



Skeptical notes on the Machine Learning for Ecology

using Orange3

<https://ai.devvinay.in/>





OutPut Box

Hello students! I am excited to be teaching a new Machine Learning in Ecology course using the software orange3. This is a 4-hour course that will cover the basics of machine learning, including how to use algorithms to predict patterns in data. The goal of this course is to provide students with the tools they need to start using machine learning in their own research. The course will be held on Thursday, September 28th from 9am-1pm EST. During the course, students will learn how to: - Understand what machine learning is and how it can be used in ecology -Use algorithms to predict patterns in data - Understand and interpret the results of machine learning models -Evaluate the performance of machine learning models Registration is now open! I hope you will join me for this exciting new course.

Try Out Regem AI Writer - Improved Version of ReAI - [Go to Writer](#)


[Contact For Project](#)

Ask Me Anything?

New ReAI Version - [Try Now](#)

You Awesome Question

welcome email to my students for a Machine Learning in Ecology course of 4 hours using the software orange3



[Answer Me ReAI](#)

[Try New AI Writer](#)

Learn How to use ReAI?[Learn More](#)

Sending a Review Motivates the Developer to Create More Cool Stuff Like this. ReAI Community Version is Free Forever.

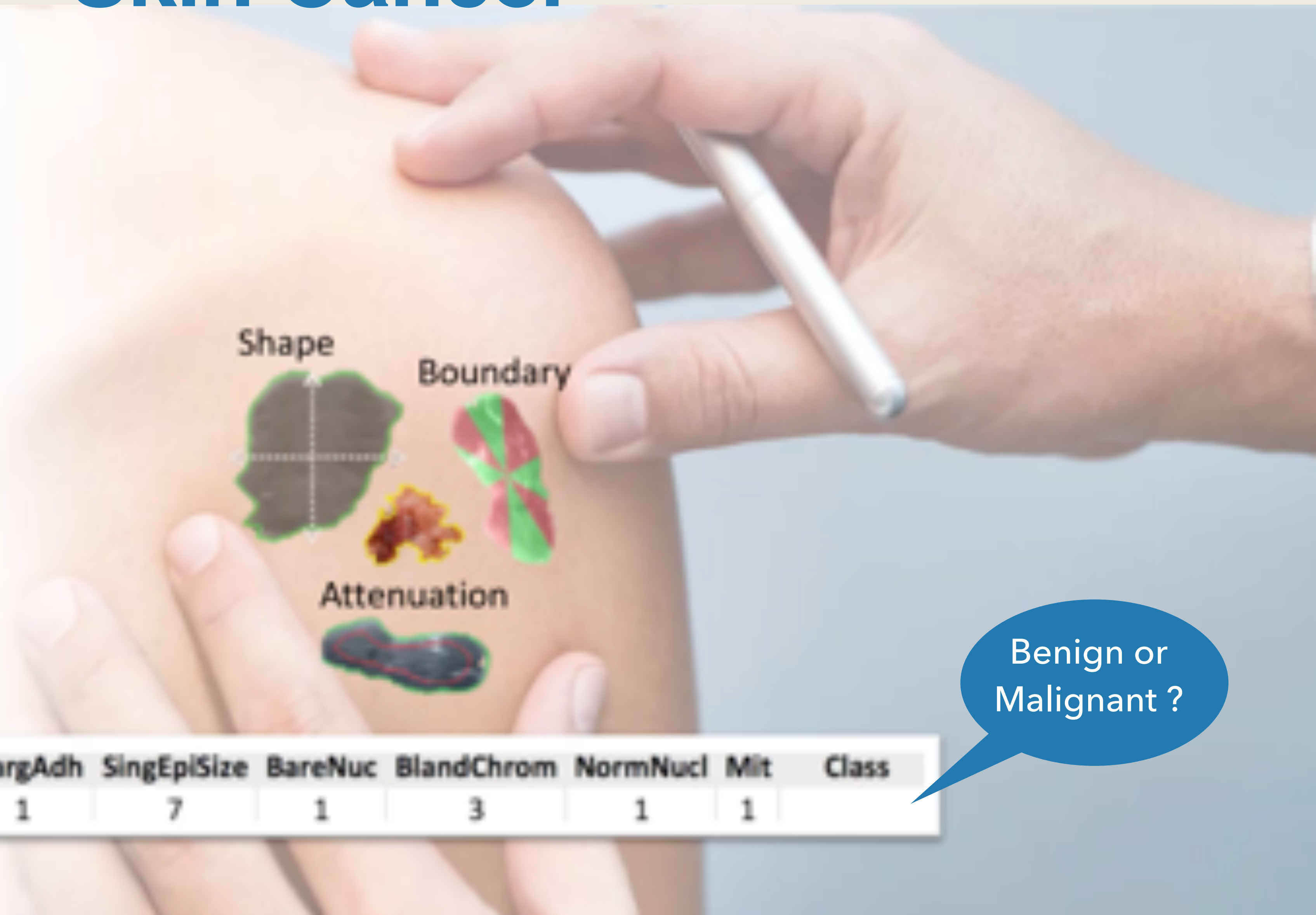
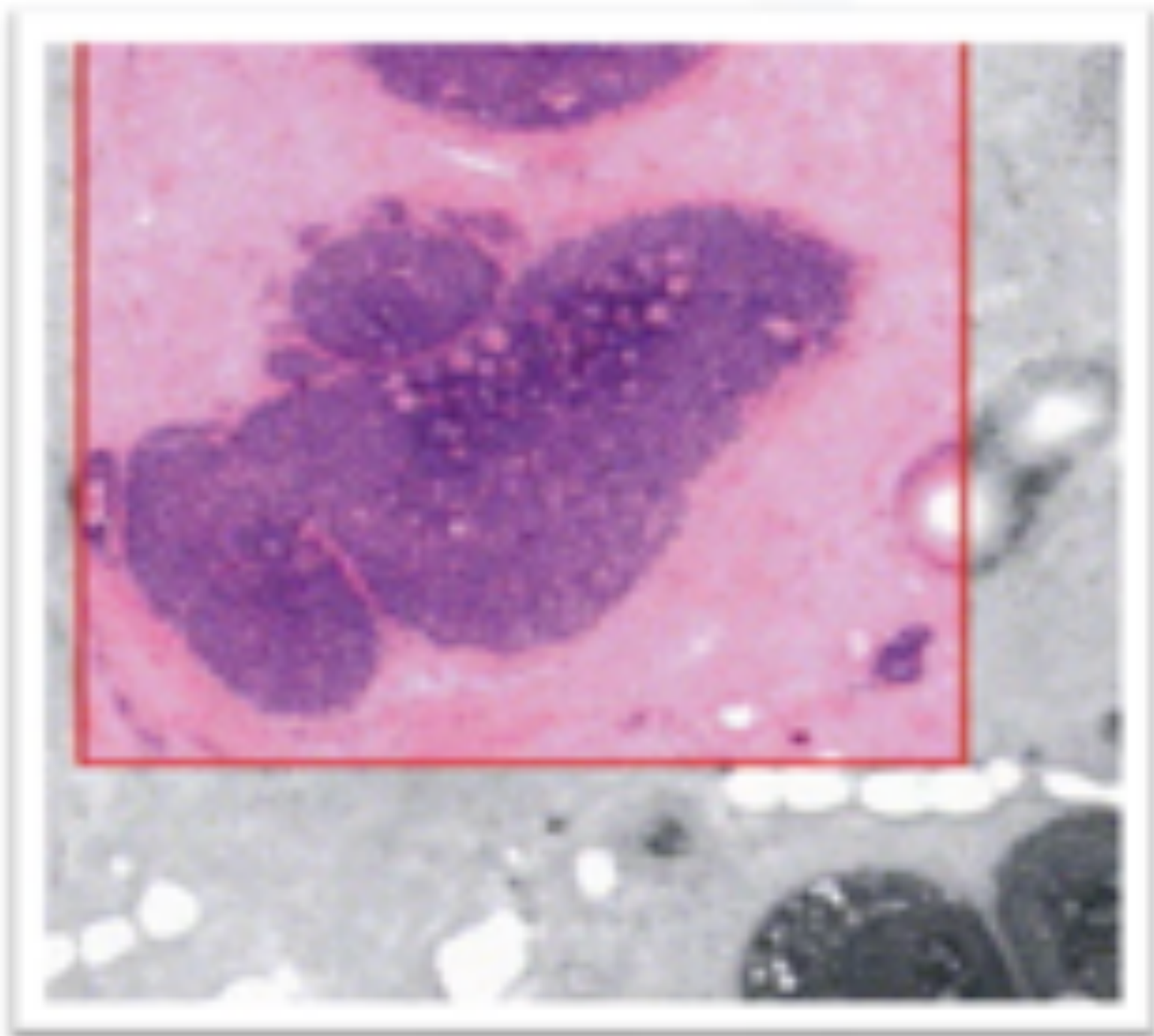
I'm a Freelance Developer! - [Contact via Whatsapp](#)

Crafted and Designed By [Vinay Chaudhary](#)

© ReAI - All Rights Reserved.

First example in Machine Learning

Skin Cancer



ID	Clump	UnifSize	UnifShape	MargAdh	SingEplSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	

Benign or
Malignant ?

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benien

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	Benign


Modeling

Prediction



Accuracy = 89%

First definition of ML

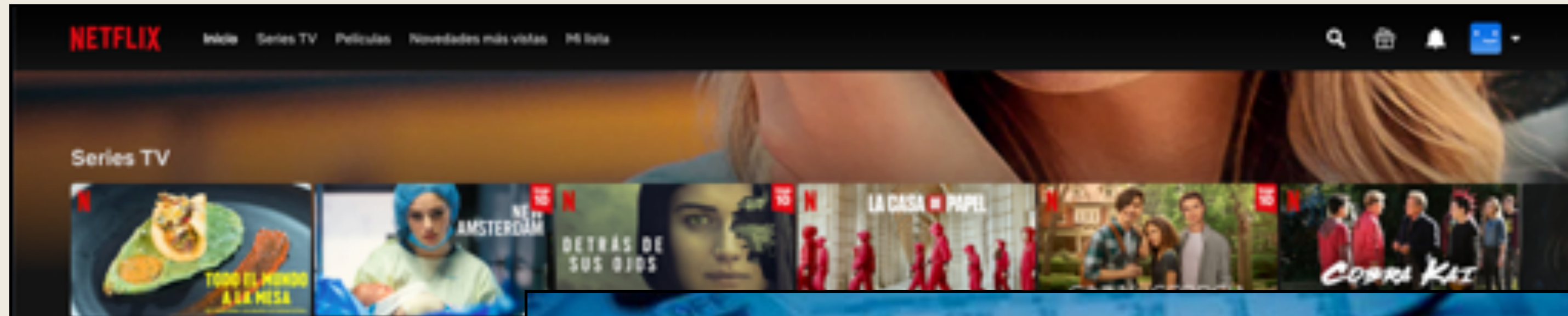


*Machine Learning is the subfield
of computer science that gives
computers the ability to learn
without being explicit
programmed*

**Arthur Samuel, coin the term
machine learning in 1959 at IBM**

Examples in our daily life.

Examples of ML today



Some ideas



Differences between

- Artificial Intelligence
- Machine Learning
- Deep Learning

AI tries to make computers intelligent so that they mimic the cognitive functions of humans.

Artificial Intelligence



Autopilot in Tesla

Automatic car driving





Pattern recognition

Using computer vision

Playing games

Playing go



ML is the branch of AI that covers the statistical part of AI. ML teaches the computers to solve problems by looking at hundreds of examples.





What problem in ML do I have?

Classical Machine Learning

Task Driven

Supervised Learning

(Pre Categorized Data)

Data Driven

Unsupervised Learning

(Unlabelled Data)

Classification

(Divide the
socks by Color)

Eg. Identity
Fraud Detection

Regression

(Divide the
Ties by Length)

Eg. Market
Forecasting

Clustering

(Divide by
Similarity)

Eg. Targeted
Marketing

Association

(Identify
Sequences)

Eg. Customer
Recommendation

Dimensionality
Reduction

(Wider
Dependencies)

Eg. Big Data
Visualization

Obj:

Predications & Predictive Models

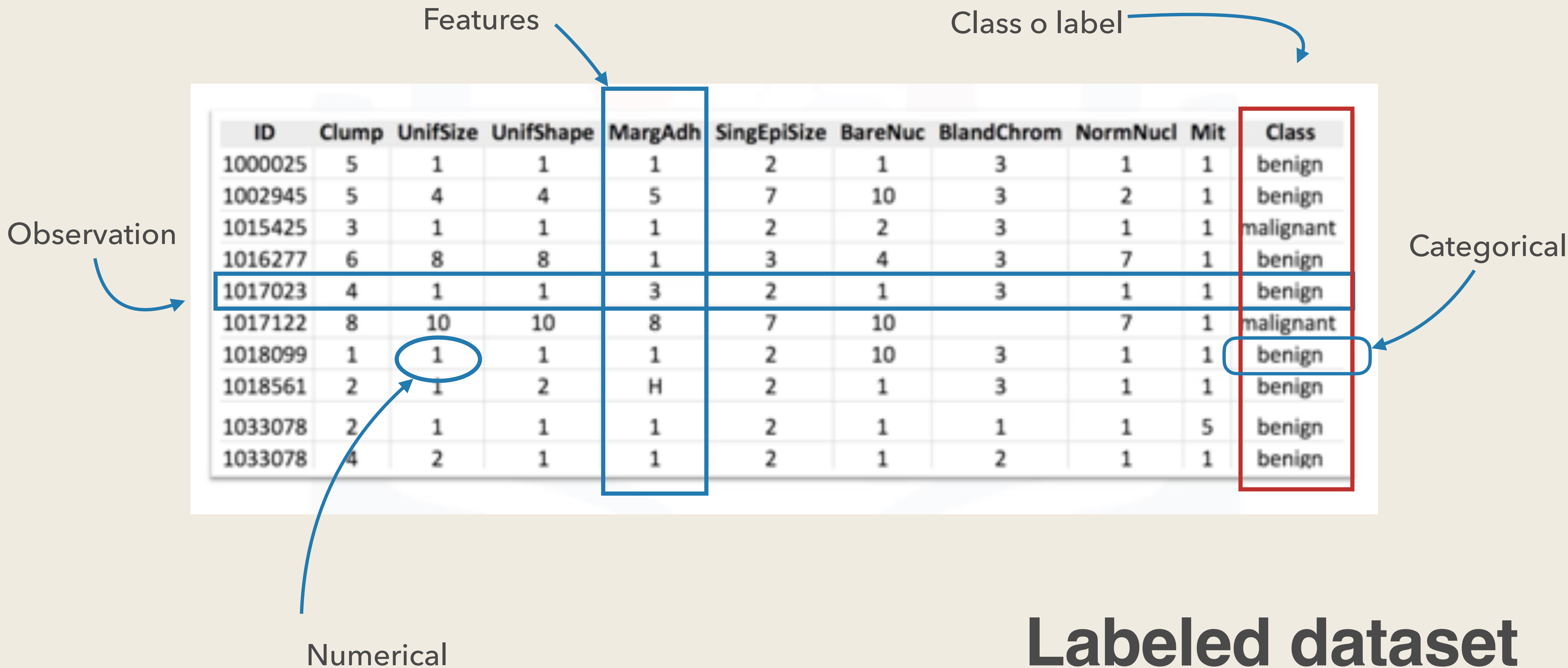
Pattern/ Structure Recognition



ID	Clump	UnitSize	UnitShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benian

Supervised model

In supervised models, we teach the model by training it with some data from labeled dataset



Labeled dataset

I am going to teach you

1 \rightarrow 2

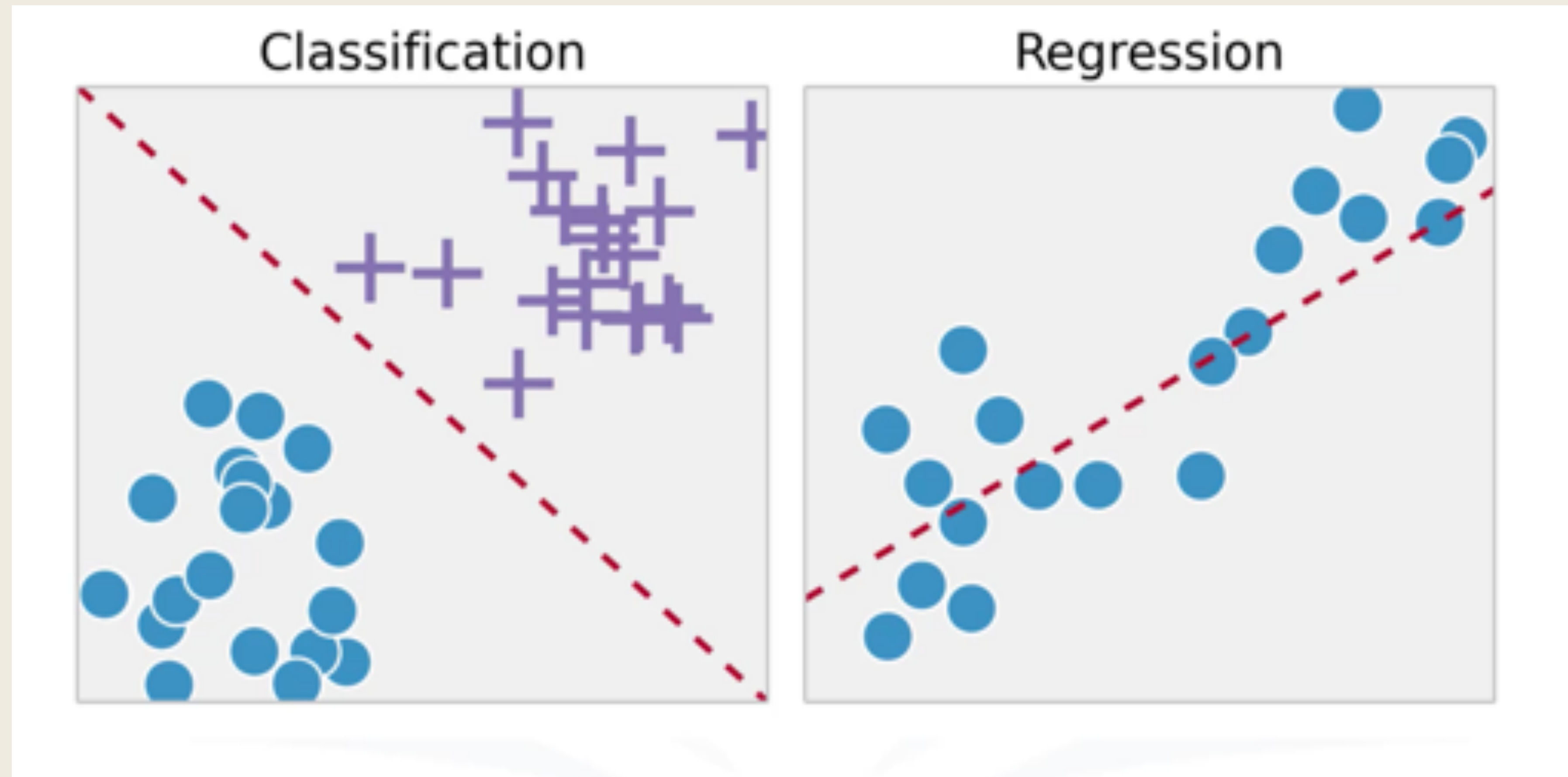
2 \rightarrow 4

5 \rightarrow 10

6 \rightarrow 12

10 \rightarrow ?

Types of supervised techniques

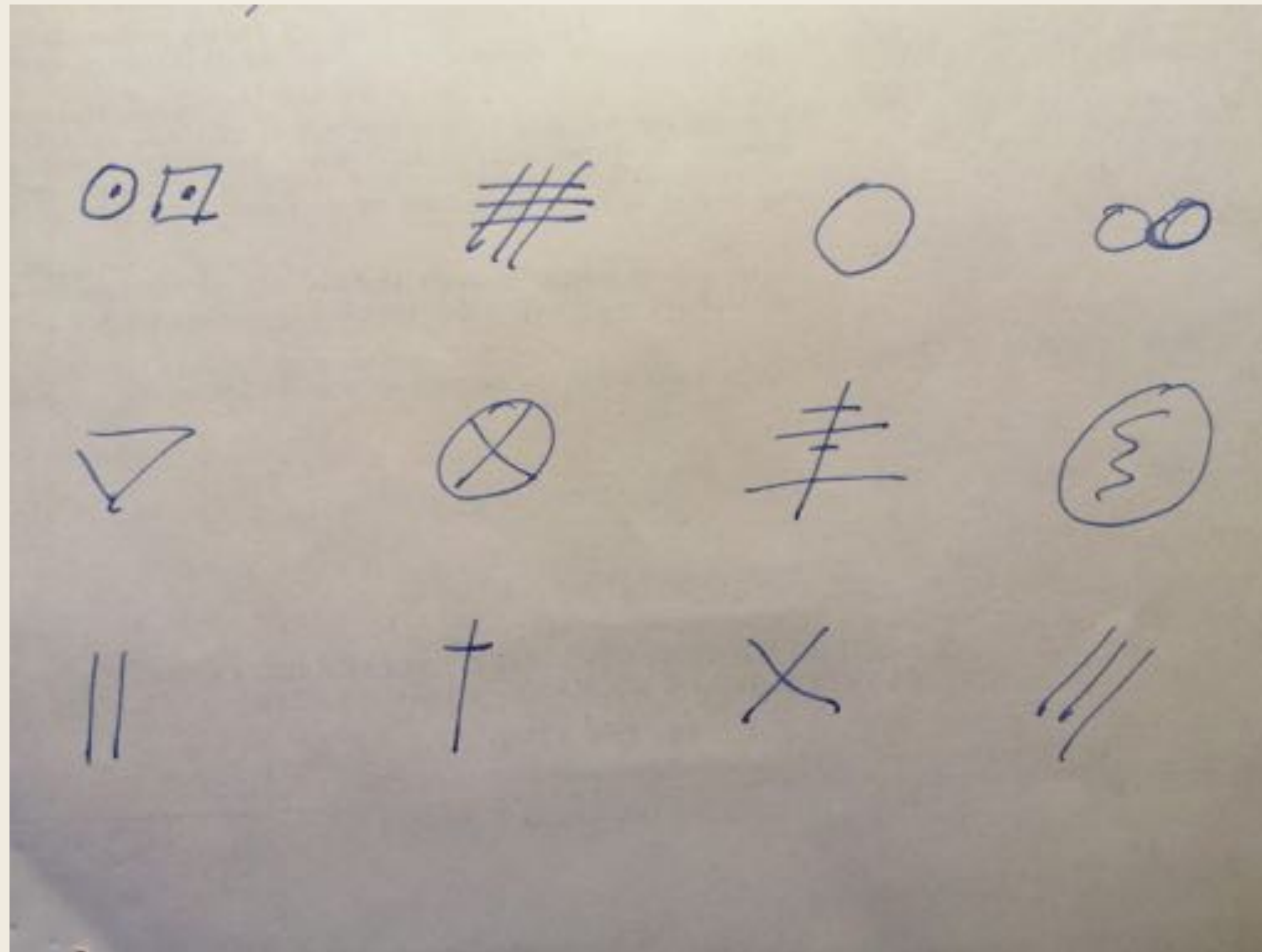


Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

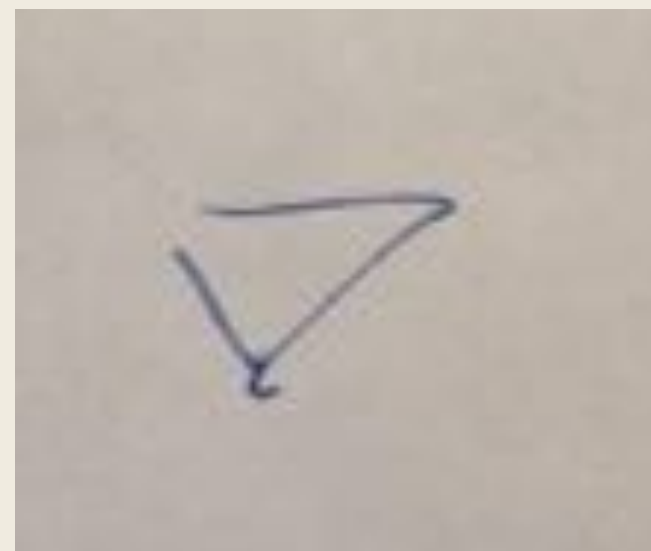
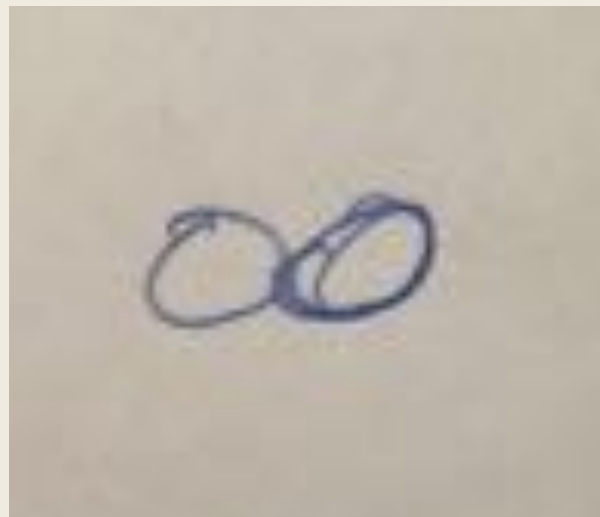
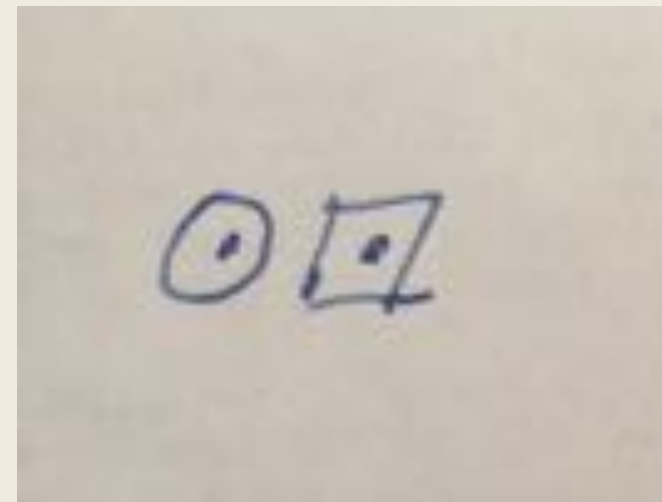
Unsupervised model

We do not supervise the model, but we let the model work on its own to discover information that may not be visible to the human eye.

I am going to teach you as a unsupervised dataset

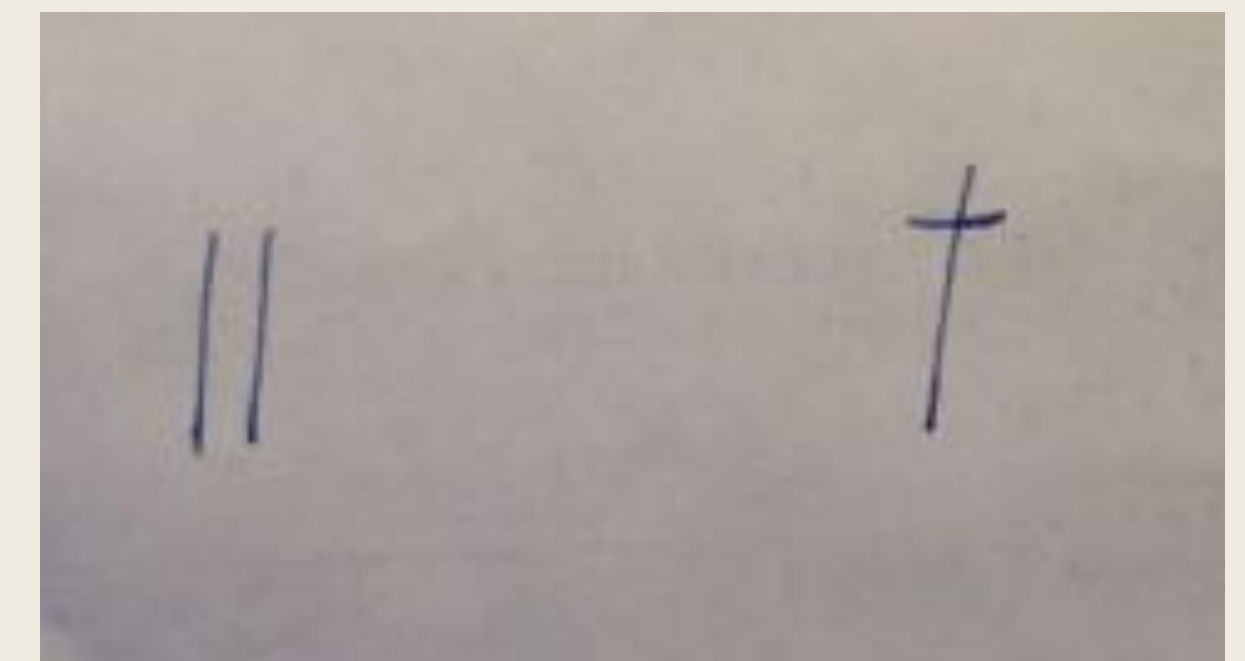
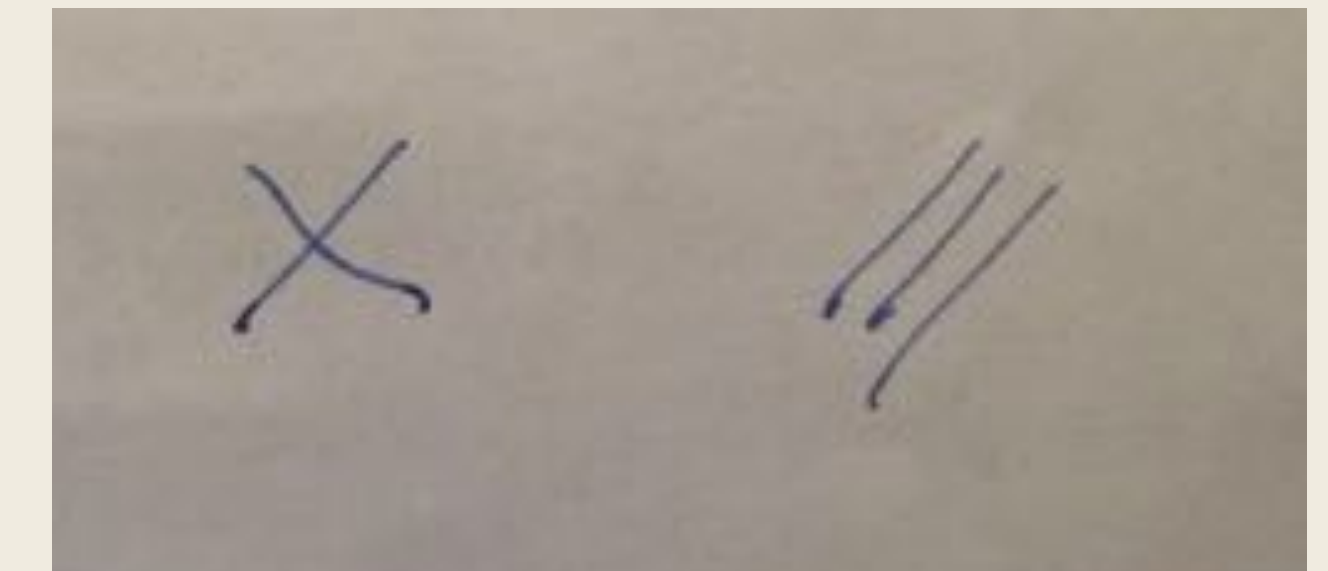


Polygonal language?

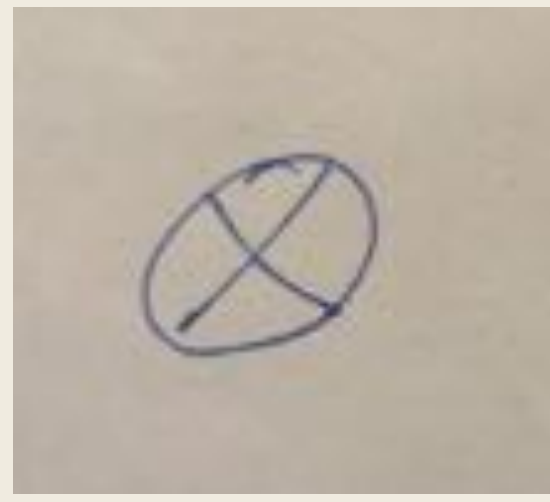
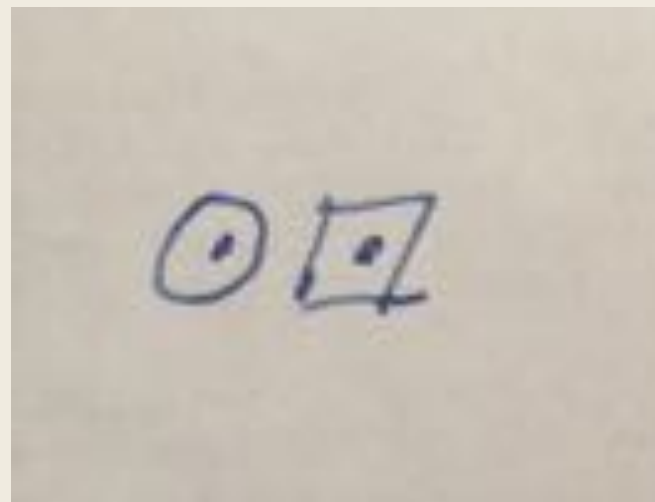


Two clusters

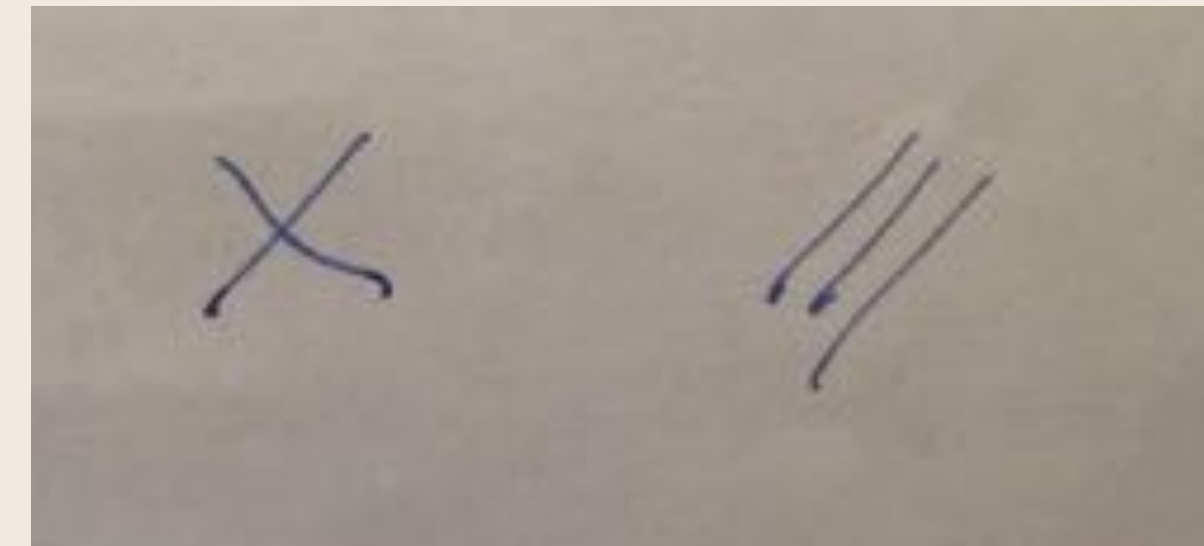
Linear language?



Fulled polygonal language?

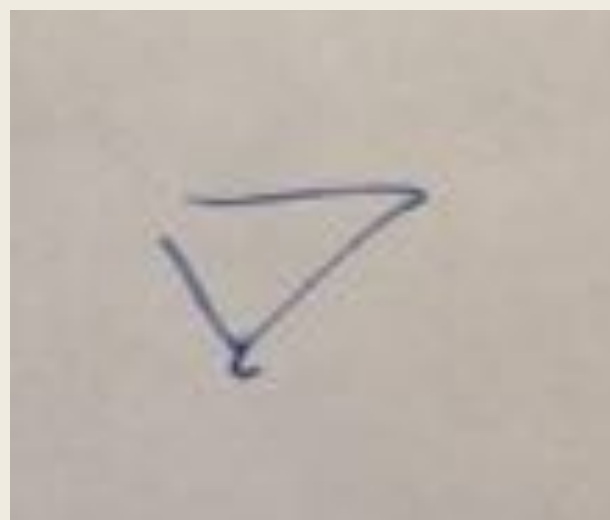
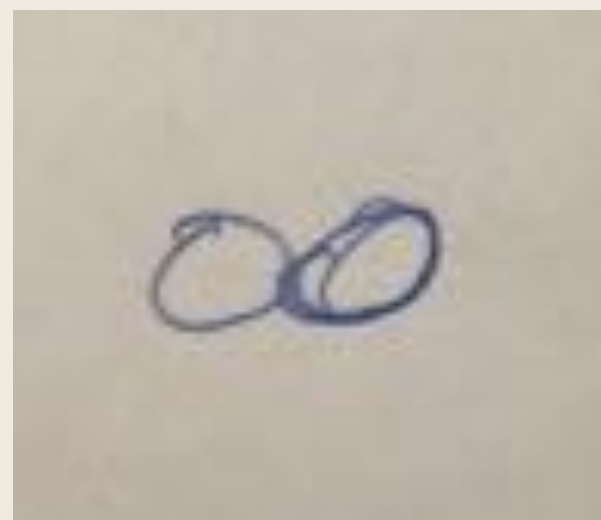


Diagonal Linear language?

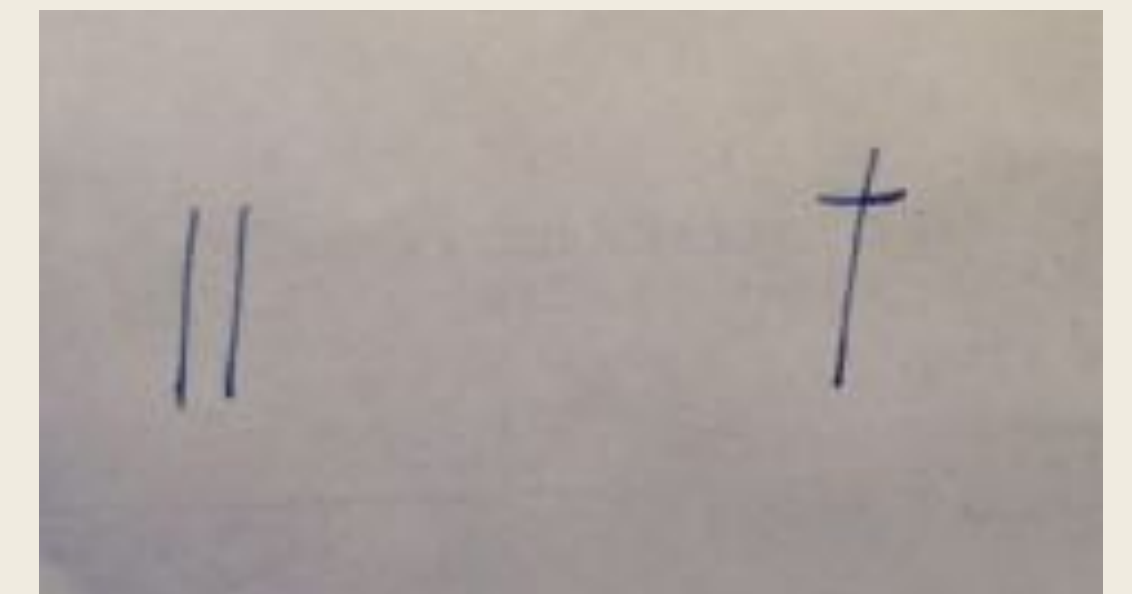


Four clusters?

Empty polygonal language?



Straight Linear language?



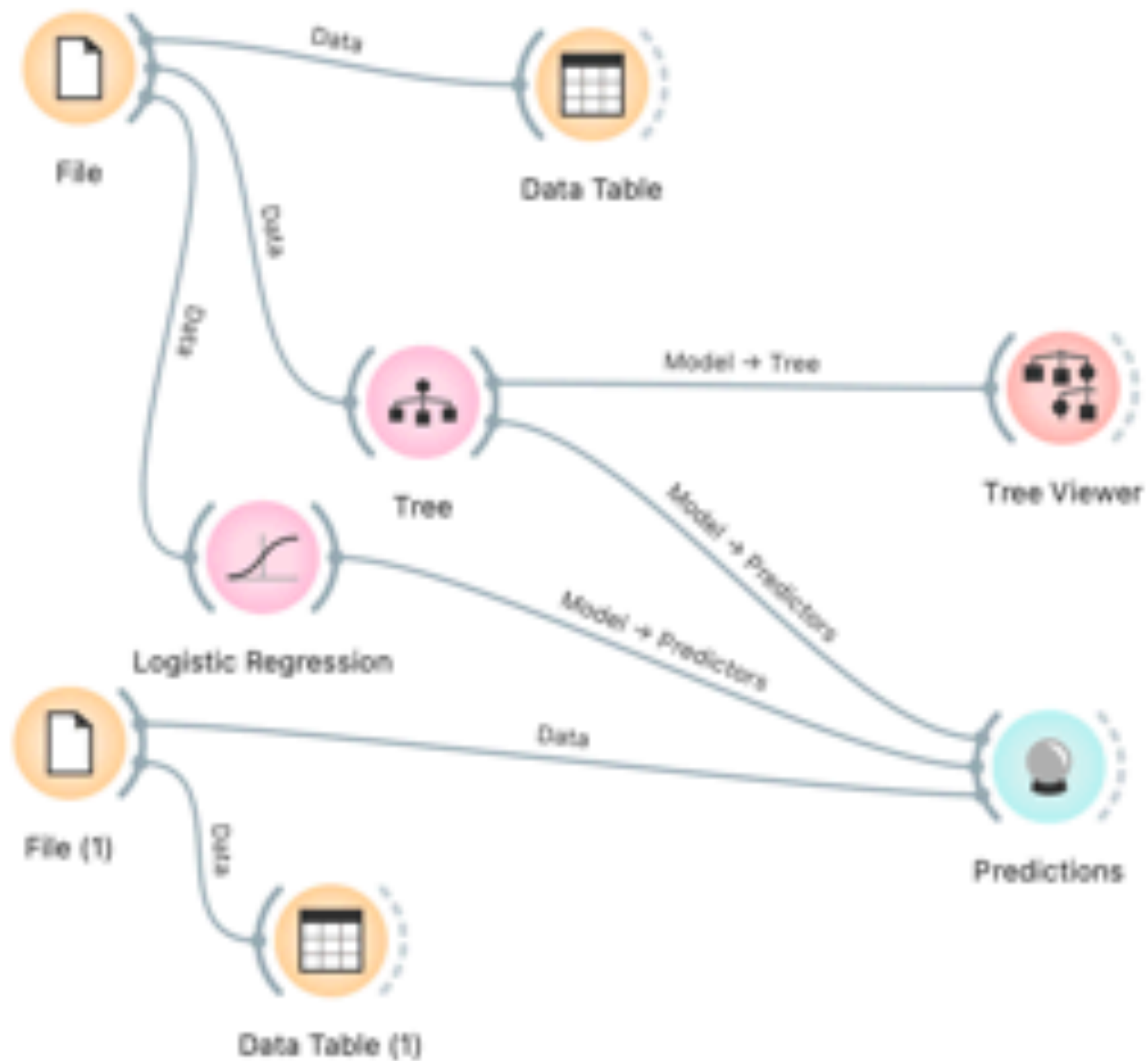
First steps using Orange3



ECOFLOr

First program

- File → Iris
- Data Table
- Scatter plot
- Selected data table



ECOFLO

First Predictions

- Download vegetables and fruits dataset
- Tree → Tree viewer
- Test file
- Prediction
- Logistic Regression

Classification problem in ML

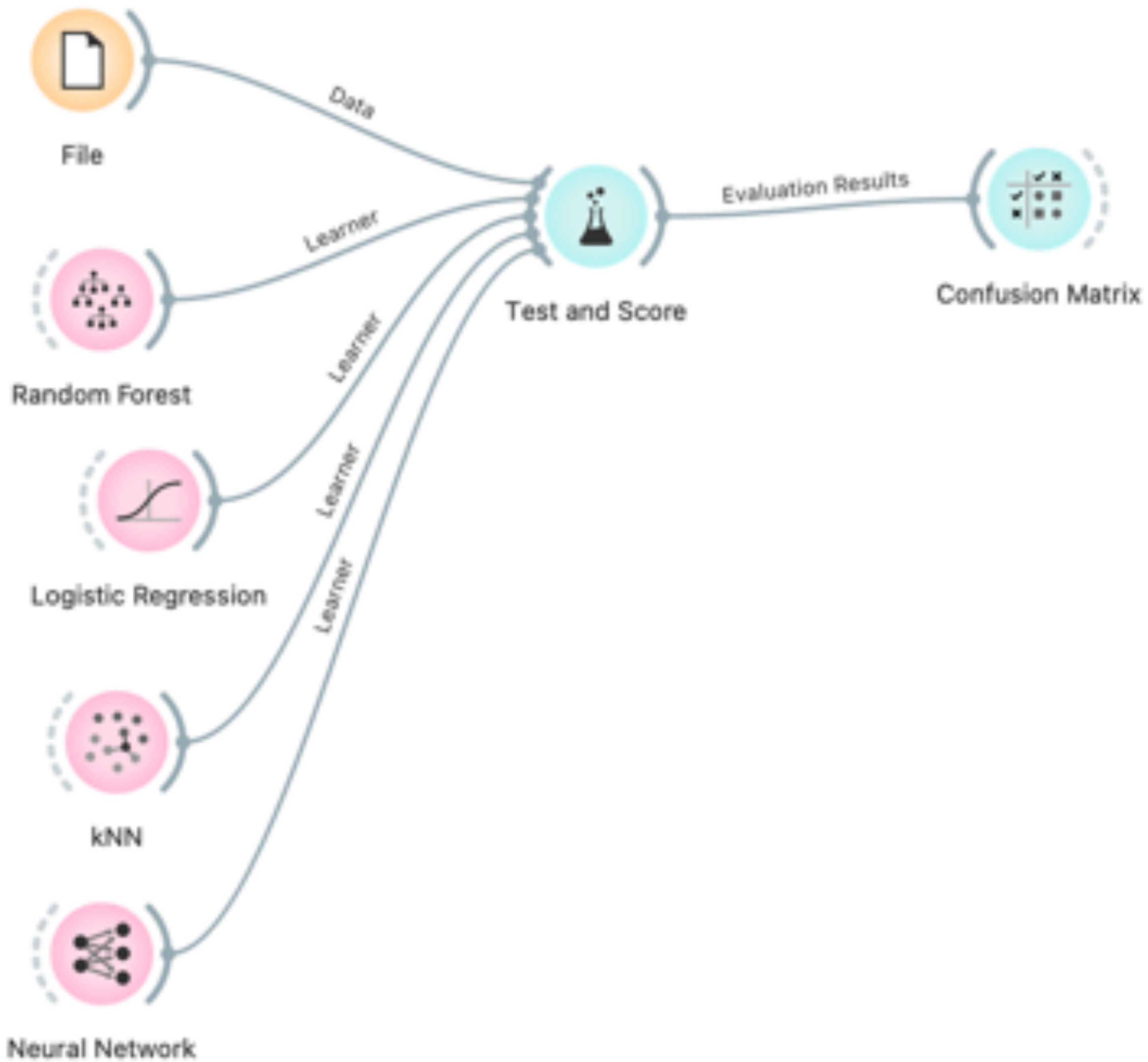
ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

ECOFLOR

Classification

In classification, the goal is to predict a *class label*, which is a choice from a predefined list of possibilities.

- Binary Classification
- Multiclass Classification



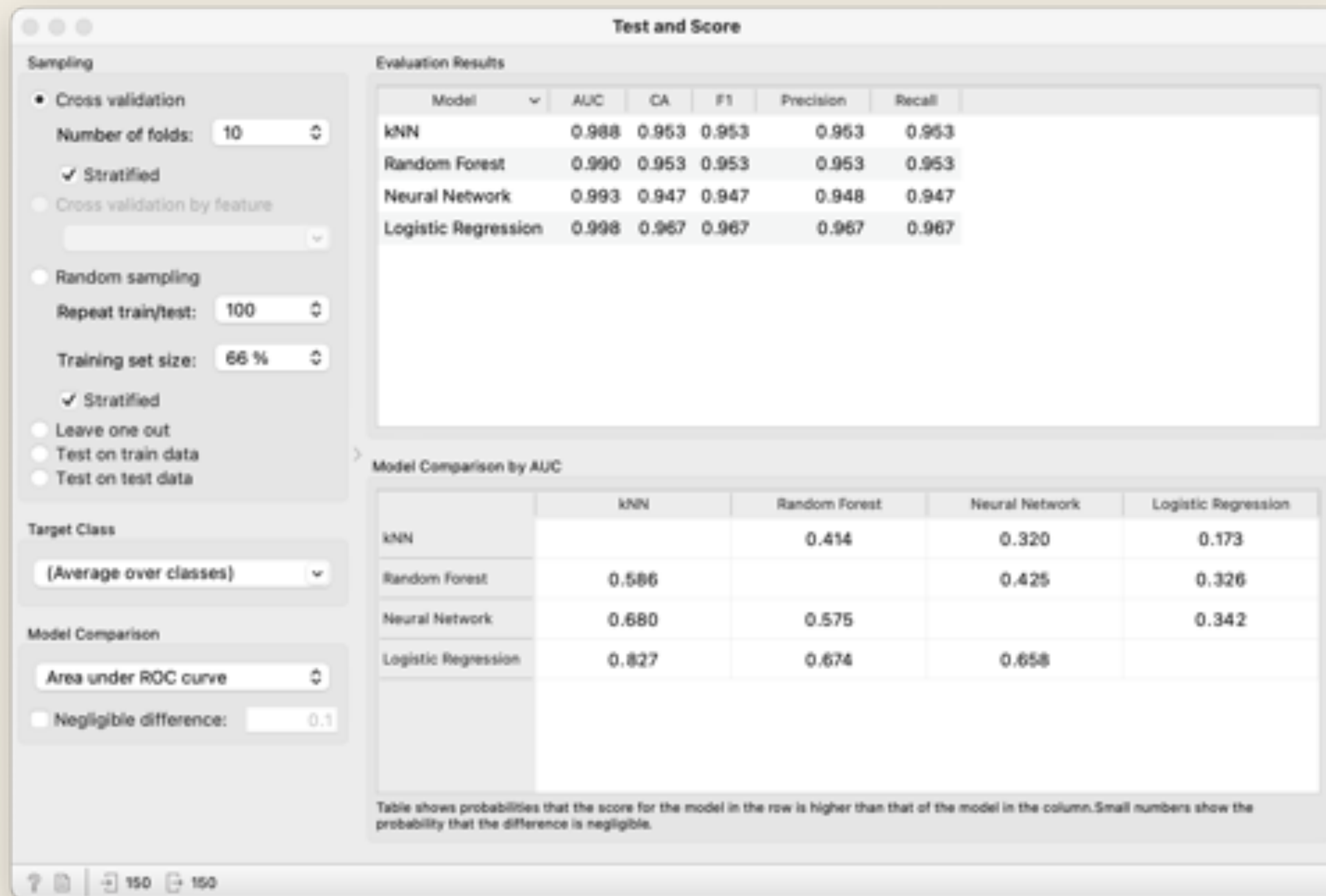
ECOFLOr

Scoring

- K-cross validation
- Test & Score
- Confusion Matrix

Evaluation metrics in classification

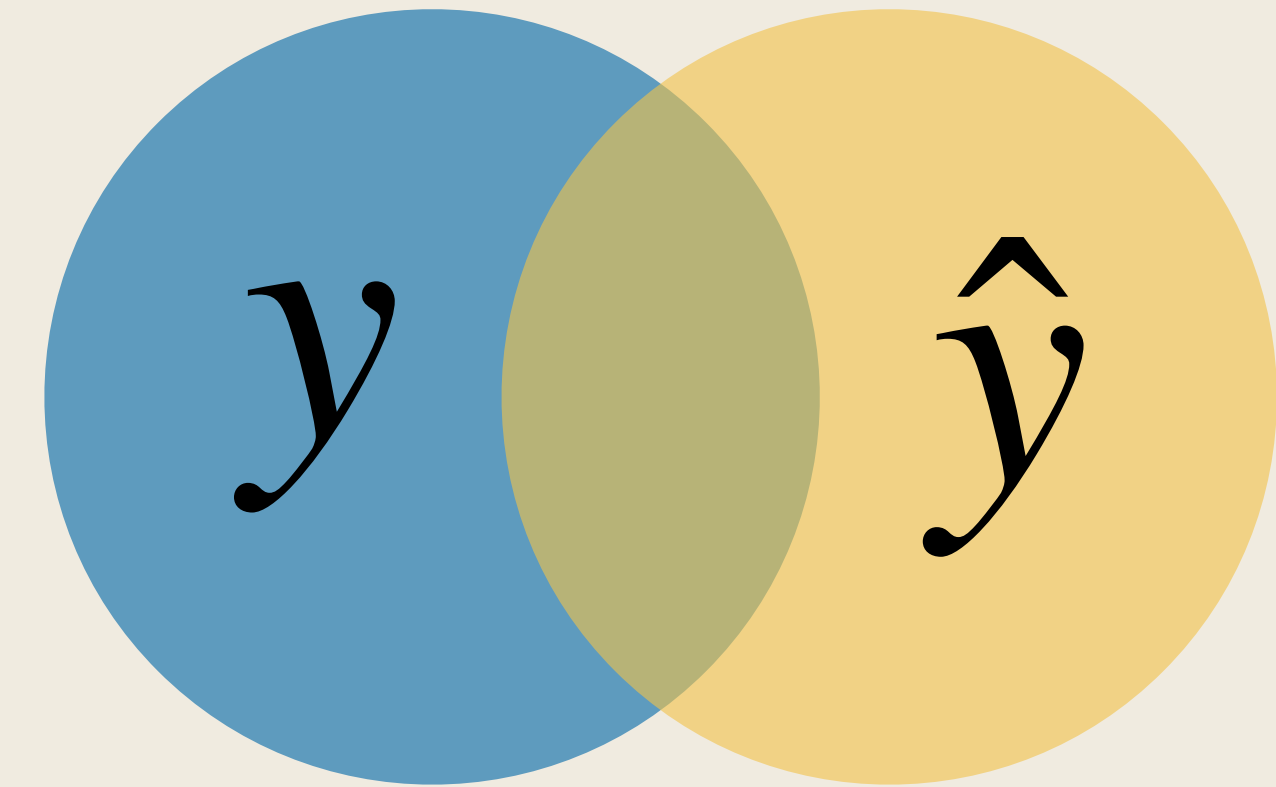
- We are talking about some metrics:
- Jaccard index,
- Confusion matrix



Jaccard index

y = True values (actual values)

\hat{y} = Predicted values of our model



Jaccard index $[0,1]$

$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$

Ejemplo

$$y = [0,0,0,0,0,0,1,1,1,1,1] \quad \hat{y} = [1,1,0,0,0,0,1,1,1,1,1]$$

$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|} = \frac{8}{10 + 10 - 8} = 0.66$$

Confusion Matrix

TP = True Positives

FN = False Negatives

FP = False Positives

TN = True Negatives

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

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

$$\textbf{F1-score} [0,1] = \frac{2 * \textit{Precision} * \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

Confusion Matrix					
		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Actual	Iris-setosa	50	0	0	50
	Iris-versicolor	0	47	3	50
	Iris-virginica	0	2	48	50
Σ		50	49	51	150

Classification Accuracy is the proportion of correctly classified examples

$$CA = \frac{TP}{Total}$$

Classification Algorithms

Machine Learning Classification Algorithms

Logistic Regression

Naive Bayes

Decision Tree



Support Vector Machines

Random Forest

K-Nearest Neighbours

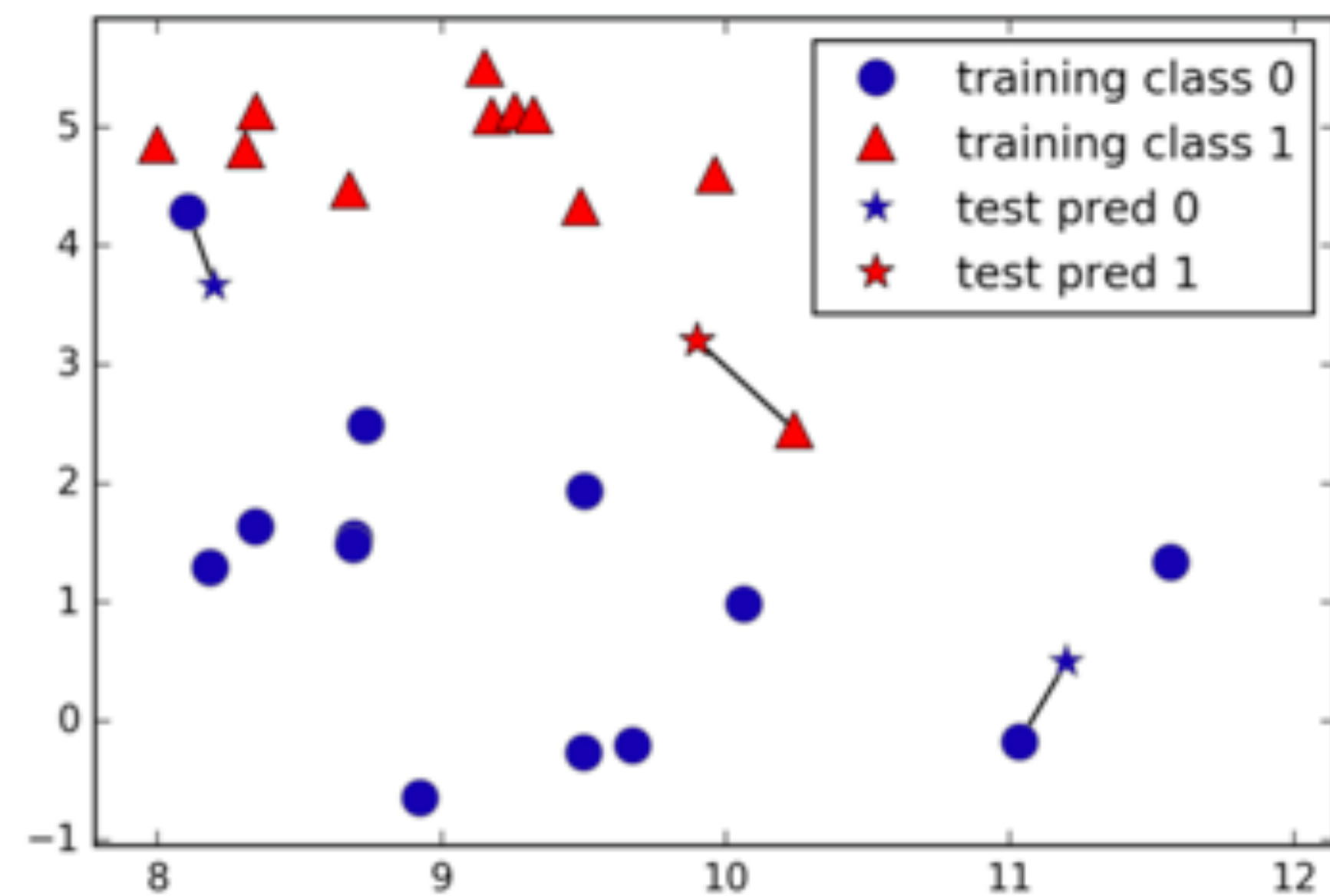


Figure 2-4. Predictions made by the one-nearest-neighbor model on the forge dataset

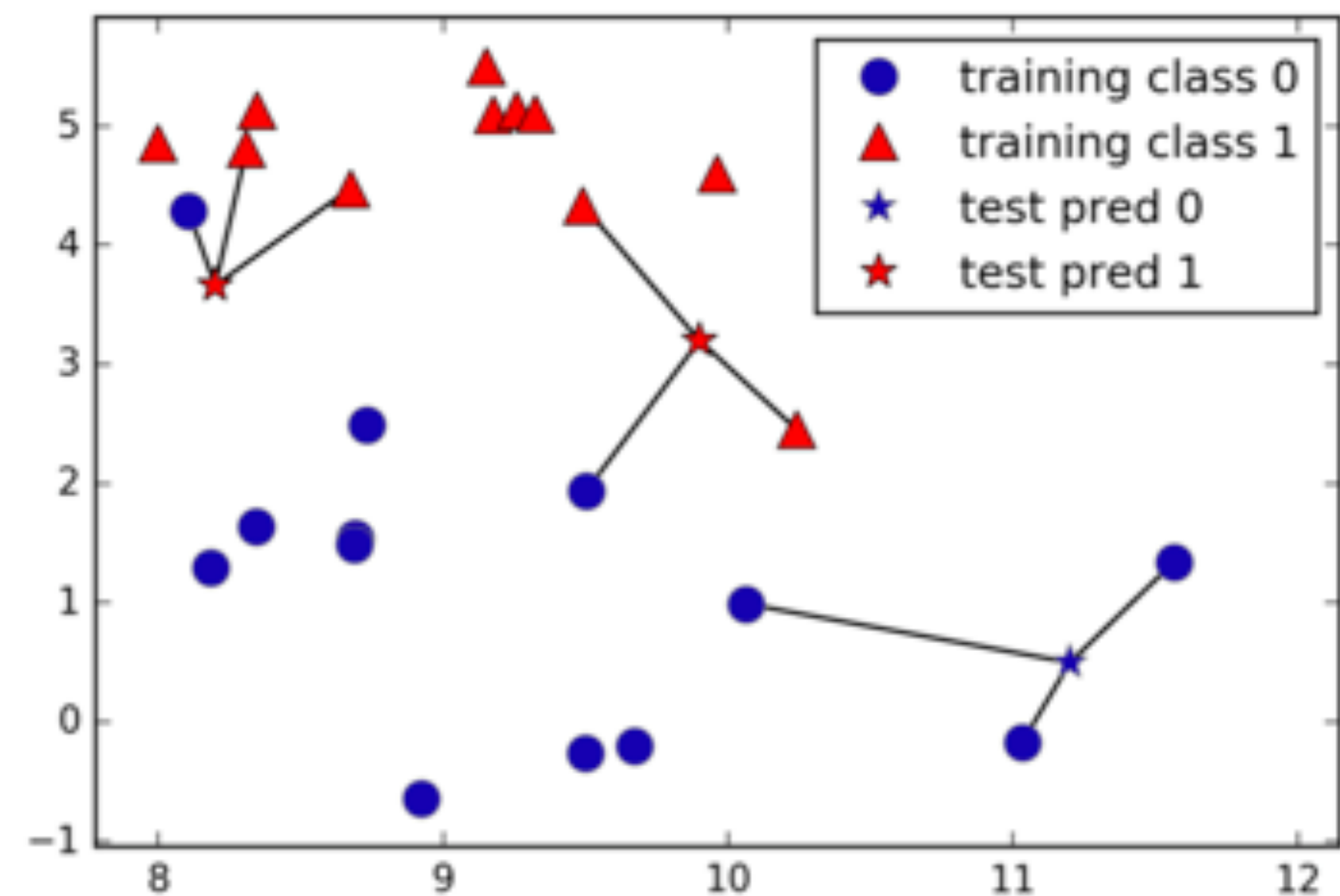
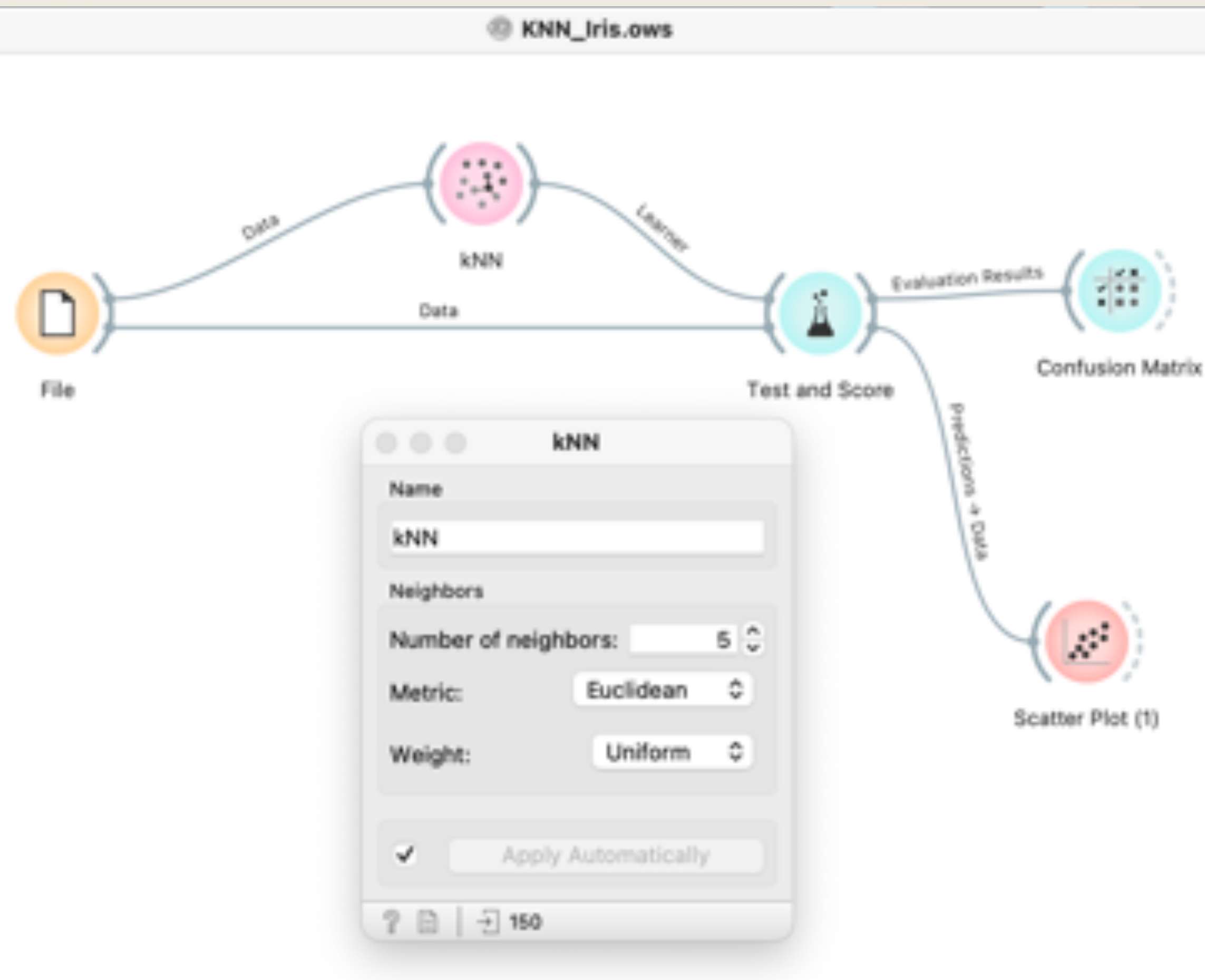


Figure 2-5. Predictions made by the three-nearest-neighbors model on the forge dataset

ECOFLOr

K-Nearest Neighbors

- Building the k-NN algorithm consists only of storing the training dataset.
- To make a prediction for a new data point, the algorithm finds the closest data points in the training dataset—its “nearest neighbors.”



ECOFLOr

K-nearest Neighbors

- Iris
- KNN → Test Score → Confusion Matrix
- Constant → Test Score → Confusion Matrix
- Scatter Plot (KNN)

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

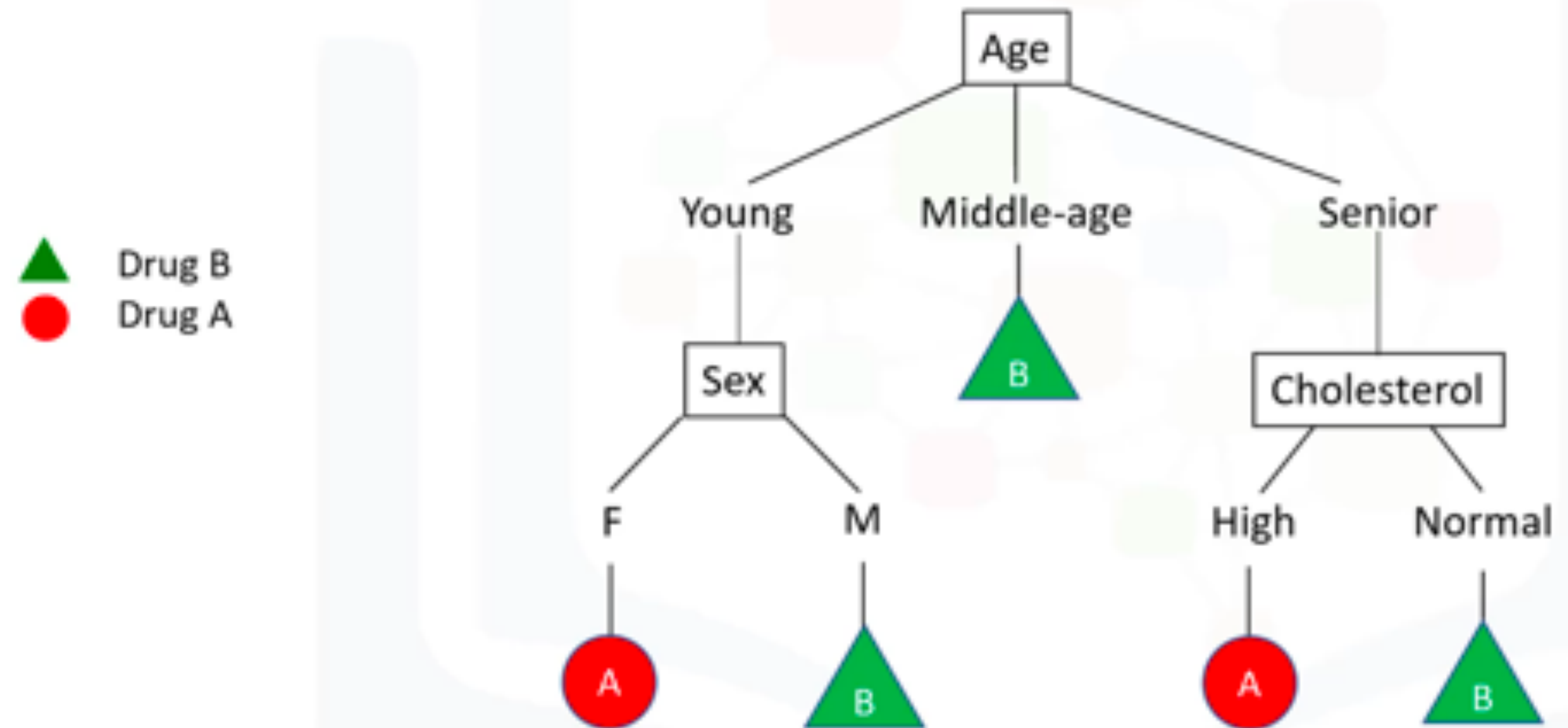
ECOFLO

Decision Tree

- What exactly is a decision tree?

How to build a decision tree?

1. Choose an attribute from your dataset.
2. Calculate the significance of attribute in splitting of data.
3. Split data based on the value of the best attribute.
4. Go to step 1.



How to calculate the significance?

Information gain is the information that can increase the level of certainty after splitting.

Information Gain = (Entropy before split) – (weighted entropy after split)

In every step, we want to dismiss the impurity of the nodes. Impurity is calculated by Entropy of data in the node

$$E = - \sum_i p_i \log_2(p_i)$$

Entropy is the amount of information disorder. If the samples are completely homogeneous the entropy is zero.

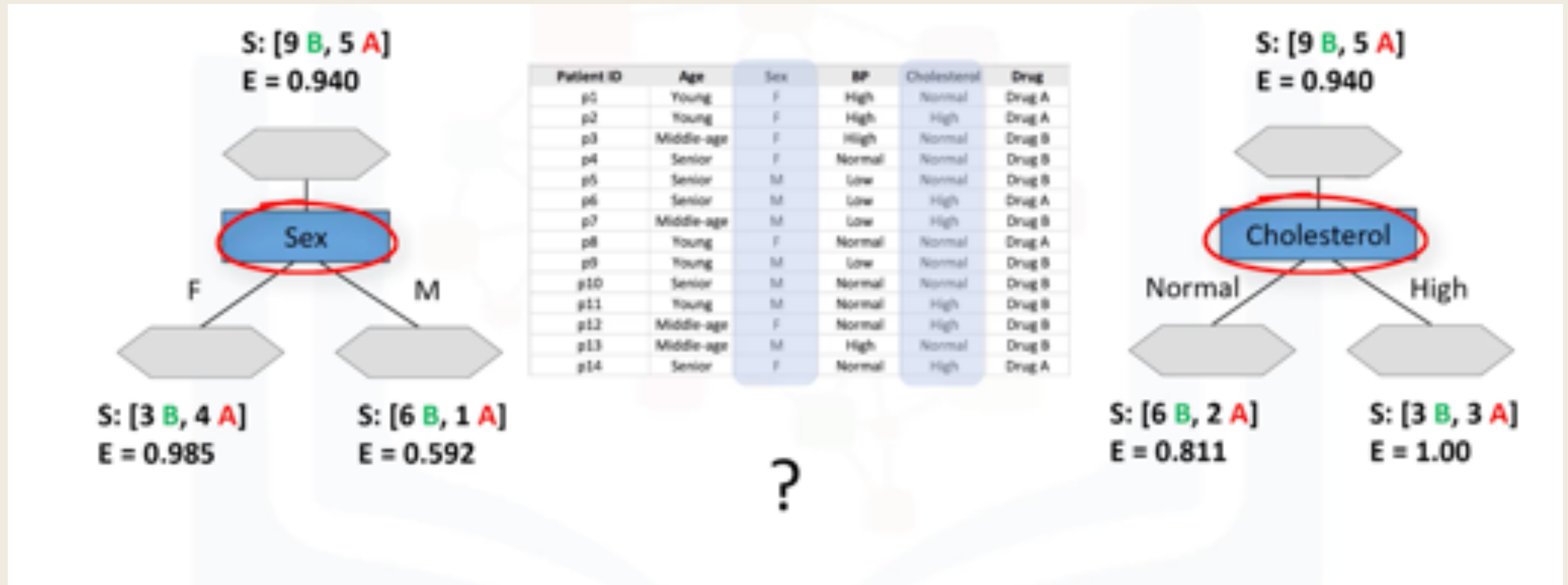
How to calculate the significance?

To compute Gini impurity for a set of items with J classes, suppose $i \in \{1, 2, \dots, J\}$, and let p_i be the fraction of items labeled with class i in the set.

$$I_G(p) = \sum_{i=1}^J \left(p_i \sum_{k \neq i} p_k \right) = \sum_{i=1}^J p_i (1 - p_i) = \sum_{i=1}^J (p_i - p_i^2) = \sum_{i=1}^J p_i - \sum_{i=1}^J p_i^2 = 1 - \sum_{i=1}^J p_i^2$$

This is the other way to measure the significance

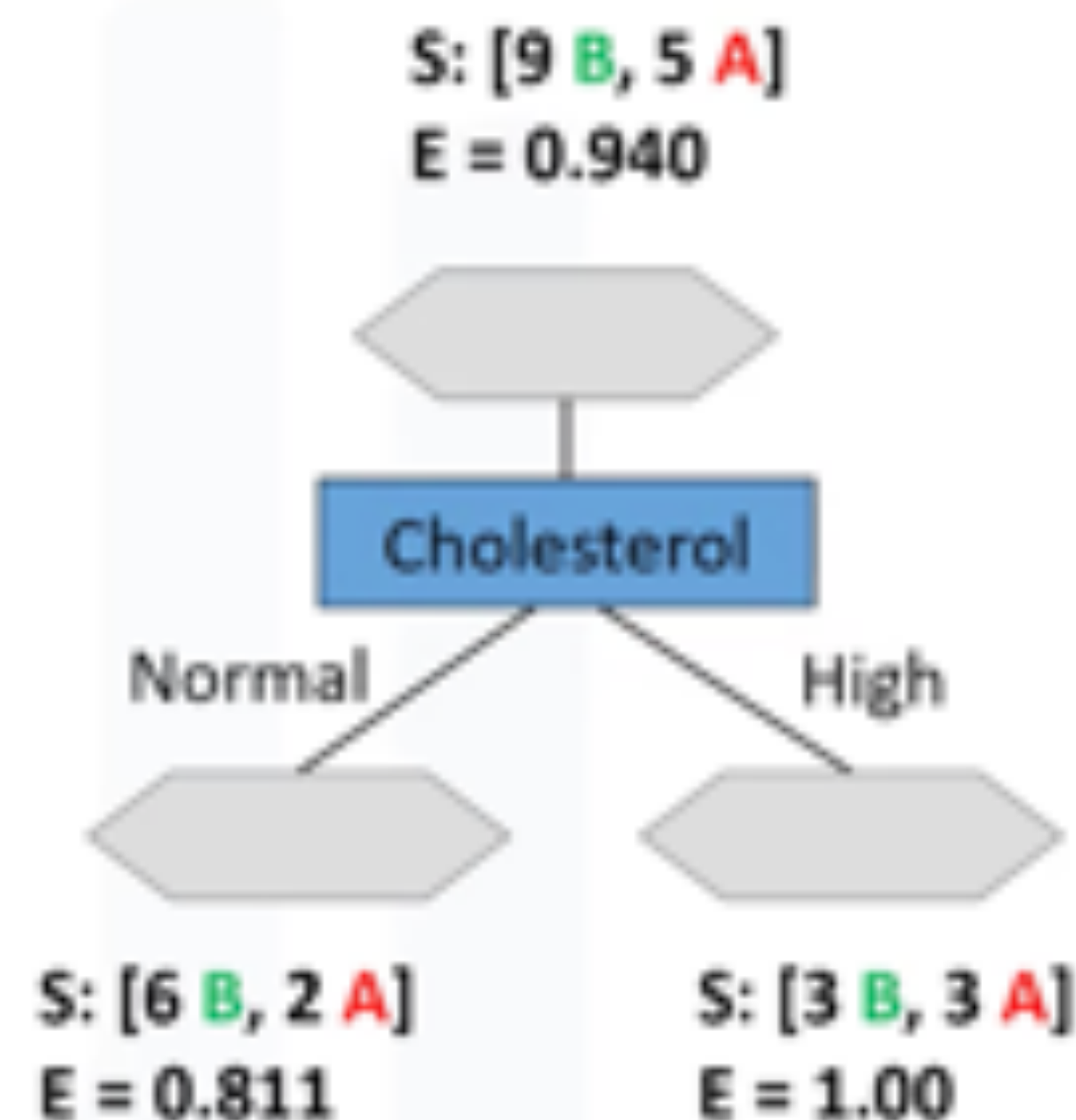
Which attributes is the best?



$$E = - \sum_i p_i \log(p_i)$$



Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A



Gain (s, Sex)
 $= 0.940 - [(7/14)0.985 + (7/14)0.592]$
 $= 0.151$

Gain (s, Cholesterol)
 $= 0.940 - [(8/14)0.811 + (6/14)1.0]$
 $= 0.048$

?

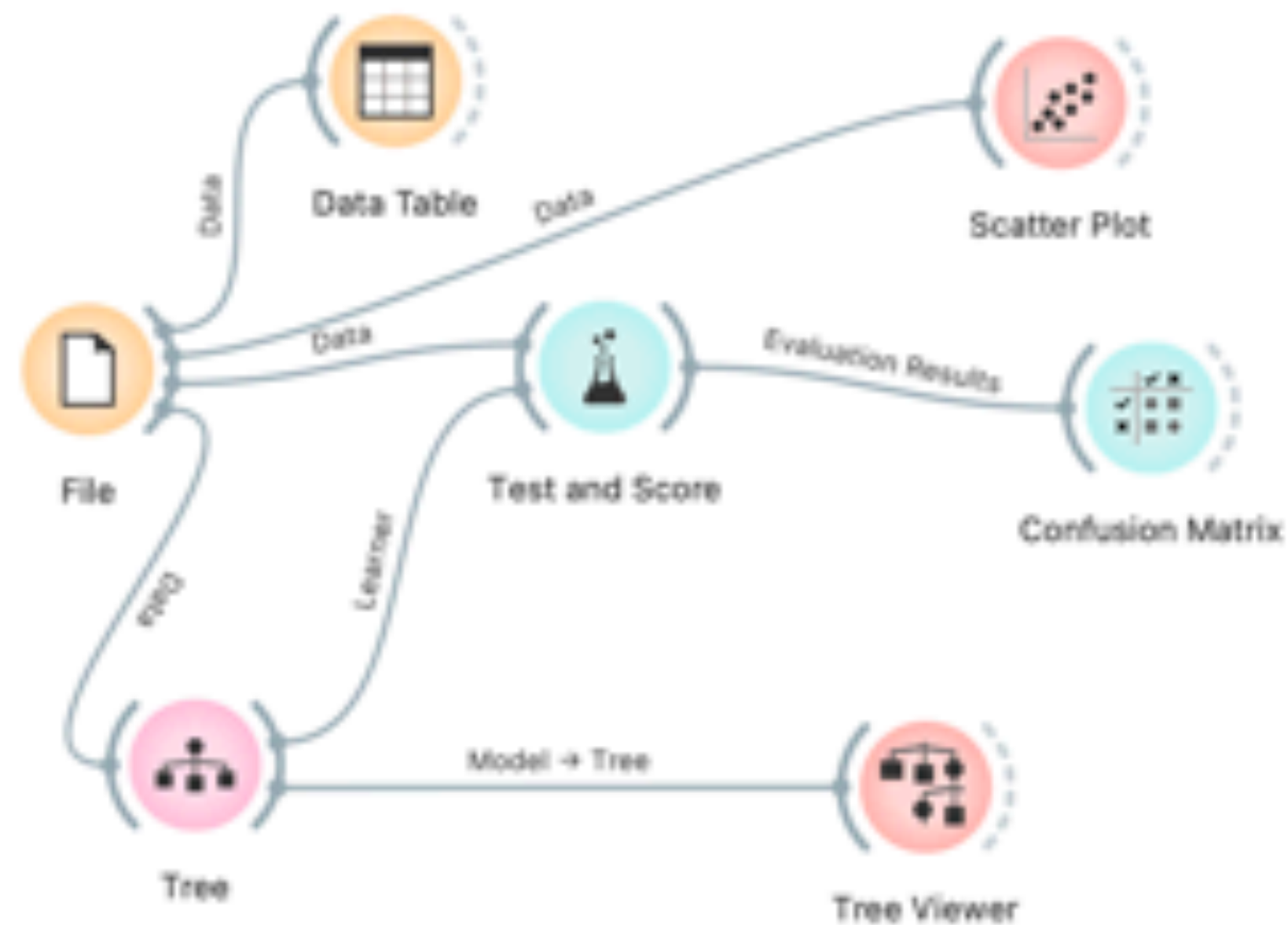
Information gain is the information that can increase the level of certainty after splitting.

Information Gain = (Entropy before split) – (weighted entropy after split)

ECOFLOR

Decision Tree

- Iris dataset
- Test and Score → confusion Matrix
- Tree → Tree Viewer



Linear Models

For classification

- Linear models are a class of models that is widely used in practice.
- Linear models make a prediction using a linear function of the input features

Linear Binary classification

In this case, the prediction is made using the formula:

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + \dots + w[p] * x[p] + b > 0$$

Where:

X [0,..p] are the features (columns)

W[0,..,p] and b are the parameters of the model

Y is the prediction of the model

Linear Models

For classification

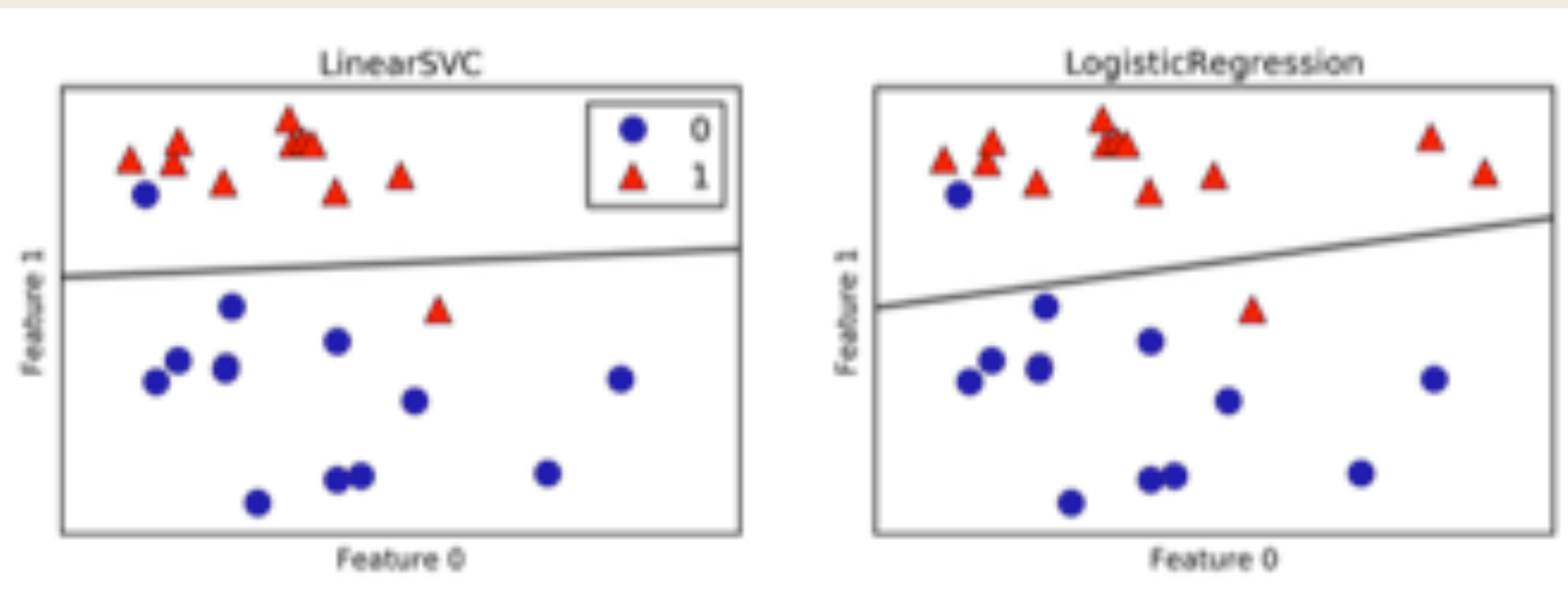
A linear classifier is a classifier that separates two classes using a line, a plane or a hyperplane

The most common models:

- Logistic Regression
- Support Vector Machine

Linear models

- Linear SVC
- Logistic Regression.



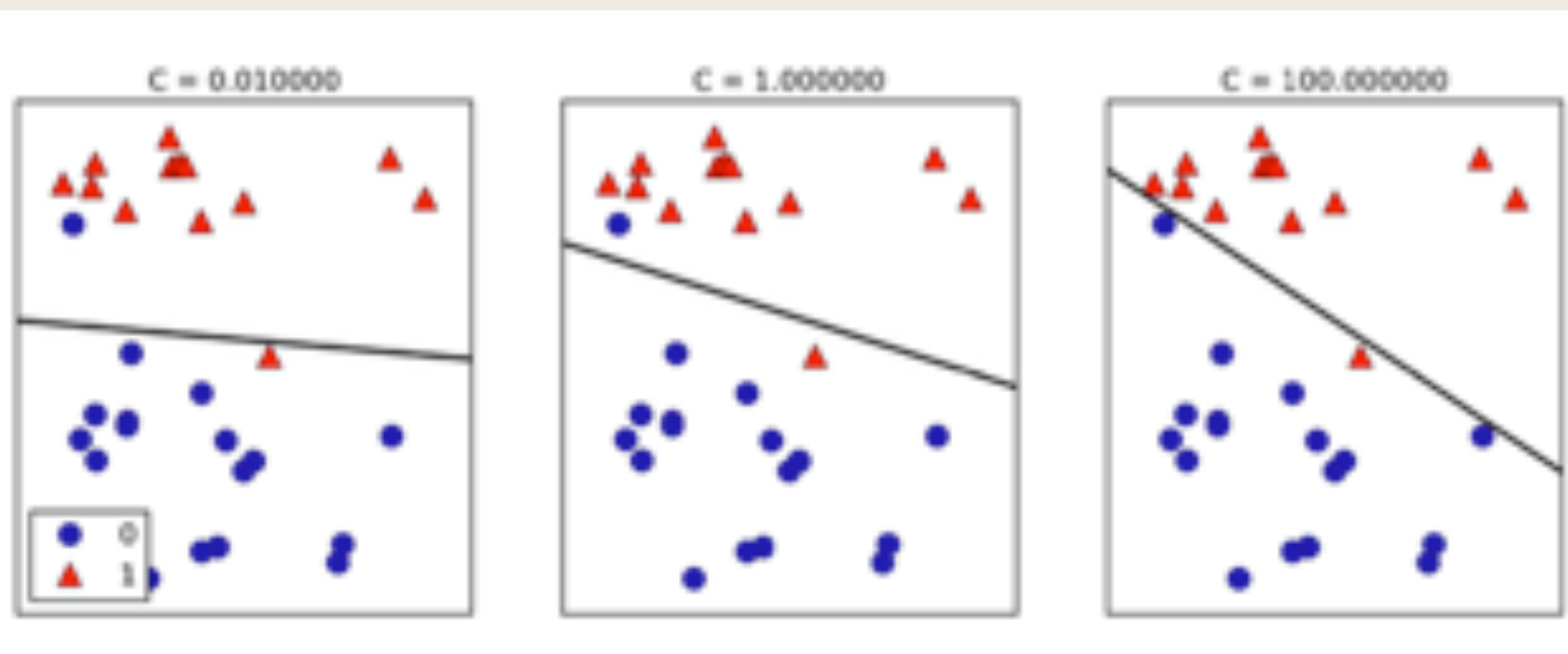


ECOFLO

Linear Models

- Iris dataset
- SVM -> Test and Score -> Confusion M.
- LR -> Test and Score -> Confusion M.
- C values
- Regularization type: L1, L2

Linear models



- Studying different C values
- Studying different regularizations.



ECOFLO

Naive Bayes models

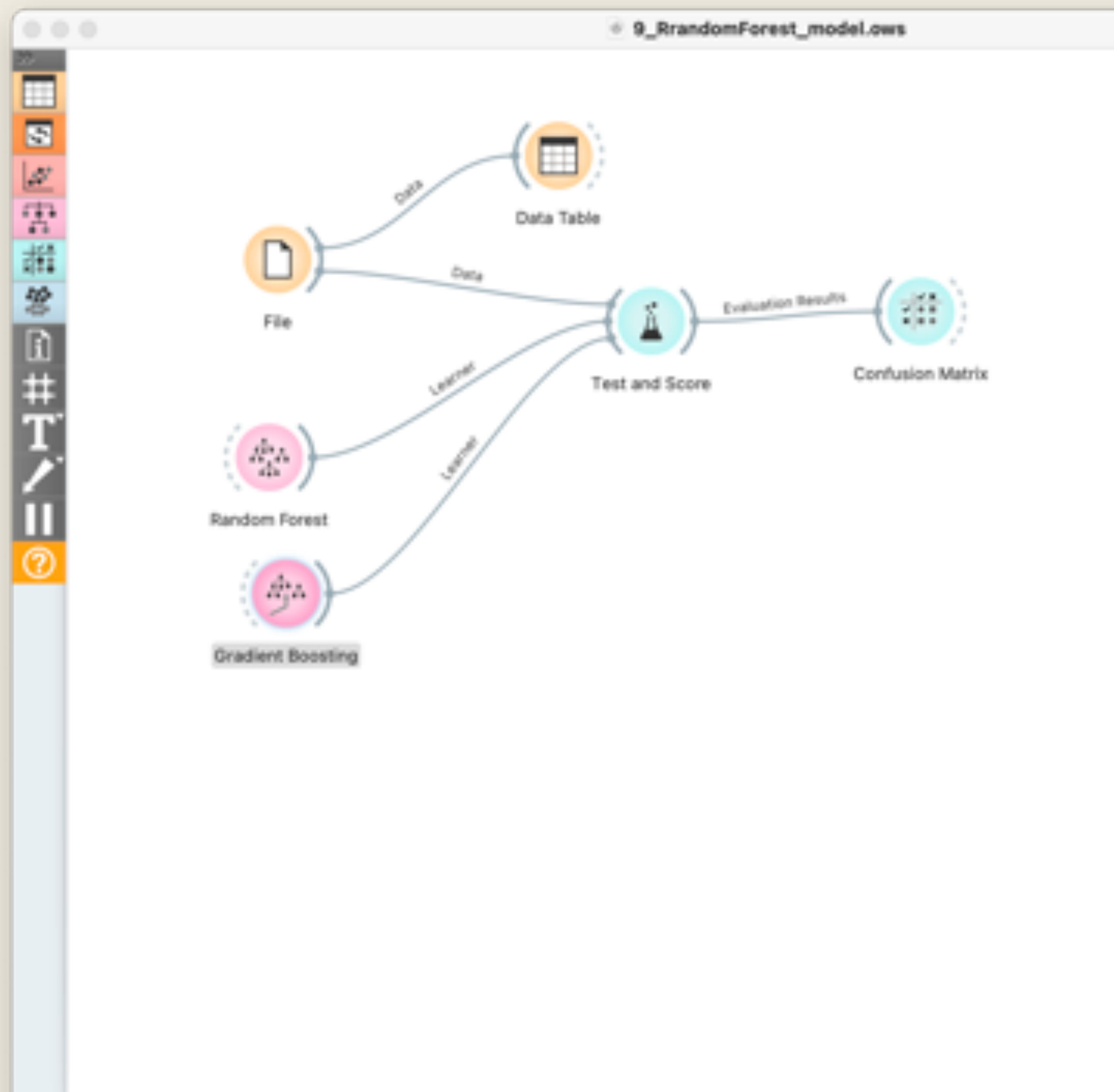
- Naive Bayes classifiers are a family of classifiers that are quite similar to the linear models.
- However, they tend to be even faster in training.
- The price paid for this efficiency is that naive Bayes models often provide generalization performance that is slightly worse than that of linear classifiers



ECOFLOP

Naive Bayes models

- There are three kinds of naive Bayes classifiers: **GaussianNB**, **BernoulliNB**, and **MultinomialNB**.
- GaussianNB can be applied to any continuous data, while BernoulliNB assumes binary data and MultinomialNB assumes count data
- Naive Bayes models are great baseline models and are often used on very large datasets, where training even a linear model might take too long.



ECOFLO

Ensembles of Decision Trees

- Ensembles are methods that combine multiple machine learning models to create more powerful models.
- There are two ensemble models that have proven to be effective, both of which use decision trees as their building blocks: **random forests** and **gradient boosted decision trees**.

Random Forest

- The main drawback of decision trees is that they **tend to overfit** the training data. Random forests are one way to address this problem.
 - A RF is essentially **a collection of decision trees**, where each tree is slightly different from the others. The idea behind random forests is that each tree might do a relatively good job of predicting, but will likely overfit on part of the data. If we build many trees, all of which work well and overfit in different ways, **we can reduce the amount of overfitting by averaging their results.**
-

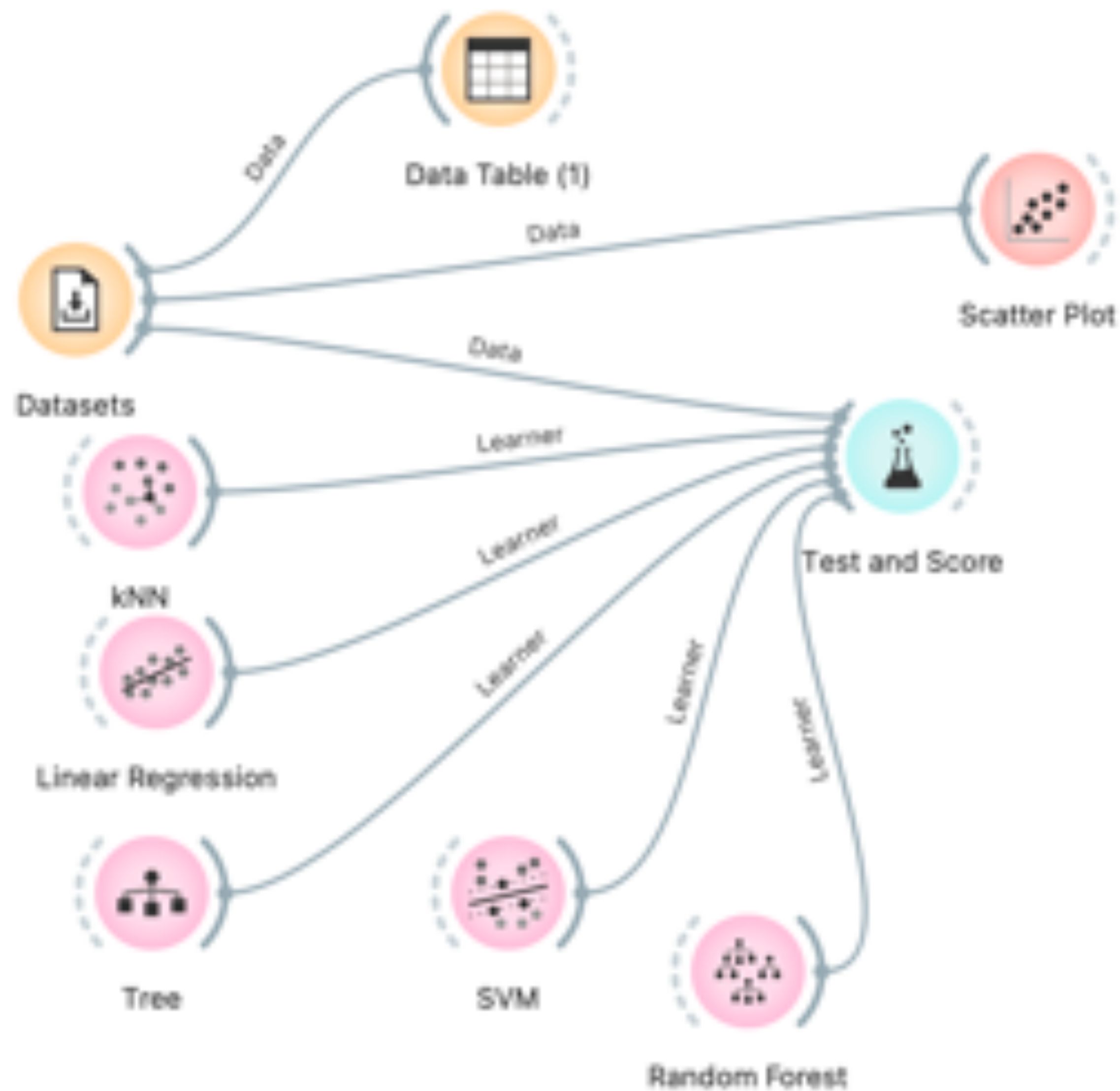
Regression problems in ML

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

ECOFLOr

Regression

In Regression, the goal is to predict a *continuos value*



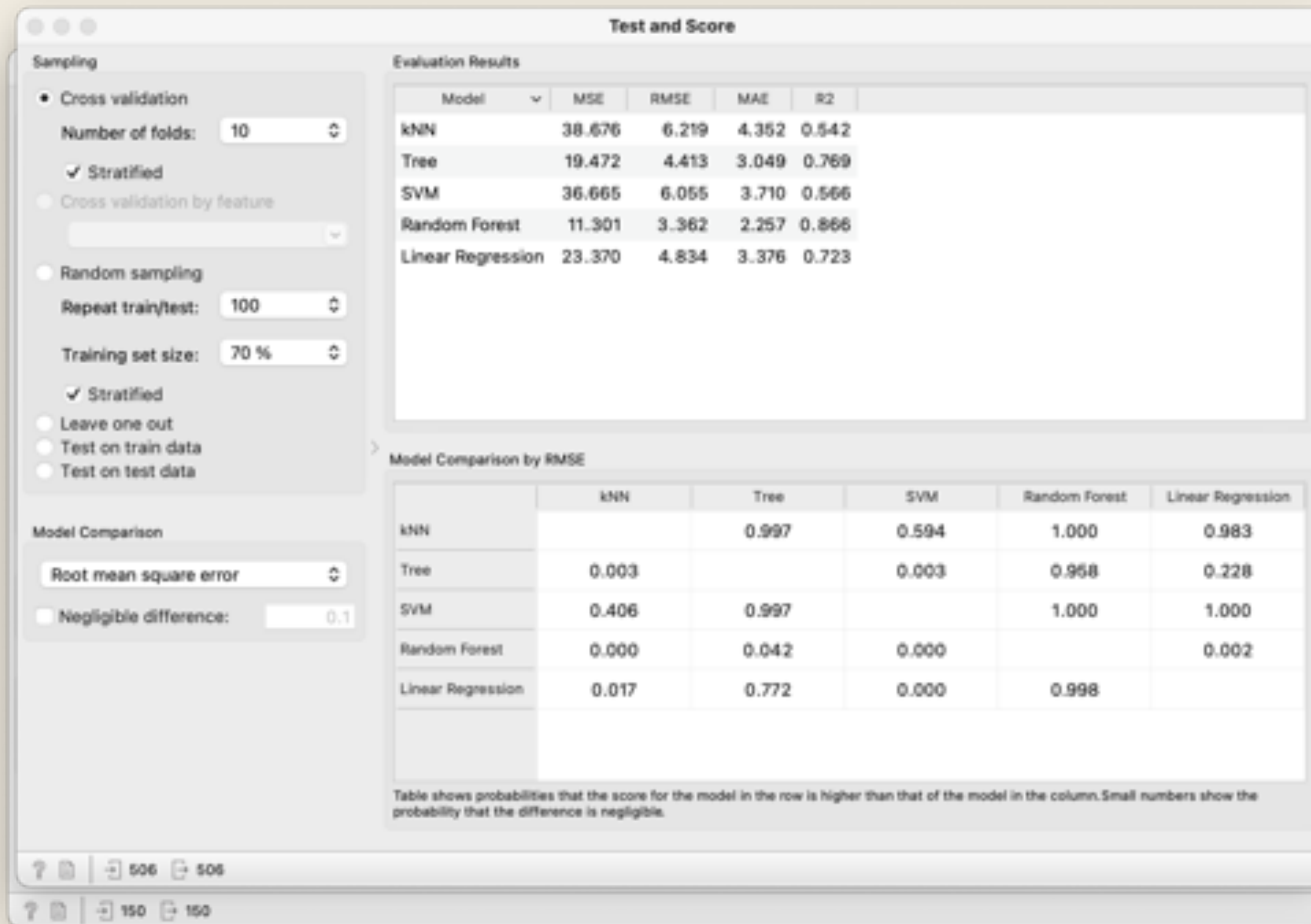
ECOFLO

Scoring

- Abalone data (numeric Target)
- Test & Score

Evaluation metrics in classification

- We are talking about some metrics:
 - MSE: Mean Squared Error
 - RMSE: Root Mean Squared Error
 - MAE: Mean Absolute Error
 - R^2



Error coefficients

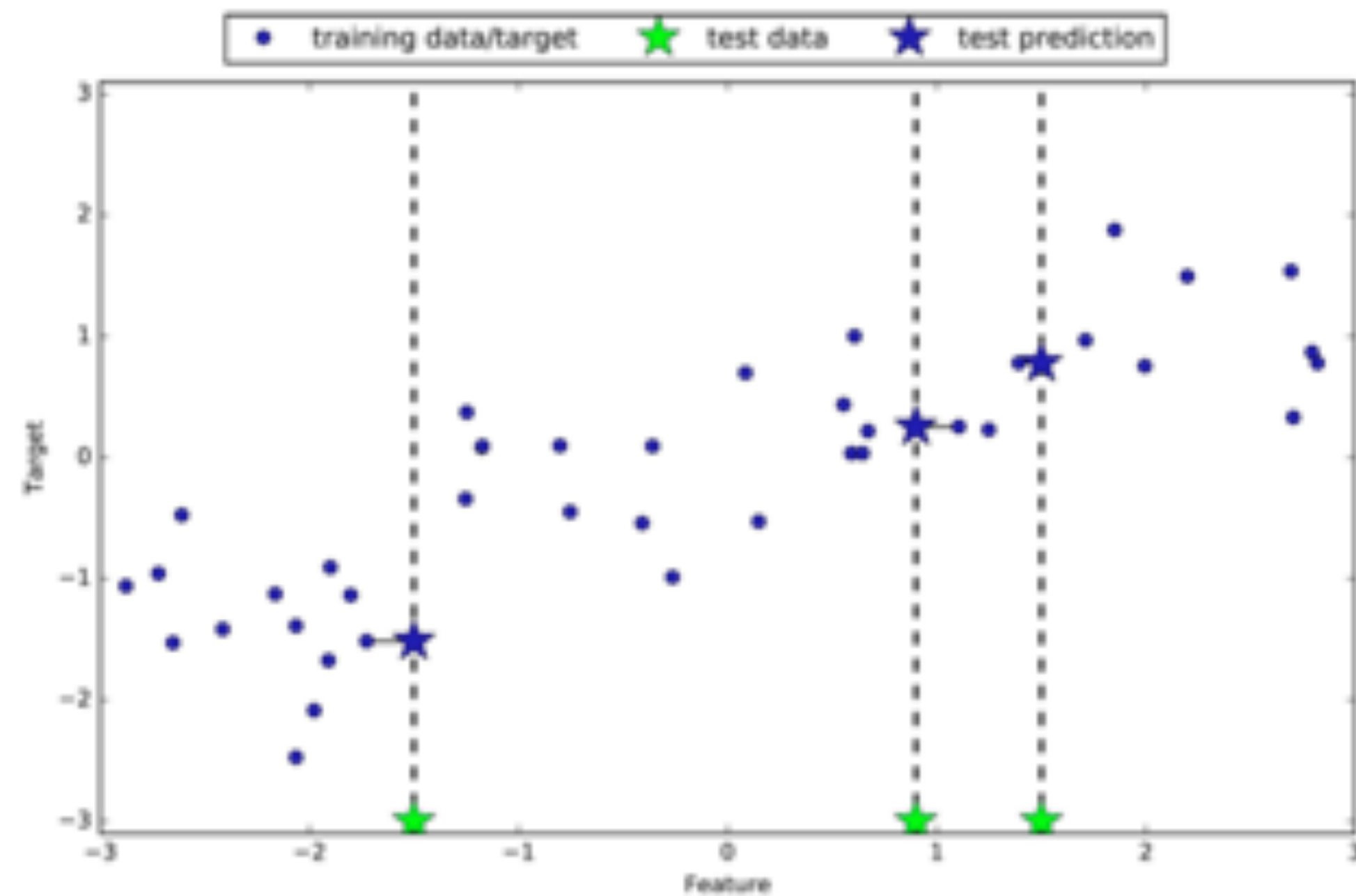
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{Y}_i \right)^2$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{Y}_i \right)^2}$$

$$\text{MAE} = \frac{\sum_{i=1}^n \left| y_i - x_i \right|}{n} = \frac{\sum_{i=1}^n \left| e_i \right|}{n}$$

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

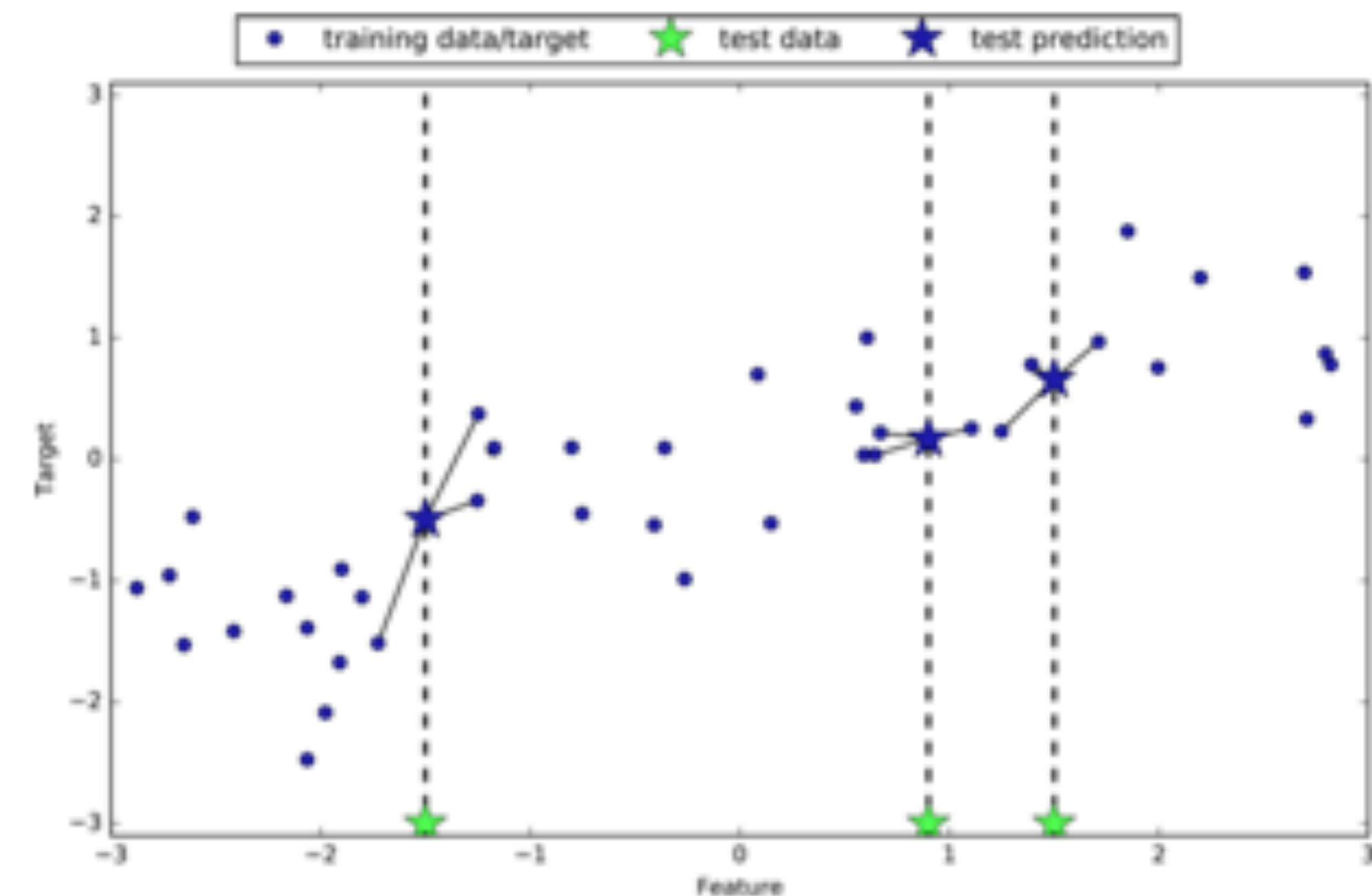
Regression Algorithms



16/03/2022

K-Nearest Neighbors

- Building the k-NN algorithm consists only of storing the training dataset.
- To make a prediction for a new data point, the algorithm finds the closest data points in the training dataset—its “nearest neighbors.”



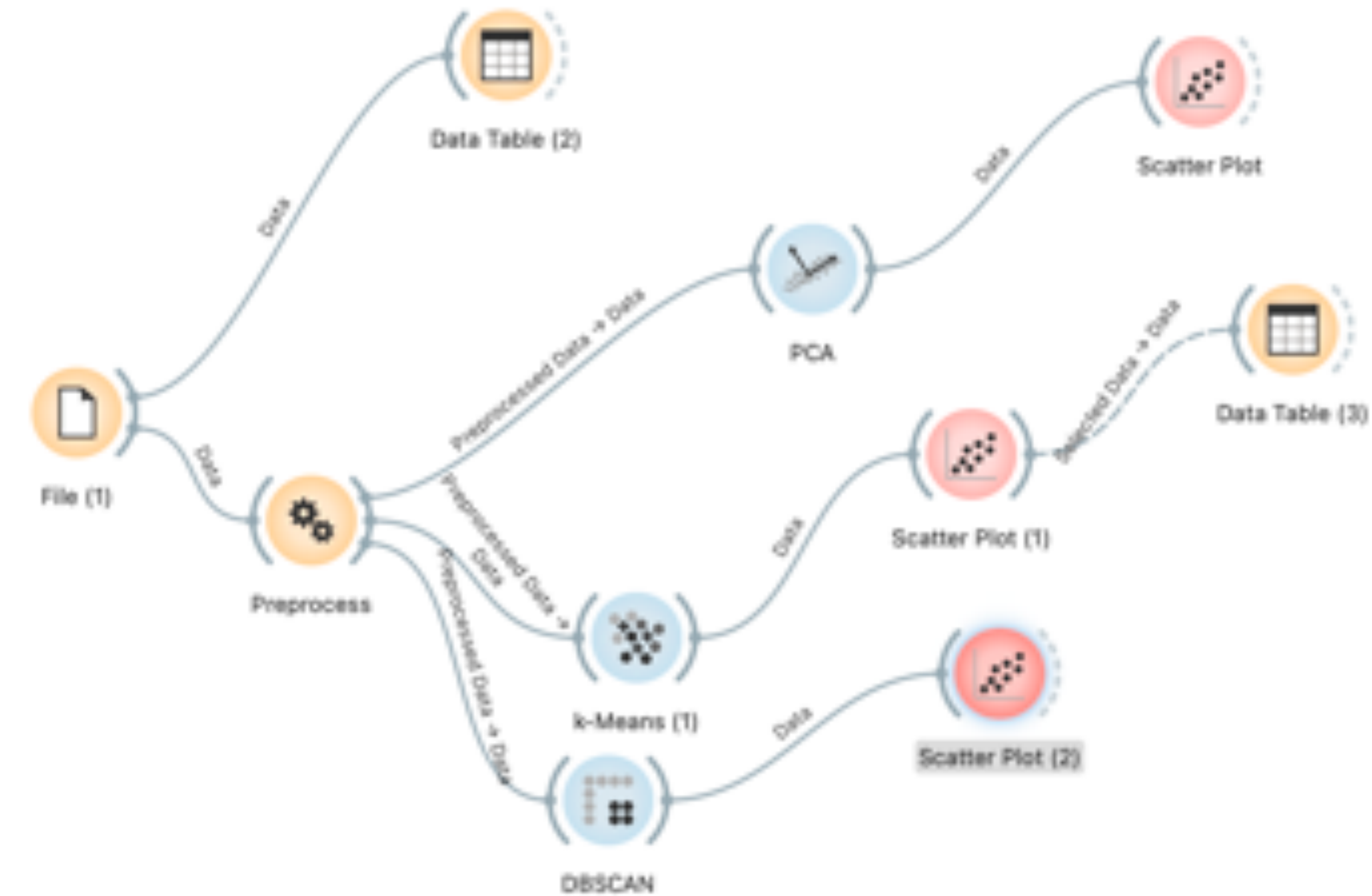
Unsupervised Learning

Challenges in Unsupervised Learning

A major challenge in unsupervised learning is evaluating whether the algorithm learned something useful. Unsupervised learning algorithms are usually applied to data that does not contain any label information, so we don't know what the right output should be. Therefore, **it is very hard to say whether a model "did well."**

Type of unsupervised learning

- **Unsupervised transformations** of a dataset are algorithms that create a new representation of the data which might be easier for humans or other machine learning algorithms to understand compared to the original representation of the data (dimensionality reduction)
- **Clustering**, on the other hand, partition data into distinct groups of similar items .



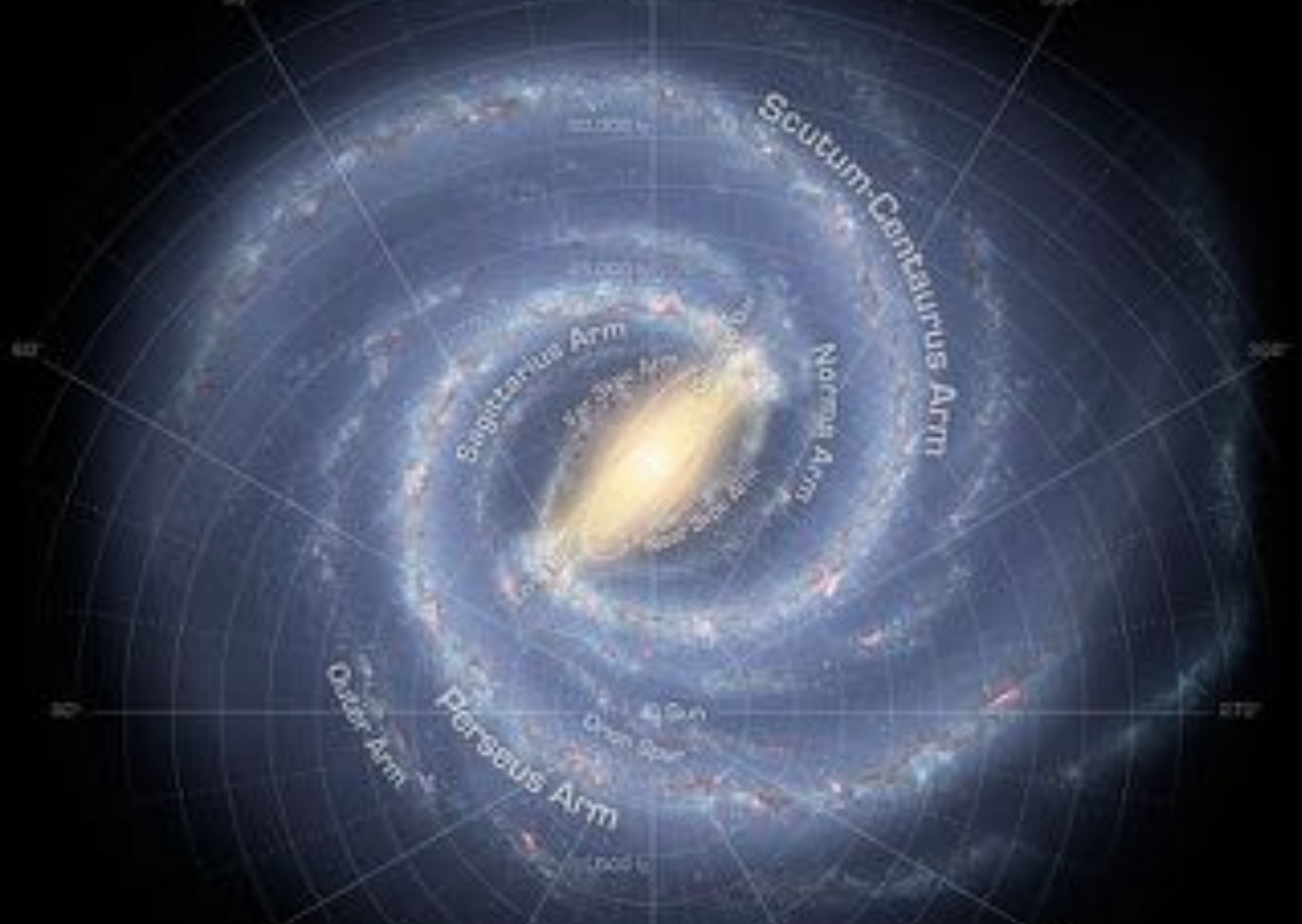
ECOFLOr

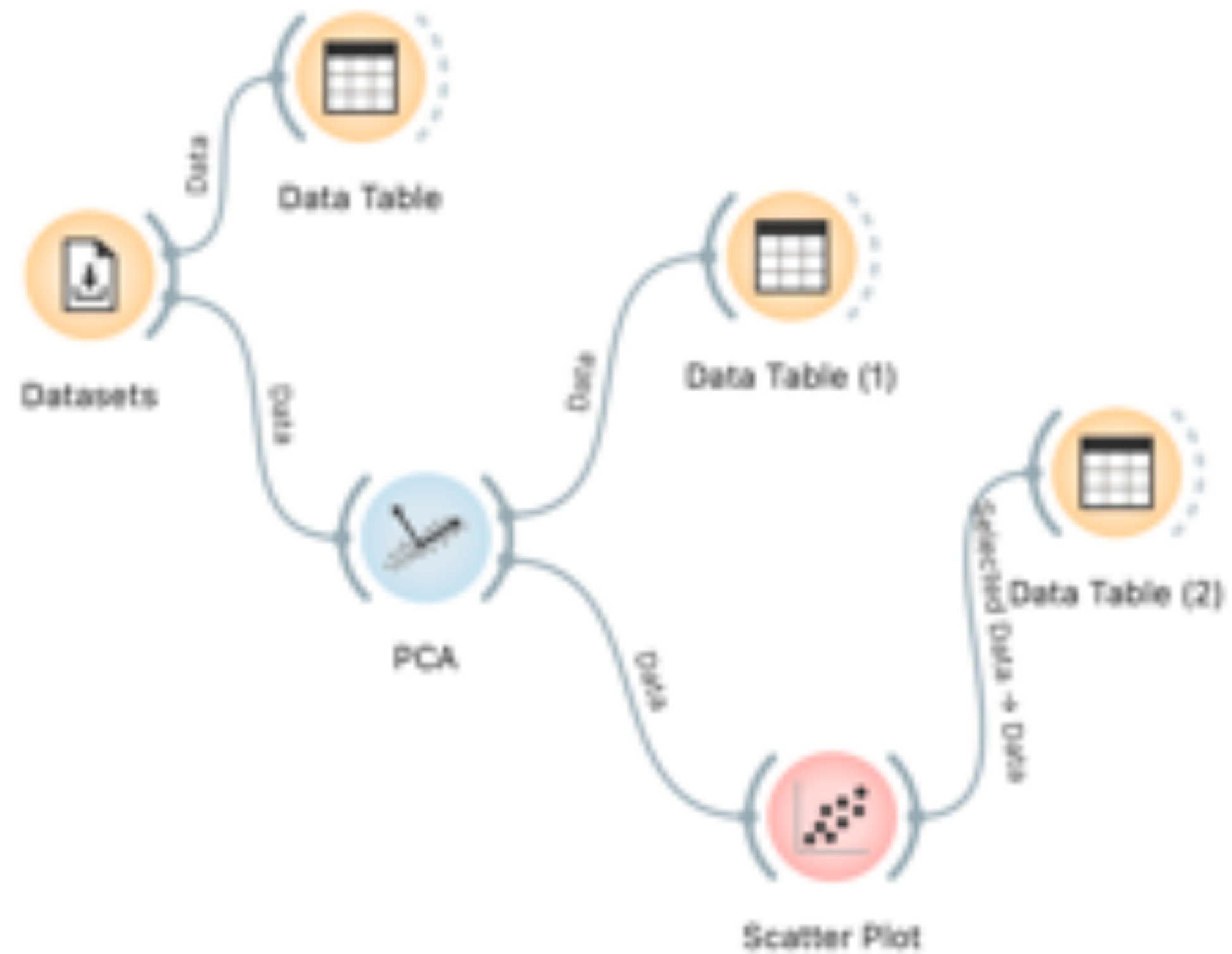
Preprocessing and Scaling

- StandardScaler
- RobustScaler
- MinMaxScaler

A long-exposure photograph of the night sky. The Milky Way galaxy is the central focus, appearing as a dense, glowing band of light and dust that stretches diagonally from the upper left towards the lower right. The dust lanes are visible as darker, reddish-brown regions within the brighter blue and white star fields. The sky is filled with countless individual stars of varying brightness. In the foreground, the dark, jagged silhouettes of mountain peaks and ridges are visible against the bright, glowing horizon of the galaxy. The overall color palette is dominated by deep blues, purples, and oranges, with the bright white and yellow of the stars providing contrast.

Let's talk about stars





ECOFLOr

PCA

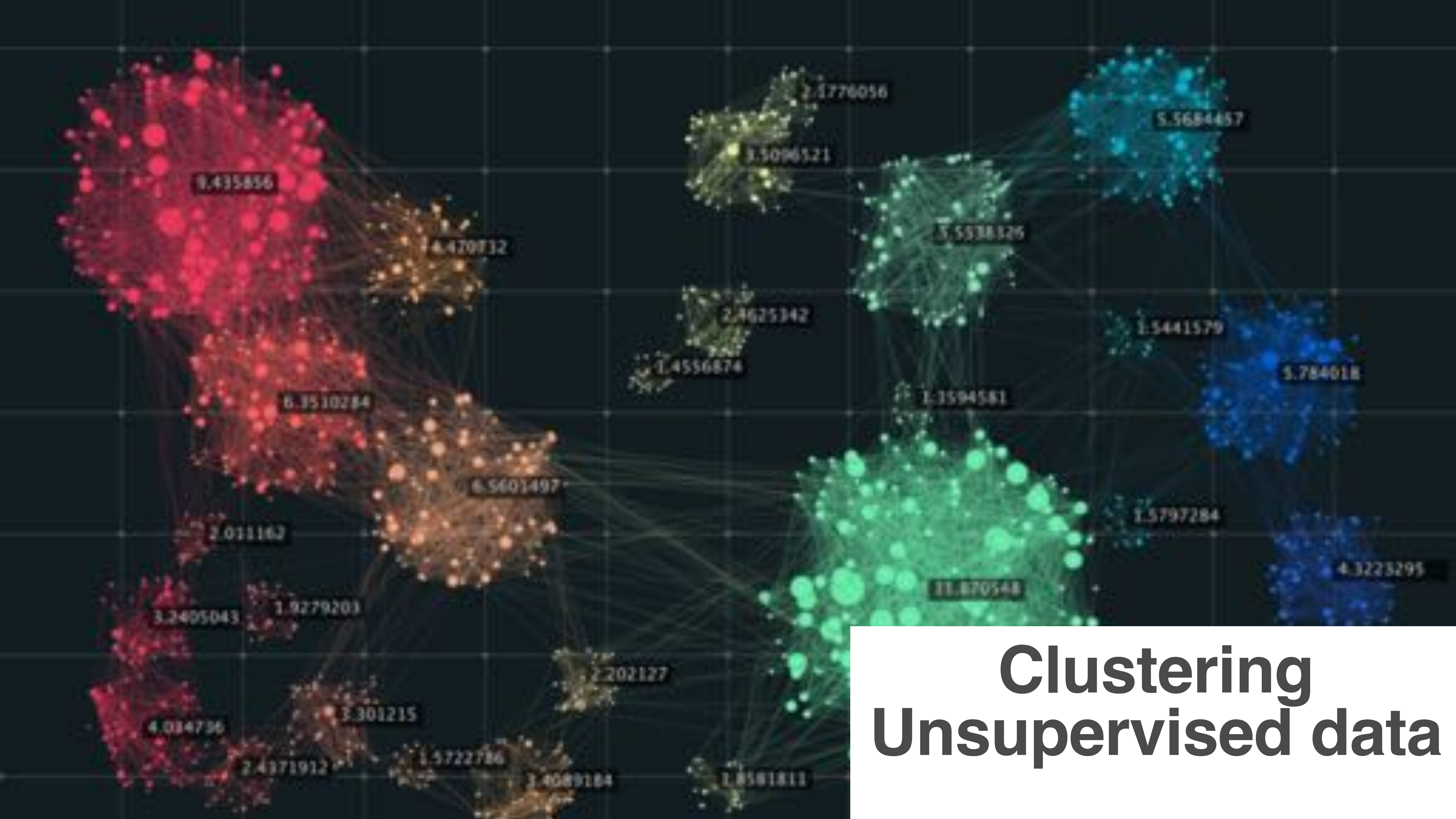
- Datasets->Zoo
- PCA
- Scatter Plot



ECOFLO

MDS and t-SNE

- Datasets->Zoo
- Distances
- MDS and t-SNE



**Clustering
Unsupervised data**



ECOFLOR

Hierarchical clustering

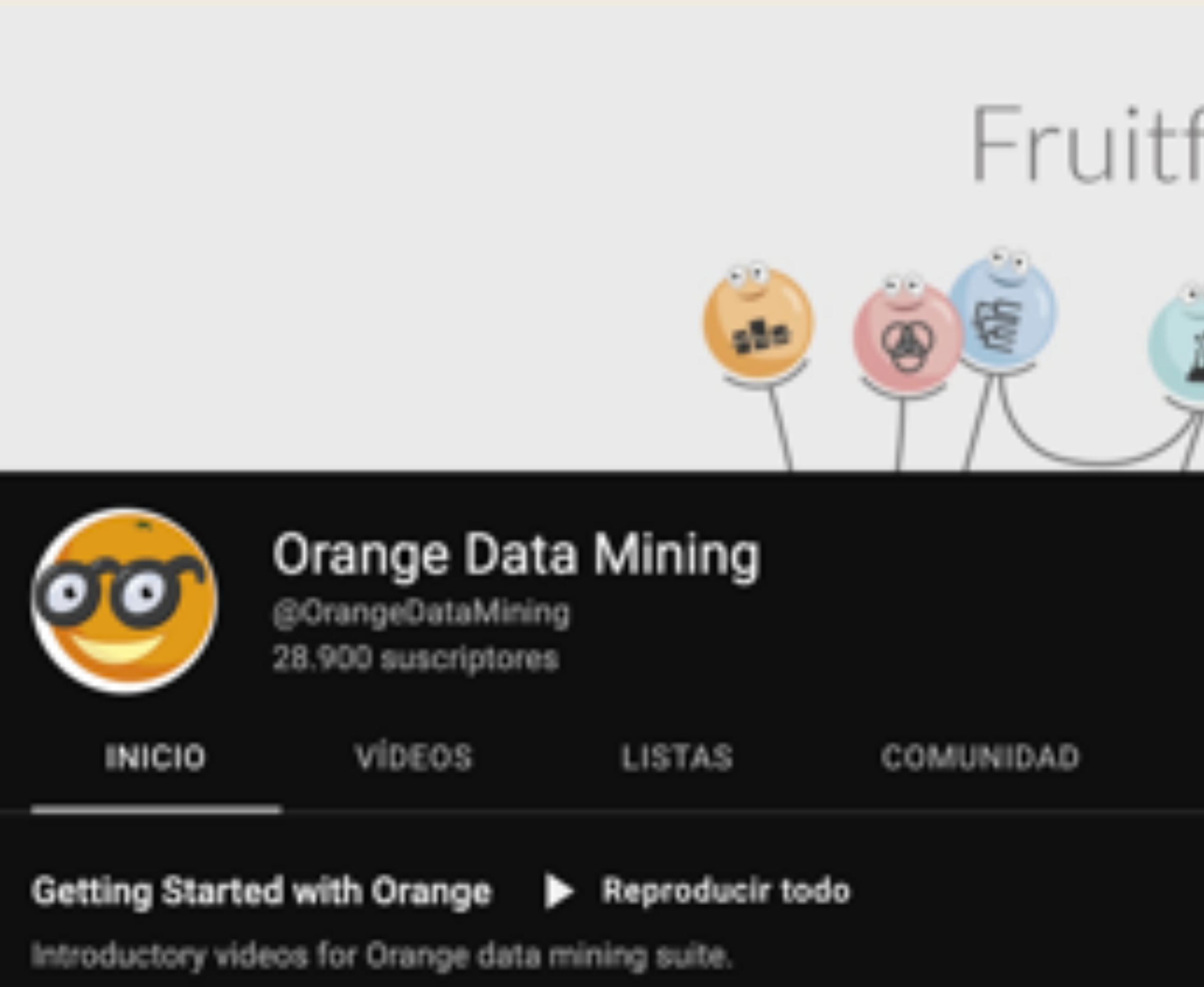
- Grades → Data Table
- Distances → hierarchical clustering
- Scatter plot
- Box plot



ECOFLO

K-means

- Paint Data
- K-means → Scatter plot
- Interactive K-Means
- 2 example:
- Iris Data



ECOFLO

Orange Data Mining

YouTube Chanel

@galeanojav
