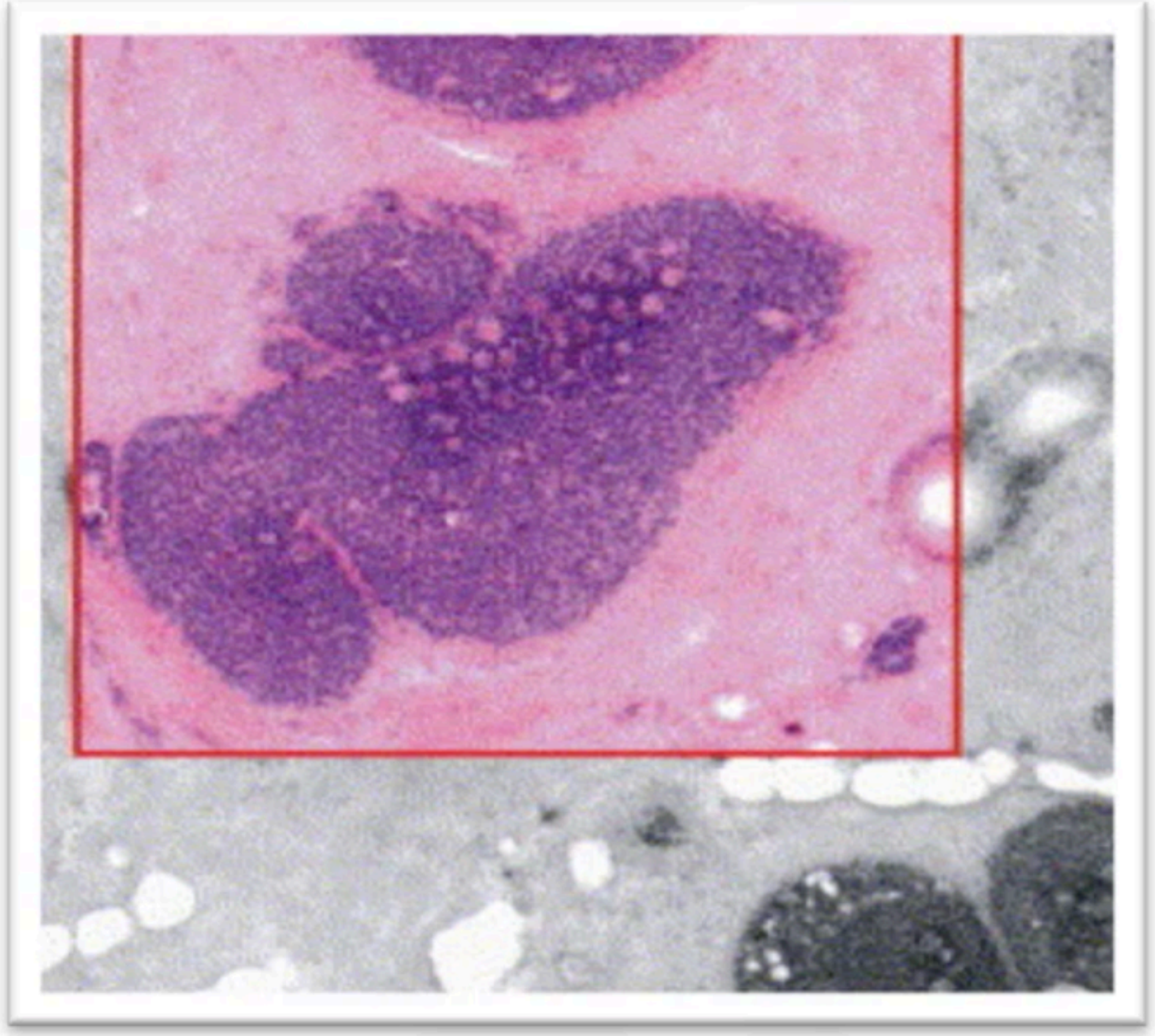
JAVIER GALEANO, MARZO DE 2022

Notas escépticas sobre el Machine Learning

usando Orange3 y otros programas

First example in Machine Learning

Skin Cancer



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	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	benign

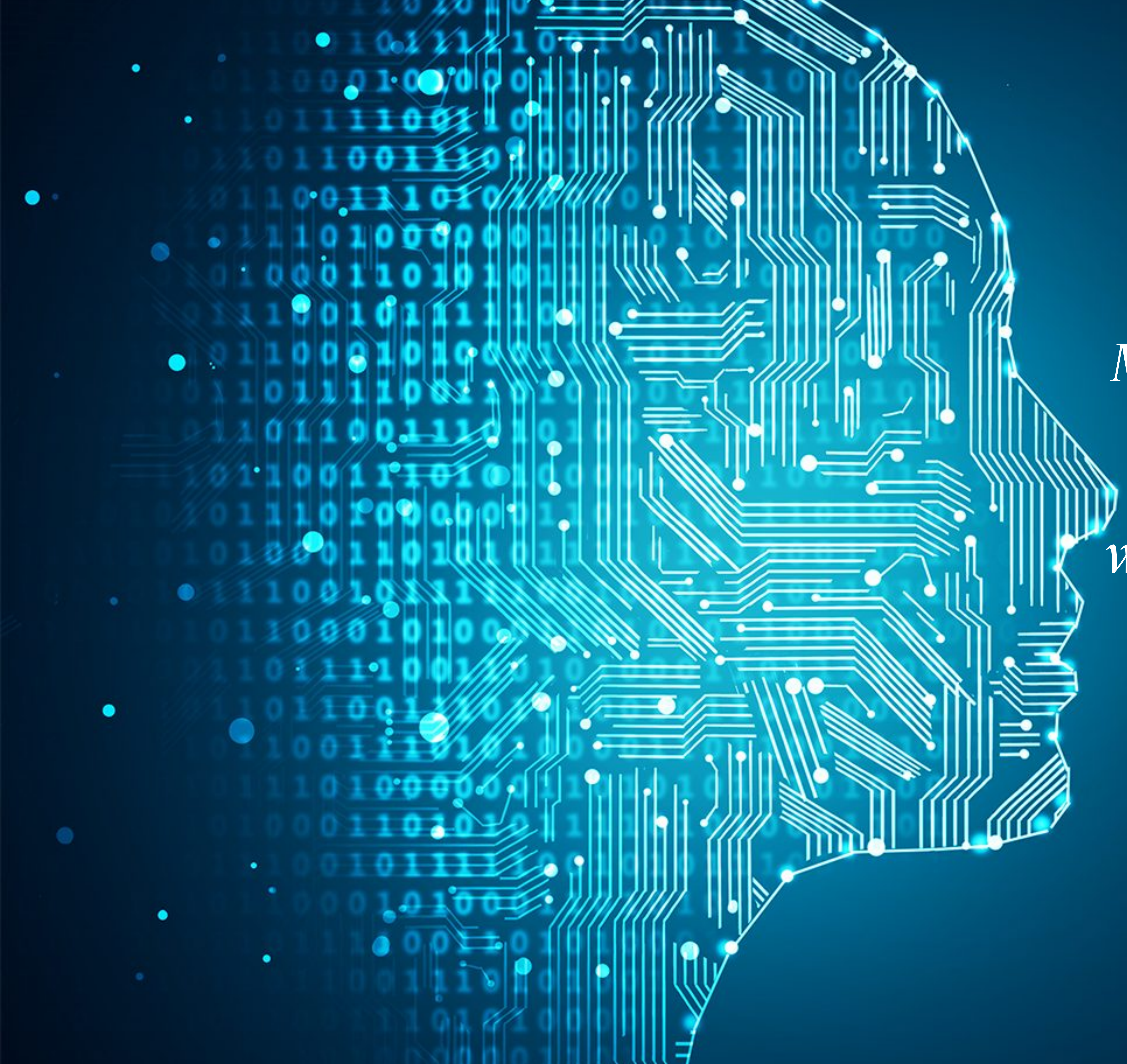
Modeling

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

Prediction

Accuracy = 89%

First definition of ML



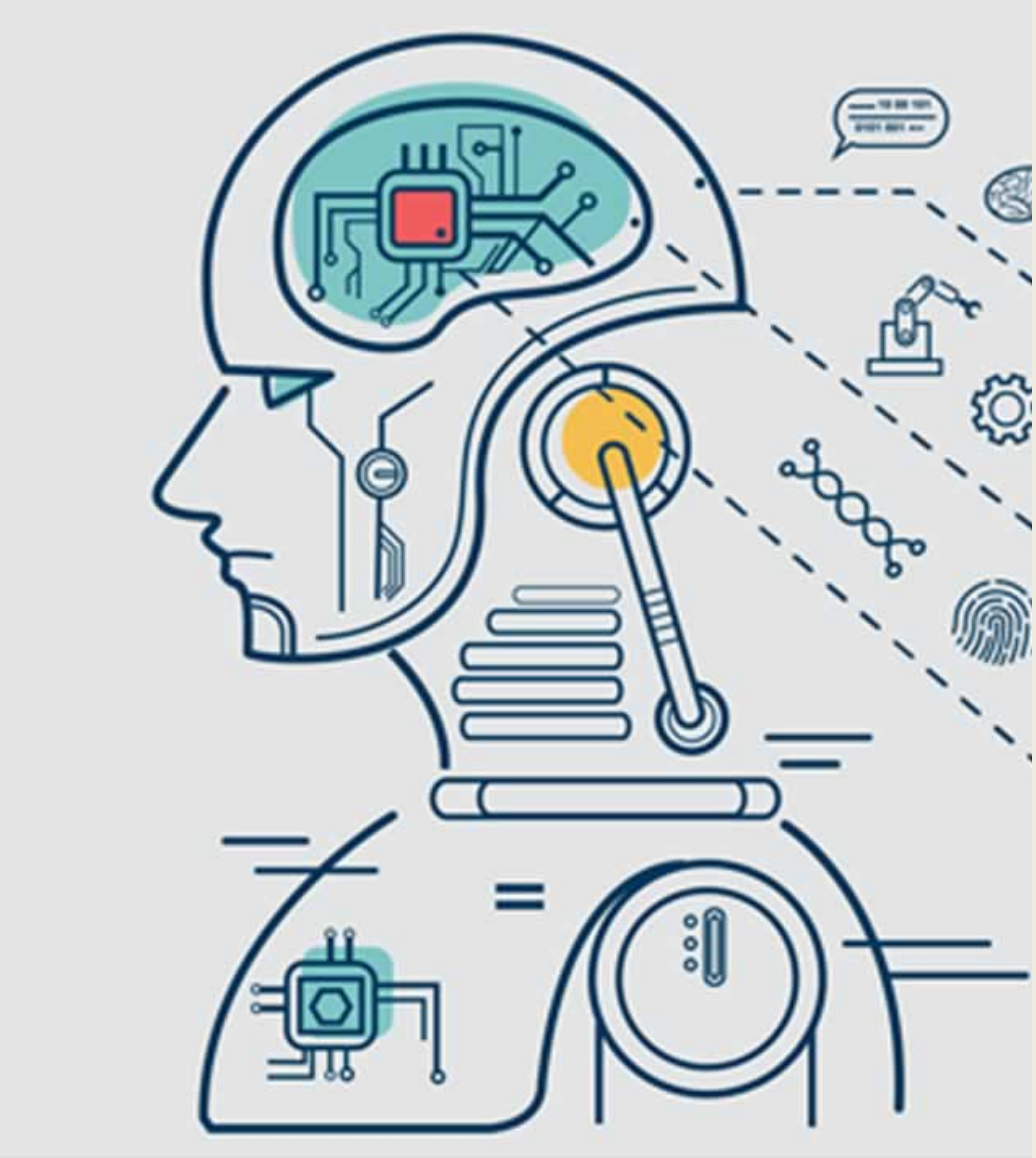
Machine Learning is the subfield of computer science that gives computers the ability to learn without being explicit programmed

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

Examples in our daily life.

Examples of ML in our daily life

Some ideas



Differences between

- Inteligencia Artificial
- Machine Learning
- Deep Learning

Artificial Inteligence

Artificial Intelligence attempts to make computers intelligent in ways that mimic the cognitive functions of humans.

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.


11001
DEEP
LEARNING
WWW

Classification in ML

Classical Machine Learning

Task Driven

Supervised Learning

(Pre Categorized Data)

Classification

(Divide the socks by Color)

Eg. Identity Fraud Detection

Regression

(Divide the Ties by Length)

Eg. Market Forecasting

Data Driven

Unsupervised Learning

(Unlabelled Data)

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	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

Supervised model

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

Features

Class o label

Observation

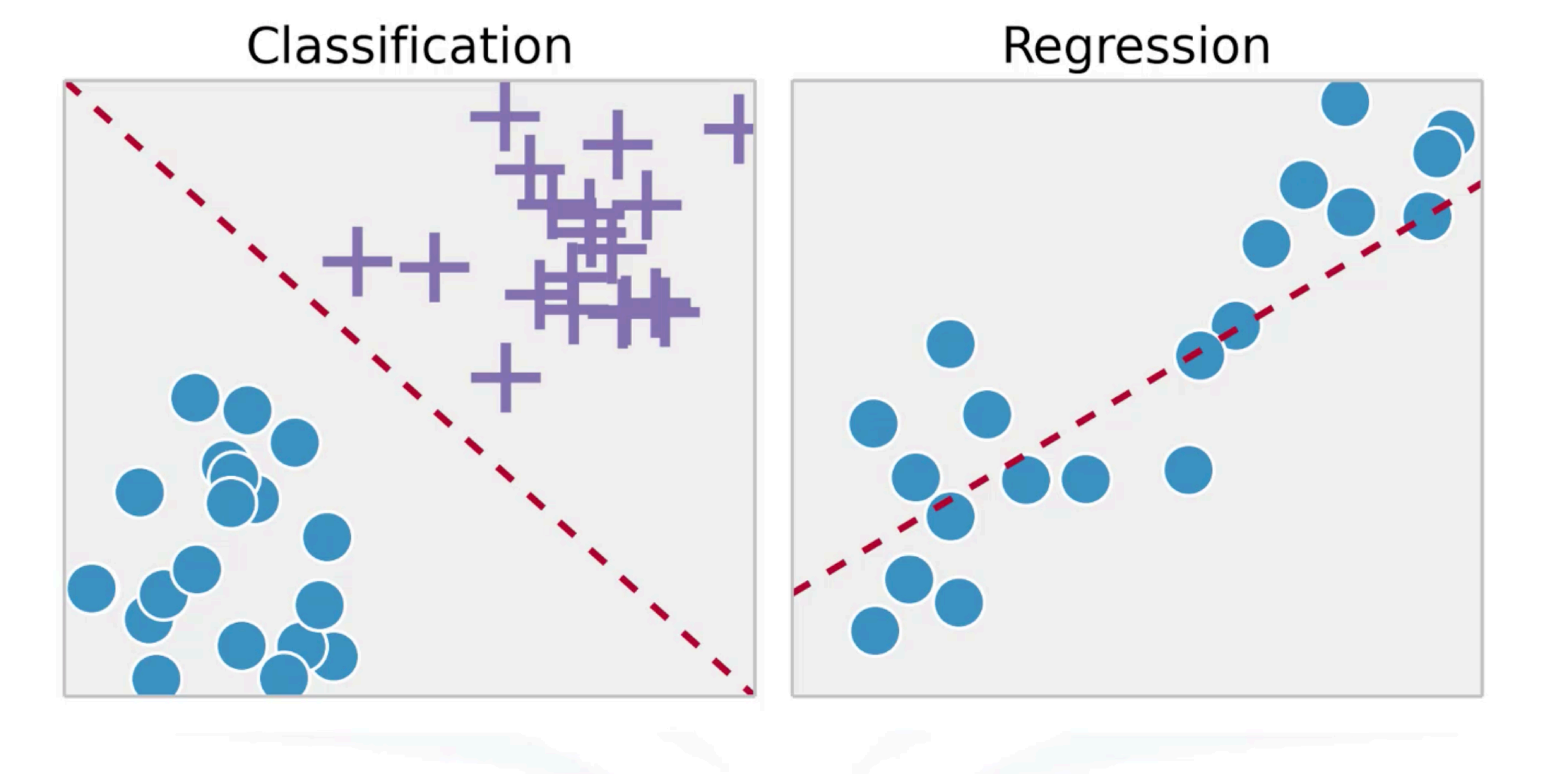
ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099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

Categorical

Numerical

Labeled dataset

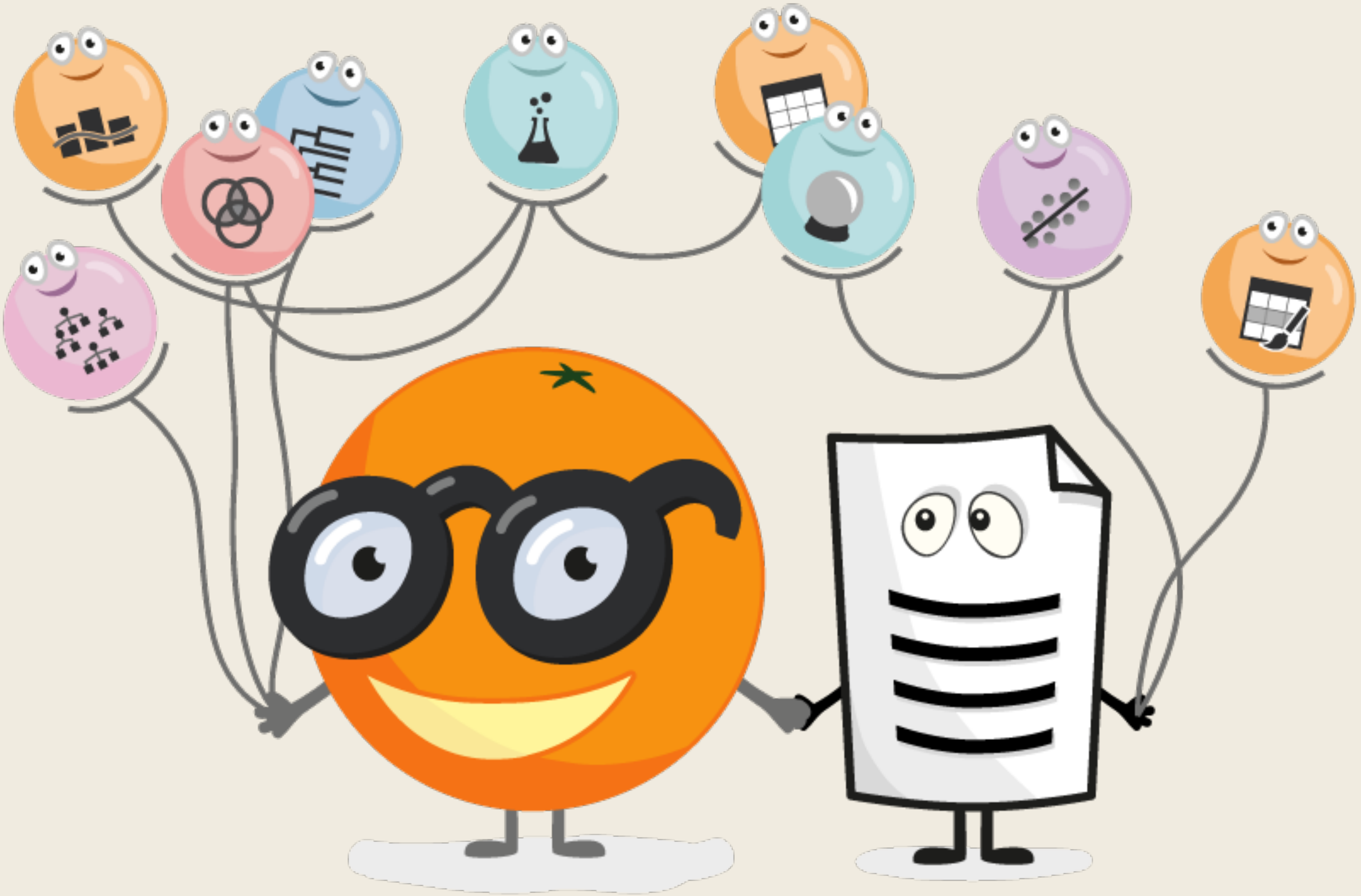
Types of supervised techniques



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

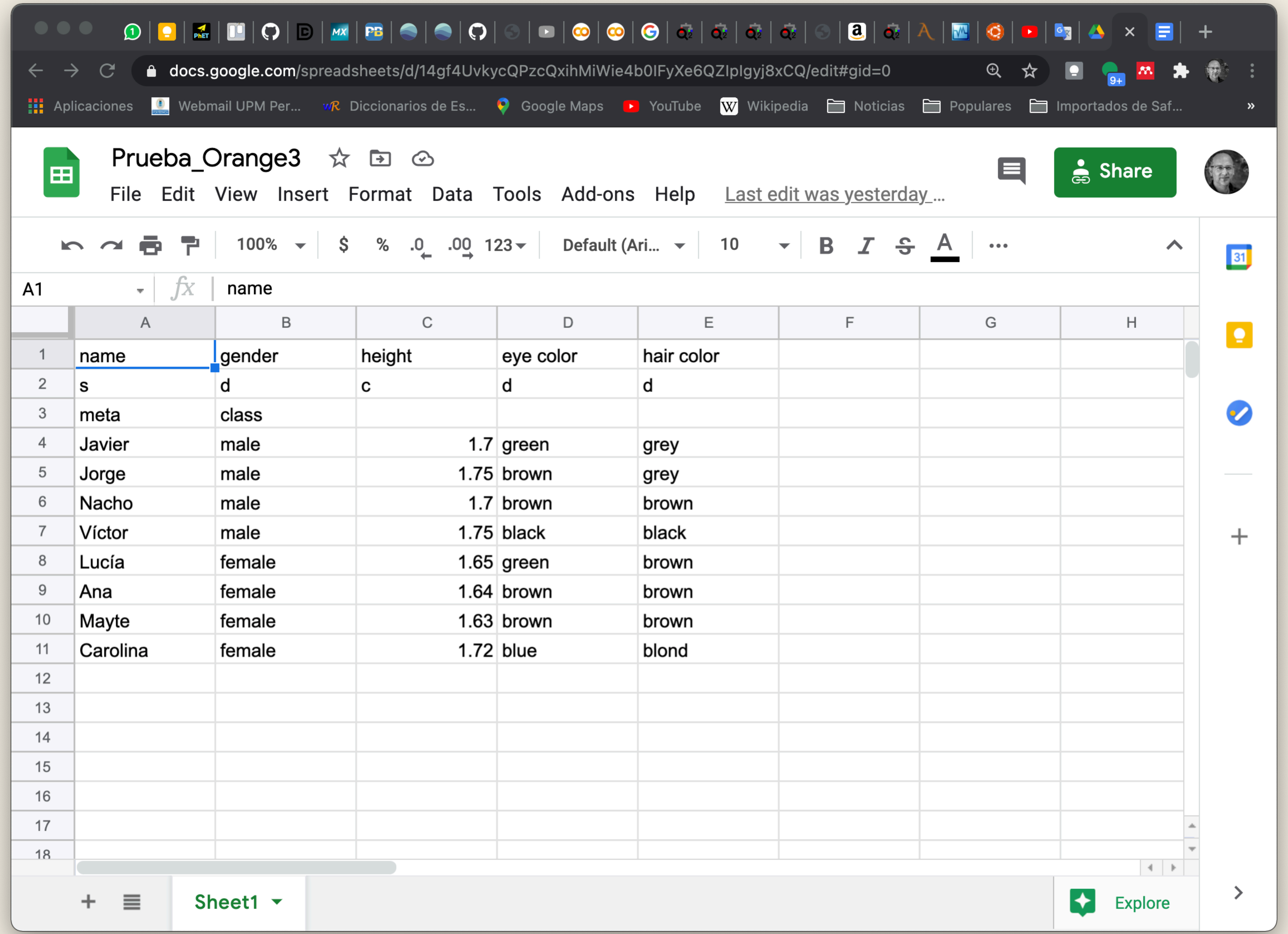
We do not supervise the model, but we let the model work on its own to discover information that may not be visible to the human eye.



9/MARZO/2022

Primer programa

- File –> Iris
- Data Table
- Scatter plot
- Selected data table

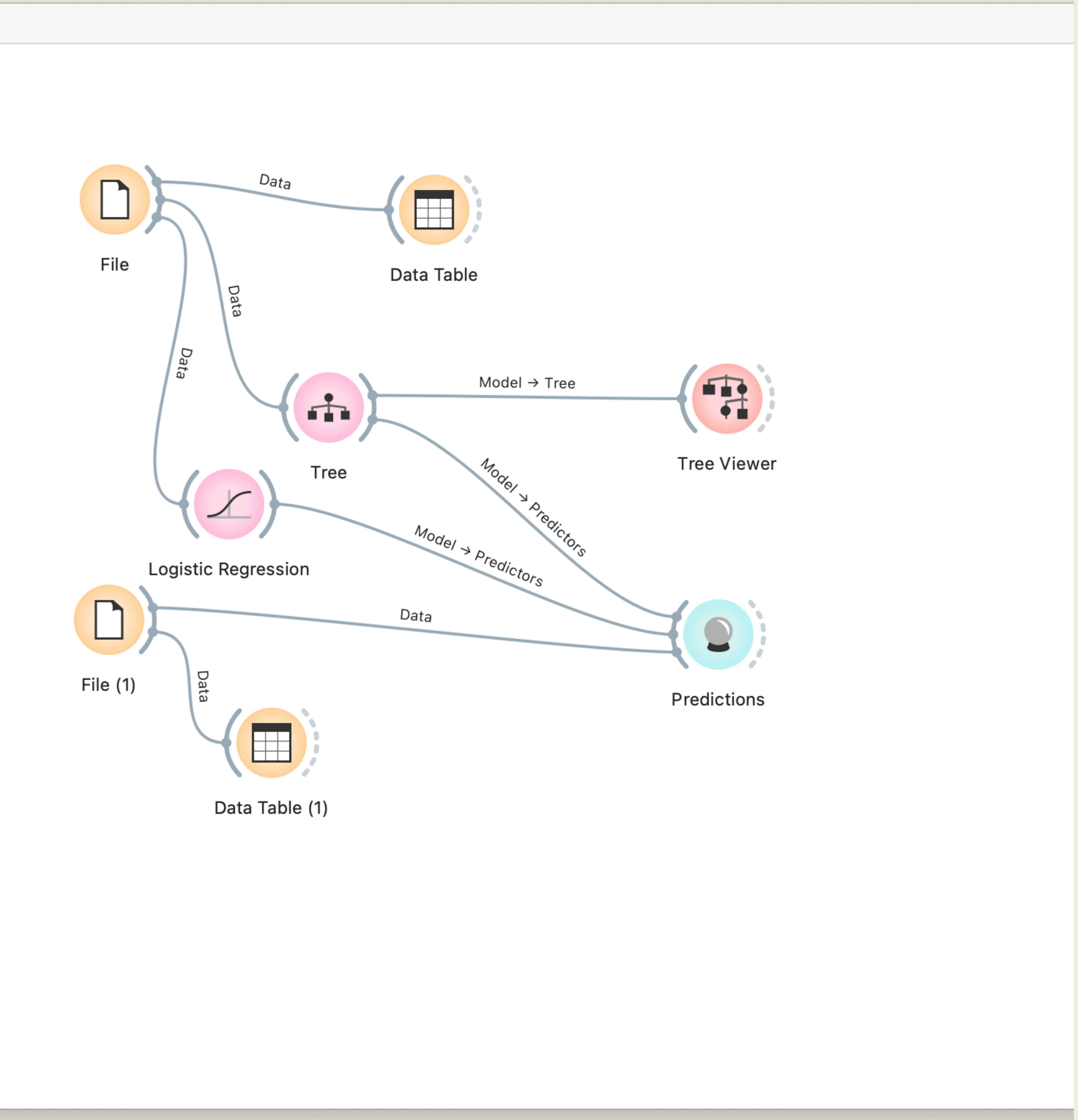


	A	B	C	D	E
1	name	gender	height	eye color	hair color
2	s	d	c	d	d
3	meta	class			
4	Javier	male	1.7	green	grey
5	Jorge	male	1.75	brown	grey
6	Nacho	male	1.7	brown	brown
7	Víctor	male	1.75	black	black
8	Lucía	female	1.65	green	brown
9	Ana	female	1.64	brown	brown
10	Mayte	female	1.63	brown	brown
11	Carolina	female	1.72	blue	blond

9/MARZO/2022

Datos propios

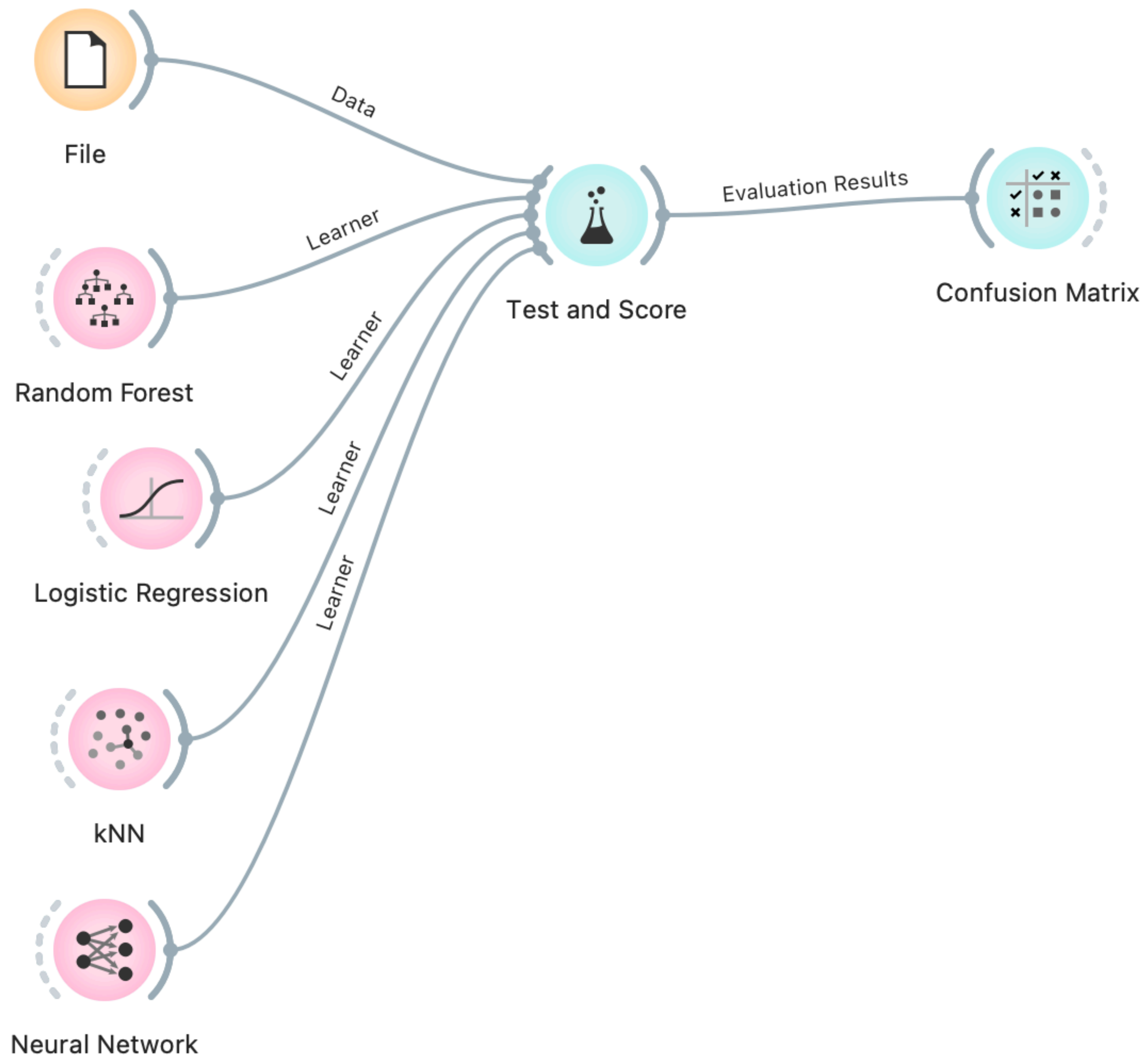
- Construir una base de datos en Google sheet.
- Importar los datos
- Adecuar los datos
- Save data



9/MARZO/2022

Primeras predicciones

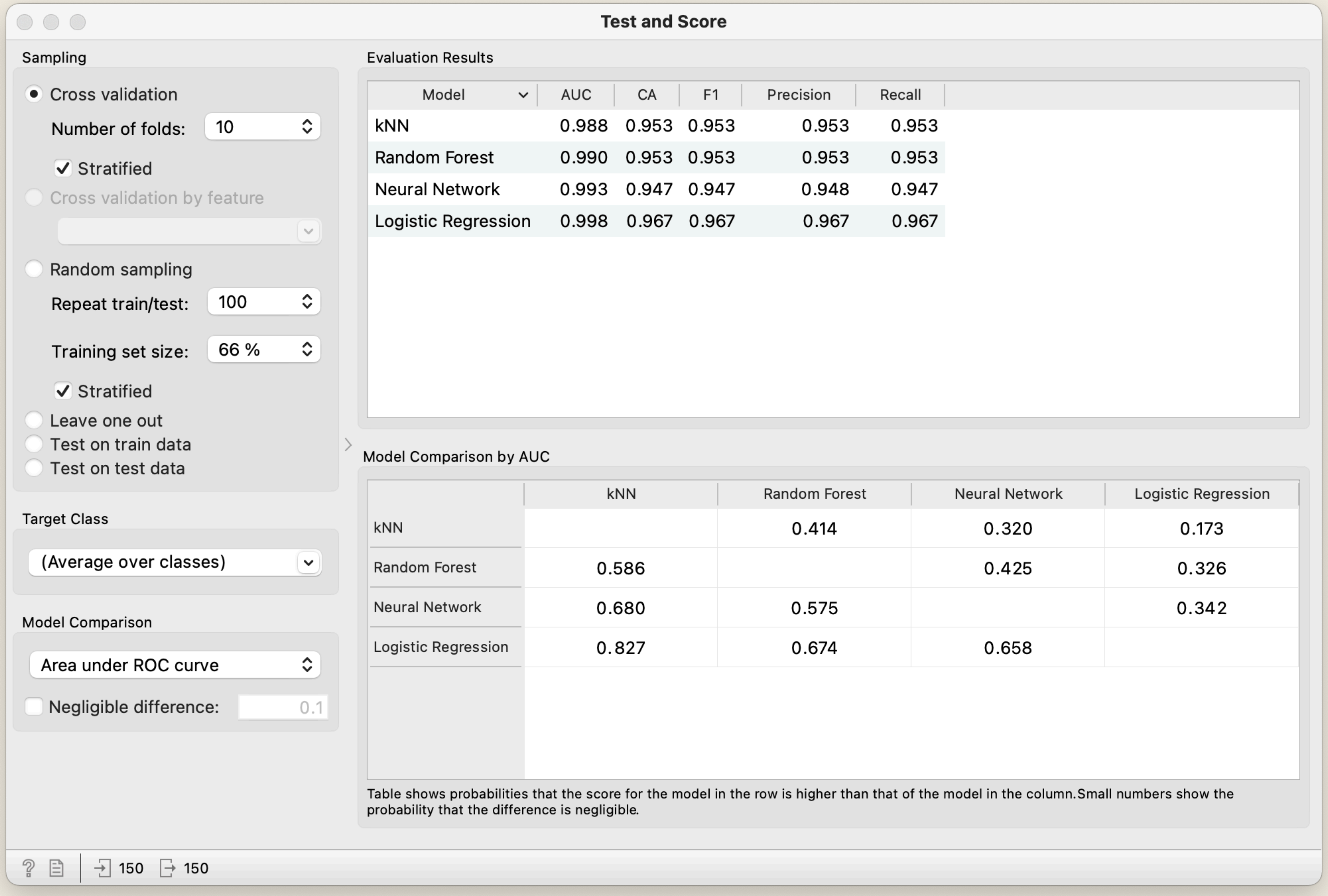
- Datos de frutas y vegetales
- Tree –> Tree viewer
- Test file
- Prediction
- Logistic Regression



9/MARZO/2022

Scoring

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



Model	AUC	CA	F1	Precision	Recall
kNN	0.988	0.953	0.953	0.953	0.953
Random Forest	0.990	0.953	0.953	0.953	0.953
Neural Network	0.993	0.947	0.947	0.948	0.947
Logistic Regression	0.998	0.967	0.967	0.967	0.967

	kNN	Random Forest	Neural Network	Logistic Regression
kNN		0.414	0.320	0.173
Random Forest	0.586		0.425	0.326
Neural Network	0.680	0.575		0.342
Logistic Regression	0.827	0.674	0.658	

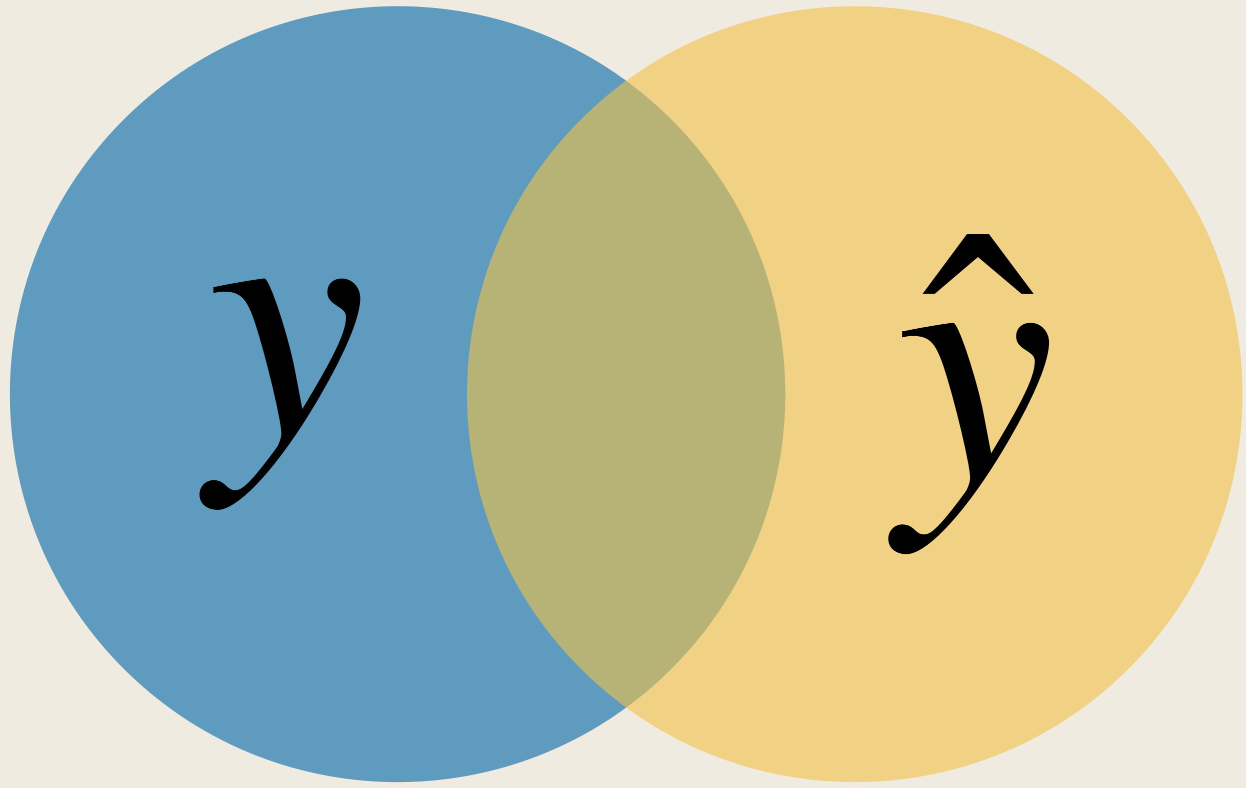
Evaluation metrics in classification

- 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

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

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

Confusion Matrix

TP = True Positives

FN = False Negatives

FP = False Positives

TN = True Negatives

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

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

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

Confusion Matrix

Predicted

Actual	Iris-setosa	Iris-versicolor	Iris-virginica	Σ
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

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