How To: AI

Presenters: Justin Wong, ZhuoFan Chen, Li KeHan

14 October 2022

SEED Student Interest Group

Sharing . Exploration . Enriching . Development

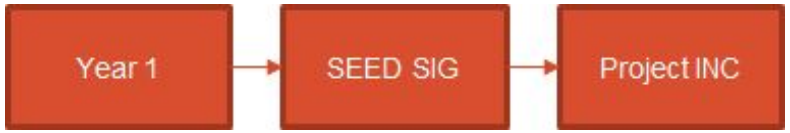
Provide platform to practice SDL

Build up technical and communication skills

Meet and bond with like-minded friends

Cultivate successful habits for life

What we do?



Speaker Profile

Who Are We?

Justin Wong

DAAA Year 2

ZhuoFan Chen

DAAA Year 2

Li KeHan

DIT Year 2

Workshop Outline

1. What is AI and Machine Learning? (20 mins)
2. Building Basic ML Models with Orange (30 mins)
3. Practical (20 mins)
4. Blooket (20 mins)
5. QnA and Conclusion (15 mins)

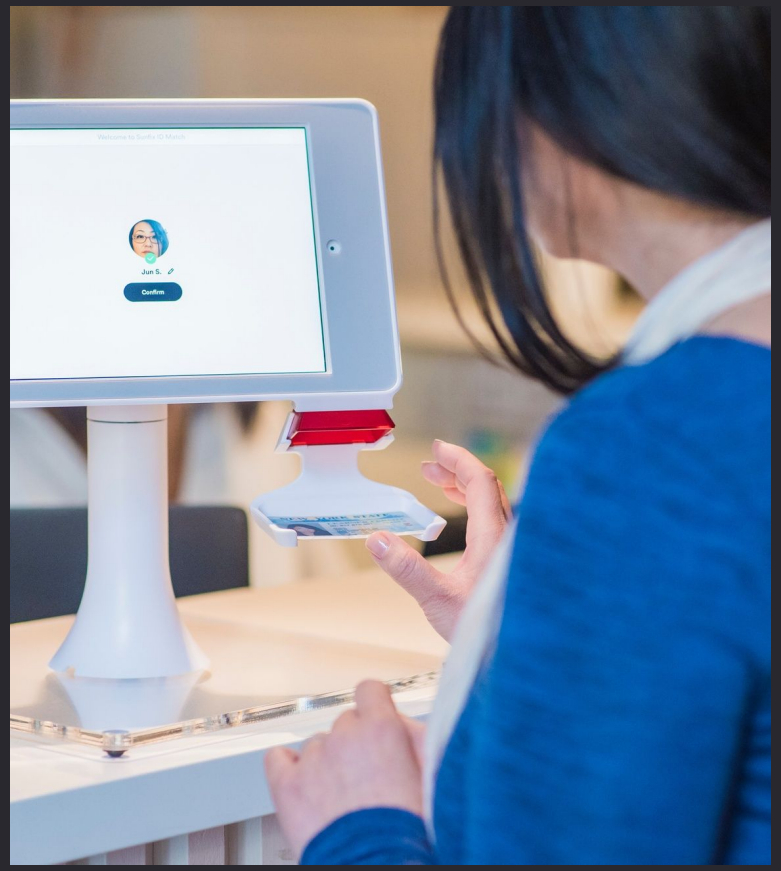
QnA
(Padlet)

Section 1

What is AI and Machine Learning?

Artificial Intelligence

AI is the ability of a computer to perform human tasks



Facial Recognition

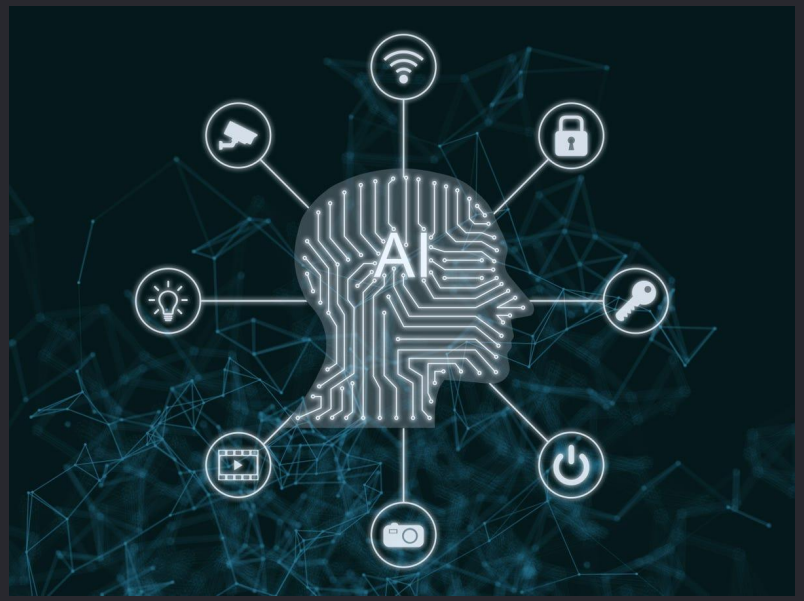
Voice Assistants

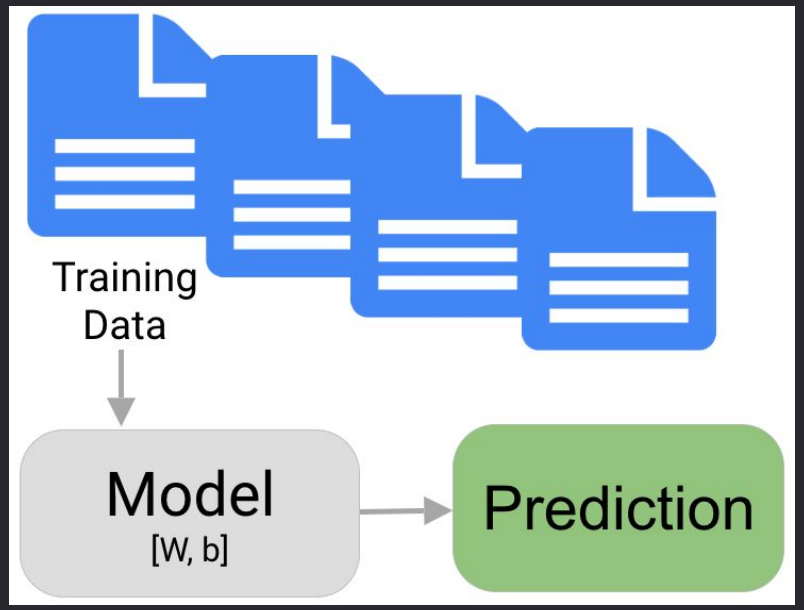
Autonomous Driving

Guess the AI Application?

What is Machine Learning?

- Machine Learning is defined by Mitchell (1997) as the study of computer algorithms that improves automatically through experience
- ML models learn from data, and make predictions.



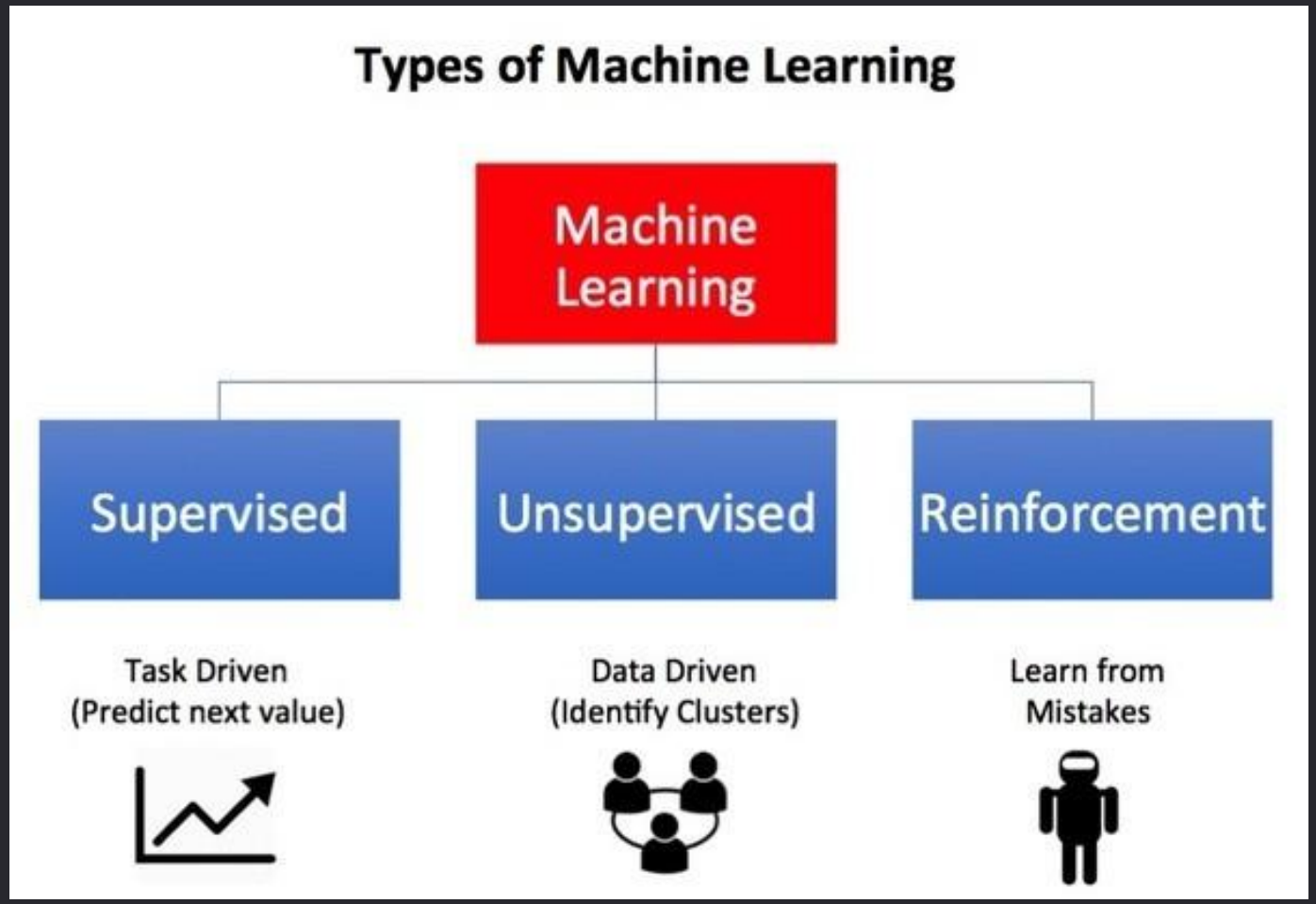


Traditional Programming

If else switch case……

Machine Learning

Type of machine learning

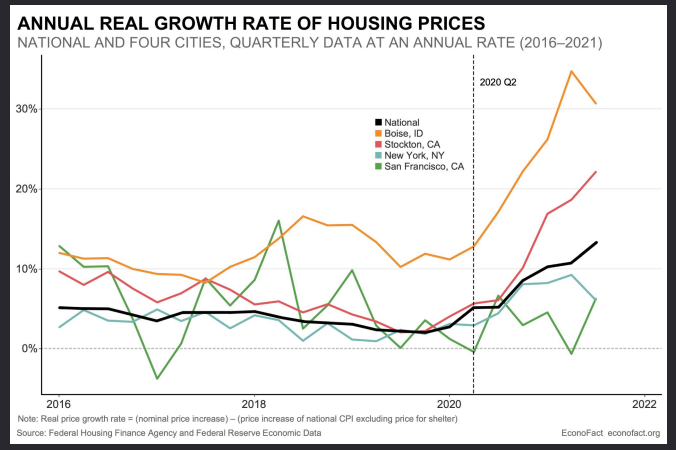


Supervised Learning

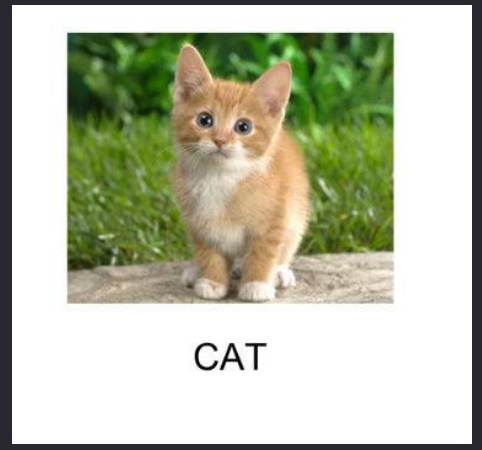
- Data with clearly defined output is given
- Direct feedback is given
- Predicts outcome/future
- Resolves classification and regression problem

Classification

Regression



Classification



Unsupervised Learning

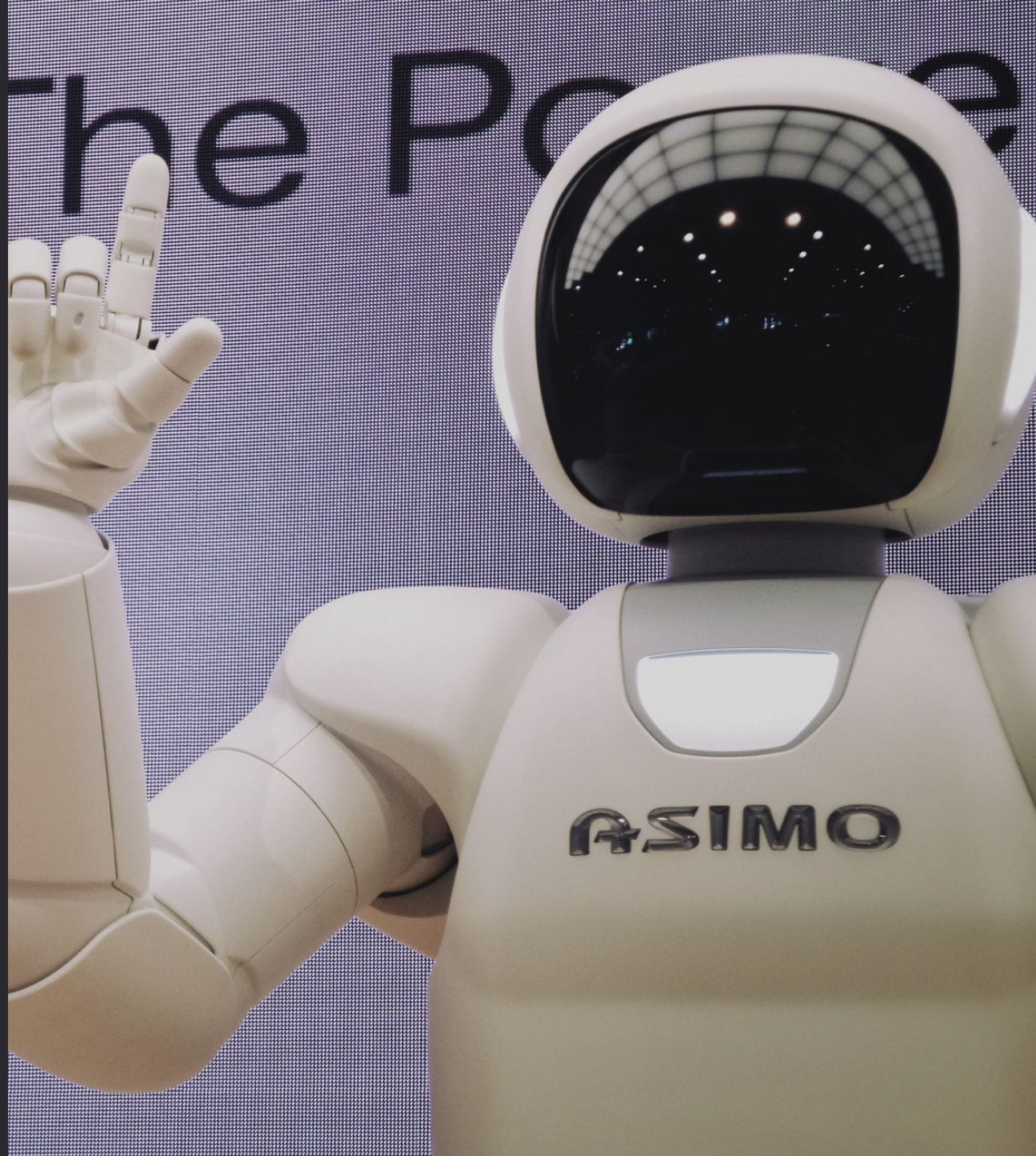
- Machine understands the data (Identifies patterns structures)
- Evaluation is qualitative or indirect
- Does not predict/find anything specifically
- Clustering solutions

Clustering

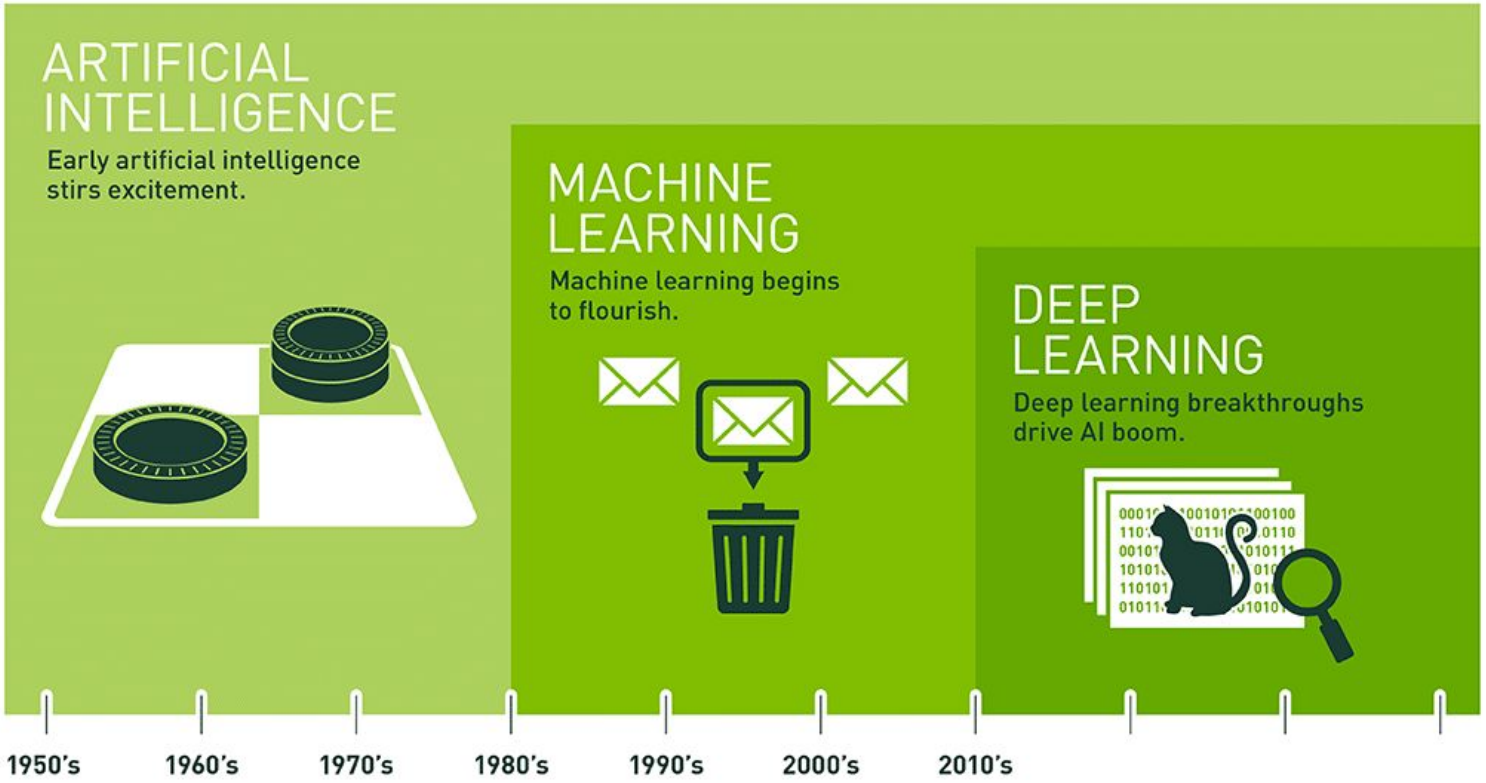
Clustering

Deep Learning

Subfield of Machine Learning using Neural Networks



Relationship between AI, ML and DL



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Section 2

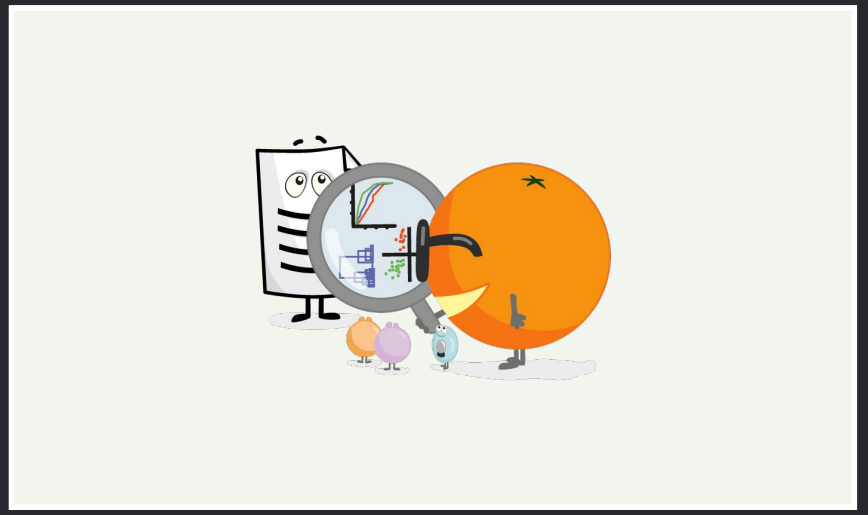
Building Machine Learning Models with Orange

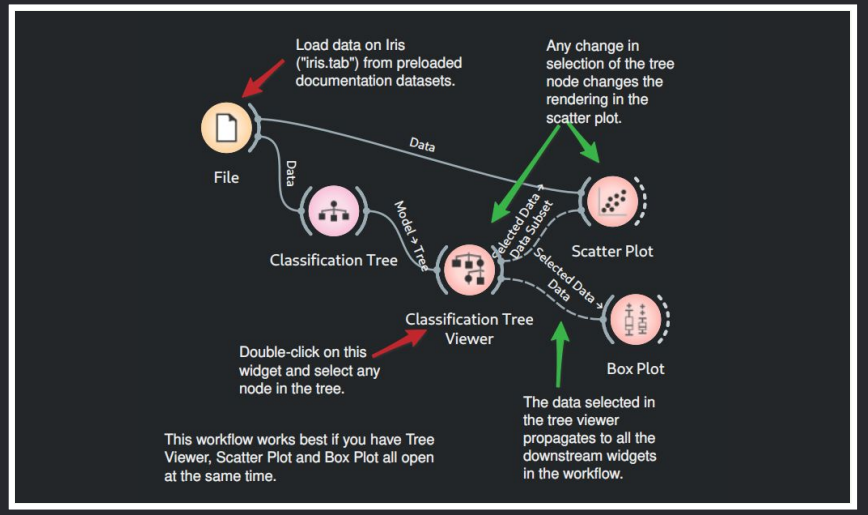
What is Orange?

- Introduction

  A component based visual programming software for data mining

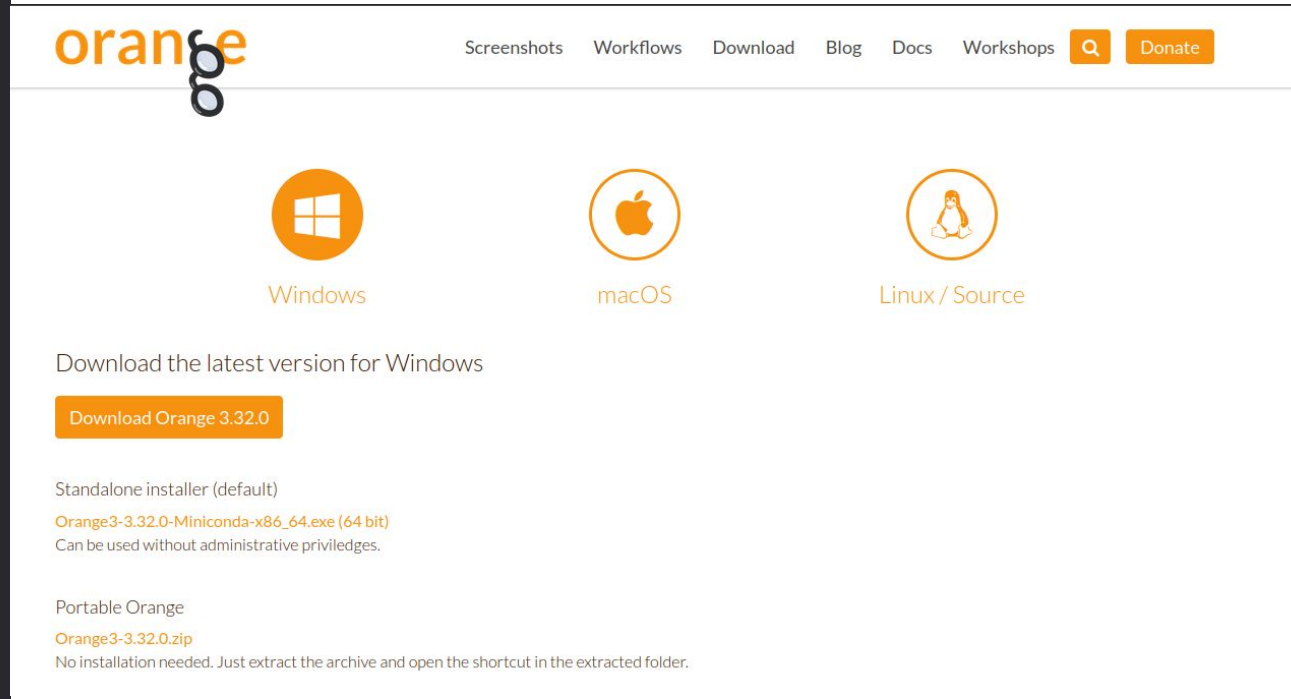
- Why Orange?
  - No Programming Needed
  - Open-Source Software





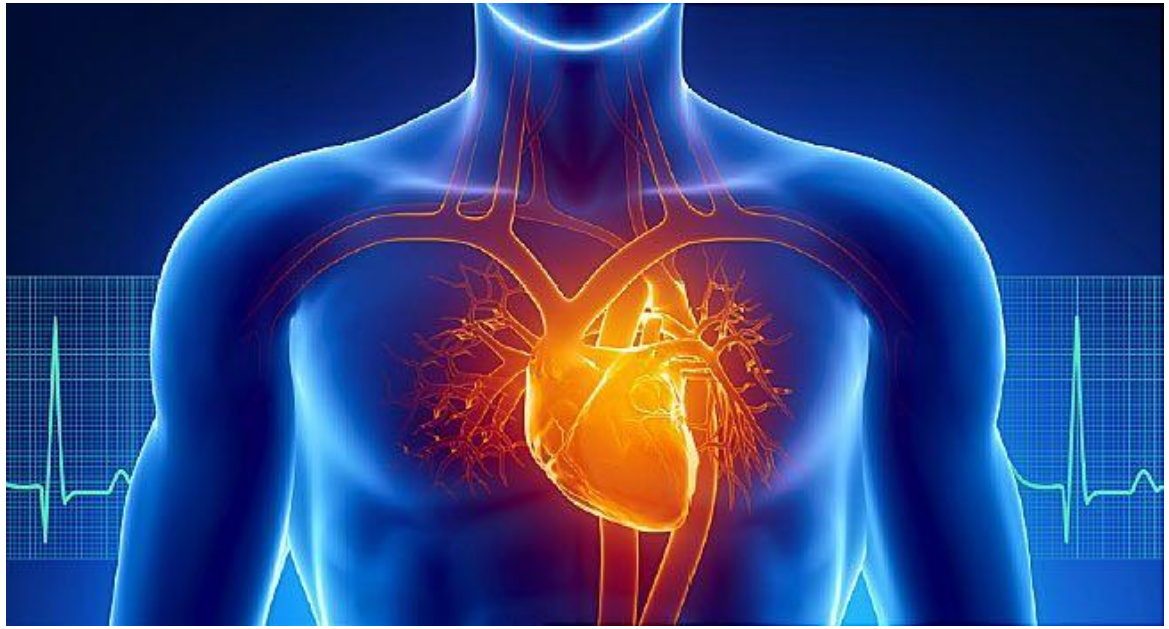
Installing Orange

https://orangedatamining.com/download



Heart Disease Prediction

UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/index.php



Heart disease is one of the leading causes of death in today's world. According to the World Health Organization (WHO), there is an estimated 17.9 million death every year due to heart disease. Furthermore, the number of cases of heart disease was discovered to be rising rapidly in recent years.

Therefore, the diagnosis of the disease became especially important. An accurate prediction can allow practitioners to diagnose heart disease in early-stage and make a more informed decision regarding the patient's treatment.

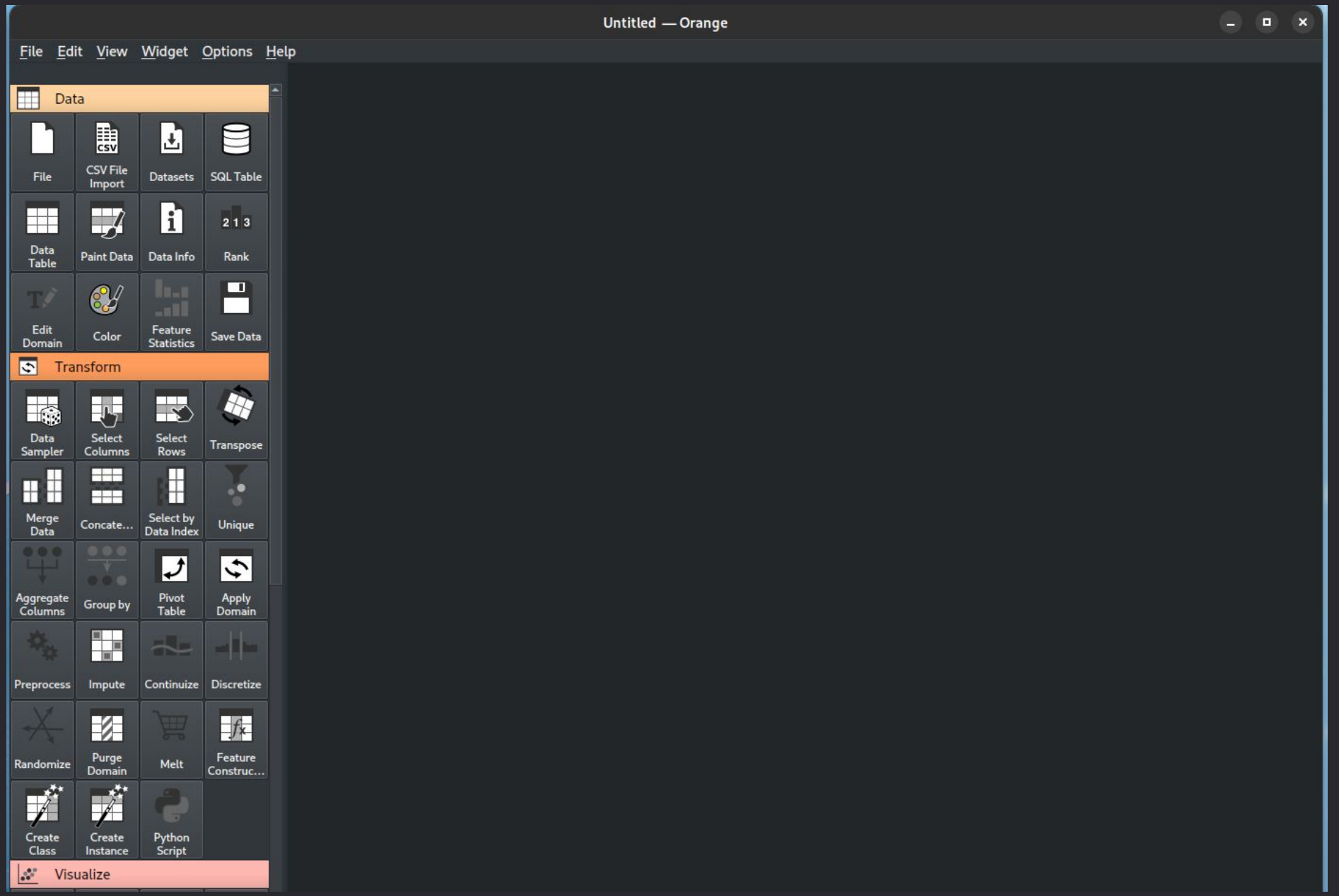
6 Steps of Machine Learning

1. Data Ingestion
2. Data Exploration
3. Data Pre-processing
4. Train Model
5. Evaluate model
6. Summary

Step 1. Data Ingestion

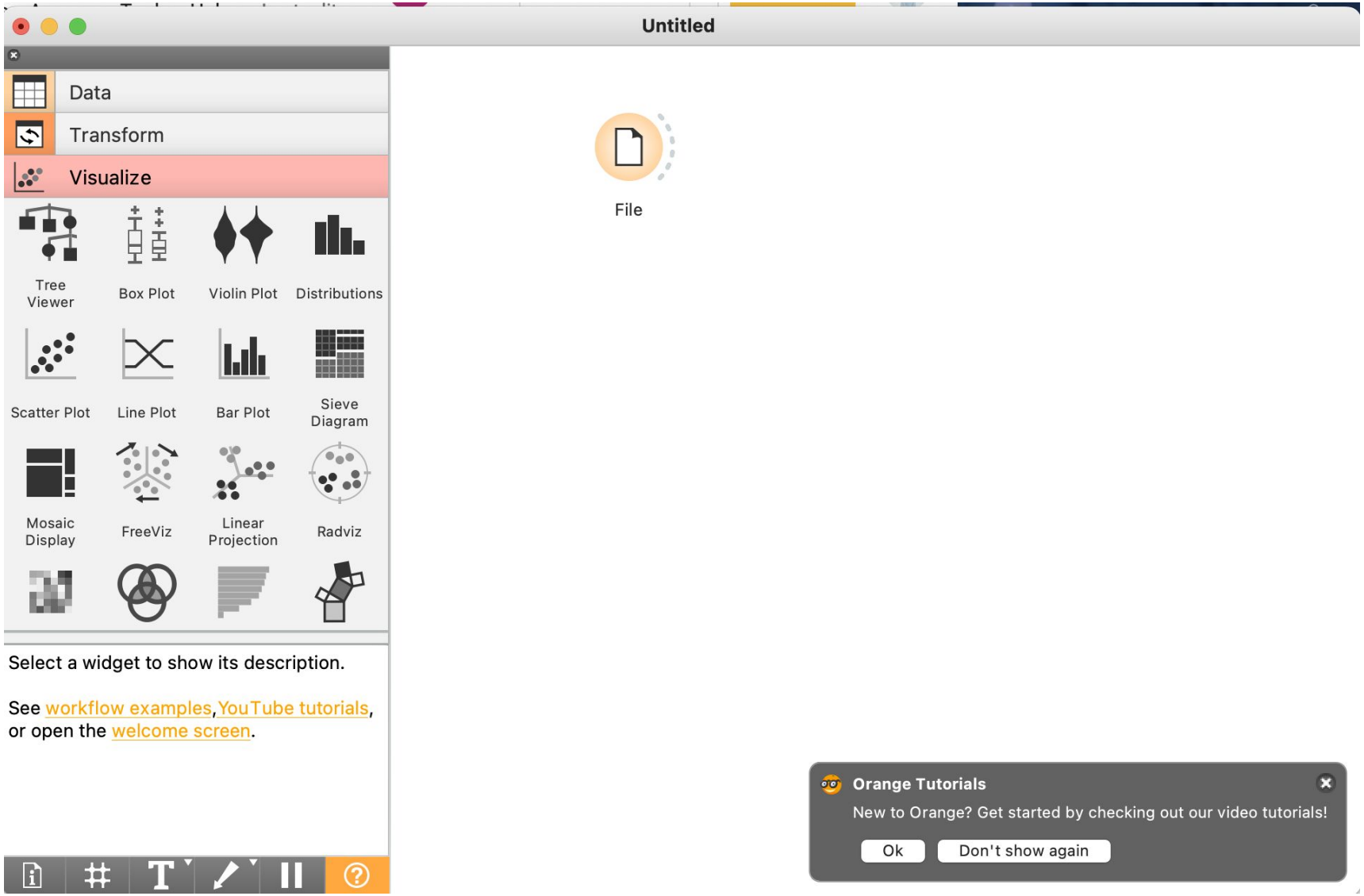
Starting Orange

After you install Orange DM successfully, start the application and navigate around.



Step 1 Data Ingestion

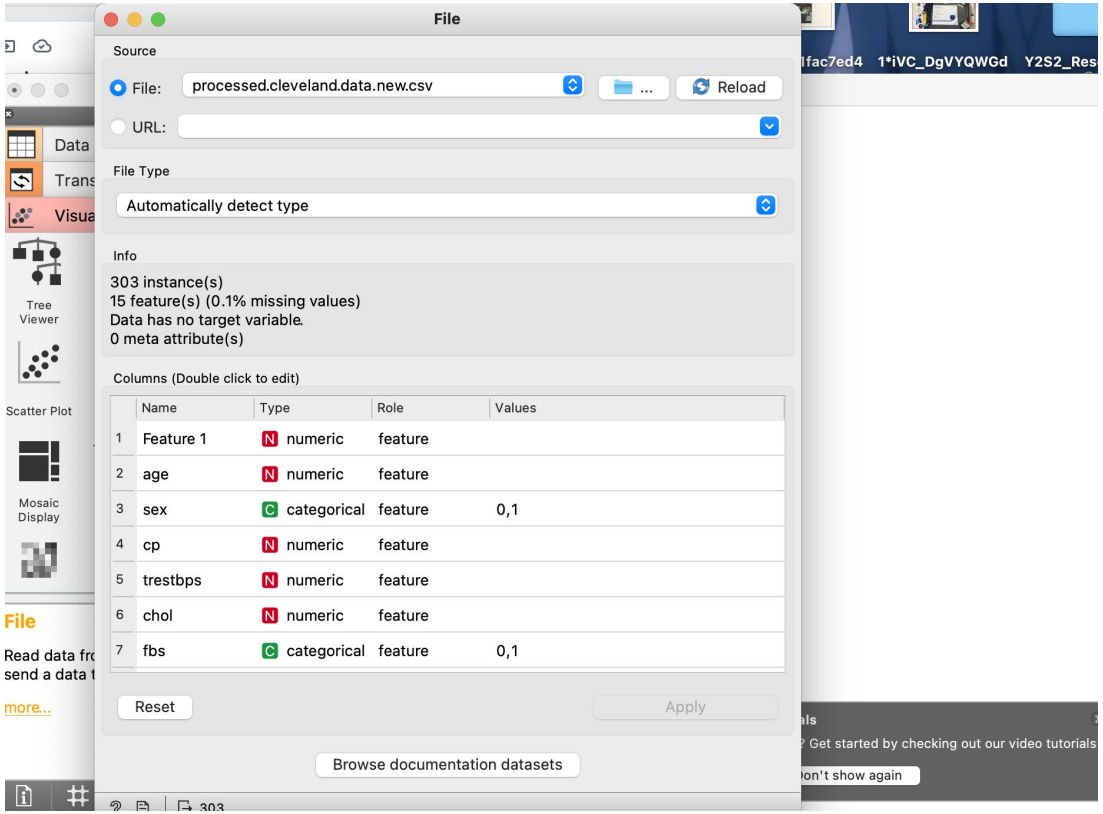
Click on the “Data” tab on the widget selector menu, and drag the widget **”File”** to our blank workflow.



Step 1 Data Ingestion

Double click the “File” widget and select the file that we want to load into the workflow.

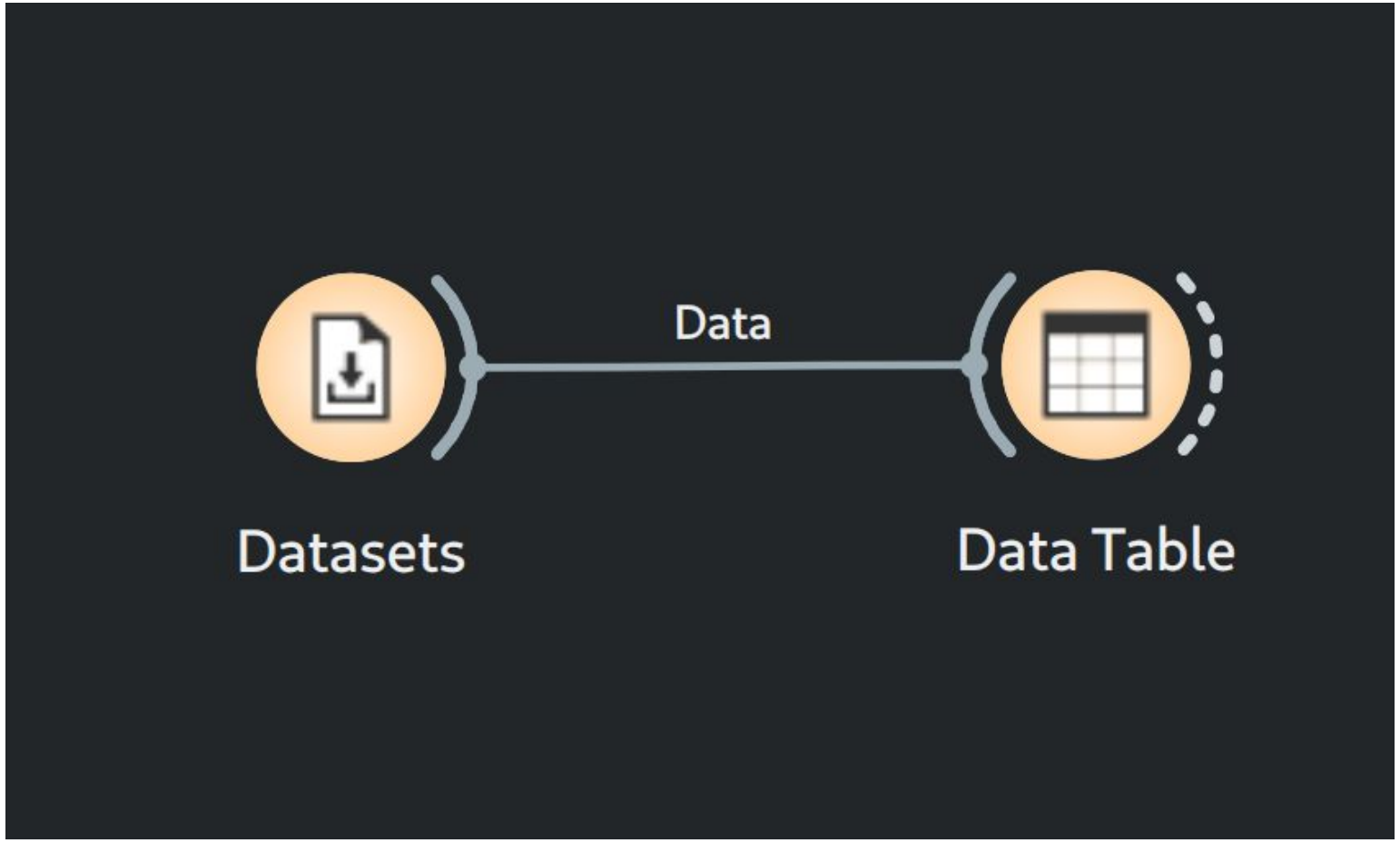
Choose the heart disease dataset that you have downloaded previously.



Step 2. Data Exploration

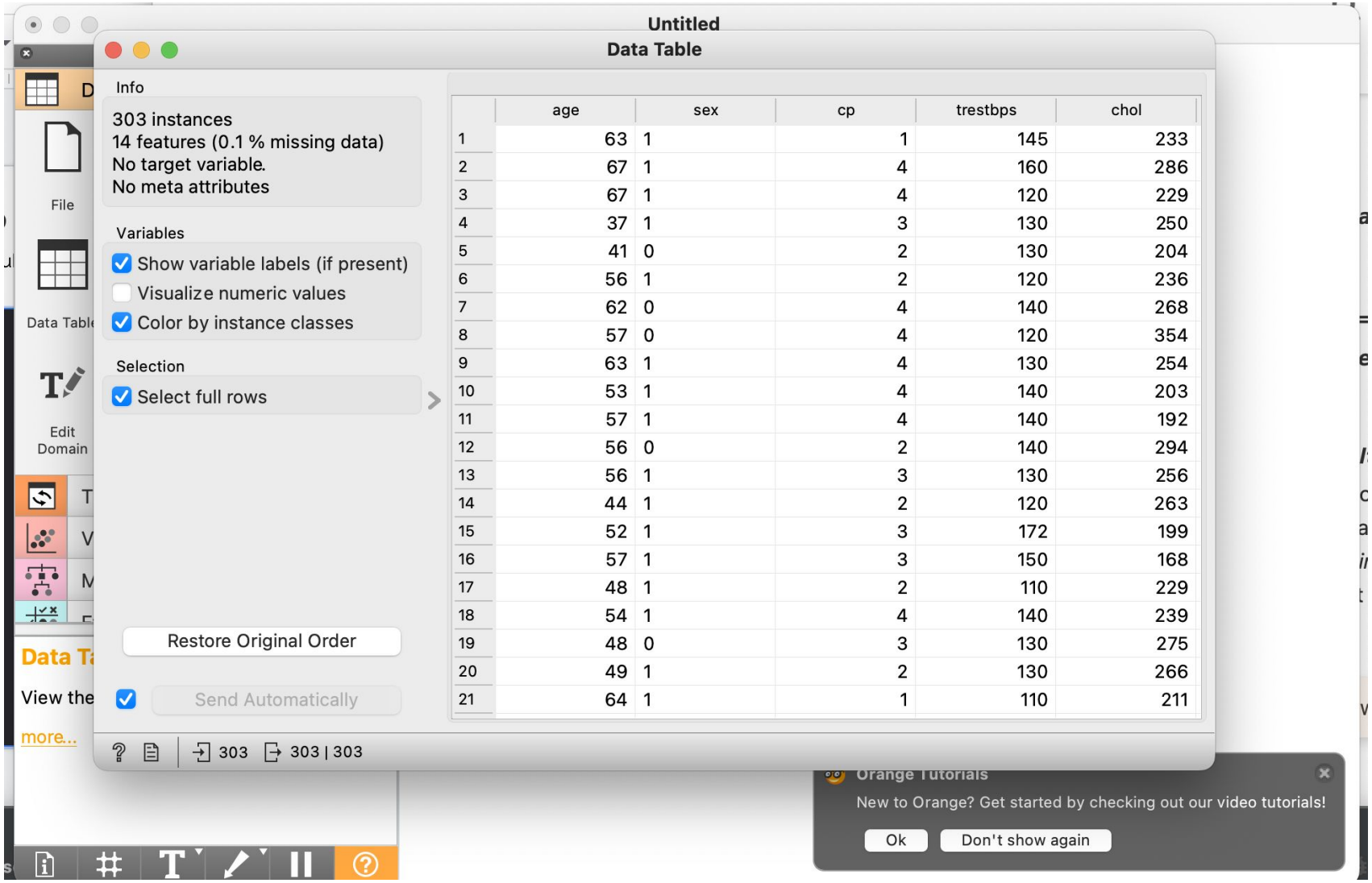
Step 2 Data Exploration

As we need a data table to better visualize our data, select the “Data Table” widget, and connect **Datasets** to **Data Table**.



Step 2

Now double click “Data Table”, you can see the data in a table format.



	age	sex	cp	trestbps	chol
1	63	1	1	145	233
2	67	1	4	160	286
3	67	1	4	120	229
4	37	1	3	130	250
5	41	0	2	130	204
6	56	1	2	120	236
7	62	0	4	140	268
8	57	0	4	120	354
9	63	1	4	130	254
10	53	1	4	140	203
11	57	1	4	140	192
12	56	0	2	140	294
13	56	1	3	130	256
14	44	1	2	120	263
15	52	1	3	172	199
16	57	1	3	150	168
17	48	1	2	110	229
18	54	1	4	140	239
19	48	0	3	130	275
20	49	1	2	130	266
21	64	1	1	110	211

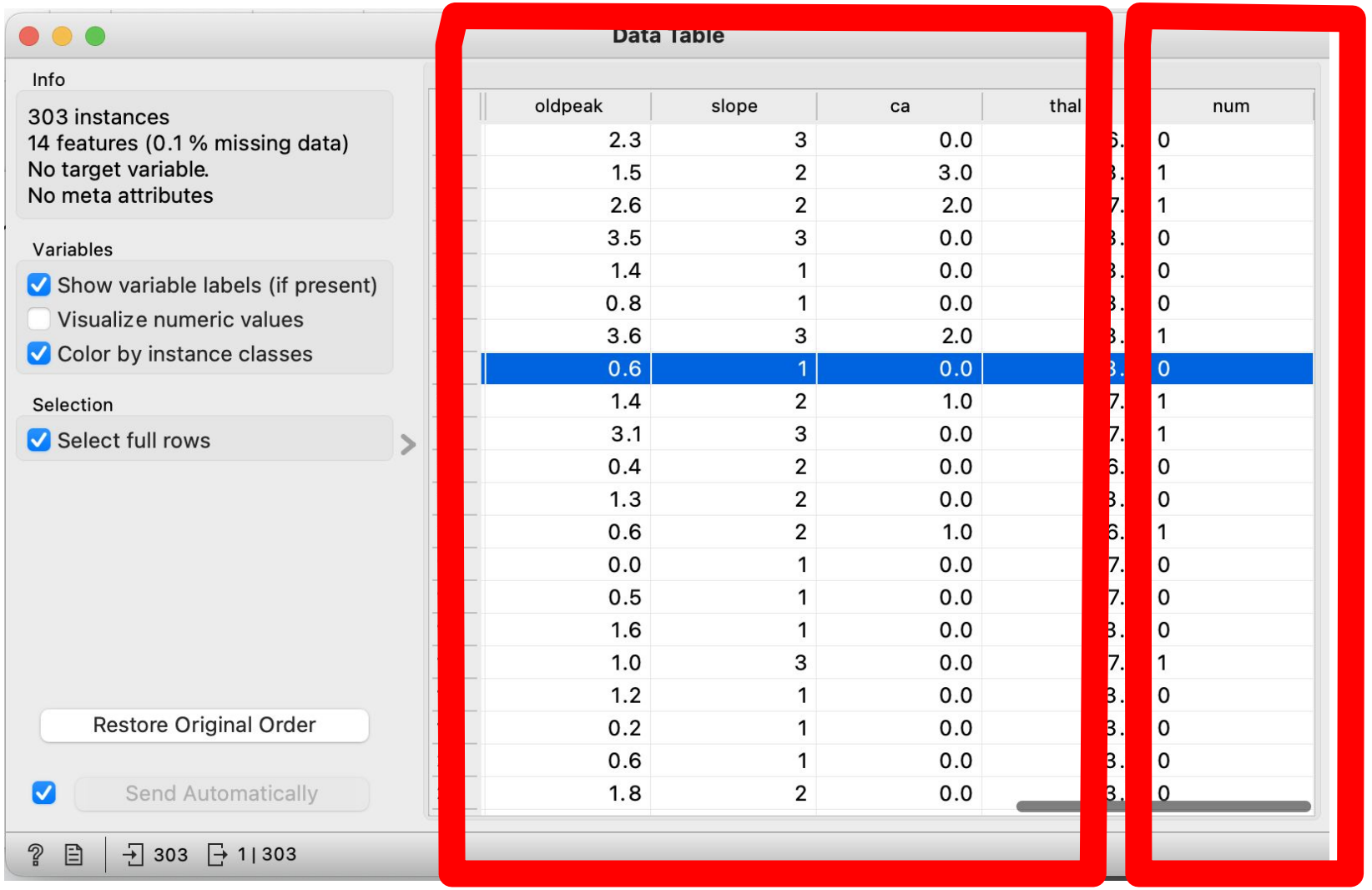
Age(age)	Integer value
Sex (sex)	0 – female; 1 – male
Chest pain type (cp)	0 – asymptomatic; 1 – atypical angina; 2 – non-anginal pain; 3 – typical angina
Resting blood pressure (trestbps)	Integer value
Cholesterol	Integer value
Fasting Blood Pressure >120 mg/dl (fps)	0 – no; 1 – yes
Electrocardiogram in rest condition (restecg)	0 – normal; 1 – having aberrant ST-T wave; 2 – showing probable or definite left ventricular hypertrophy
Maximum heart rate (thalach)	Integer value
Exercise induced angina	0 – no; 1 – yes
ST depression induced by exercise ST segment (oldpeak)	Integer value
The slope of the peak exercise ST segment	1 – upsloping; 2 – flat; 3 – downsloping
Number of colored vessels by fluoroscopy (ca)	Integer value ranged 0-3
Thalassemia status (thal)	3 – normal; 6 – fixed defect; 7 – reversible defect
Angiographic disease status (num)	0 – absence heart disease; 1 – presence of heart disease

Feature and Label

This is call a **label**
- An outcome
- Dependent/ response variable

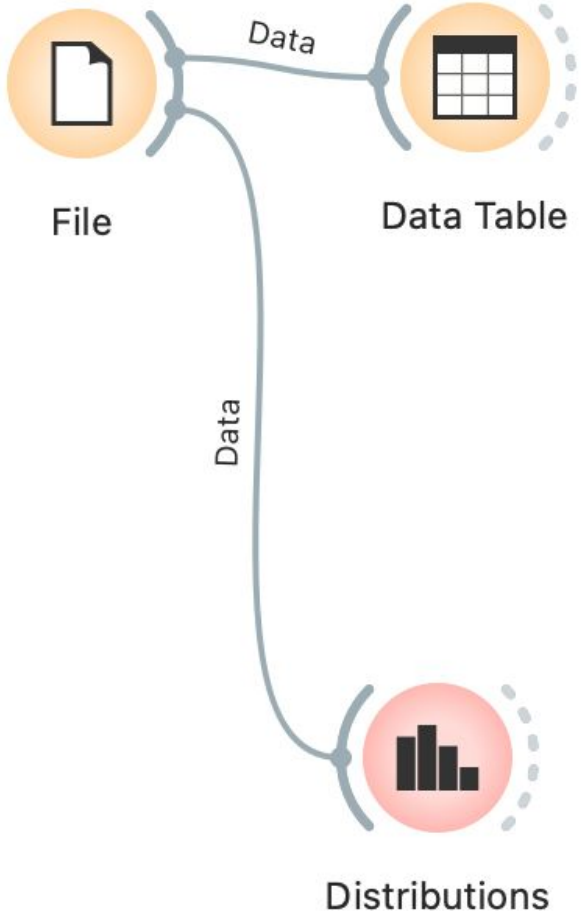
This is called **feature**
- Characteristic

or property



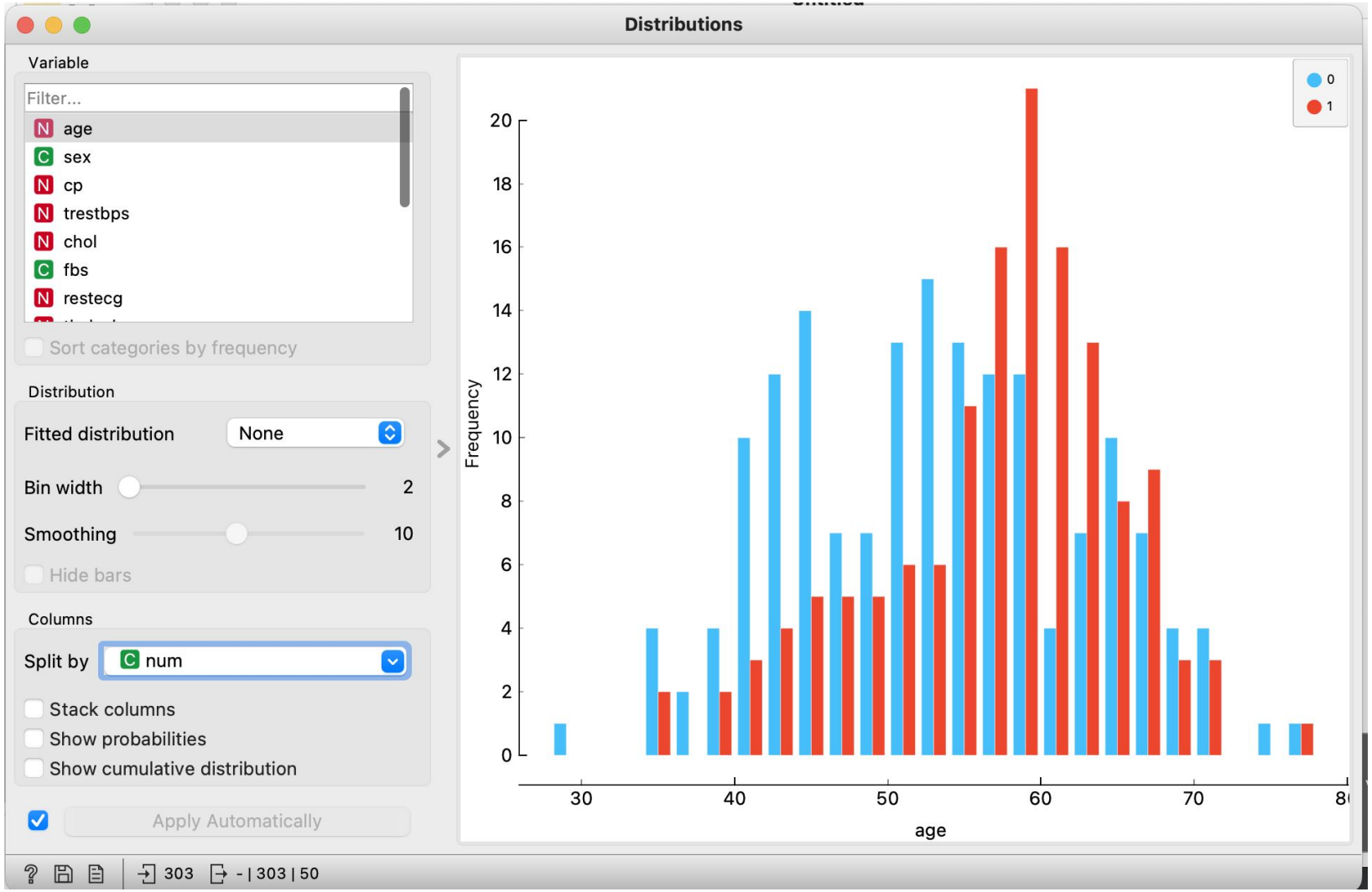
Step 2: Data Visualisation

Click on the semicircle in front of the “Datasets” widget and drag it to an empty space in the workflow and select the ”Distribution Plot” widget.



Step 2: Distribution Plot

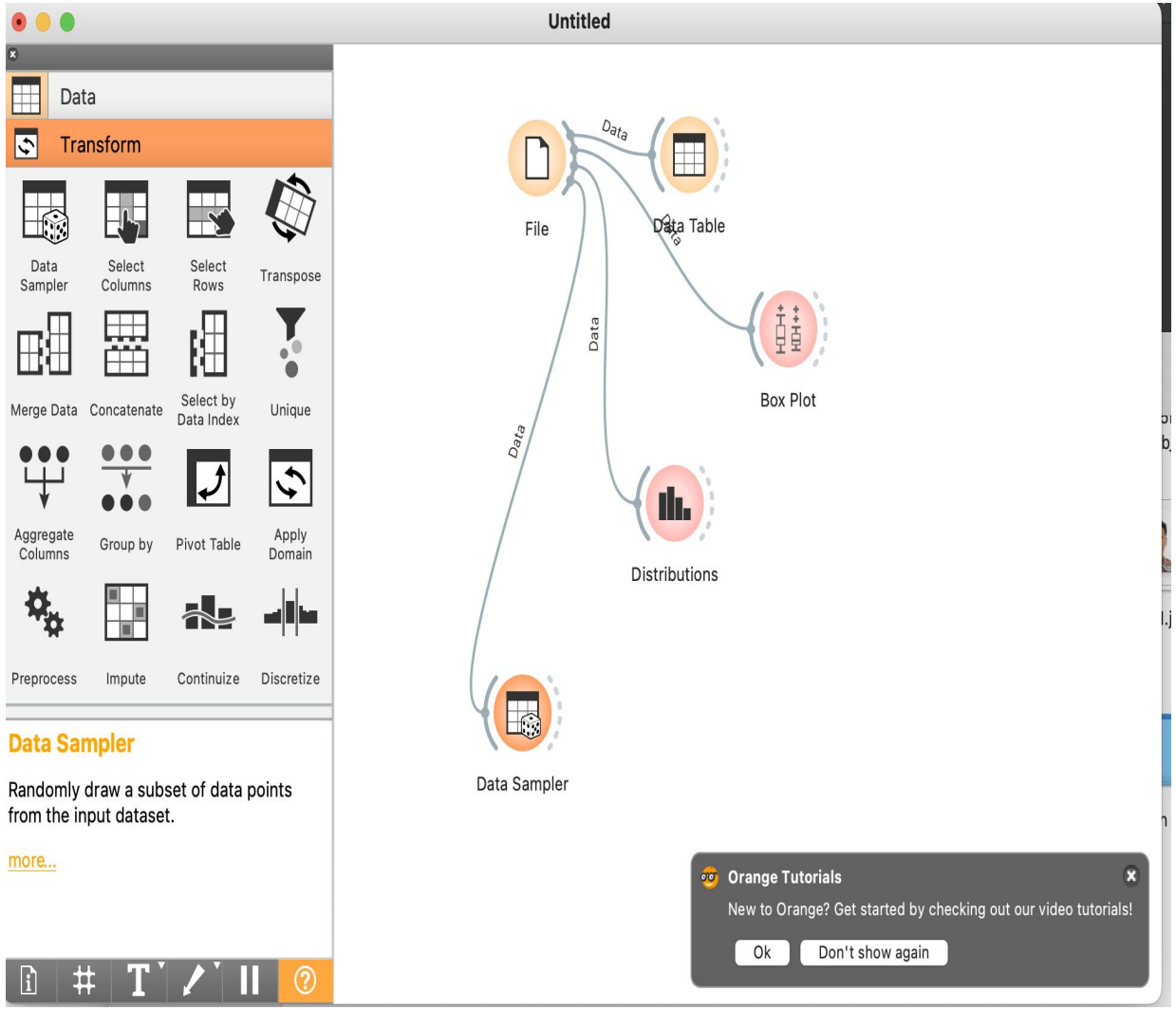
Once you create a Distribution Plot widget, double click it and explore your data! You can select X and Y axis, colors, shapes, and a lot of other manipulations.

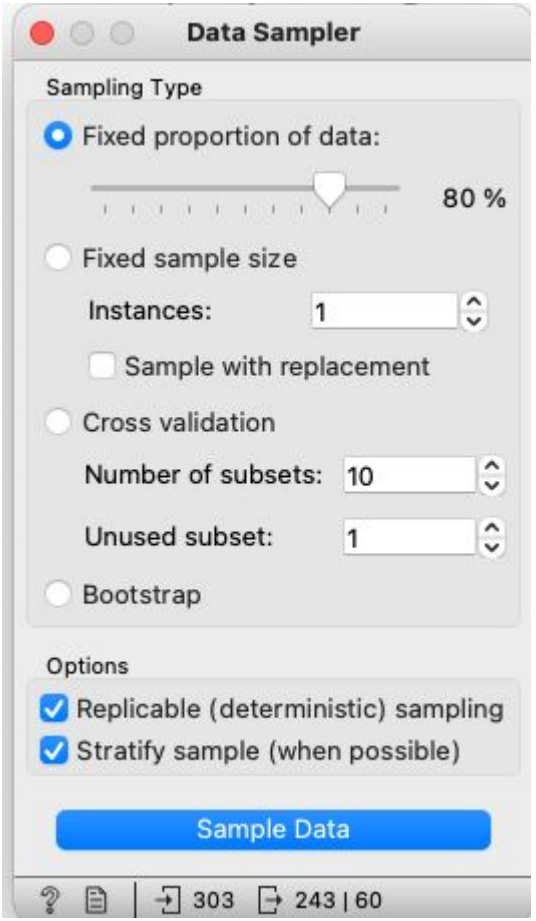


Step 3. Data Pre-processing

Step 3: Data Split

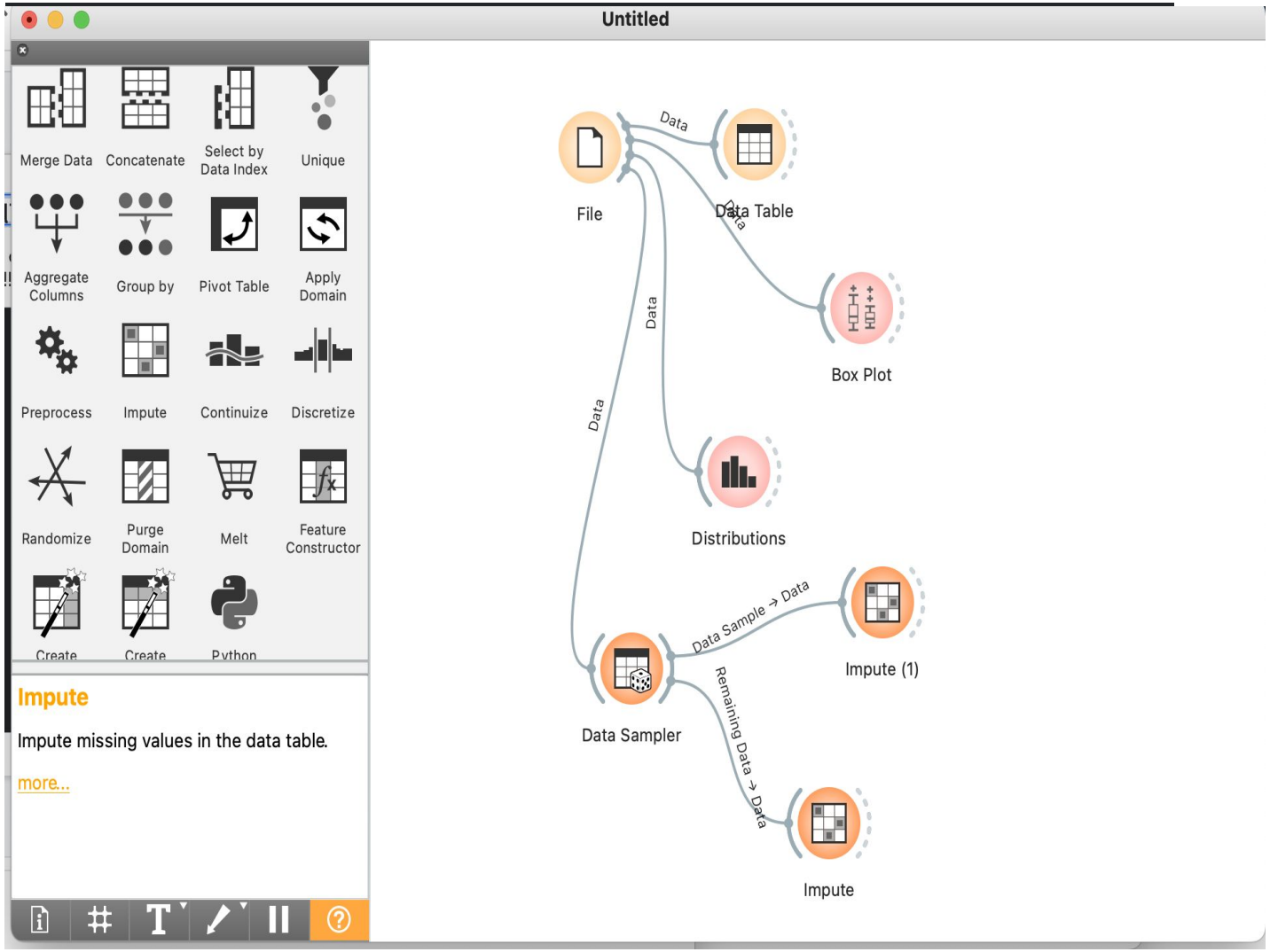
Now, use the "**Data sampler**" to split the data into training set and testing set





Step 3: Data Imputation

Remember that there is missing data previously?
Now, it is time to manage them !!!

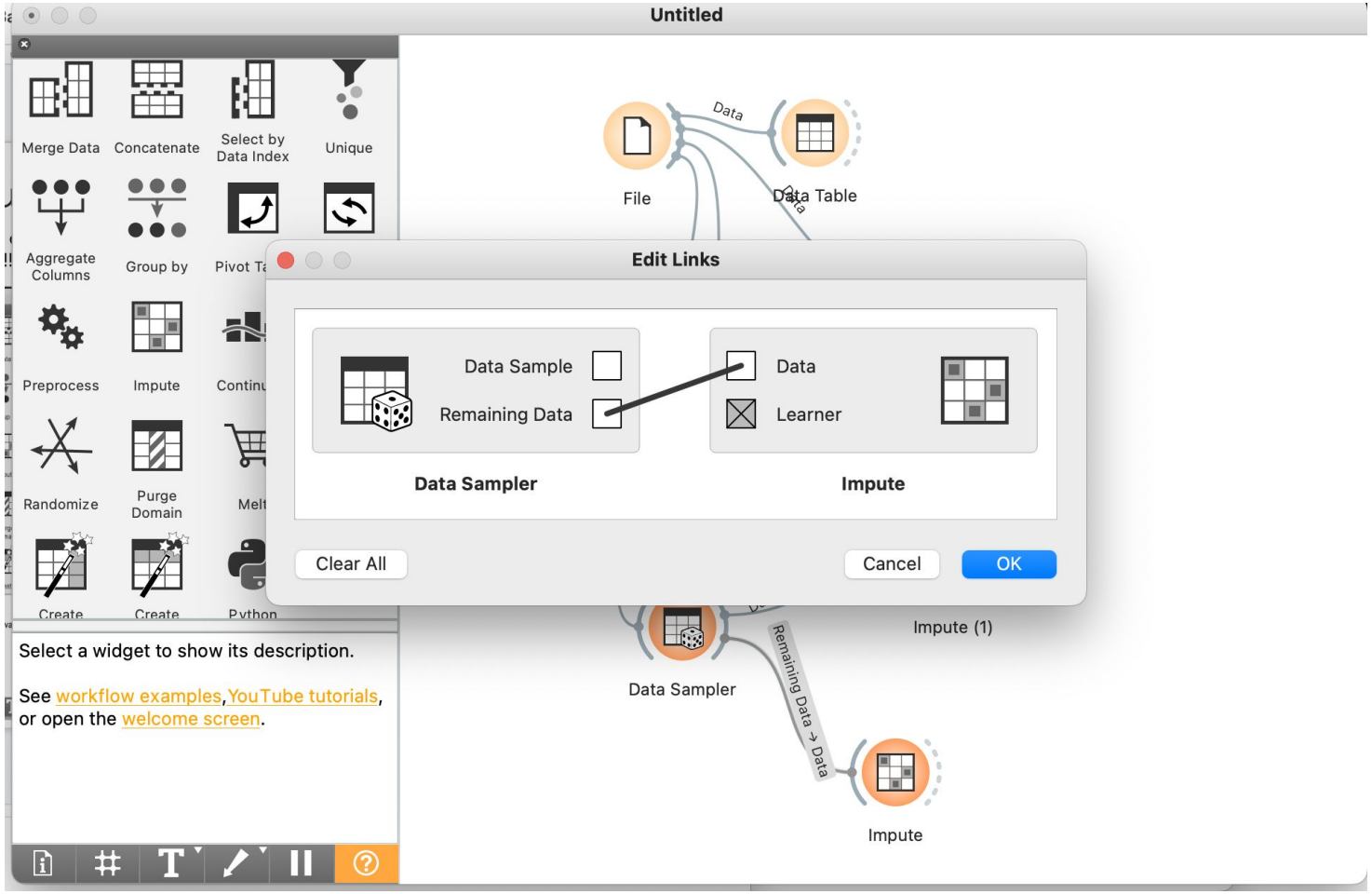


Data Leakage

Data leakage is when information from outside the training dataset is used to create the model. This additional information can allow the model to learn or know something that it otherwise would not know and in turn invalidate the estimated performance of the mode being constructed.

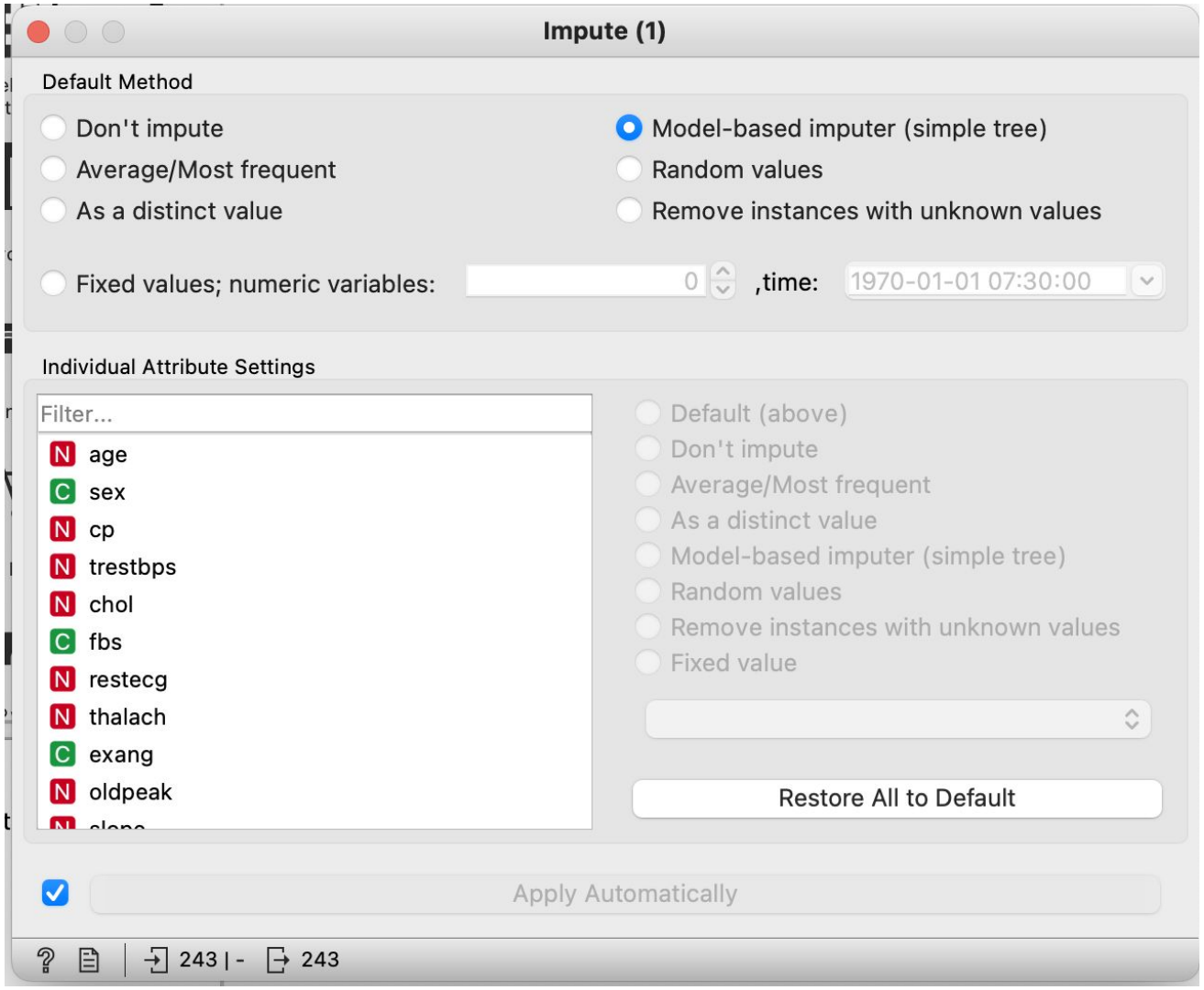
Step 3: Data Imputation

Make sure the linking is correct !!



Step 3: Data Imputation

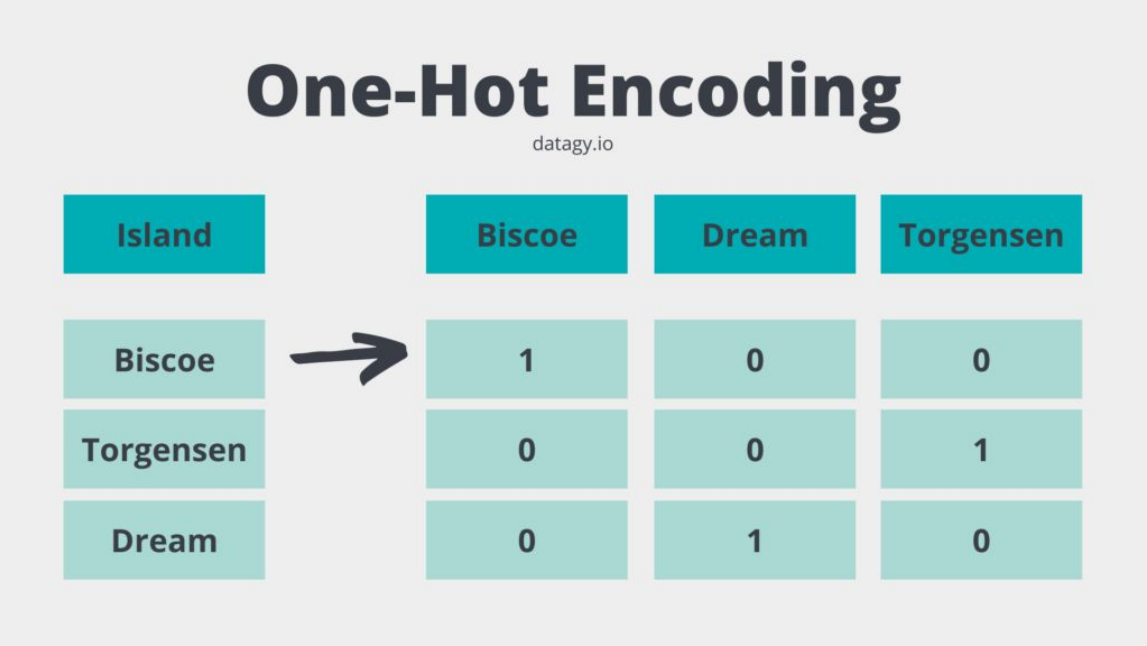
Let's use model-based imputer to impute the missing values.



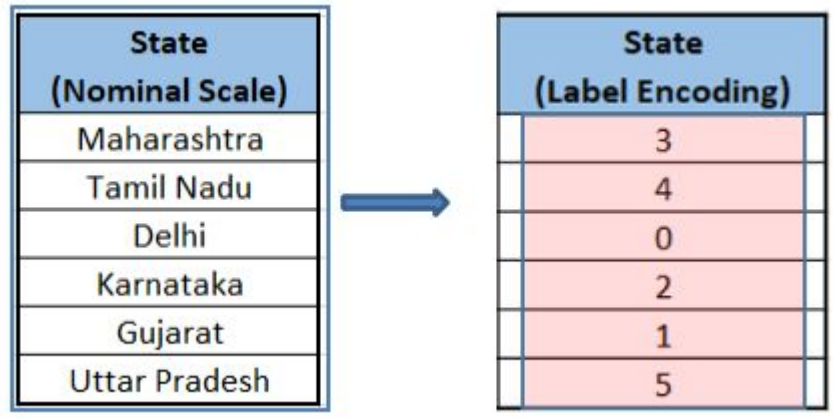
Step 3: Categorical Encoding

Categorical data are usually encoded with string values. However, most machine learning works better with integer values. Therefore, there is a need for us to encode these categorical data into numerical data.

One Hot Encoder



Label Encoder



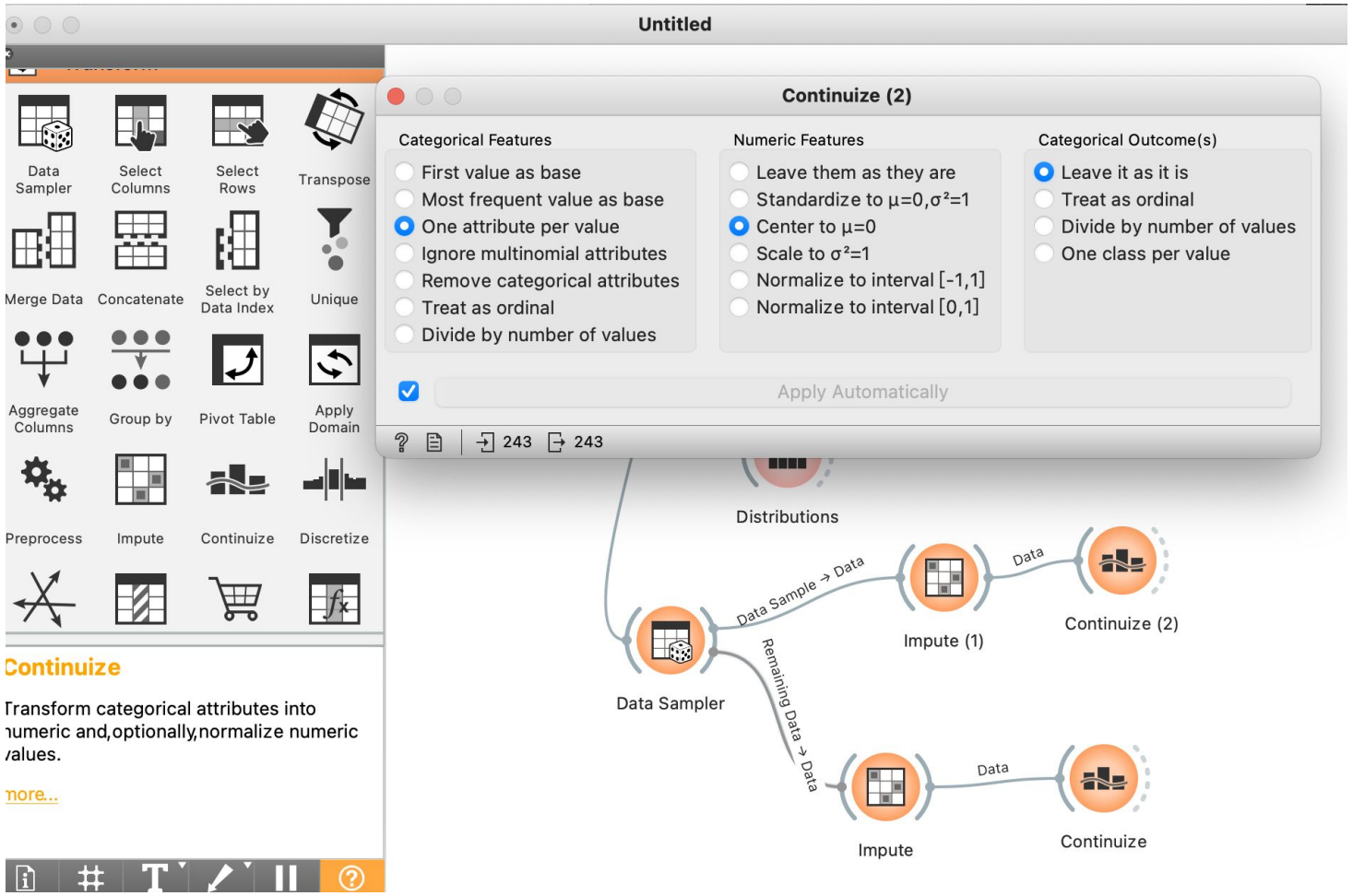
Step 3: Feature Scaling

Feature Scaling is an approach used to standardize and normalize the range of data. This helps to ensure that all features are weight equally in their representation.

- **Min Max Scaler -** will transform each value in the column proportionally within the range [0,1]. Use this as the first scaler choice to transform a feature, as it will **preserve the shape of the dataset**
- **Standard Scaler** - will transform each value in the column to range about the mean 0 and standard deviation 1. **Use if you know the data distribution is normal.**
- **Robust Scaler -** will removes the median and scales the data according to the quantile range. **Use if there is outliers**

Step 3: Encoding and Scaling in Orange

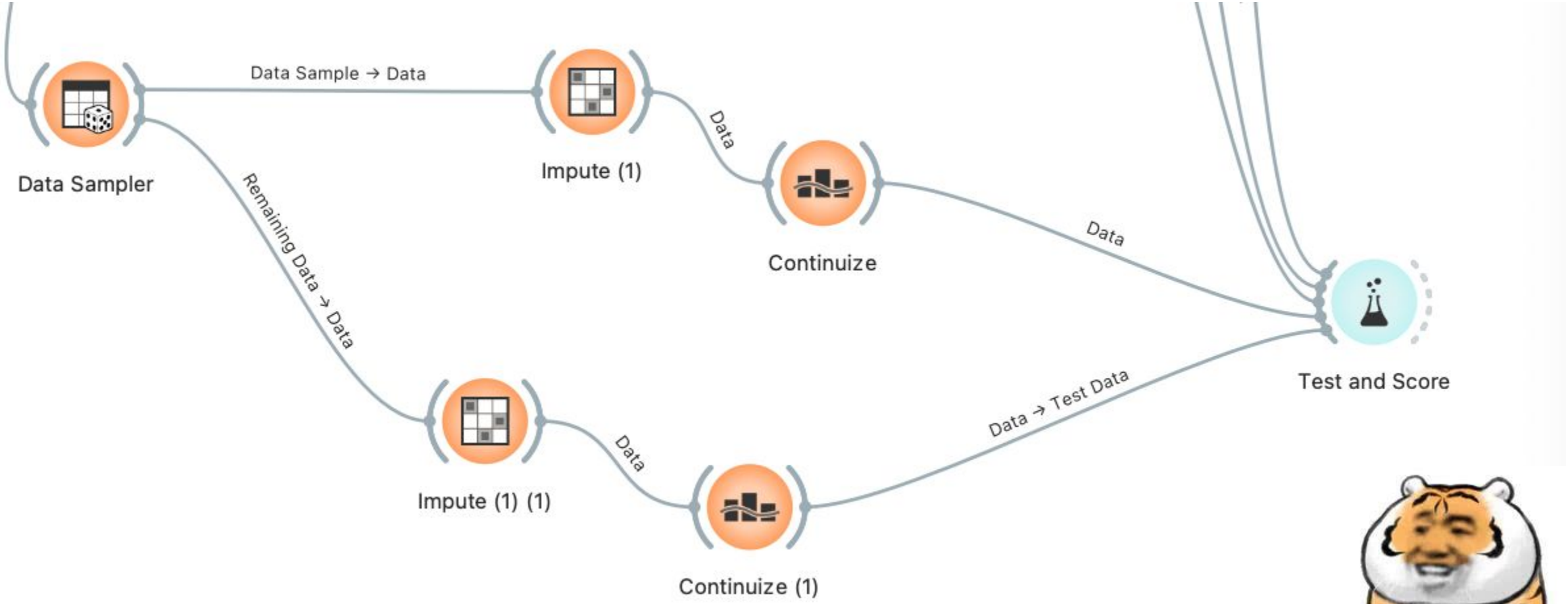
Let do it in orange !!



Step 4. Train Model

Step 4: Modelling

Now let’s go to modelling !!! Drag the “test and score” widget

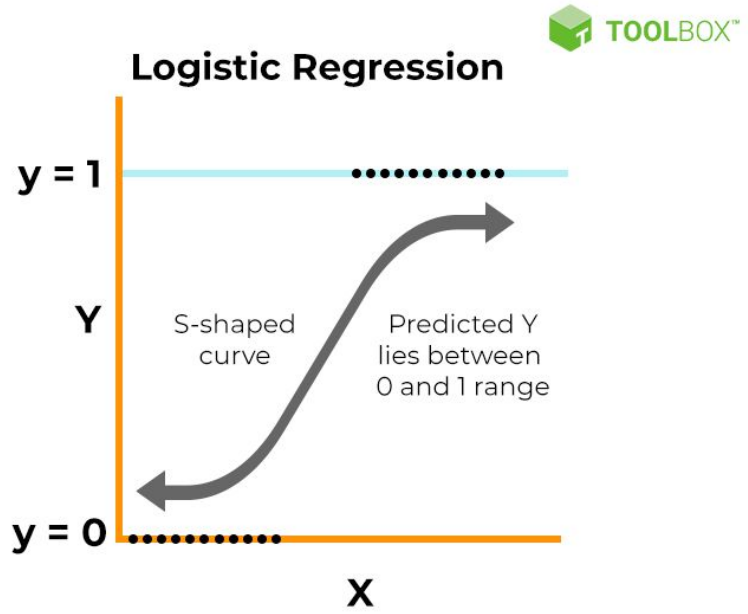


Step 4: Logistic Regression

Logistic regression is a statistical model proposed by Berkson (1944).

The model evaluates the probability of one event by using the logarithm of the odds for the event to be linearly combined with 1 or more variables. Input variables are linearly combined to predict an output binary value 0 or 1.The logistic function is shown below.

$$\rho(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

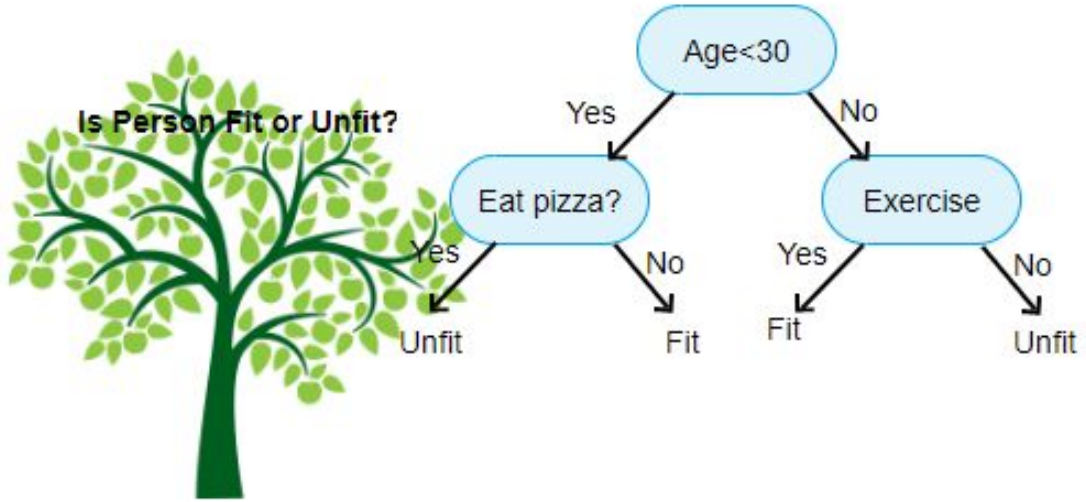


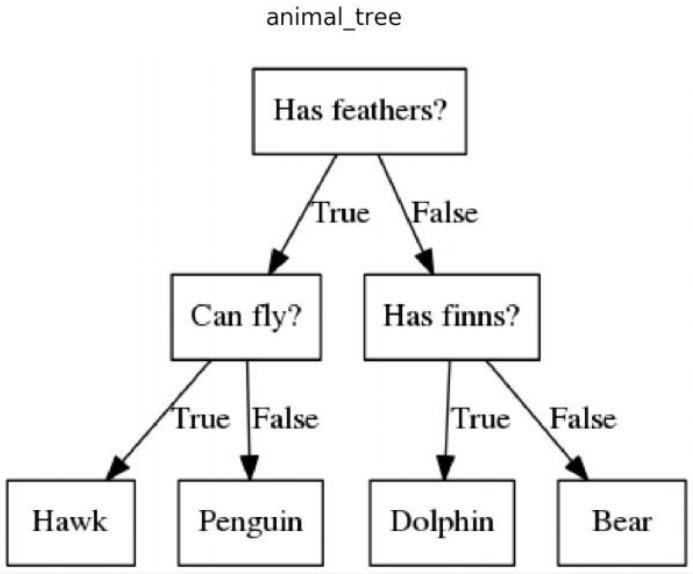
Step 4: Decision Tree

Decision tree is a rule-based model with a flowchart-like structure.

- Each branch represents a possible outcome.
- Each leaf node may be marked by a class or a probability distribution.

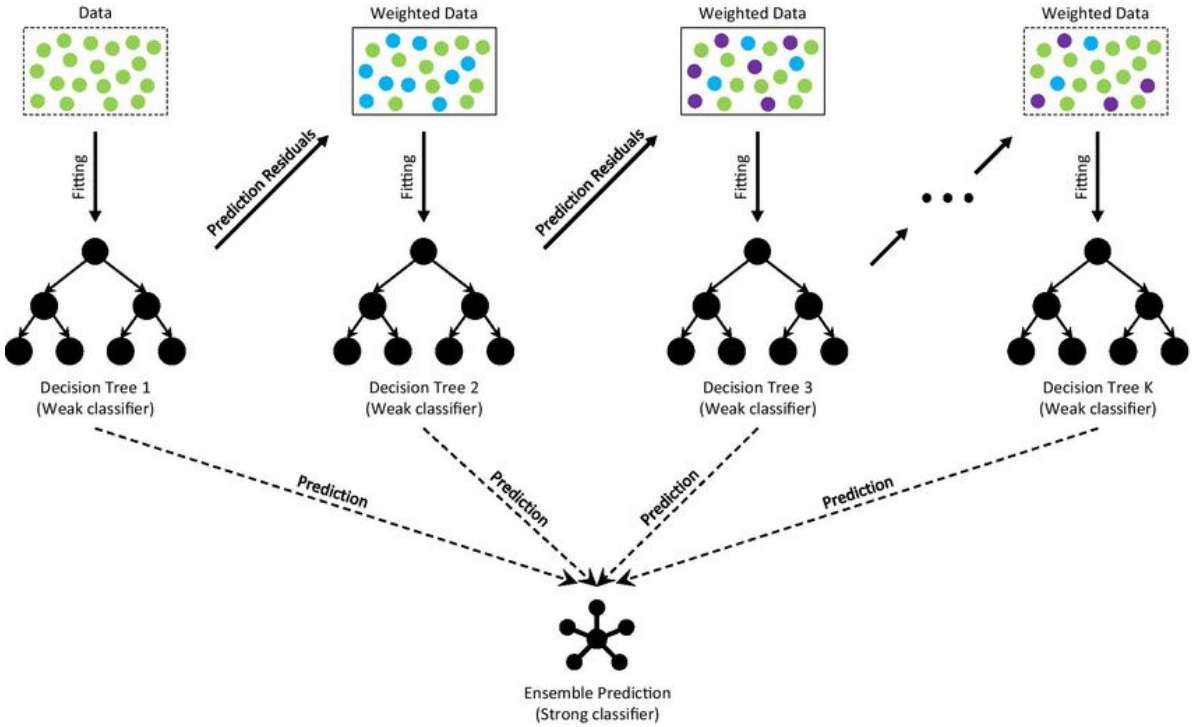
The algorithm evaluates the best feature for the roots of the tree, it splits the sets into disjoint sets and introduces branches and nodes to the tree respectively. The feature space of the dataset will be narrowed down with every split.



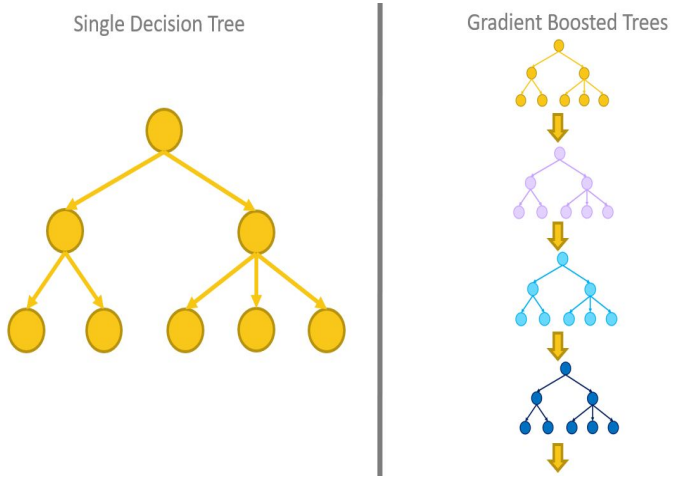


Step 4: Gradient Boosting Tree

Gradient boosting tree was first brought up by Friedman (1999). The algorithm usually combines an ensemble of weak decision trees to predict the target. The gradient boosting tree algorithm assumes the target variable as a true value and forms an estimated in the form of a weighted sum of function from weak learners.

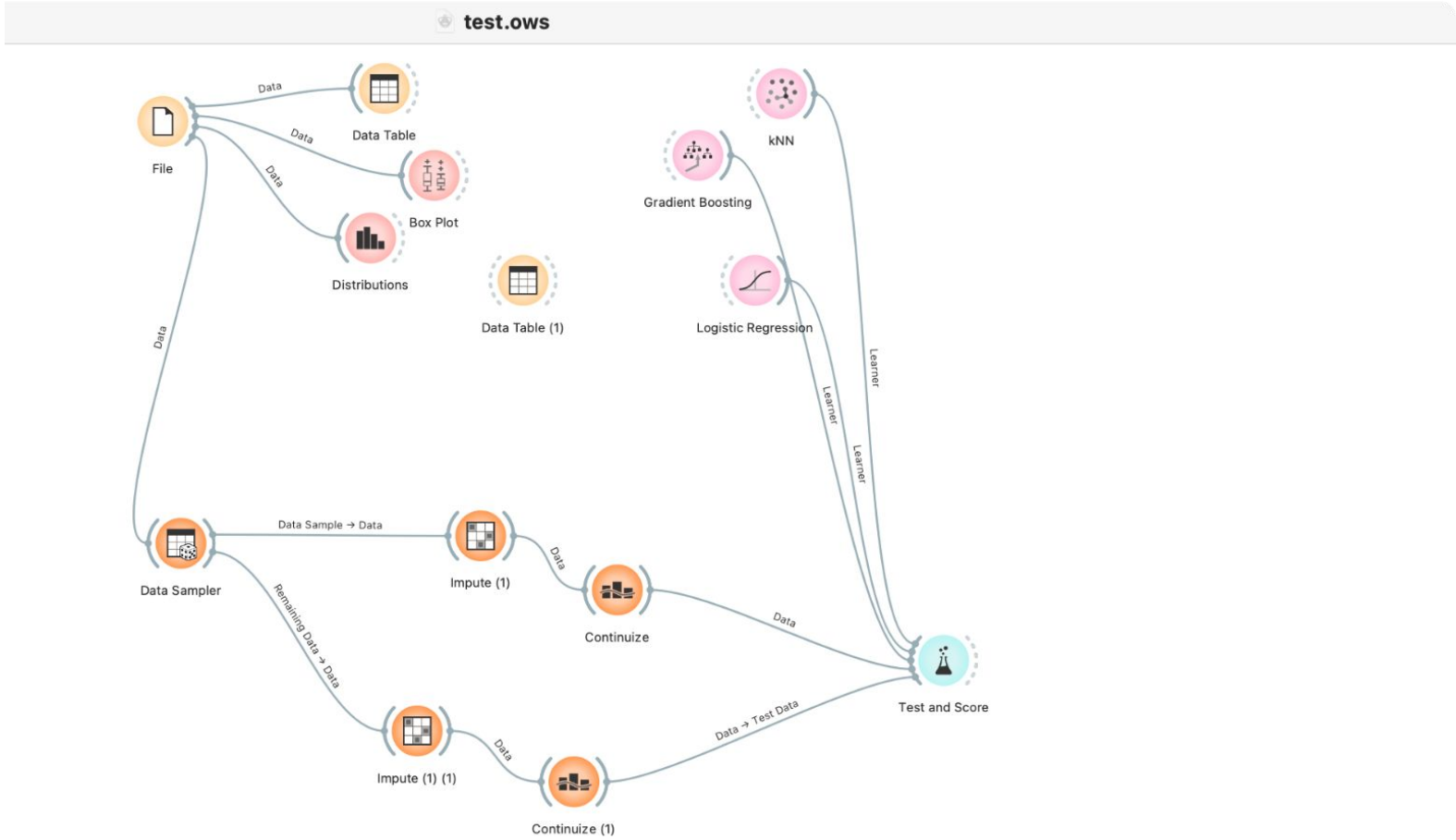


$$\hat{F}(x) = \sum_{t=1}^{M} \Upsilon_t h_t(x) + const$$



Step 4: Train Model

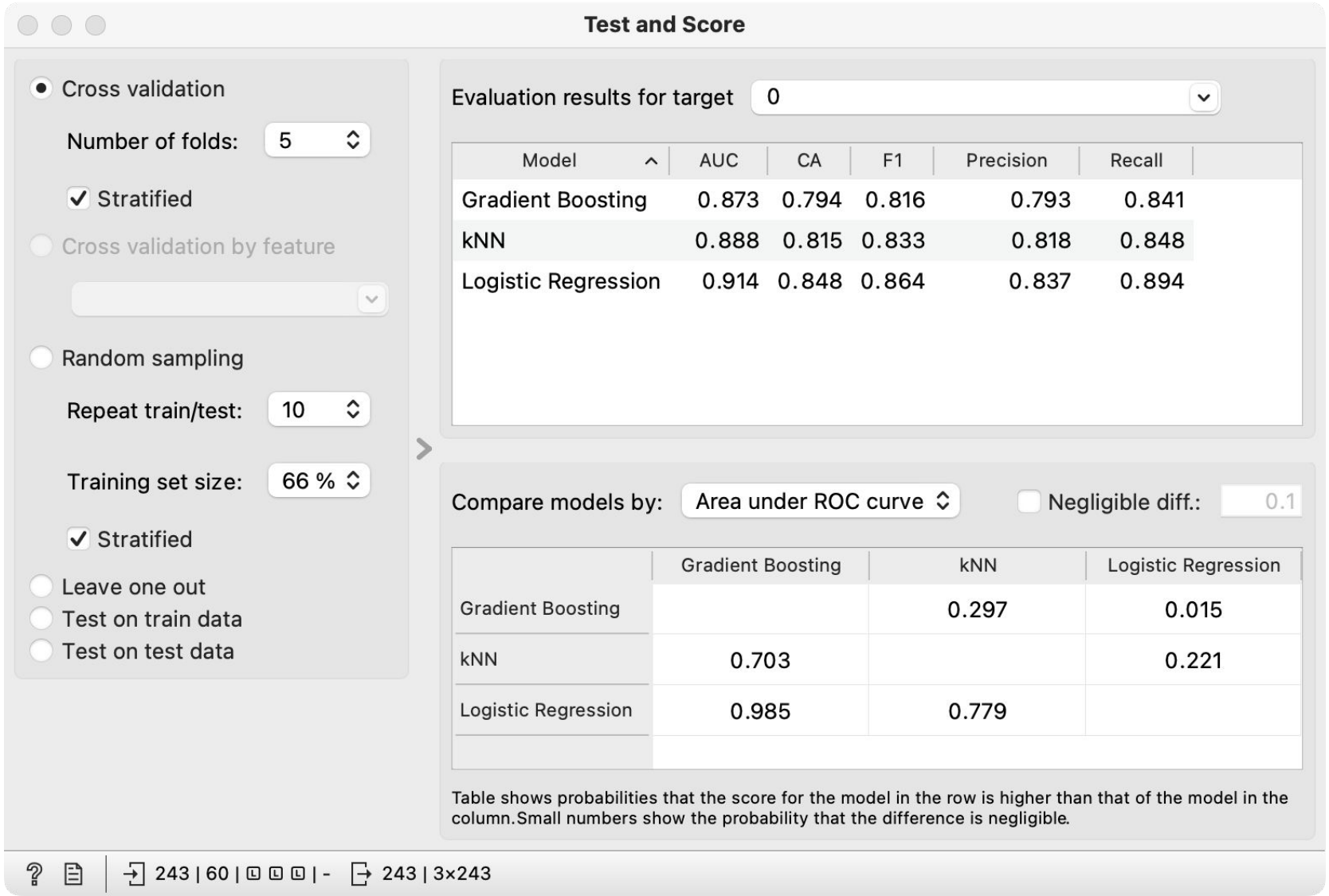
Drag in the model you want to test out !!!



Step 5. Evaluate Model

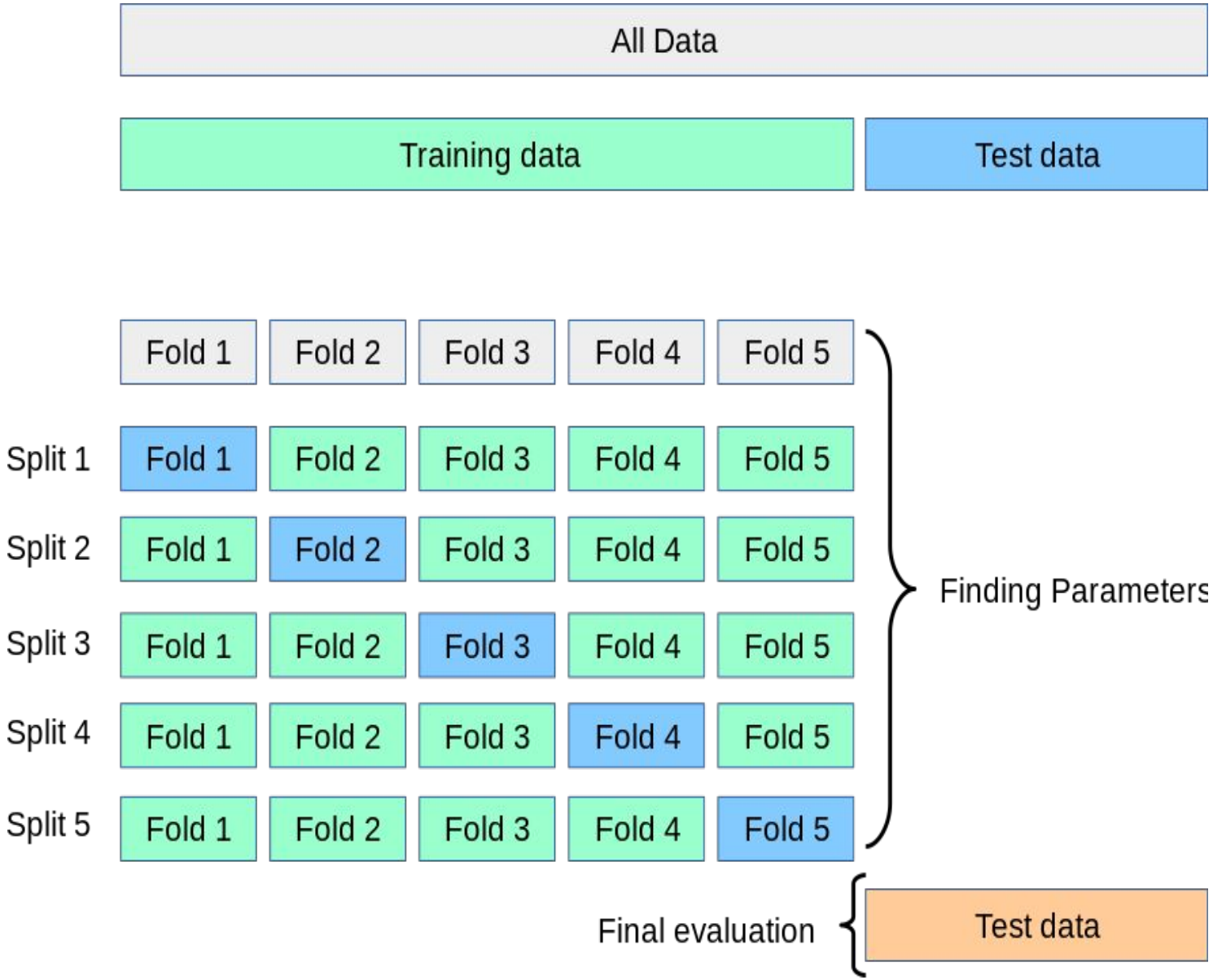
Step 5

Now, double click on the “Test and Score” widget, choose cross validation and see the score !!!

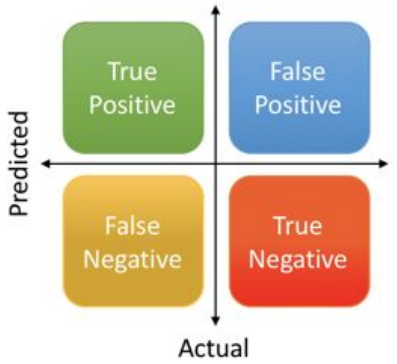


Using Cross Validation

- Cross-validation is used to protect a model from **overfitting**, especially if the amount of data available is limited.
- It's also known as rotation estimation or out-of-sample testing and is mainly used in settings where the model's target is prediction.



Step 5: Different Metrics



TP = True positive

TN = True negative

FP = False positive

FN = False negative

- **Accuracy-** Accuracy is a type of scoring metric that generally describes how well the model performed over all classes. It highlights the percentage of correct predictions over the total number of predictions.

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

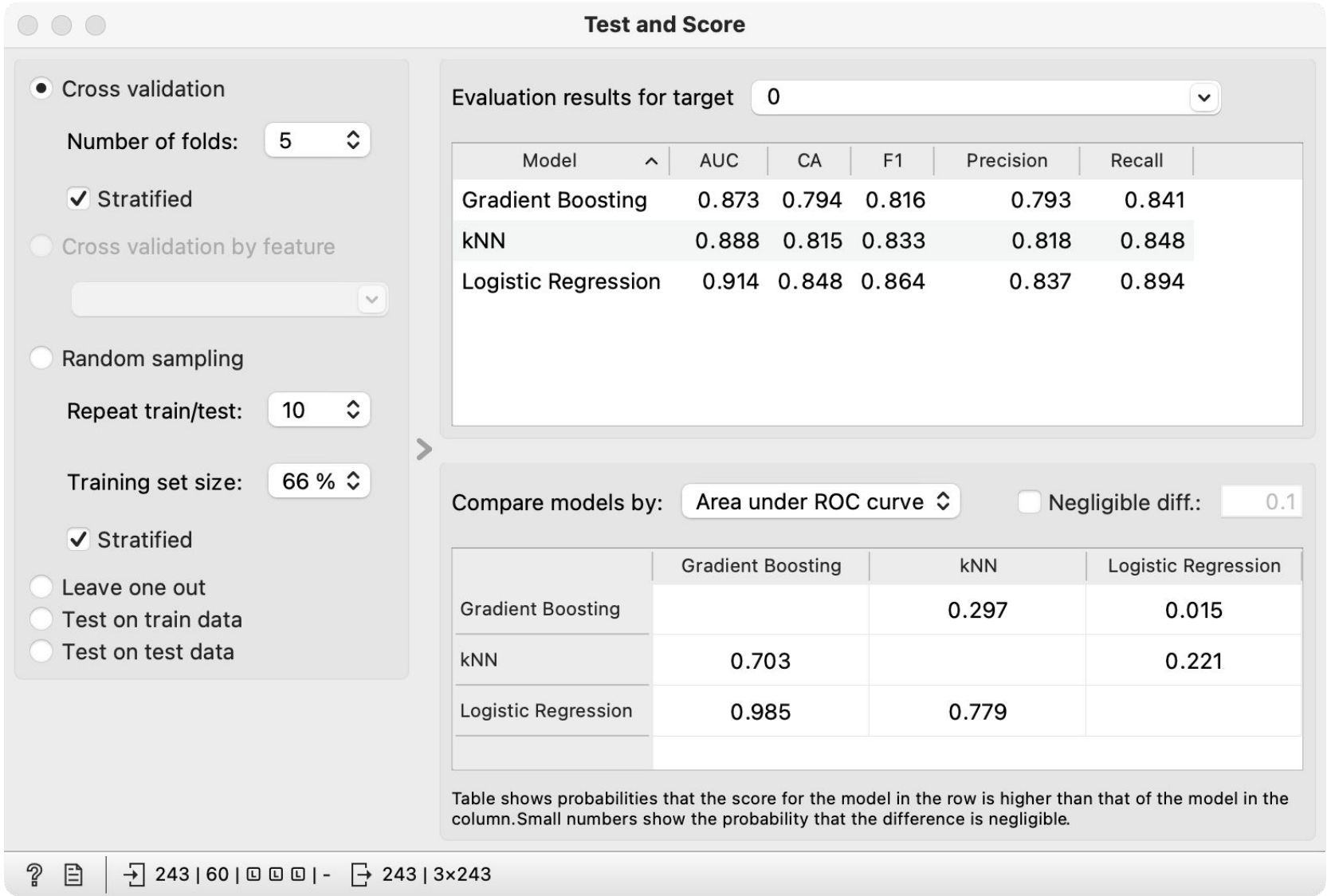
- **Precision-** Precision is a type of scoring metric which measure the positive predictive value (PPV). It highlights the ability of model labelling a sample true positive to total positive prediction made.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall-** Recall is calculated by the ratio between the count of true positives to the summation of the count of true positives and false positives. The recall rate will reduce if there are more false negatives

$$Recall = \frac{TP}{TP + FN}$$

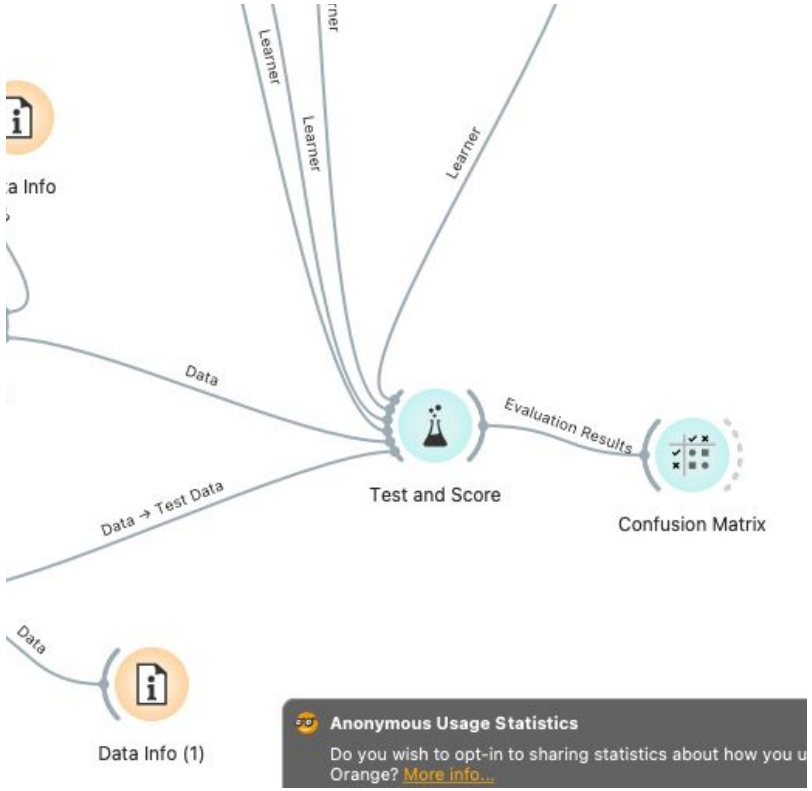
Step 5


Test and Score
Cross validation
Number of folds: 5
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 10
Training set size: 66 %
Stratified
Leave one out
Test on train data
Test on test data
Evaluation results for target 0
Model AUC CA F1 Precision Recall
Gradient Boosting 0.873 0.794 0.816 0.793 0.841
kNN 0.888 0.815 0.833 0.818 0.848
Logistic Regression 0.914 0.848 0.864 0.837 0.894
Compare models by: Area under ROC curve
Negligible diff.: 0.1
Gradient Boosting kNN Logistic Regression
Gradient Boosting 0.297 0.015
kNN 0.703 0.221
Logistic Regression 0.985 0.779
Table shows probabilities that the score for the model in the row is higher than that of the model in the column.Small numbers show the probability that the difference is negligible.
243 | 60 | - 243 | 3×243

Step 5

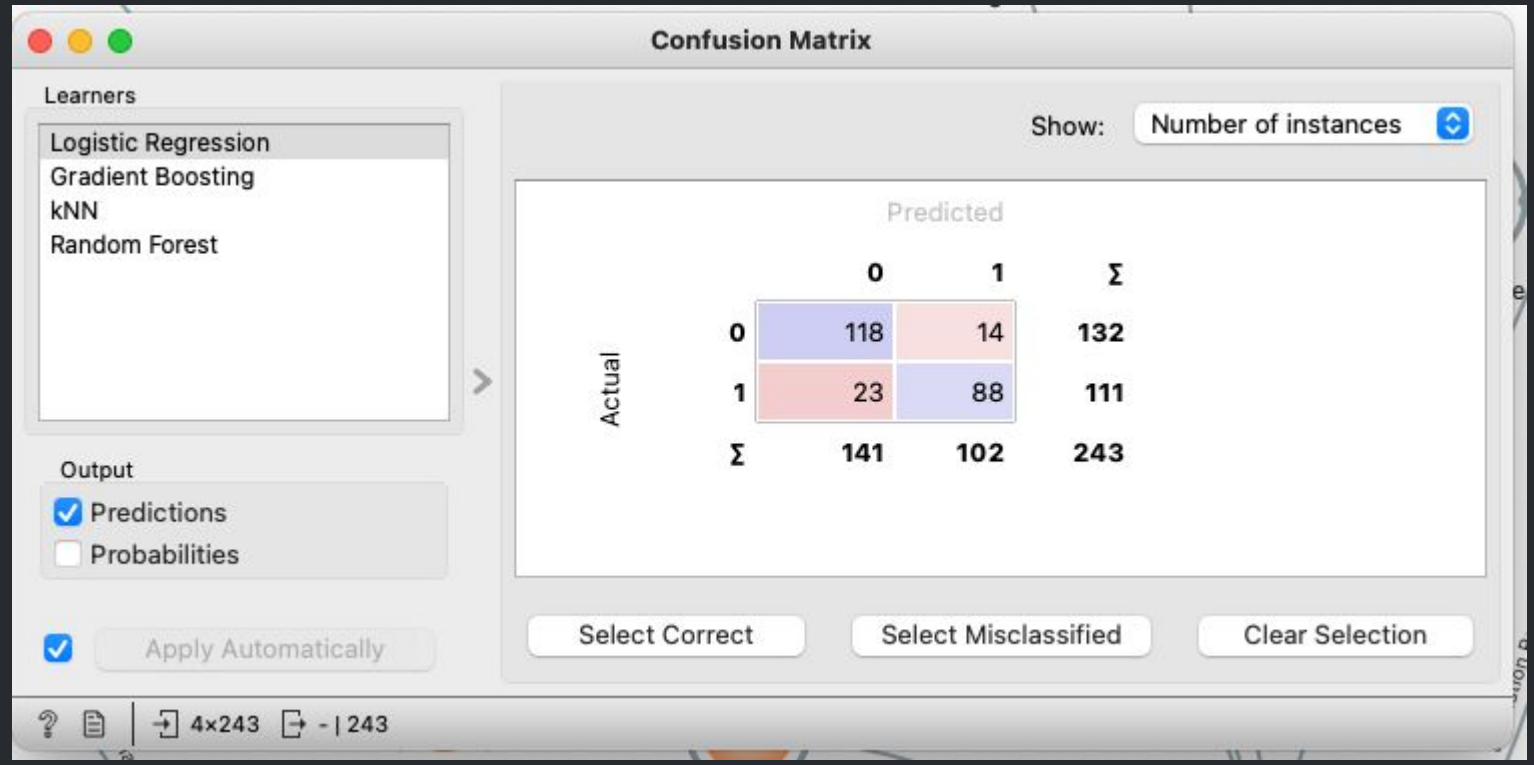
To visualize the results better, drag and drop the “Test and Score” widget to find ”Confusion Matrix”.

The confusion matrix was first introduced by Pearson (1904). It is a 2-dimensional contingency table that highlights the count of true negative, true positive, false negative and false positive. The columns display the predicted results while the rows display the actual results. It offers more information regarding the performance of the classification.



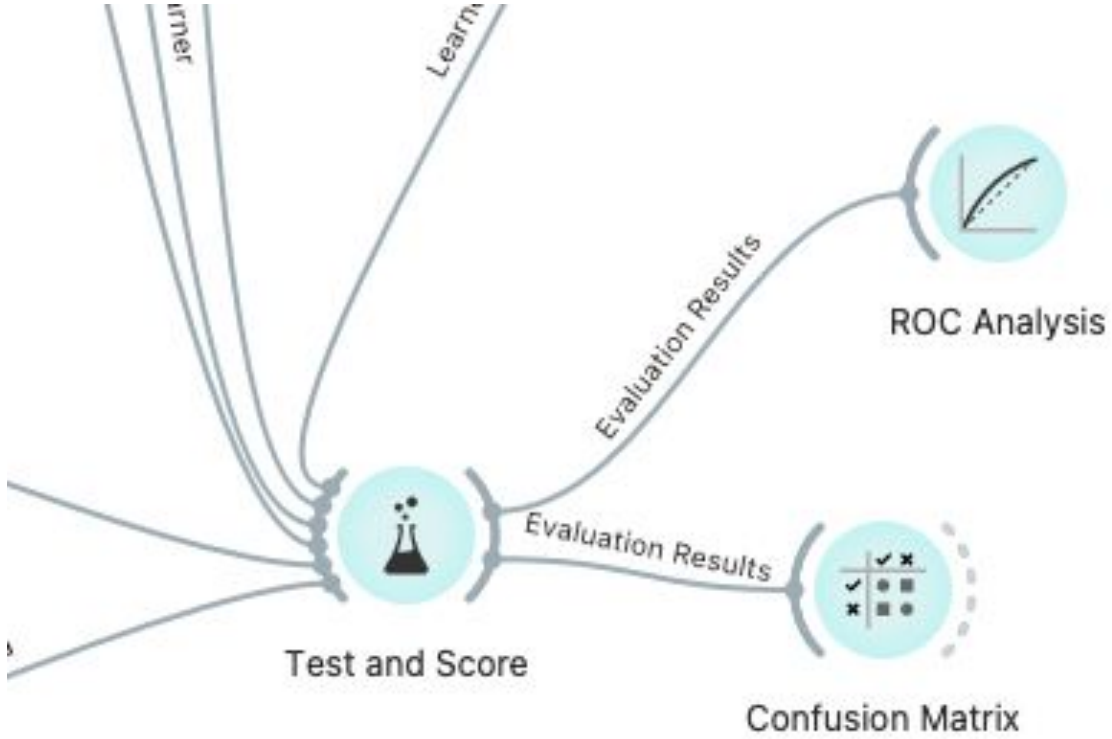
Step 5

Once you've placed it, double click "Confusion Matrix" to visualize your findings!



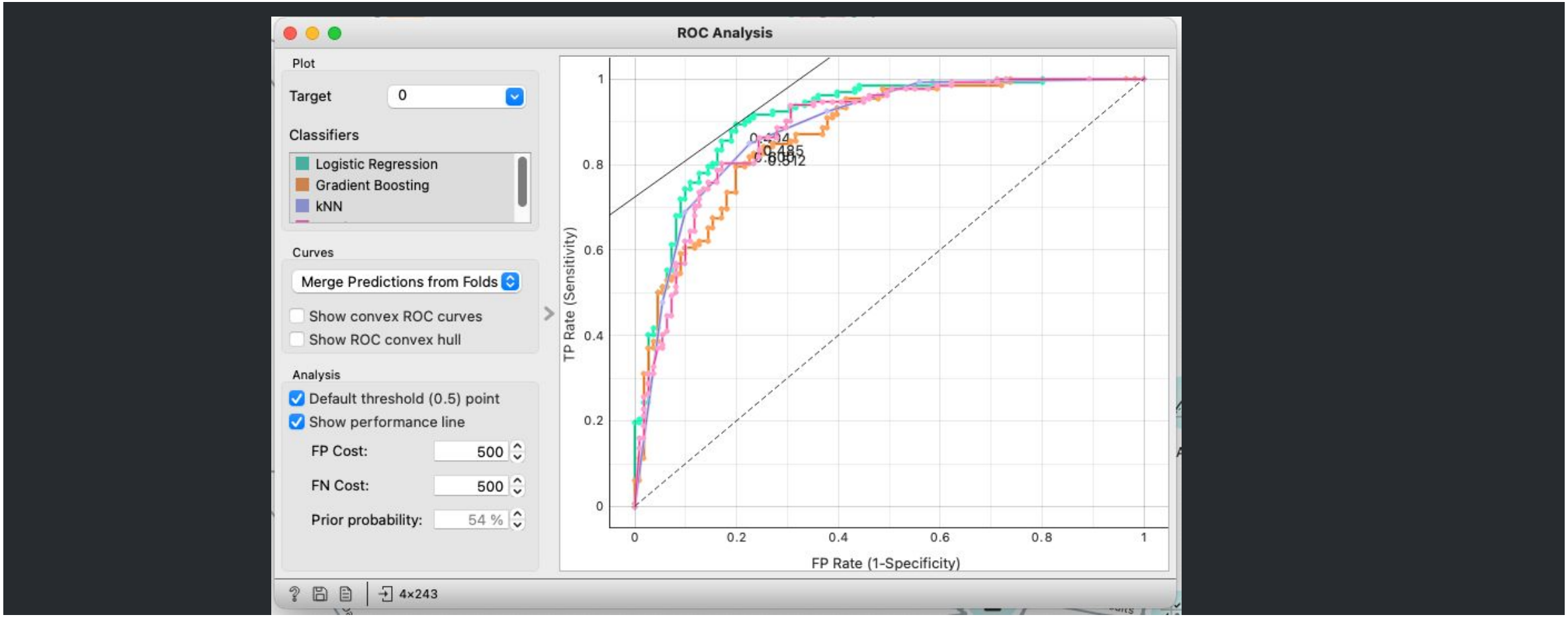
Step 5

To visualize the results better, drag and drop the “Test and Score” widget to find ”ROC Curve”.



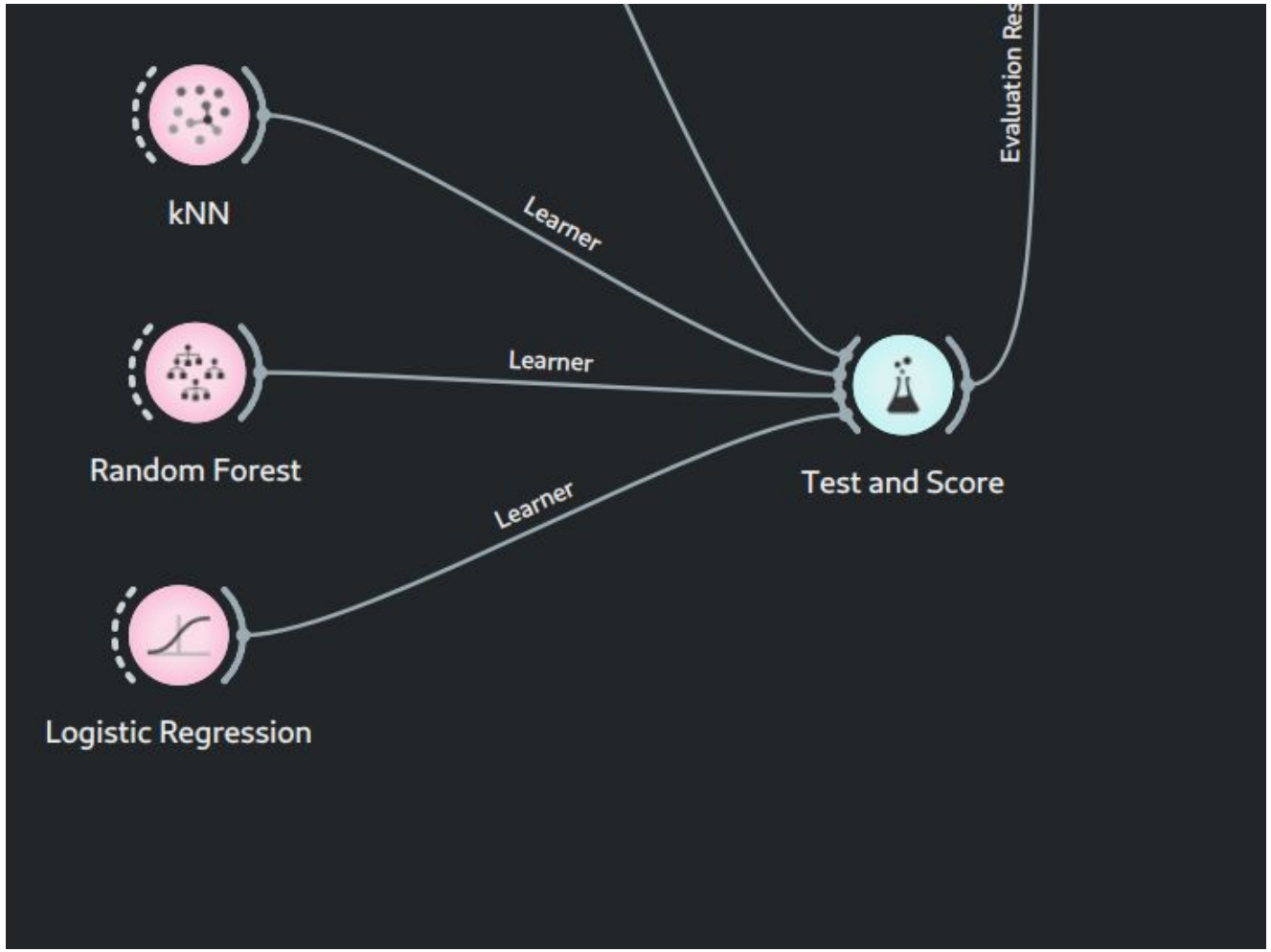
Step 5

Once you've placed it, double click "ROC Curve" to visualize your findings!



Step 5

You can now feel free to add different machine learning models and compare their performance.



Congratulations!

You have made your first ML model

Section 3

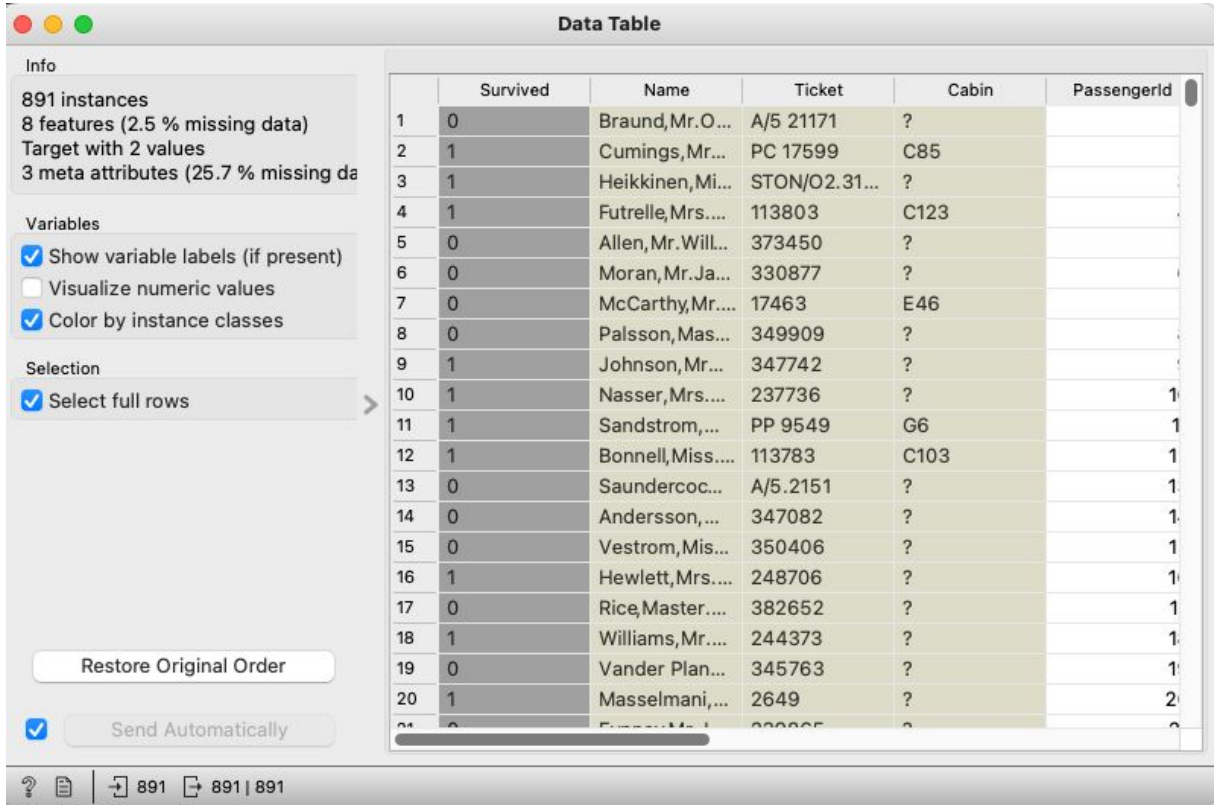
Practical Time!!

Try out the titanic dataset yourself !!

Dataset Information

This practical is simple: use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.

- 891 observations and 12 attributes, with Survived being the target attribute.
- SipSp = No. of Siblings and Spouses
- Parch = No. of Parents and Children
- Embarked = Port in which the passenger boarded the Titanic



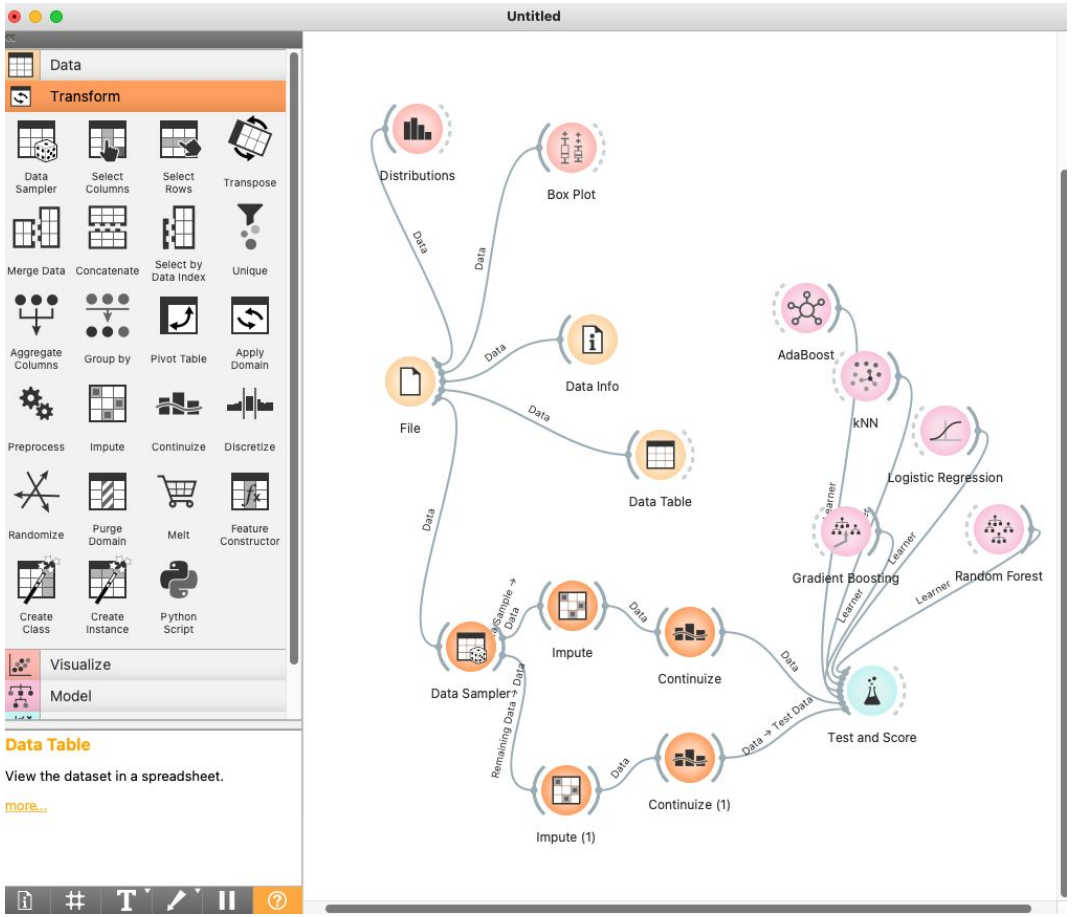
	Survived	Name	Ticket	Cabin	PassengerId
1	0	Braund,Mr.O...	A/5 21171	?	
2	1	Cumings,Mr...	PC 17599	C85	
3	1	Heikkinen,Mi...	STON/O2.31...	?	
4	1	Futrelle,Mrs....	113803	C123	
5	0	Allen,Mr.Will...	373450	?	
6	0	Moran,Mr.Ja...	330877	?	
7	0	McCarthy,Mr....	17463	E46	
8	0	Palsson,Mas...	349909	?	
9	1	Johnson,Mr...	347742	?	
10	1	Nasser,Mrs....	237736	?	1
11	1	Sandstrom,...	PP 9549	G6	1
12	1	Bonnell,Miss....	113783	C103	1
13	0	Saundercoc...	A/5.2151	?	1
14	0	Andersson,...	347082	?	1
15	0	Vestrom,Mis...	350406	?	1
16	1	Hewlett,Mrs....	248706	?	1
17	0	Rice,Master....	382652	?	1
18	1	Williams,Mr....	244373	?	1
19	0	Vander Plan...	345763	?	1
20	1	Masselmani,...	2649	?	2

Things you should consider doing

- Explore whether there is a relationship between the features
- Check whether there is missing data
- Check if there is numerical features
- Set your target variable
- Split data ? Encoding?
- Select a good primary metric

Sample Answer

In machine learning, there is no right answer. Getting high accuracy does not mean you did everything correctly. Getting low accuracy does not mean you are doing the wrong thing. Use logics when doing machine learning.



Section 4

Quiz Time!!

Go to Blooket.com/play

QnA (Answering)

Section 5

Feedback