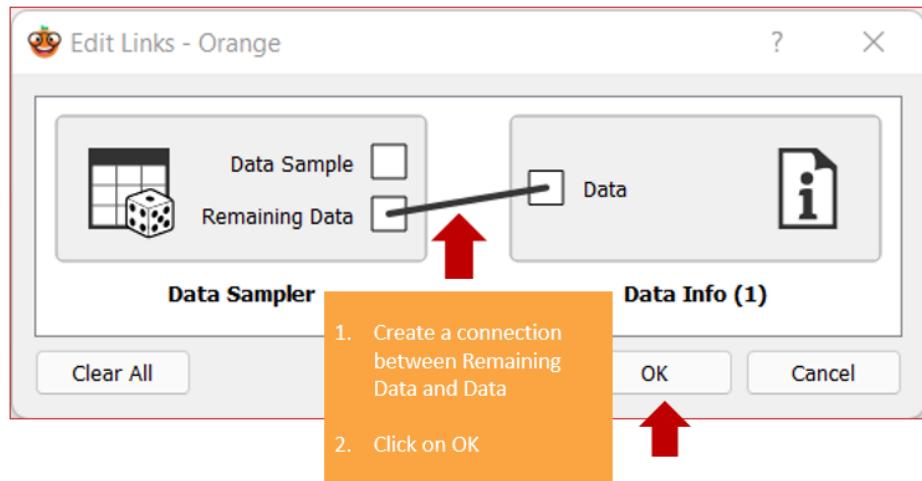
Step 4(g)



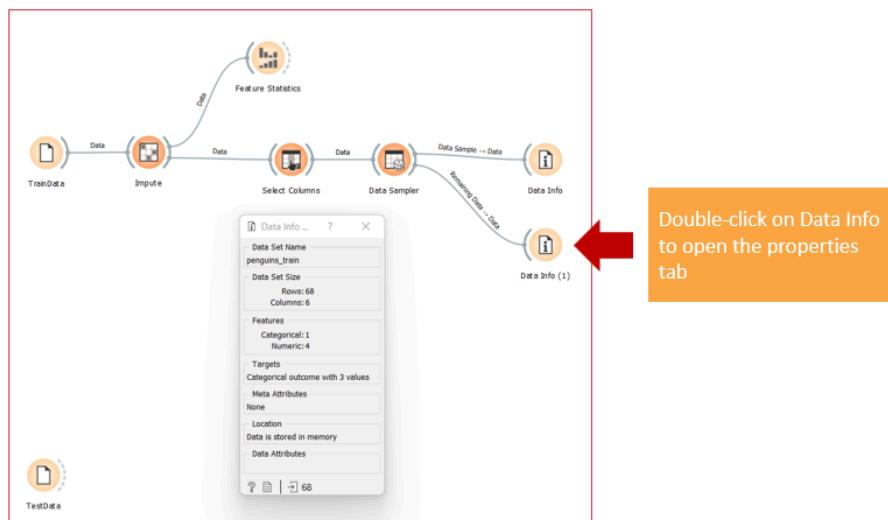
Step 4(g)



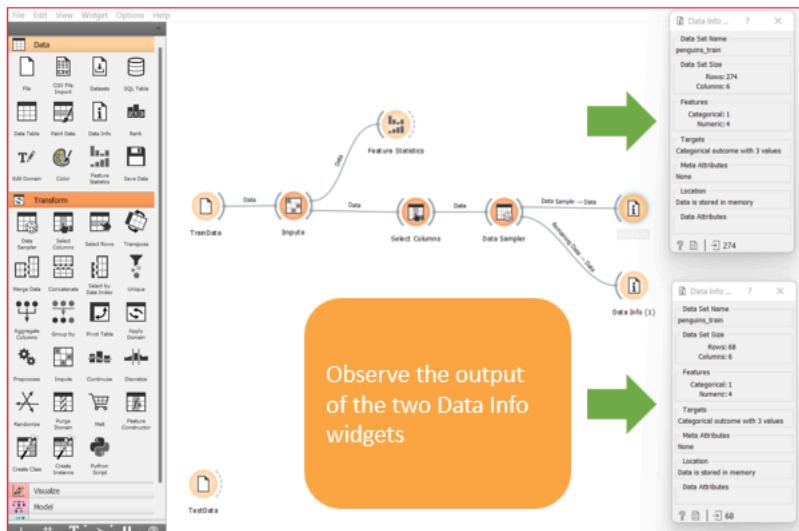
Step 4(h)



Step 4(i)



Step 4(j)

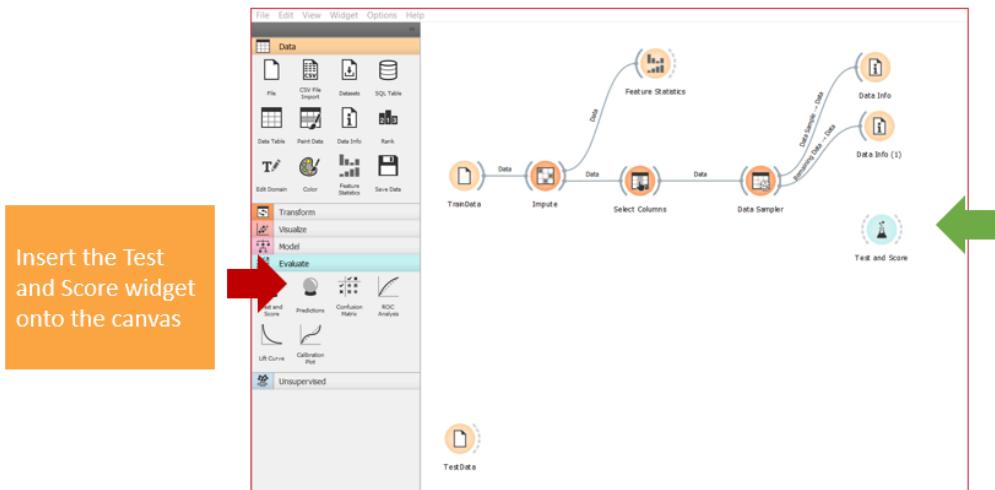


What do we do after having split the data?

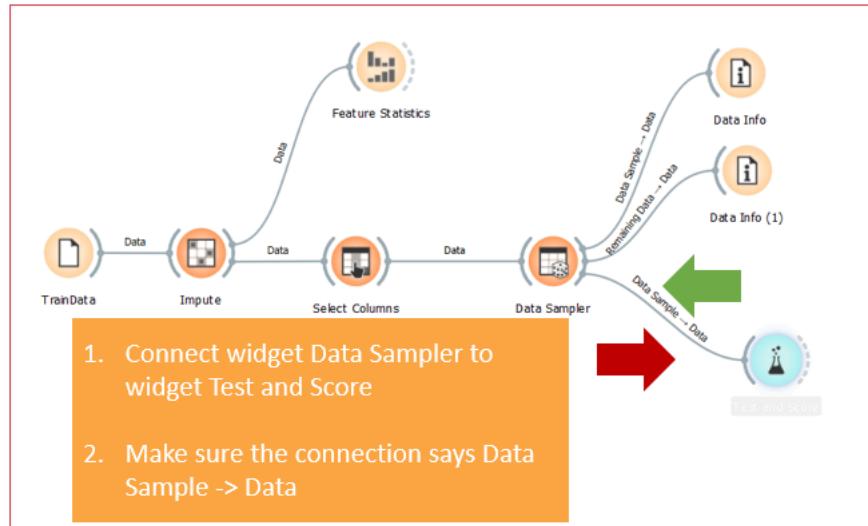
Step 5: Train Model



Step 5(a)

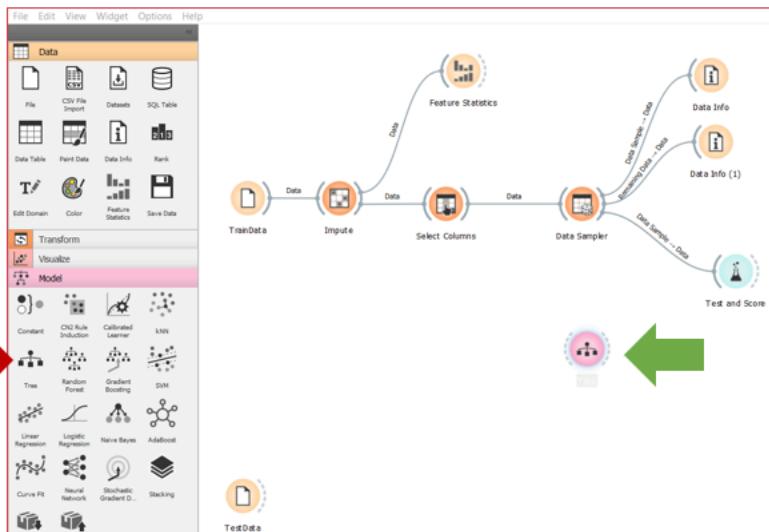


Step 5(b)



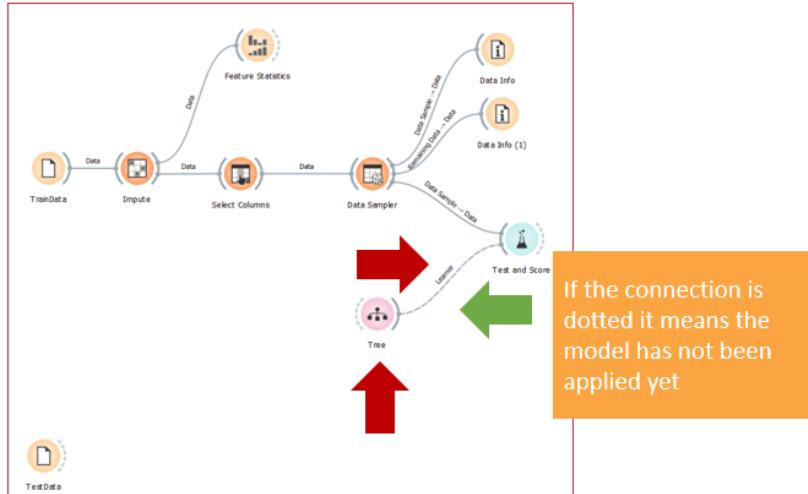
Step 5(c)

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



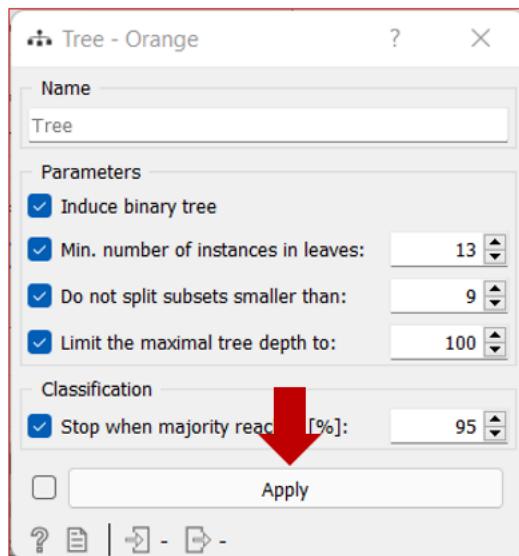
Step 5(d)

1. Connect widget Tree to widget Test and Score
2. Double click on the model icon to open its properties



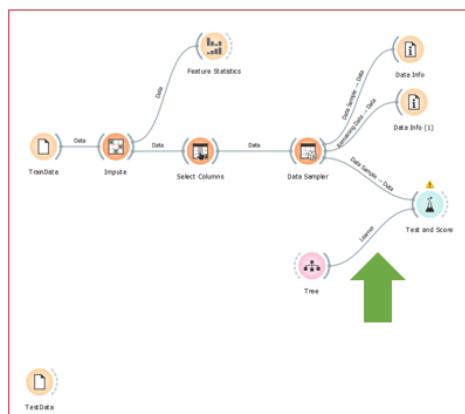
Step 5(e)

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

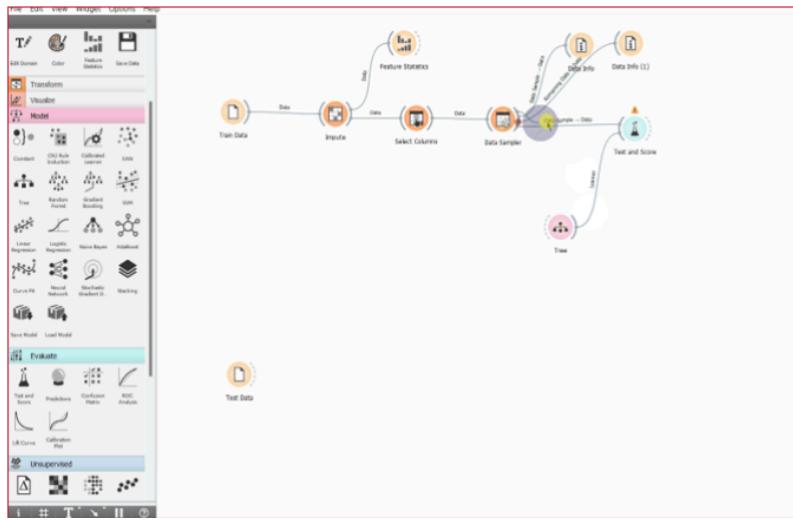


If there is a warning sign on the test and score widget, we will need to change its properties. Let's talk about it next

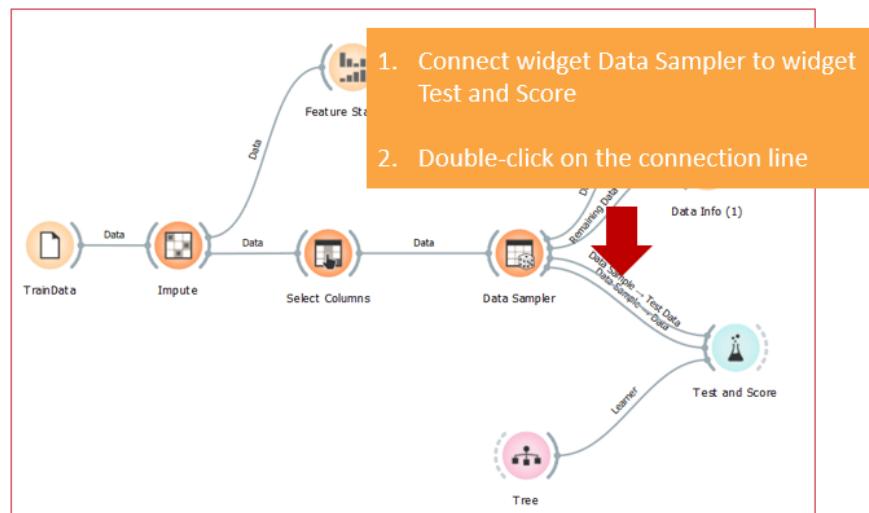
After creating a model, we need to test the model and check its accuracy

Step 6: Evaluate Model

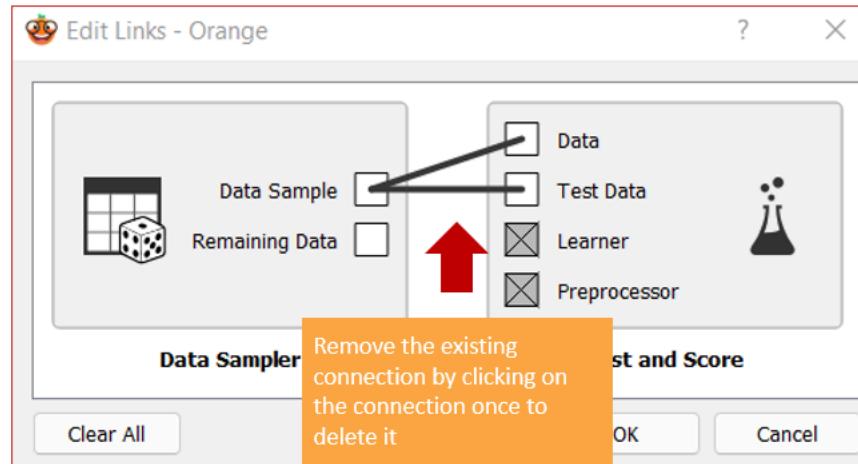
Evaluation



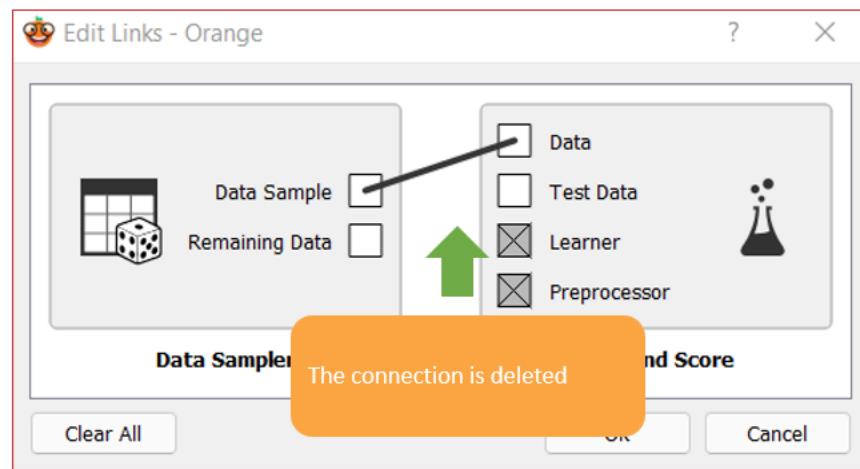
Step 6(a)



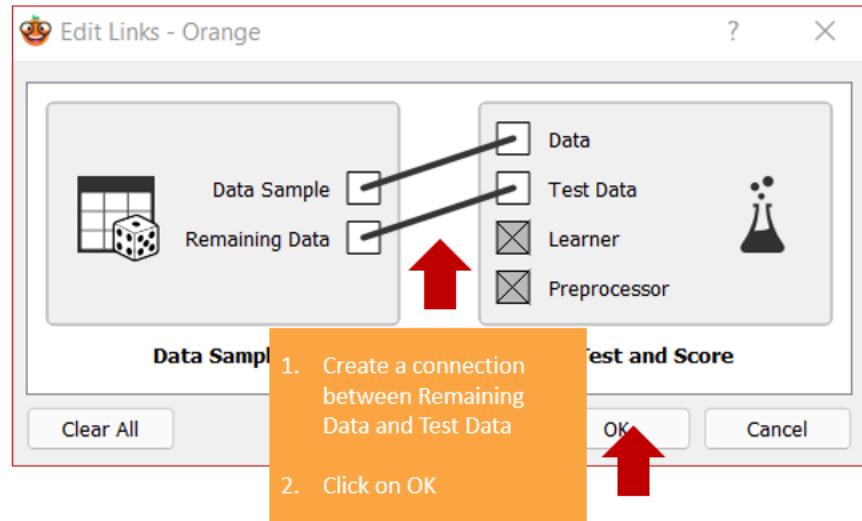
Step 6(b)



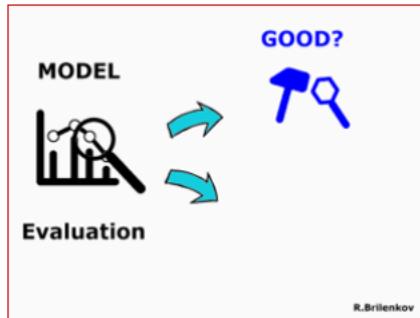
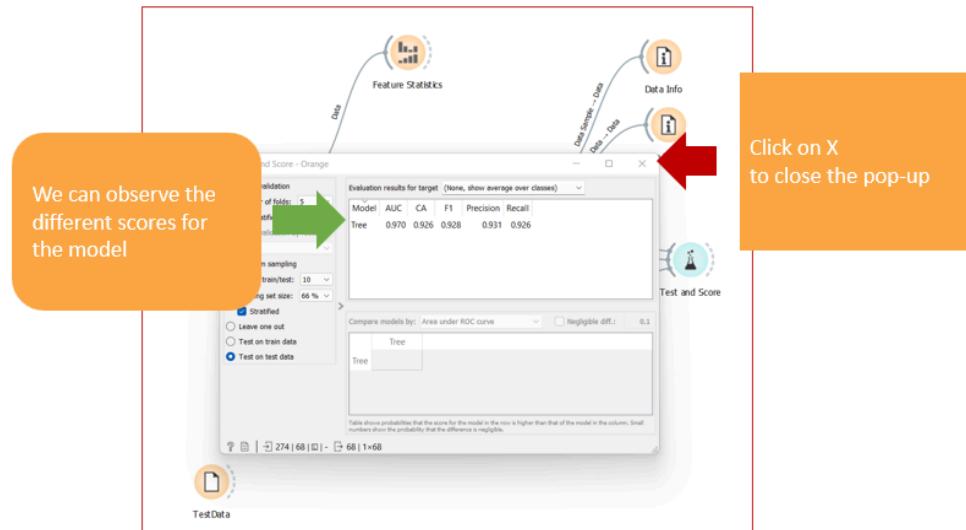
Step 6(b)



Step 6(c)



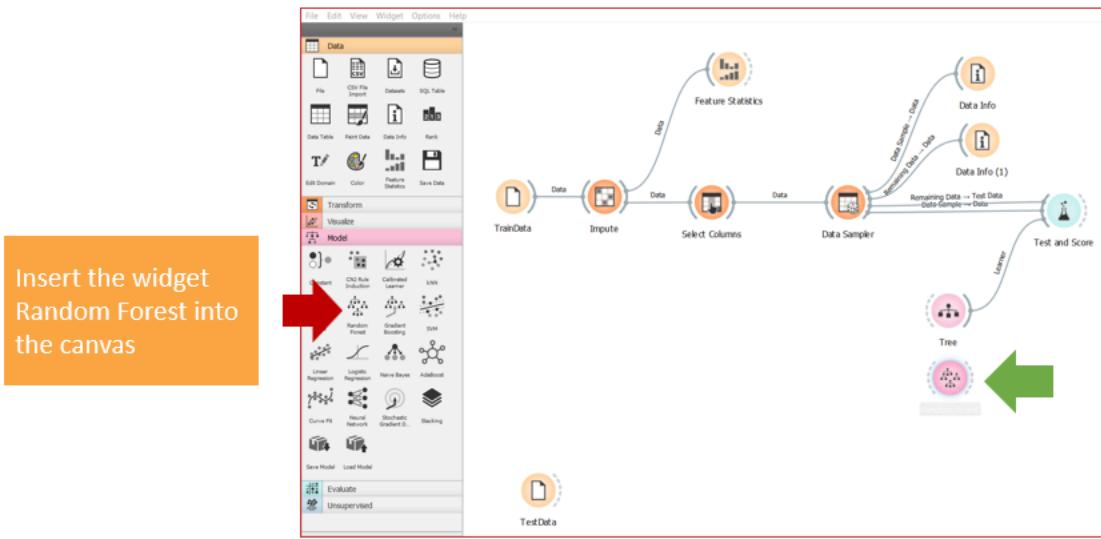
Step 6(d)



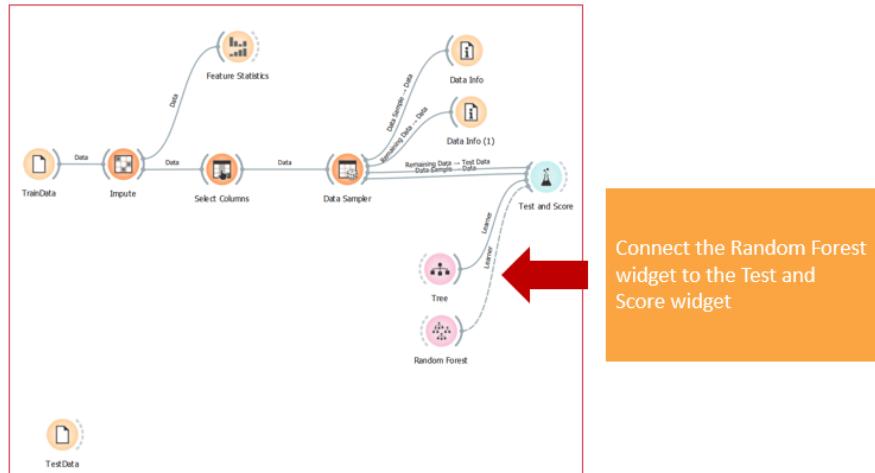
Through Evaluation we know if a model is good or bad

Let's try a couple of other classification algorithms

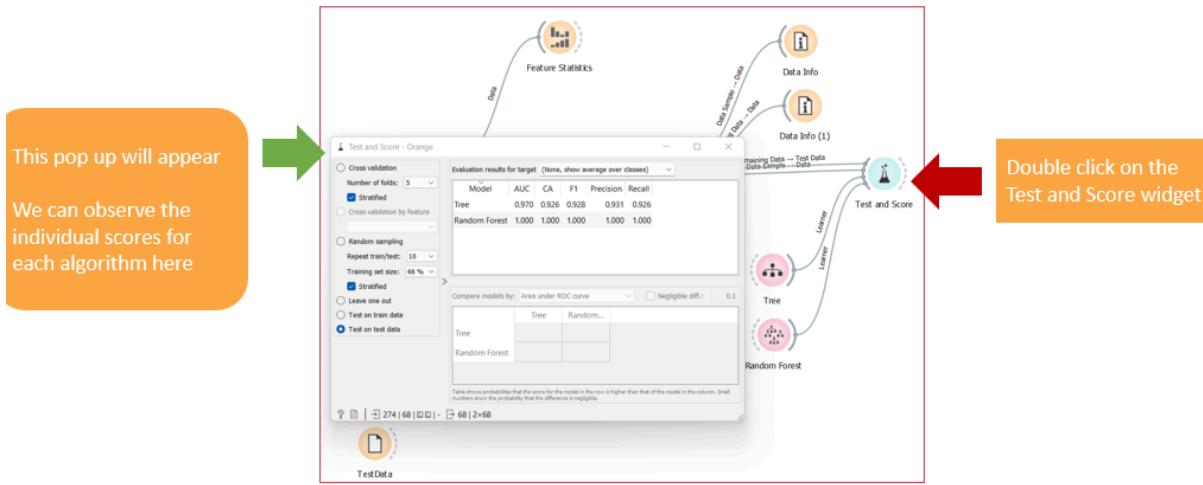
Step 6(h)



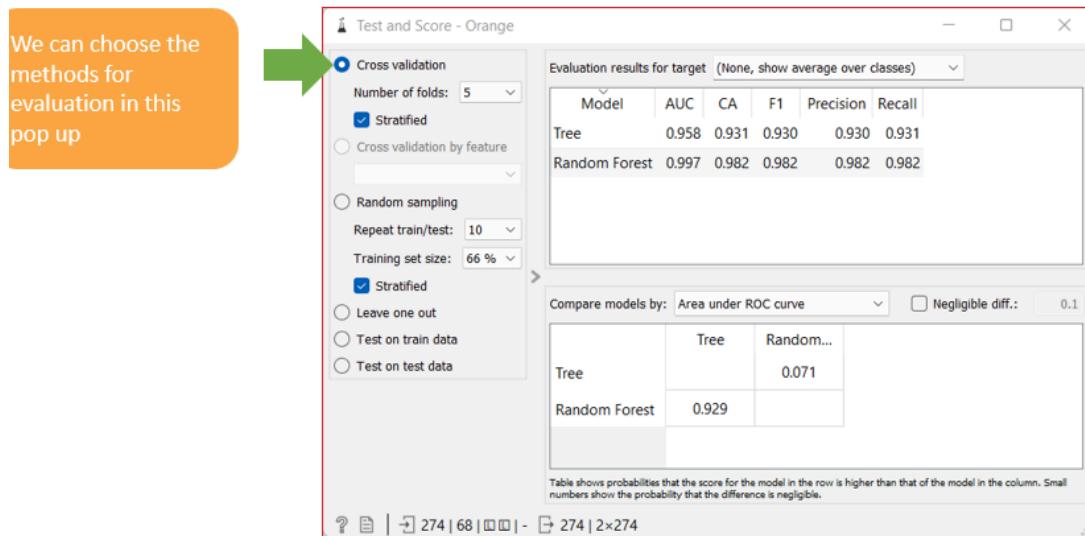
Step 6(i)



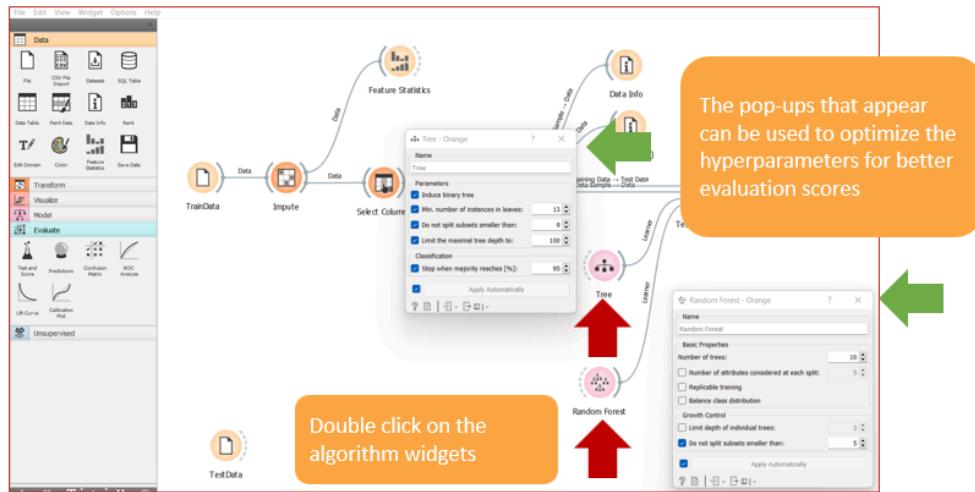
Step 6(j)



Step 6(k)



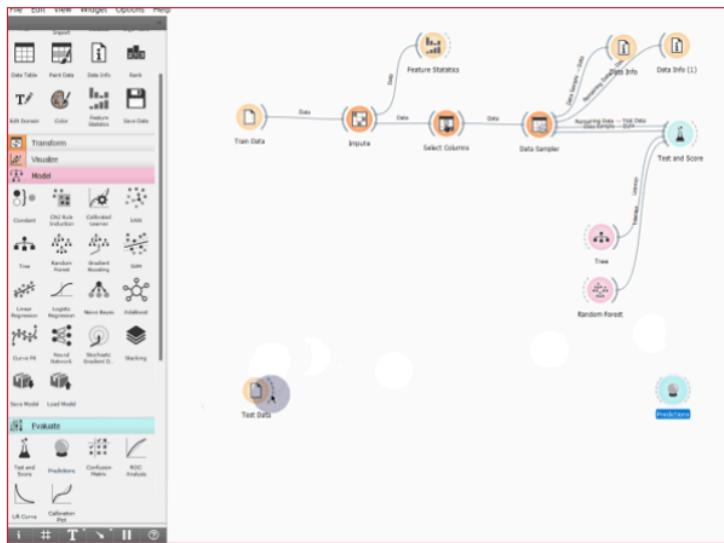
Step 6(l)



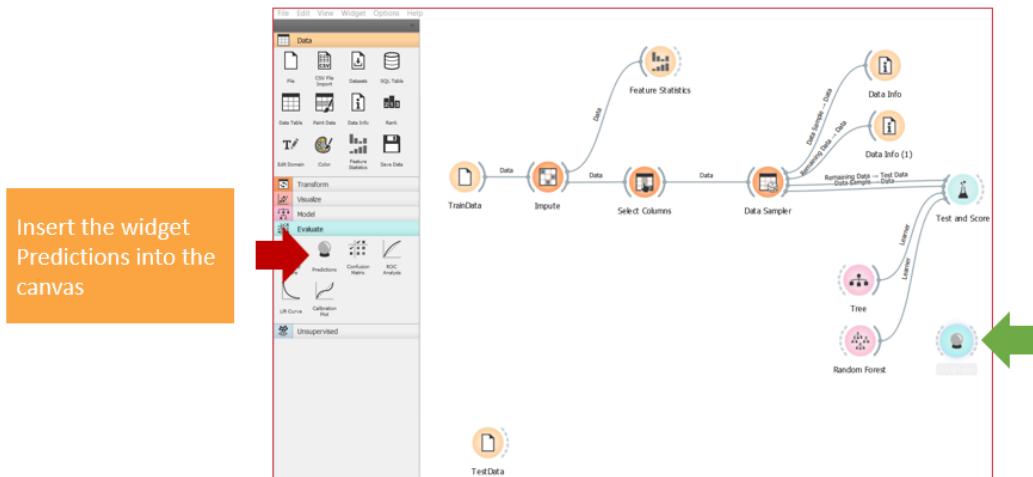
Now that we have found which model gives us the best results, we can use that one!

Step 7: Predictions

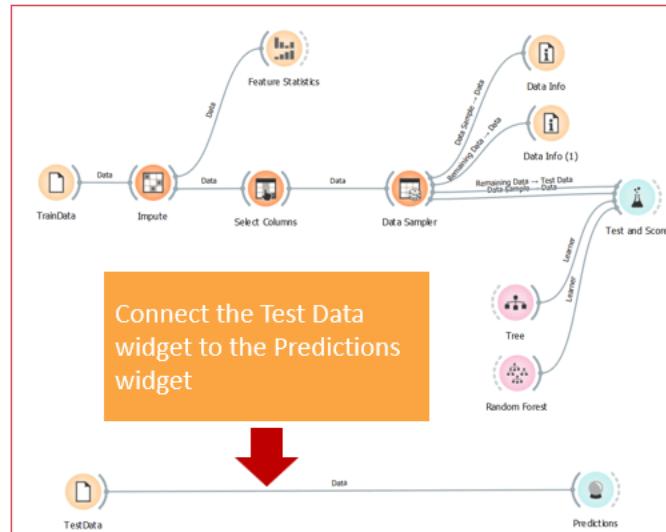
Predictions



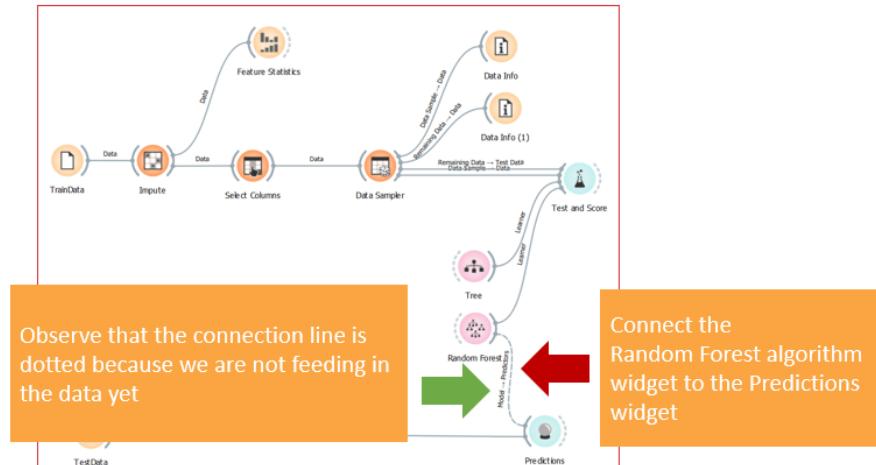
Step 7(a)



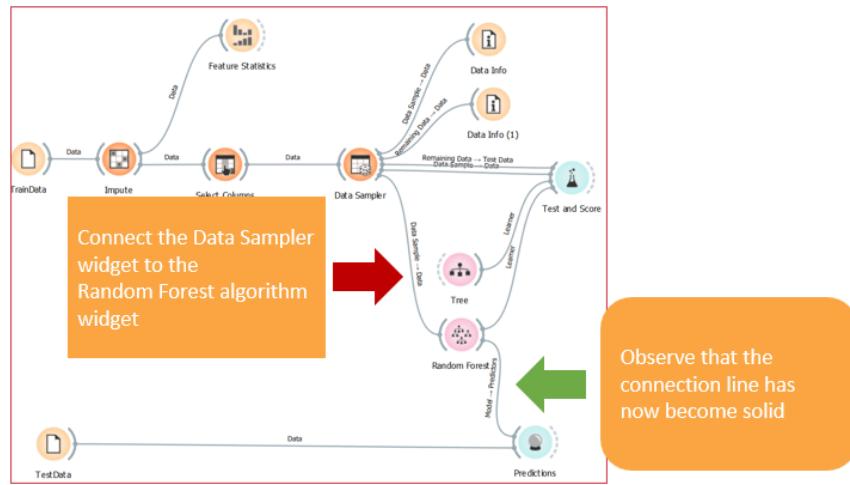
Step 7(b)



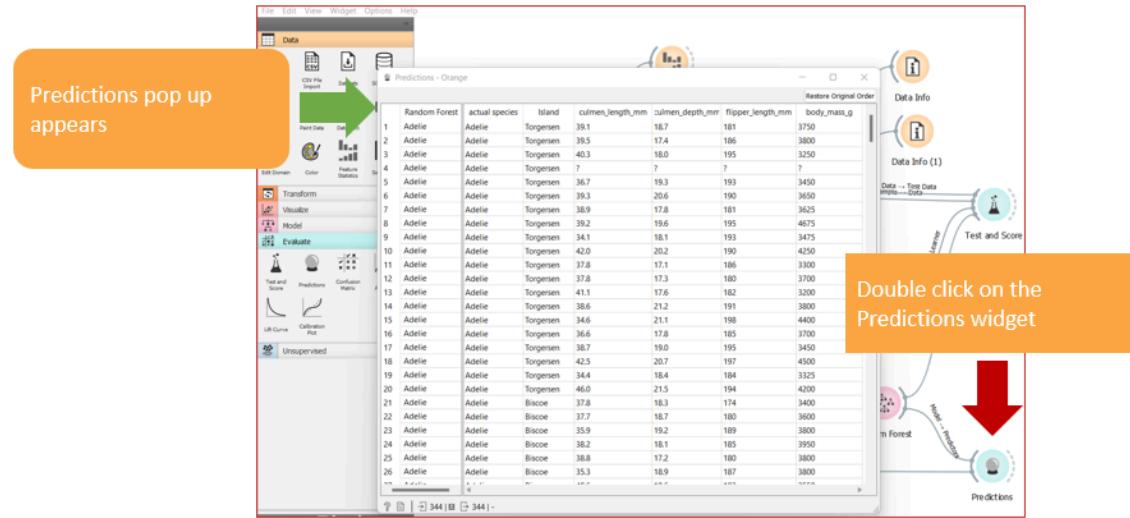
Step 7(b)



Step 7(c)



Step 7(d)



Random Forest Predictions

Actual Species

Observe that the predictions made for Chinstrap by Random Forest are false

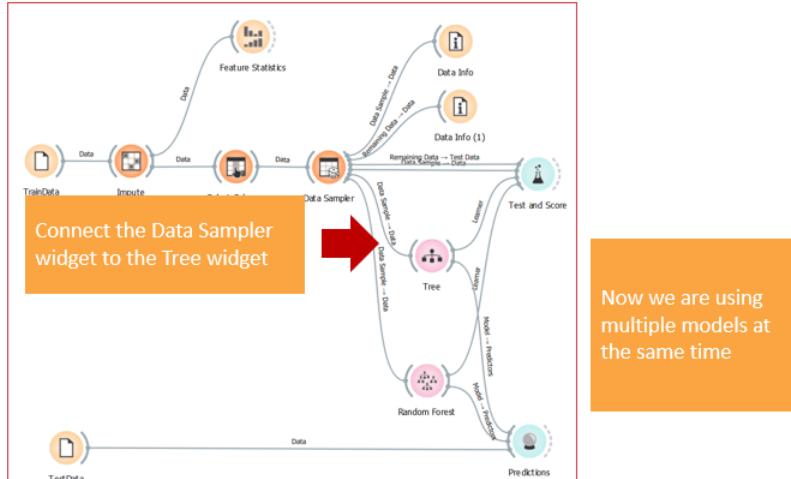
Random Forest is classifying Chinstrap as Adelie

Random Forest	actual species	Island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
156	Adelie	Chinstrap	45.4	18.7	188	3525
157	Adelie	Chinstrap	52.7	19.8	197	3725
158	Adelie	Chinstrap	45.2	17.8	198	3950
159	Adelie	Chinstrap	46.1	18.2	178	3250
160	Adelie	Chinstrap	51.3	18.2	197	3750
161	Adelie	Chinstrap	46.0	18.9	195	4150
162	Adelie	Chinstrap	51.3	19.9	198	3700
163	Adelie	Chinstrap	46.6	17.8	193	3800
164	Adelie	Chinstrap	51.7	20.3	194	3775
165	Adelie	Chinstrap	47.0	17.3	185	3700
166	Adelie	Chinstrap	52.0	18.1	201	4050
167	Adelie	Chinstrap	45.9	17.1	190	3575
168	Adelie	Chinstrap	50.5	19.6	201	4050
169	Adelie	Chinstrap	50.3	20.0	197	3300
170	Adelie	Chinstrap	58.0	17.8	181	3700
171	Adelie	Chinstrap	46.4	18.6	190	3450
172	Chinstrap	Chinstrap	49.2	18.2	195	4400
173	Adelie	Chinstrap	42.4	17.3	181	3600
174	Adelie	Chinstrap	48.5	17.5	191	3400
175	Adelie	Chinstrap	43.2	16.6	187	2900
176	Adelie	Chinstrap	50.6	19.4	193	3800
177	Adelie	Chinstrap	46.7	17.9	195	3300
178	Adelie	Chinstrap	52.0	19.0	197	4150
179	Adelie	Chinstrap	50.5	18.4	200	3400
180	Adelie	Chinstrap	49.5	19.0	200	3800
181	Adelie	Chinstrap	46.4	17.8	191	3700

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

Random Forest Predictions

Tree Predictions

Observe that the predictions made for Chinstrap by Tree are correct

This suggests that some models give better results than others

Actual Species