



MODULO 1. UNIDAD 3

Ejercicios propuestos



DIRECTRICES GENERALES

- Guardar el documento de soluciones con el siguiente formato para su entrega:
M1U3_nombre_apellido1_apellido2.pdf y **M1U3_nombre_apellido1_apellido2.ows**
- Los ejercicios 1 y 2 son teóricos. Se deberán entregar en formato PDF
- El ejercicio 3 se deberán entregar con el tipo de fichero de Orange (ows)



EJERCICIO 1

Escoge una herramienta software que se utilice para Big Data y que no se haya visto en clase y comenta brevemente sus características, utilidad y funcionamiento.



EJERCICIO 2

Investiga y comenta brevemente que tipo de bases de datos se utilizan en los proyectos de data science. Si son SQL o noSQL y porque. Y que gestores de bases de datos predominan en el sector.



EJERCICIO 3. Instalar Orange: <https://orange.biolab.si/>



Features

Screenshots

Workflows

Download

Blog

Docs

Workshops

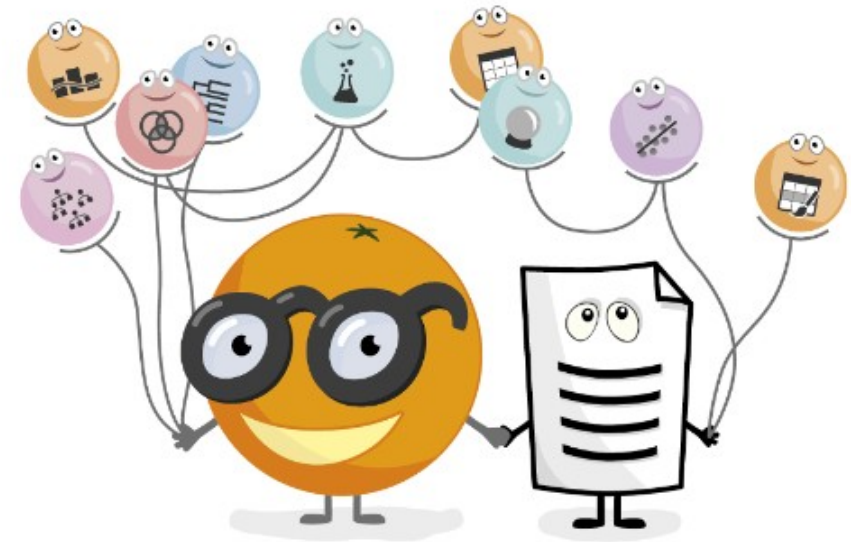
Donate

Data Mining Fruitful and Fun

Open source machine learning and data visualization.

Build data analysis workflows visually, with a large, diverse toolbox.

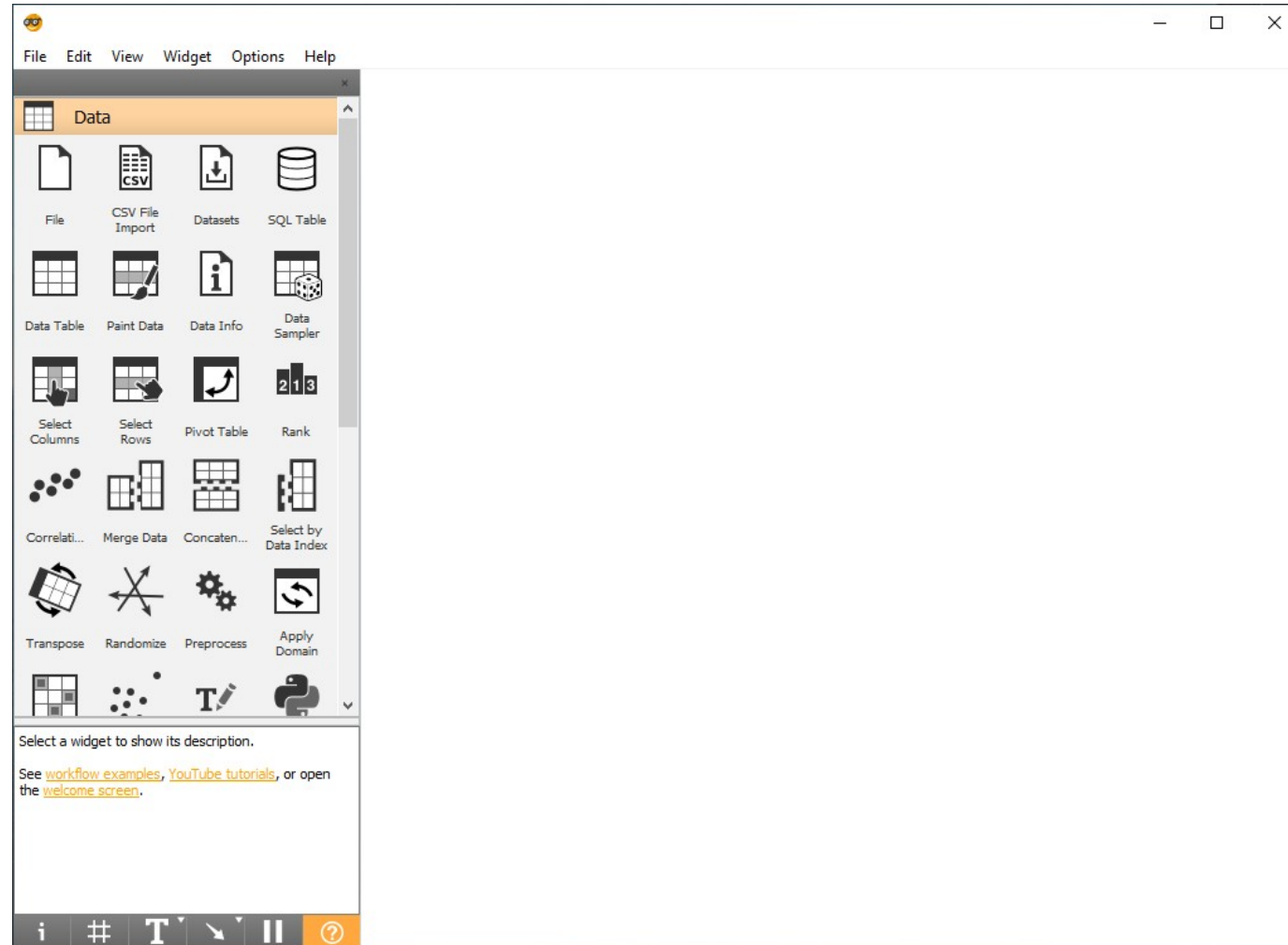
Download Orange



Tokio.



Programa instalado





En la sección Docs de la web se accede a videotutoriales y manuales de uso básico (widget catalog)



Features

Screenshots

Workflows

Download

Blog

Docs

Workshops

Donate

Documentation



Visual Programming

Getting started

YouTube tutorials

Loading your data

Widget catalog



Development

Widget development

Example addon



Python Library

Tutorial

Reference

Orange 2.7 documentation



Widget Catalog

En esta sección de la web se puede acceder a manuales específicos con imágenes y pasos a seguir de todas las herramientas disponibles en Orange.

Es el punto de partida ideal para comenzar con este software.

A continuación se muestran las diferentes temáticas mostradas en esta sección.



Widget Catalog. Data (datos)

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



Impute



Outliers



Edit Domain



Python Script



Color



Continuize



Create Class



Discretize



Feature Constructor



Feature Statistics



Neighbors



Purge Domain



Save Data



Widget Catalog. Visualize (visualización de datos)

Visualize



Tree Viewer



Box Plot



Distributions



Scatter Plot



Line Plot



Sieve Diagram



Mosaic Display



FreeViz



Linear Projection



Radviz



Heat Map



Venn Diagram



Silhouette Plot



Pythagorean Tree



Pythagorean Forest



CN2 Rule Viewer



Nomogram



Widget Catalog. Model (modelos que se pueden utilizar)

Model



Constant



CN2 Rule Induction



Calibrated Learner



kNN



Tree



Random Forest



SVM



Linear Regression



Logistic Regression



Naive Bayes



AdaBoost



Neural Network



Stochastic Gradient
Descent



Stacking



Save Model



Load Model



Widget Catalog. Evaluate (evaluación)

Evaluate



Test and Score



Predictions



Confusion Matrix



ROC Analysis



Lift Curve



Calibration Plot



Widget Catalog. Unsupervised (aprendizaje no supervisado)

Unsupervised



Distance File



Distance Matrix



t-SNE



Distance Map



Hierarchical
Clustering



k-Means



Louvain Clustering



DBSCAN



Manifold Learning



PCA



Correspondence
Analysis



Distances



Distance
Transformation



MDS



Save Distance Matrix Self-Organizing Map





Widget Catalog. Spectroscopy (espectroscopia para ciencias agrarias)

Spectroscopy



Spectra



HyperSpectra



Spectral Series



Interpolate



Preprocess Spectra



Integrate Spectra



Multifile



Tile File



Average Spectra



Bin



Interferogram to
Spectrum



Reshape Map



Align Stack



Widget Catalog. Text Mining (minería de texto)

Text Mining



Corpus



Import Documents



The Guardian



NY Times



Pubmed



Twitter



Wikipedia



Preprocess Text



Corpus to Network



Bag of Words



Document
Embedding



Similarity Hashing



Sentiment Analysis



Tweet Profiler



Topic Modelling



Corpus Viewer



Word Cloud



Concordance



Document Map



Word Enrichment



Duplicate Detection



Statistics



Widget Catalog. Bioinformatics (bioinformática: secuenciación de genes, proyección, clúster, etc.)

Bioinformatics



Databases Update



GEO Data Sets



dictyExpress



Genialis Expressions



Genes



Differential
Expression



GO Browser



KEGG Pathways



Gene Set Enrichment



Gene Sets



Cluster Analysis



Volcano Plot



Homologs



Marker Genes



Annotator



Widget Catalog. Single Cell (análisis celular)

Single Cell



Load Data



Single Cell Datasets



Dropout Gene
Selection



Filter



Single Cell
Preprocess



Batch Effect Removal



Align Datasets



Score Genes



Score Cells



Dot Matrix



Widget Catalog. Image Analytics (análisis de imágenes)

Image Analytics



Import Images



Image Viewer



Image Embedding



Image Grid



Save Images



Widget Catalog. Networks (redes)

Networks



Network File



Network Explorer



Network Generator



Network Analysis



Network Clustering



Network Of Groups



Network From
Distances



Single Mode



Save Network



Widget Catalog. Geo (geografía)

Geo



Geocoding



Geo Map



Choropleth Map



Widget Catalog. Educational (educación)

Educational



Google Sheets



EnKlik Anketa



Interactive k-Means



Gradient Descent



Polynomial
Regression



Polynomial
Classification



Pie Chart



Random Data



Widget Catalog. Time Series (series de tiempo)

Time Series



Yahoo Finance



As Timeseries



Interpolate



Moving Transform



Line Chart



Periodogram



Correlogram



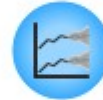
Spiralogram



Granger Causality



ARIMA Model



VAR Model



Model Evaluation



Time Slice



Aggregate



Difference



Seasonal Adjustment



Widget Catalog. Associate (asociación)

Associate



Frequent Itemsets

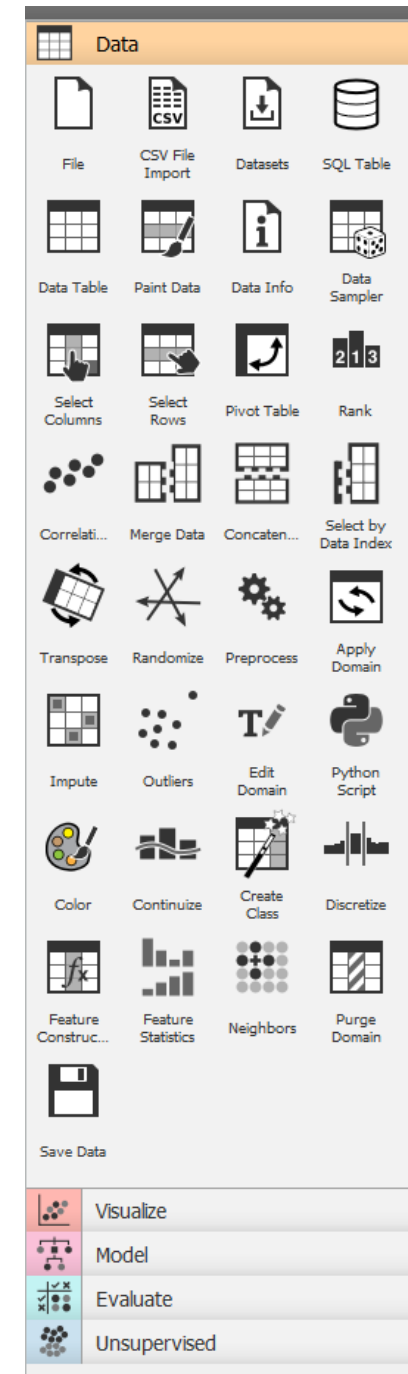


Association Rules

CURSO: IA

Por defecto, Orange solo incluye 5 de estos módulos

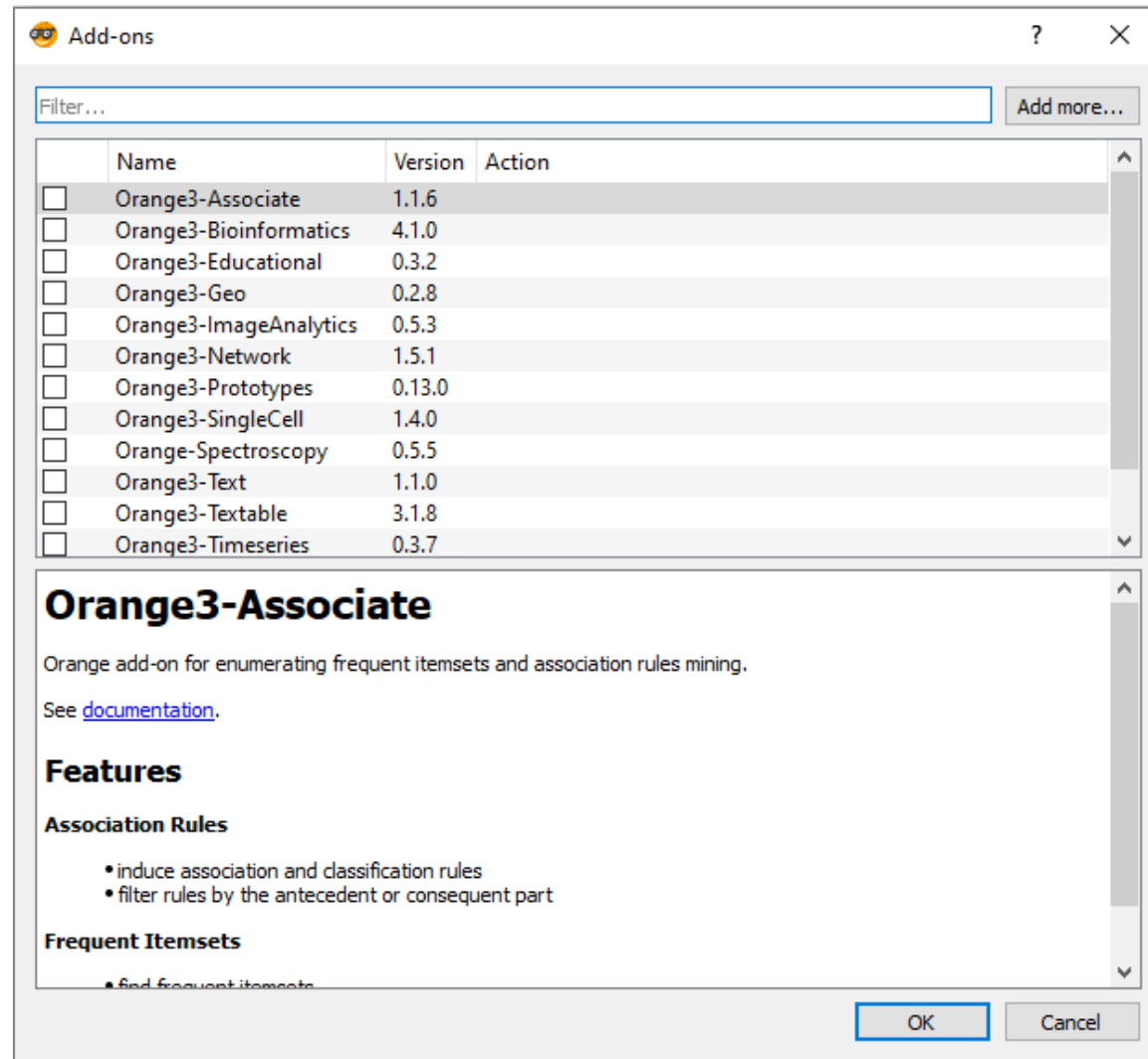
- Data
- Visualize
- Model
- Evaluate
- Unsupervised



Tokio.

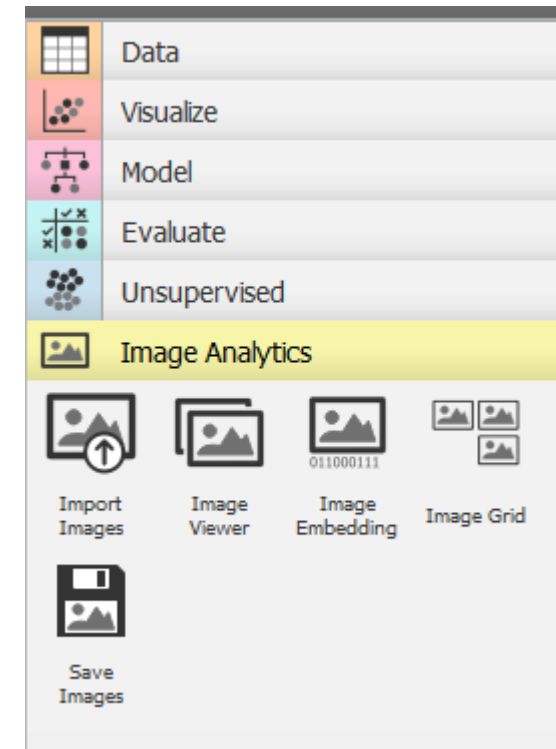
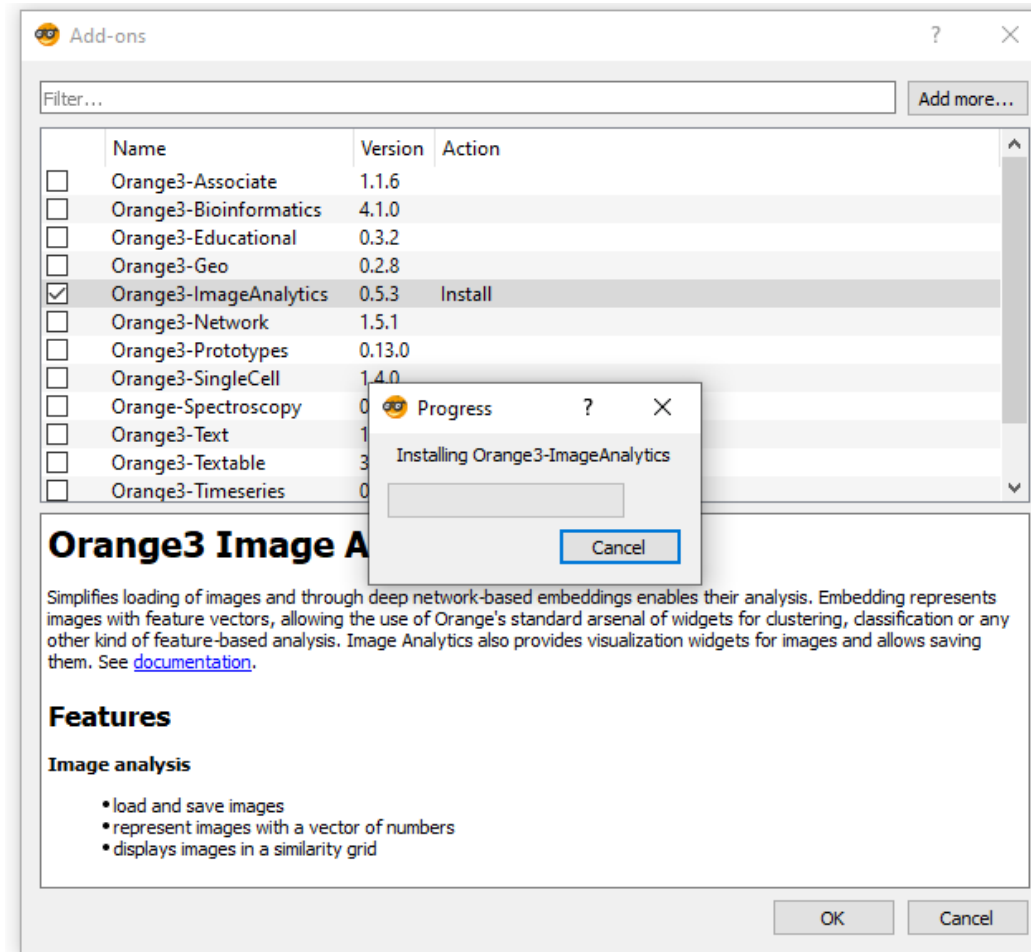


Instalar más modelos





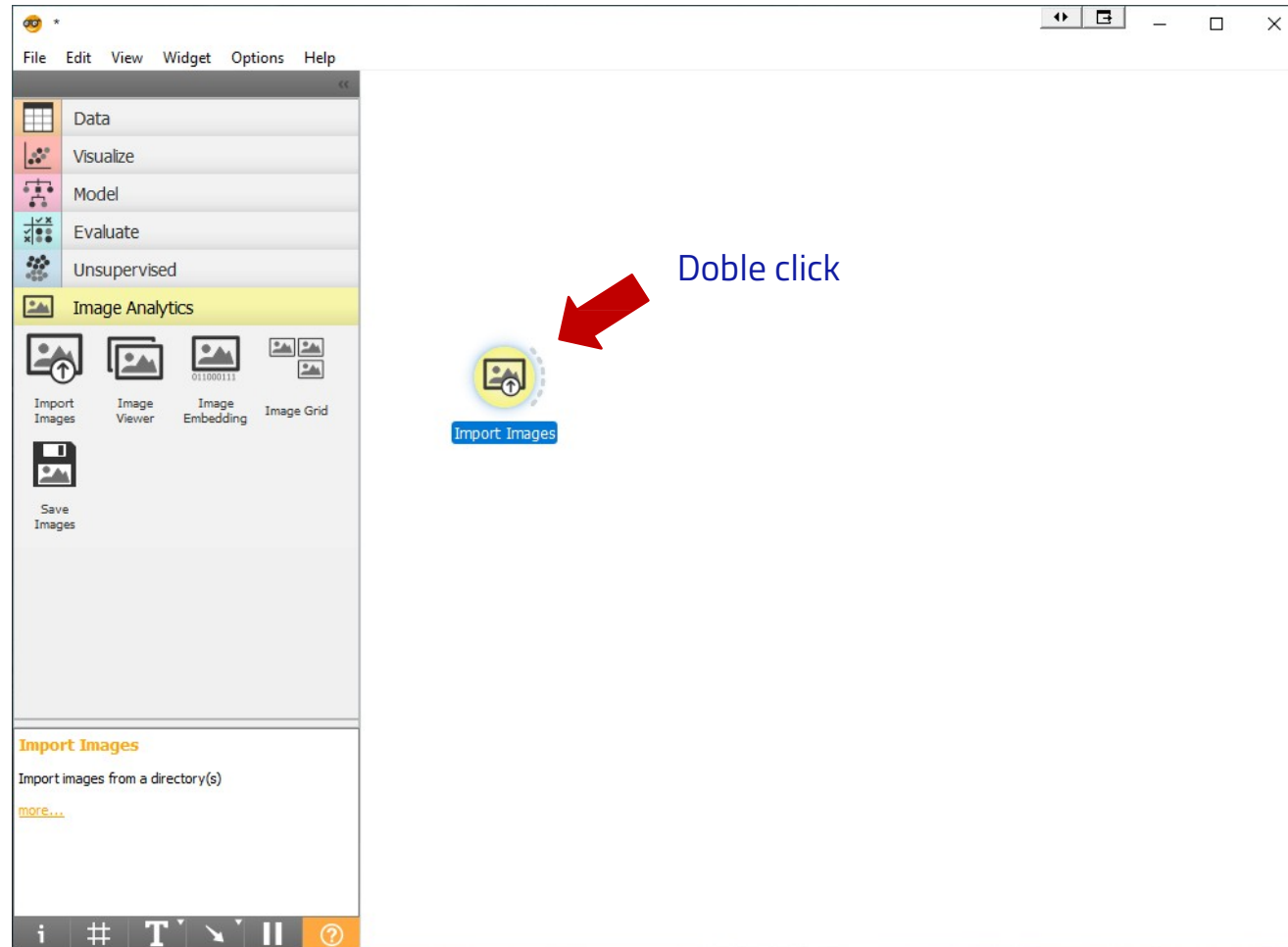
Instalar el modelo Orange3-ImageAnalytics





Realicemos un ejercicio de clustering de imágenes

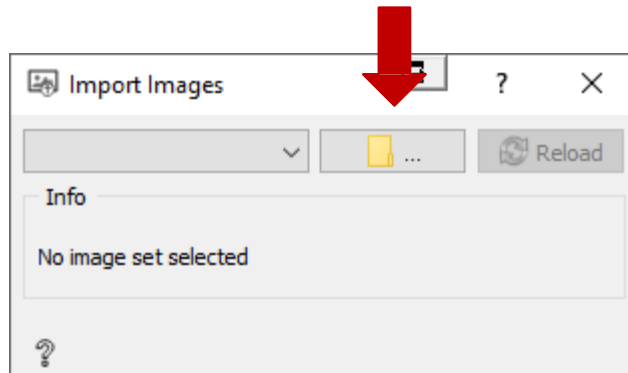
Arrastramos
Import Images a
la ventana del
proyecto



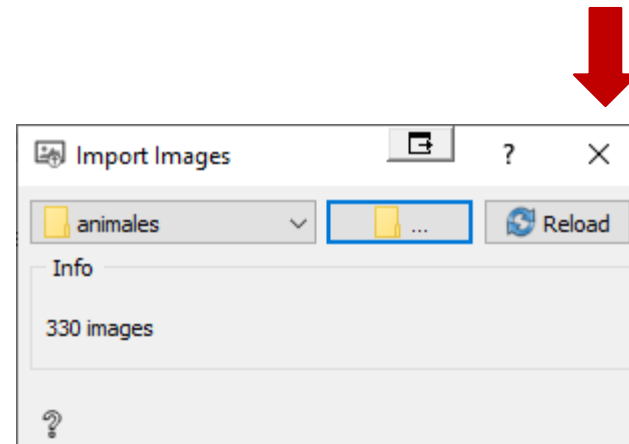


Realicemos un ejercicio de clustering de imágenes

Seleccionar la carpeta animales que se adjunta a esta practica y esperar a que se carguen todas las imágenes.



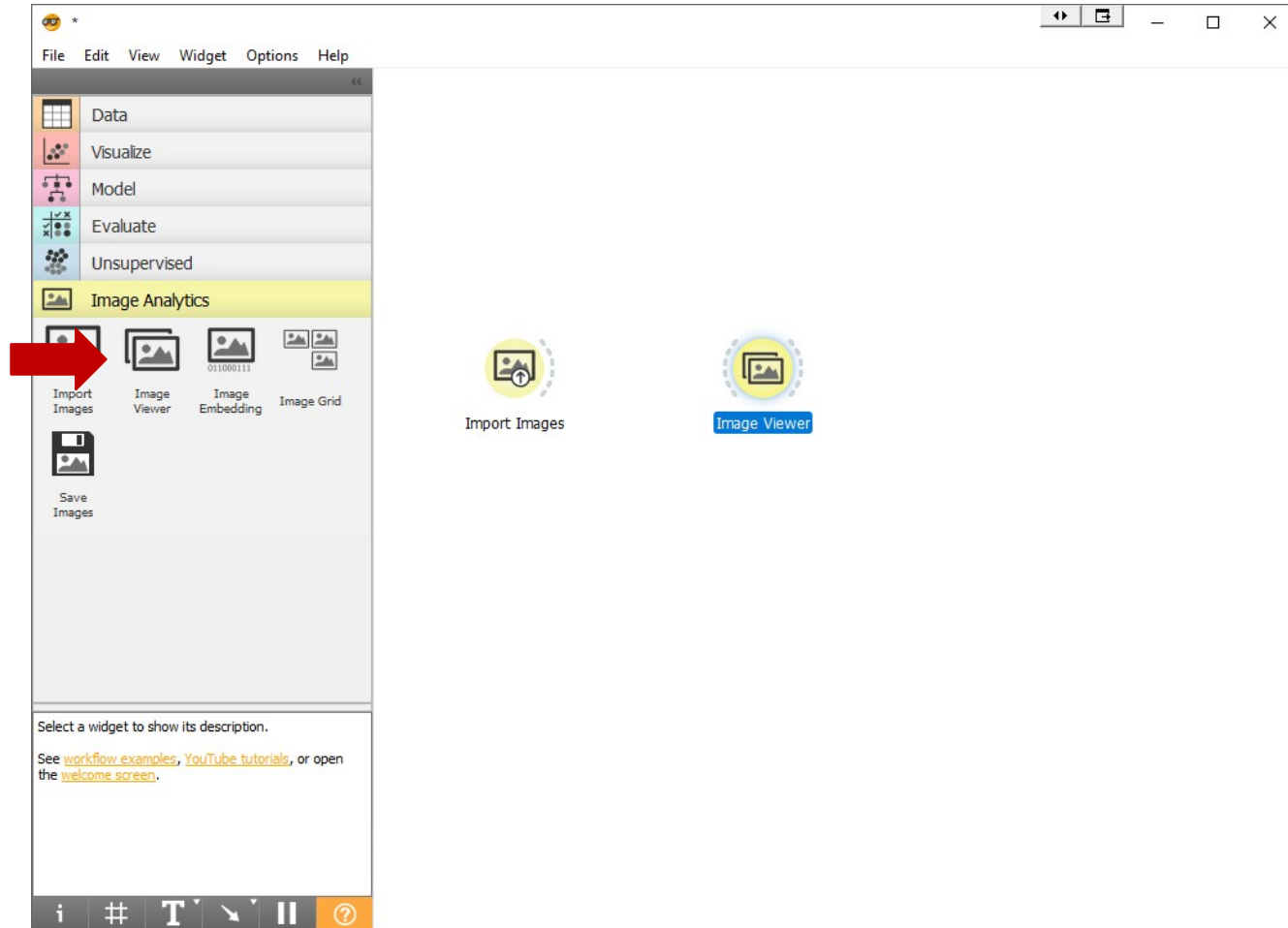
Finalmente, cerrar esta ventana.





Realicemos un ejercicio de clustering de imágenes

Arrastramos Image Viewer a la ventana del proyecto





Realicemos un ejercicio de clustering de imágenes

The screenshot shows the Orange3 software interface. On the left is a widget palette with categories: Data, Visualize, Model, Evaluate, Unsupervised, and Image Analytics. The Image Analytics section is highlighted and contains widgets: Import Images, Image Viewer, Image Embedding, Image Grid, and Save Images. The main workspace displays a workflow with two widgets, 'Import Images' and 'Image Viewer', connected by a 'Data' link. Two red arrows point to the connection line between the two widgets. Below the workflow, text in blue states: 'Conectamos los dos nodos (arrastrando con el ratón)'. To the right of the workflow, red text reads: 'Este proceso puede bloquear el programa durante varios minutos.' The bottom of the interface shows a status bar with icons for information, zoom, text, selection, and execution.

Este proceso puede bloquear el programa durante varios minutos.

Conectamos los dos nodos (arrastrando con el ratón)



Realicemos un ejercicio de clustering de imágenes

Doble click para abrir la visualización y comprobar que se hayan cargado correctamente las imágenes

Image Viewer

Image Filename Attribute: image

Title Attribute: image name

Image Size: [slider]

Send Automatically: ☒

1 10 (2) 10 11 12

13 14 16 (2) 16 17

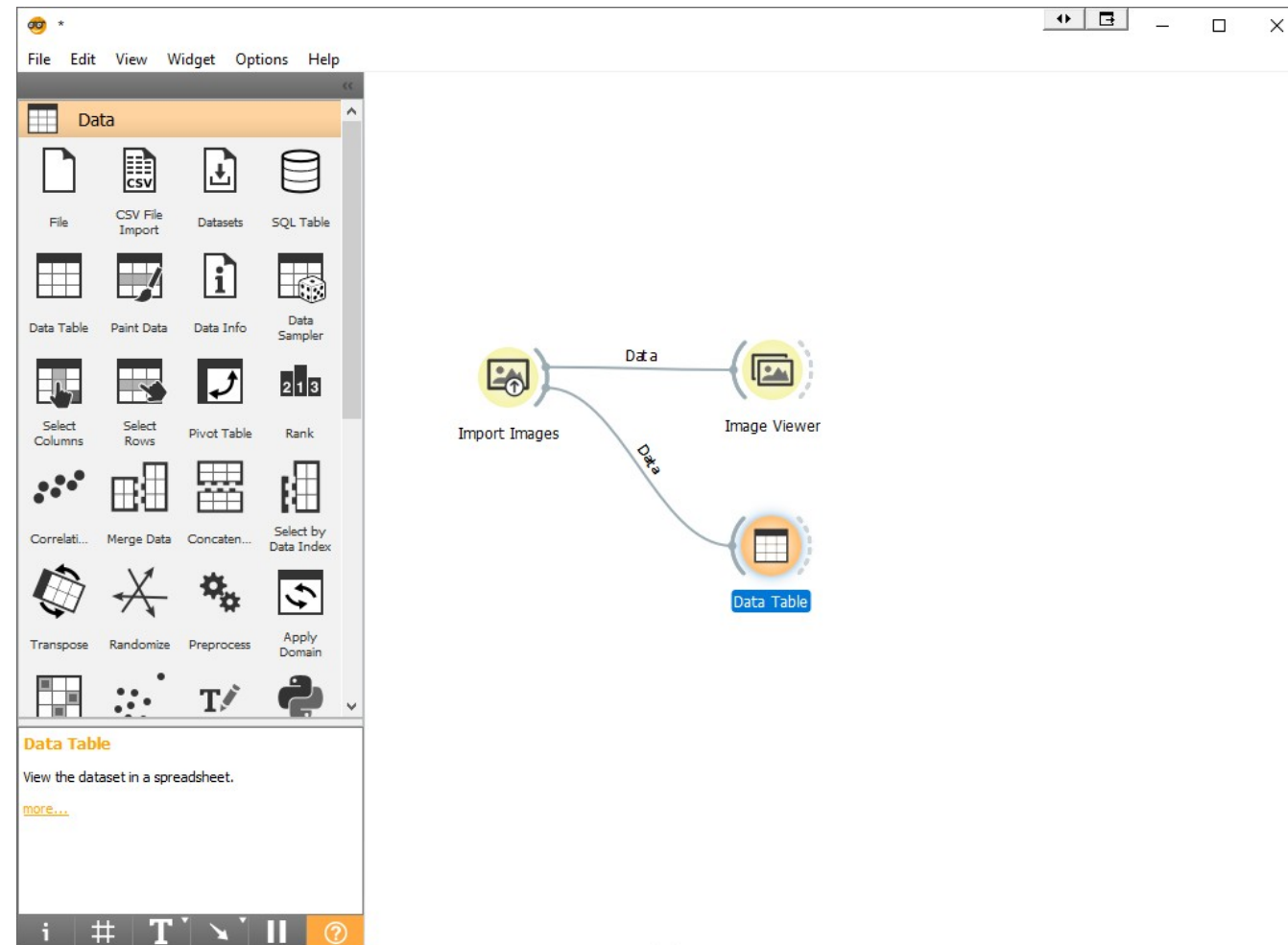
19 (2) 19 2 (2) 2 21

22 25 26 (2) 26 27



Realicemos un ejercicio de clustering de imágenes

Arrastramos Data Table a la ventana del proyecto y conectamos Import Images a Data Table





Realicemos un ejercicio de clustering de imágenes

Doble click para abrir la tabla y comprobar los datos

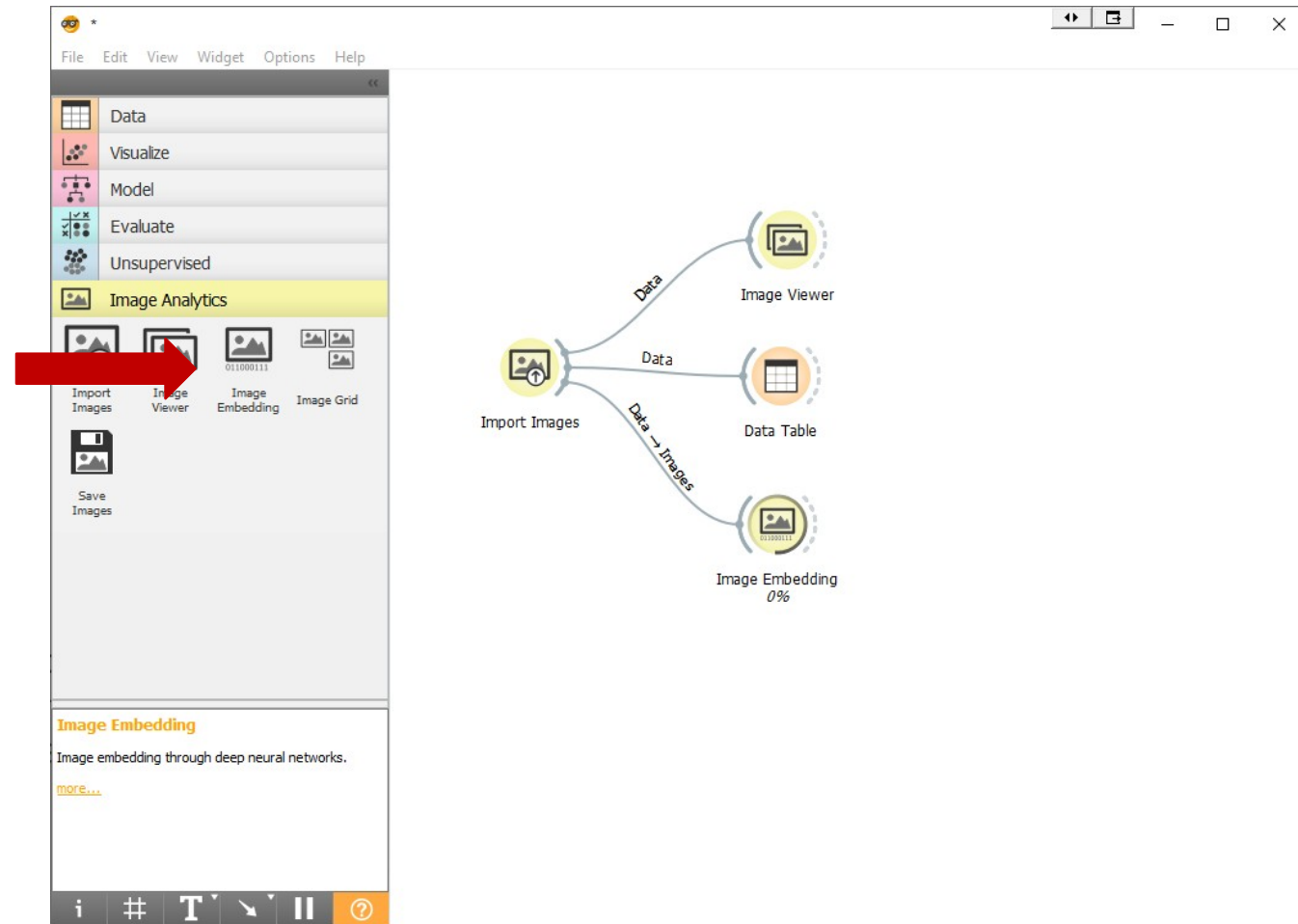
Data Table

origini type	image name	image Cristian/Desktop/image	size	width	height
1	1	1.jpeg	10715	300	210
2	10 (2)	10 (2).jpeg	8885	300	188
3	10	10.jpeg	27041	291	300
4	11	11.jpeg	14506	300	300
5	12	12.jpeg	12228	300	200
6	13	13.jpeg	14284	300	300
7	14	14.jpeg	21789	300	225
8	16 (2)	16 (2).jpeg	14034	300	225
9	16	16.jpeg	24253	300	225
10	17	17.jpeg	17011	300	225
11	19 (2)	19 (2).jpeg	6435	300	200
12	19	19.jpeg	21014	300	225
13	2 (2)	2 (2).jpeg	6582	300	169
14	2	2.jpeg	10963	182	300
15	21	21.jpeg	24456	262	300
16	23	23.jpeg	14462	300	200
17	25	25.jpeg	8587	300	126
18	26 (2)	26 (2).jpeg	14538	300	226



Realicemos un ejercicio de clustering de imágenes

Arrastramos Image Embedding a la ventana del proyecto y lo conectamos a Import Images





Realicemos un ejercicio de clustering de imágenes

Orange3 interface showing a workflow for image clustering. The workflow consists of the following widgets:

- Import Images
- Image Viewer
- Data Table
- Image Embedding

A red arrow points to the **Image Embedding** widget with the text: **Doble click para abrir cuando finalice el porcentaje**

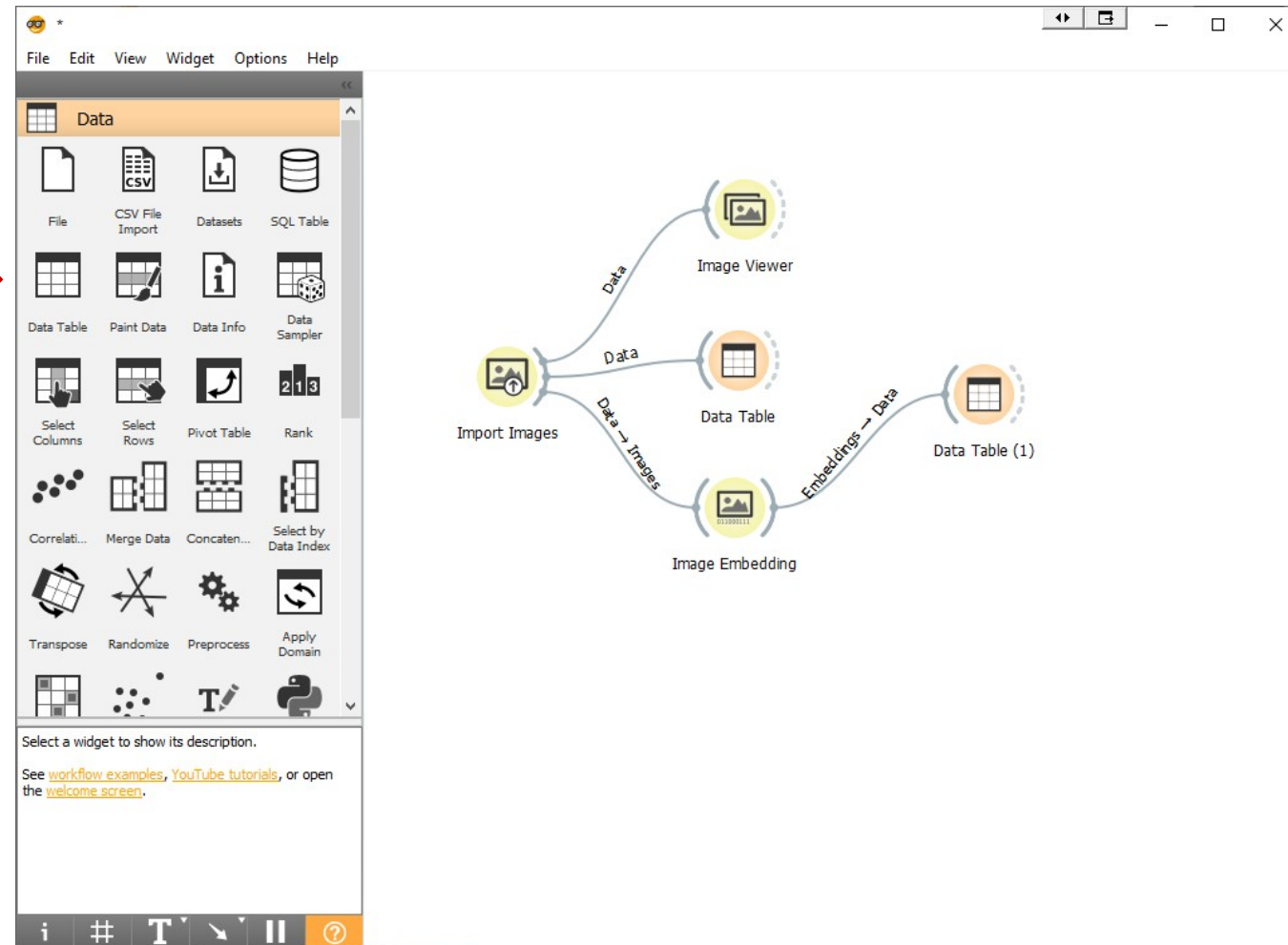
The settings window for the **Image Embedding** widget is shown, displaying the following configuration:

- Image attribute: image
- Embedder: Inception v3
- Google's Inception v3 model trained on ImageNet.
- ☒ Apply Automatically
-



Realicemos un ejercicio de clustering de imágenes

Arrastramos Data Table a la ventana del proyecto y conectamos Image Embedding con la nueva tabla





Realicemos un ejercicio de clustering de imágenes

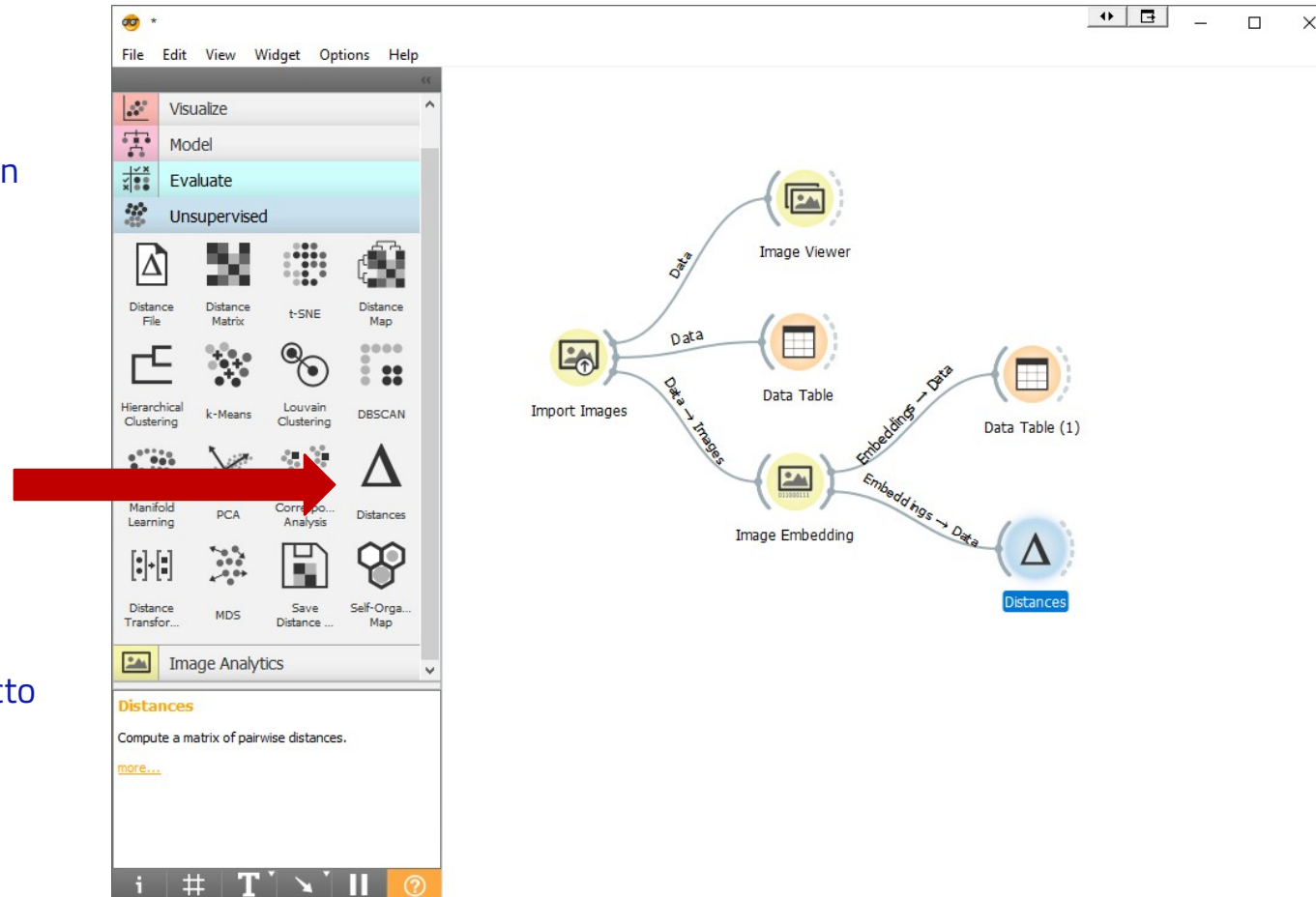
Doble click en la tabla para ver las características antiguas y las nuevas (más de 2000)

	hidden origin type	image name	image	size	width	height	n0	n1	n2	n3	n4	n5	n6
1	1	1.jpeg	Cristian/Desktop/image	10715	300	210	0.397051	0.495888	0.545028	0.109505	0.327051	0.255548	0.1561
2	10 (2)	10 (2).jpeg		8885	300	188	0.209595	0.137083	0.376282	0.191668	0.802205	0.30927	0.2386
3	10	10.jpeg		27041	291	300	0.36504	0.237076	0.109388	0.0532404	0.268079	0.261473	0.4508
4	11	11.jpeg		14506	300	300	0.381177	0.289169	0.469631	0.055005	0.193364	0.19631	0.4246
5	12	12.jpeg		12228	300	200	0.286768	0.269774	0.422119	0.0672271	0.159255	0.426304	0.08315
6	13	13.jpeg		14284	300	300	0.575785	0.330516	0.507948	0.344836	0.0988798	0.442921	0.6073
7	14	14.jpeg		21789	300	225	0.388018	0.0820228	0.168095	0.0206683	0.404229	0.0802024	0.1504
8	16 (2)	16 (2).jpeg		14034	300	225	0.312201	0.310658	0.457219	0.254739	0.536977	0.380994	0.0488
9	16	16.jpeg		24253	300	225	0.415903	0.0959218	0.214328	0.0257579	0.622985	0.279794	0.06146
10	17	17.jpeg		17011	300	225	0.259892	0.0150862	0.117327	0.0145964	0.390622	0.45269	0.317
11	19 (2)	19 (2).jpeg		6435	300	200	0.176416	0.111179	0.298867	0.0769707	0.201462	0.253343	0.2702
12	19	19.jpeg		21014	300	225	0.64976	0.110111	0.297397	0.00252096	0.28963	0.310854	0.2876
13	2 (2)	2 (2).jpeg		6582	300	169	0.131795	0.111499	0.222462	0.204399	0.126044	0.214174	0.159
14	2	2.jpeg		10963	182	300	0.464039	0.176594	0.2309	0.0154283	0.134579	0.109081	0.06799
15	21	21.jpeg		24456	262	300	0.471657	0.391995	0.129059	0.0193356	0.379822	0.095044	0.1524
16	23	23.jpeg		14462	300	200	0.27259	0.0152578	0.0955987	0	0.552066	0.0377672	0.017



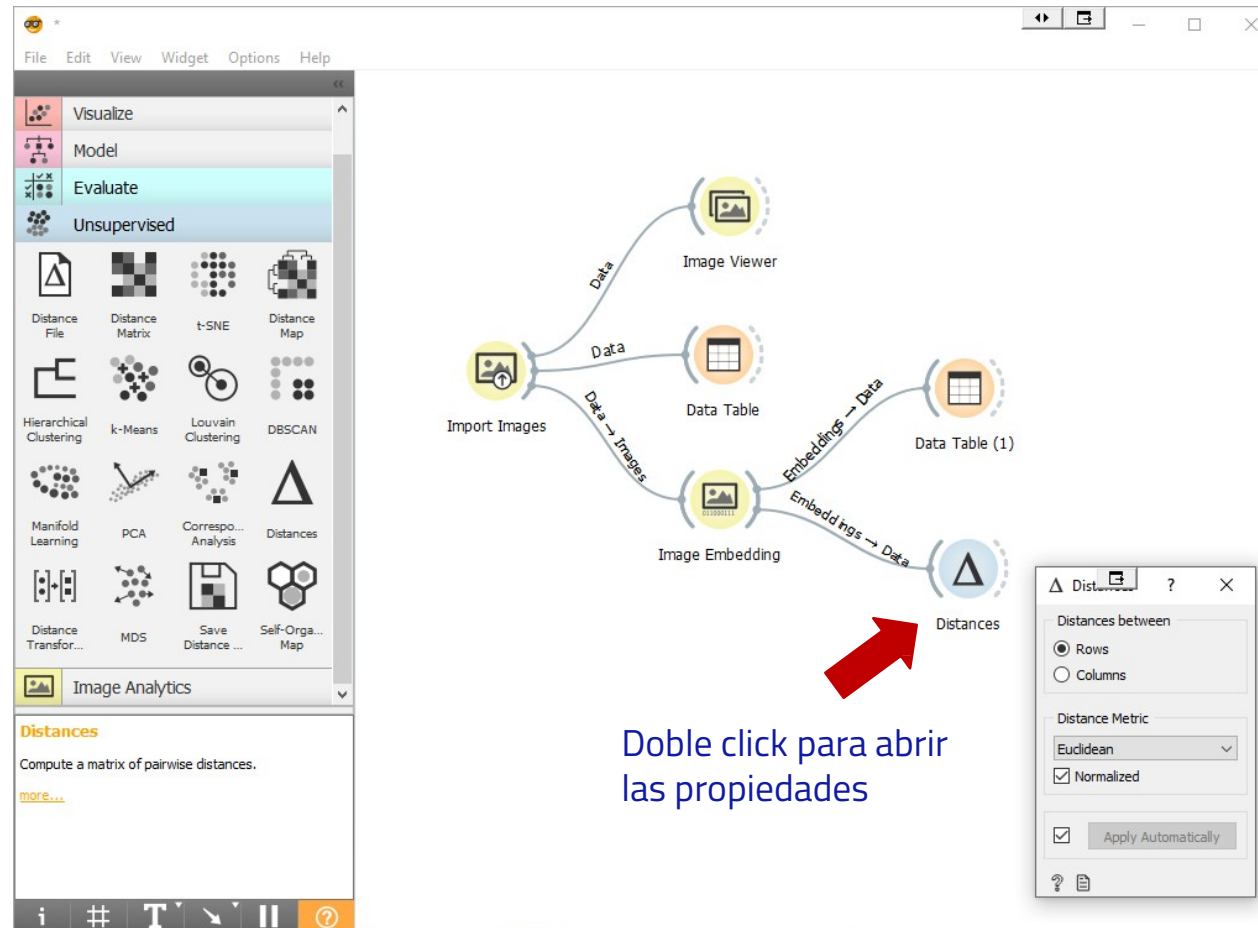
Realicemos un ejercicio de clustering de imágenes

Ahora que se tienen multitud de características nuevas para cada imagen, calcularemos la similitud de las imágenes con la herramienta Distances. Arrastramos Distances a la ventana del proyecto y conectamos a Image Embedding





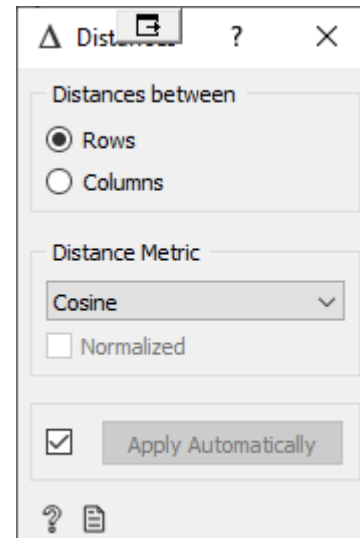
Realicemos un ejercicio de clustering de imágenes





Realicemos un ejercicio de clustering de imágenes

Pasamos las imágenes a distancias para poder comparar su similitud. Y la mejor métrica de distancia al trabajar con imágenes es "Cosine"

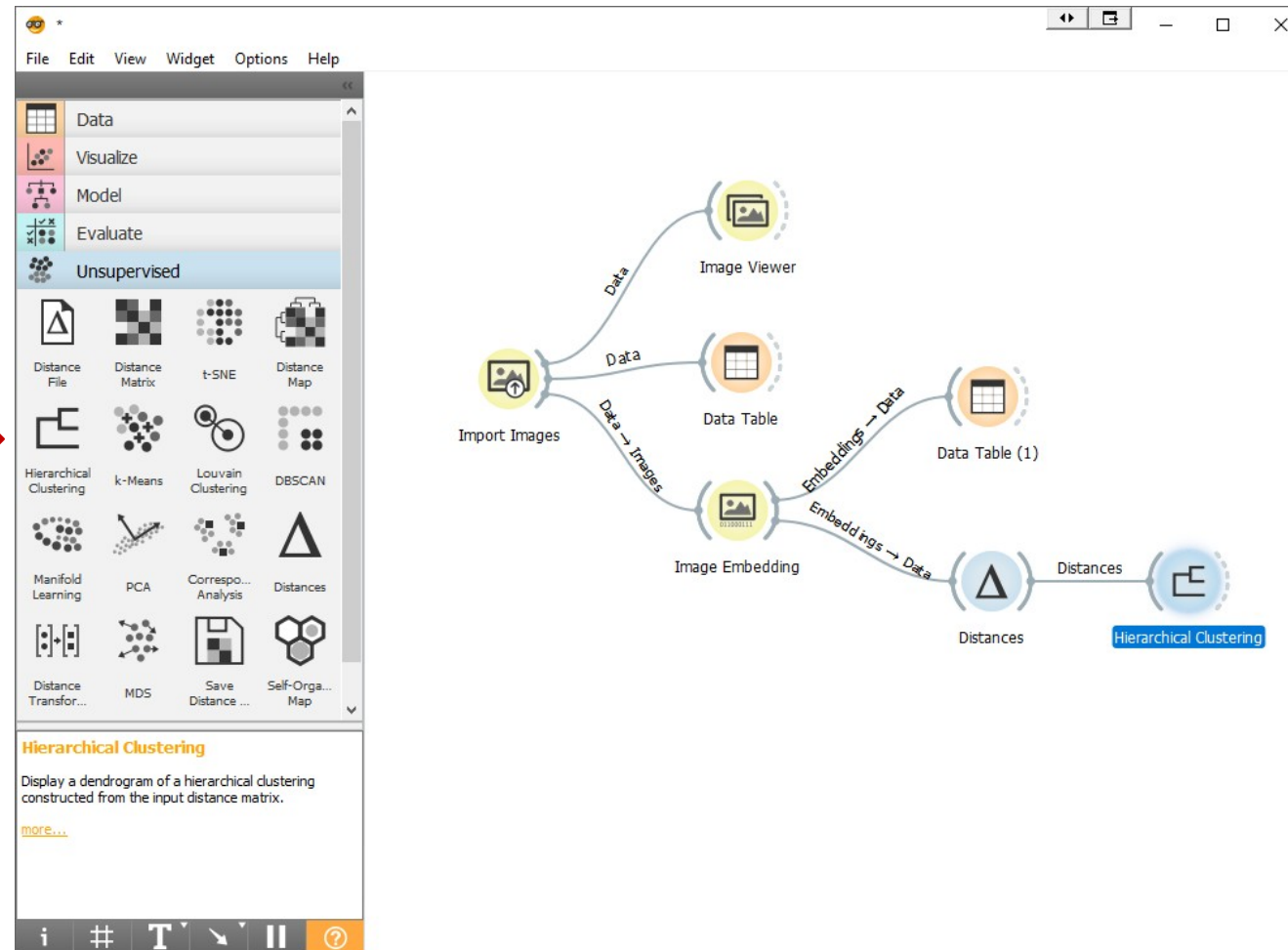




Realicemos un ejercicio de clustering de imágenes

Luego, pasamos la matriz de distancia a una agrupación jerárquica.

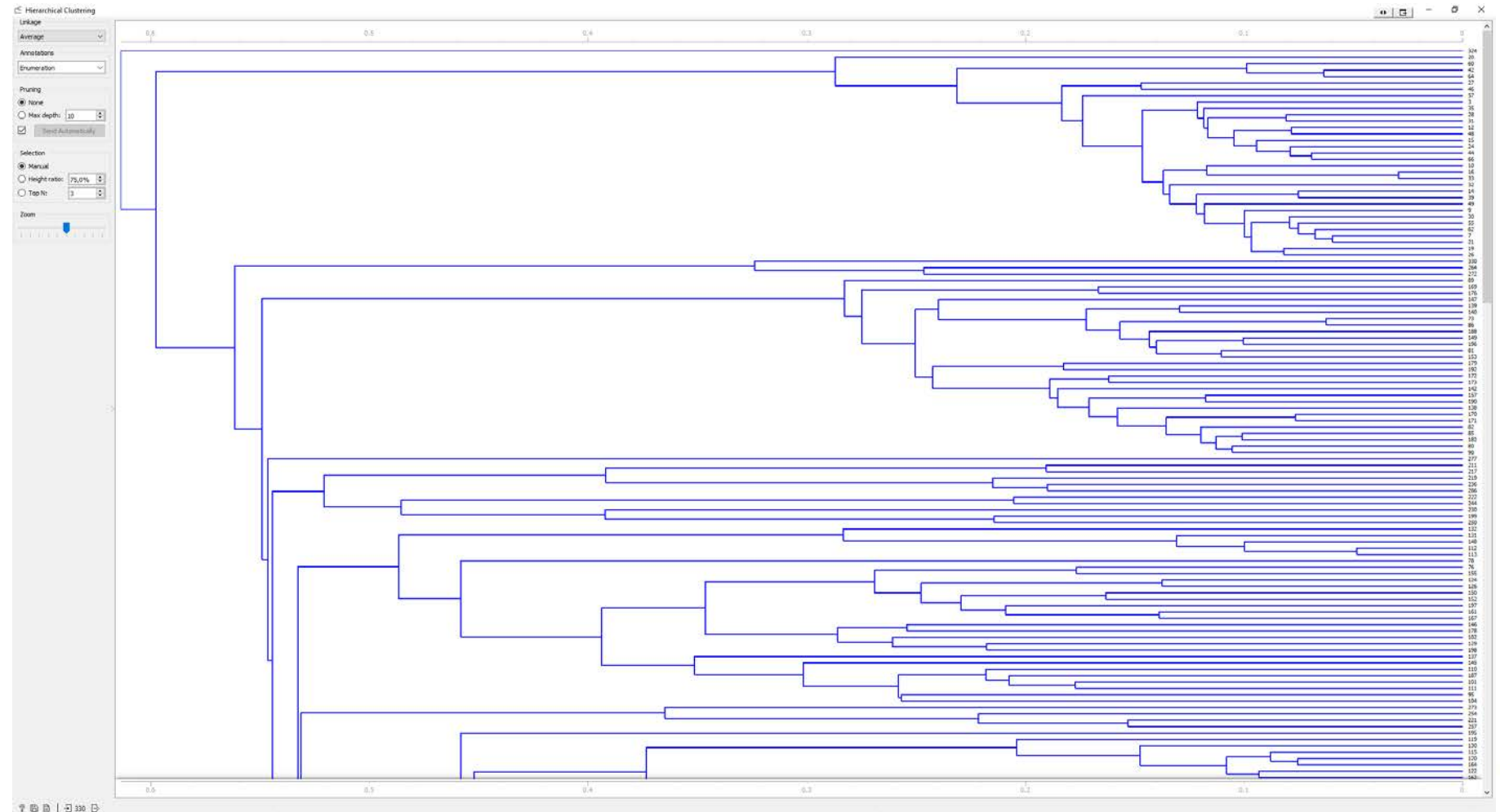
Arrastramos Hierarchical Clustering a la ventana del proyecto y conectamos a Distances





Realicemos un ejercicio de clustering de imágenes

Al abrir el
Hierarchical
Clustering
podremos ver
algo como lo
siguiente

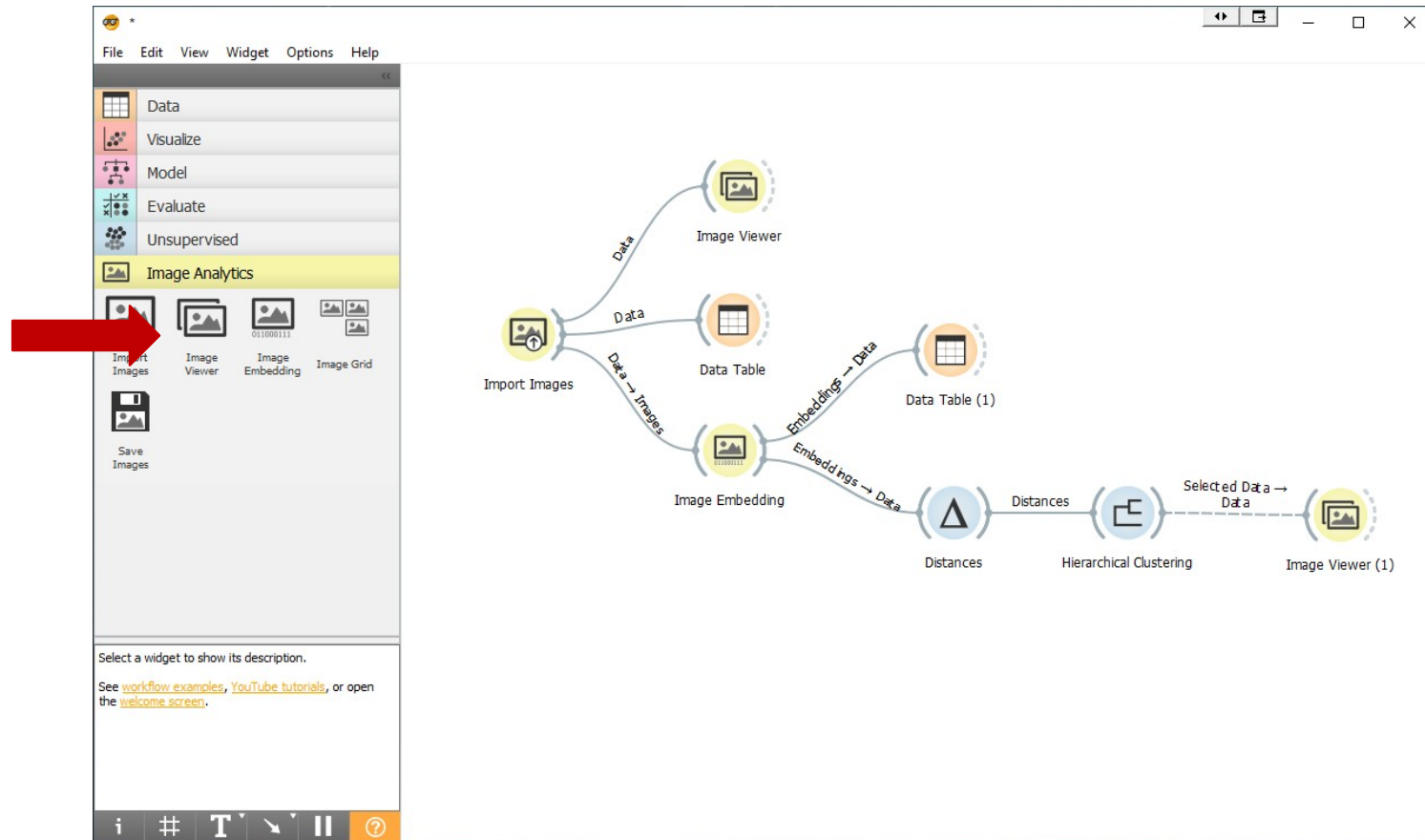




Realicemos un ejercicio de clustering de imágenes

Vamos a visualizar el resultado para comprobar si ha realizado la similitud de imágenes correctamente.

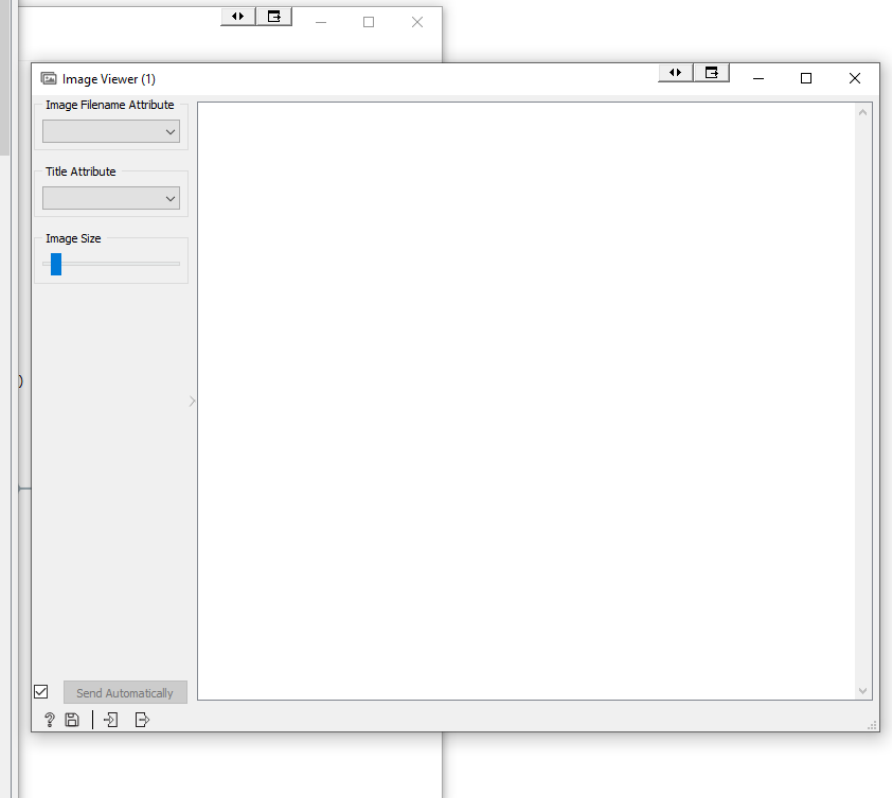
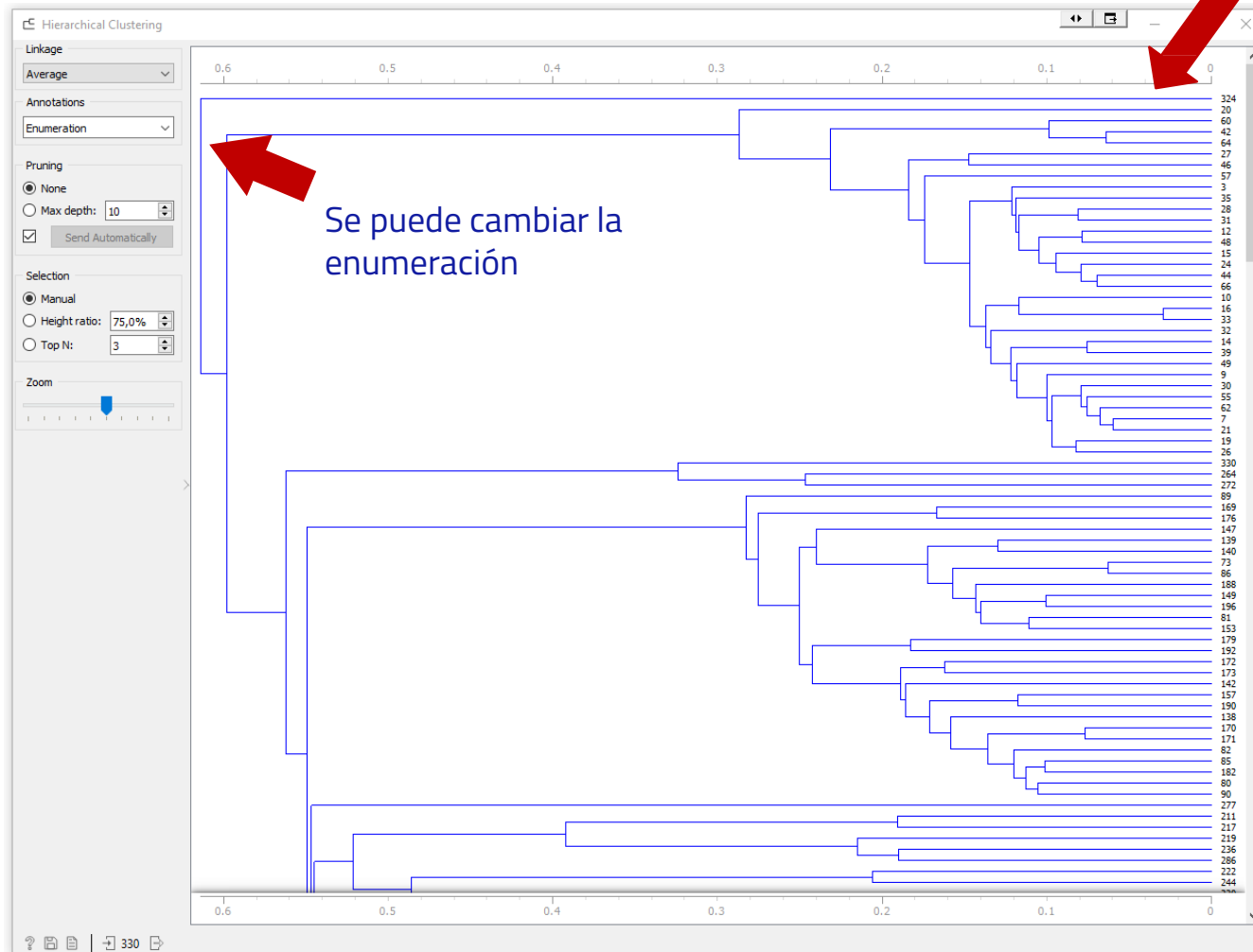
Arrastramos Image Viewer, lo conectamos a Hierarchical Clustering y abrimos Hierarchical Clustering e Image Viewer para la comprobación.





Realicemos un ejercicio de clustering de imágenes

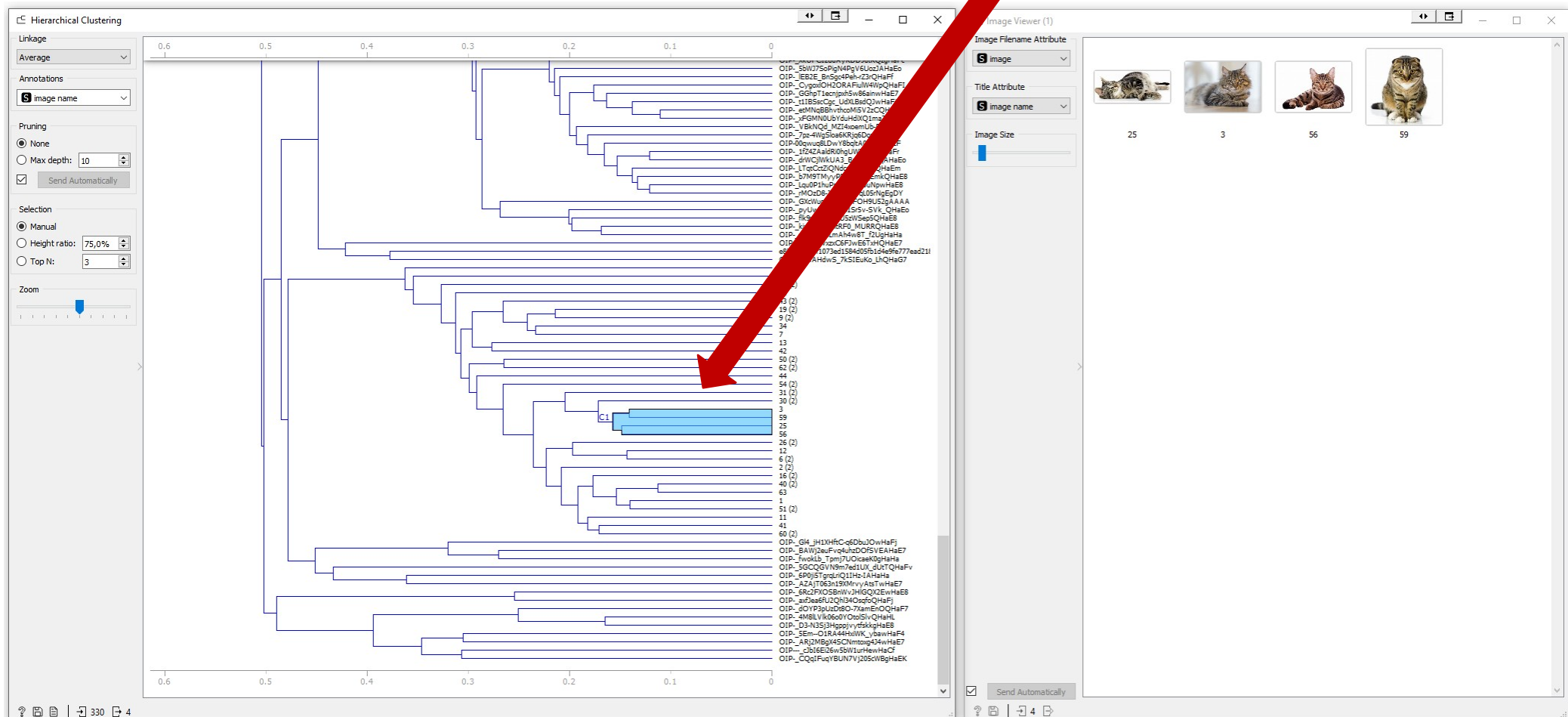
Al ir seleccionando las agrupaciones se irán viendo las imágenes en el Image Viewer





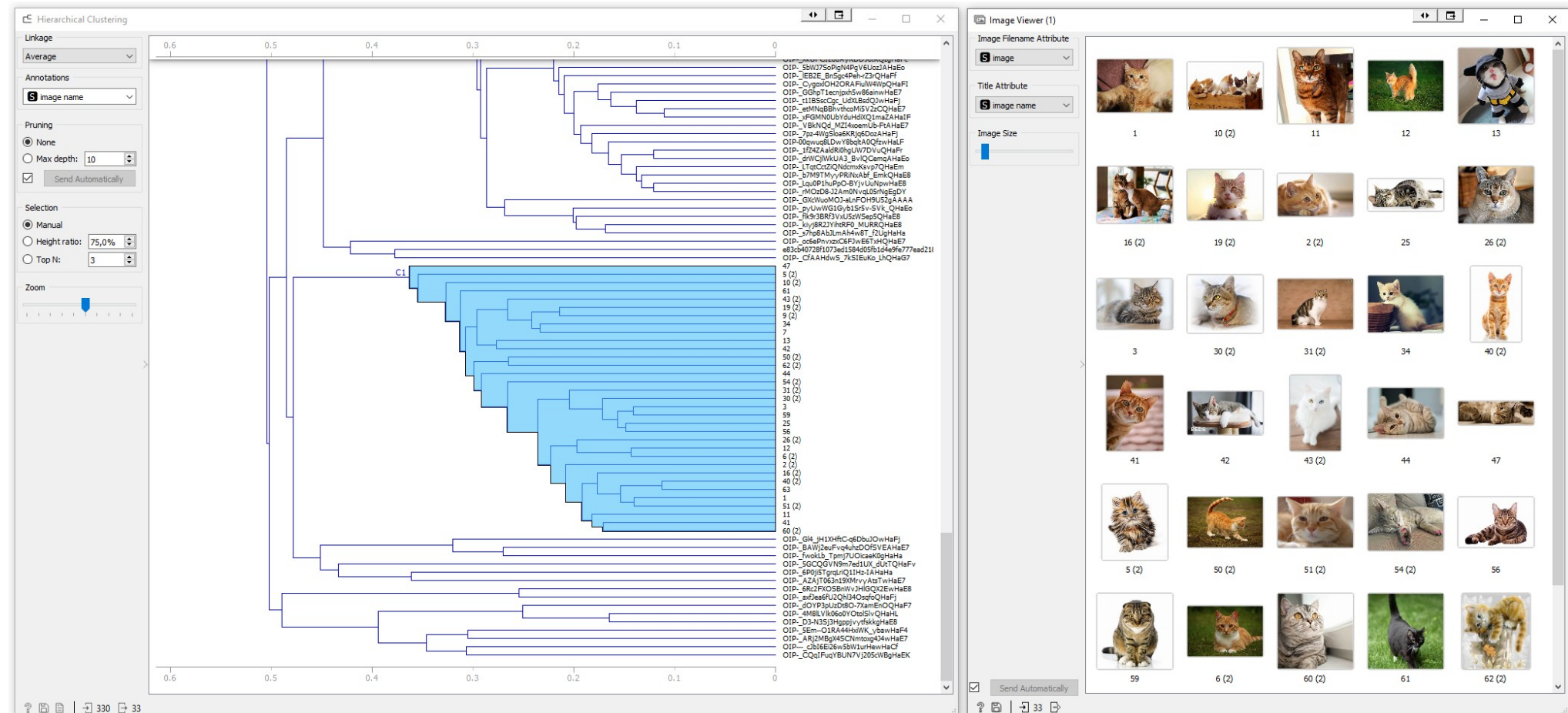
Realicemos un ejercicio de clustering de imágenes

Vamos seleccionando desde la agrupación más pequeña a la mayor para ver si somos capaces a encontrar a todos los gatos y que estén agrupados





Realicemos un ejercicio de clustering de imágenes



33 imágenes de gatos es correcto.
Hay 33 imágenes de cada animal



EJERCICIO OPCIONAL

- Has visto como realizar un ejercicio de clustering sobre un dataset de imágenes. Si quieres realizar otro ejercicio opcional, busca la forma de realizar un ejercicio de clasificación sobre el mismo dataset de imágenes u otro proyecto que prefieras para profundizar conceptos en Orange.