El Aprendiz de Scikit-Learn
(Machine Learning with Python)

TECHIDEAS
ENGINEER YOUR DREAMS

Pablo Hernández
CEO & Co-Founder
@phermar

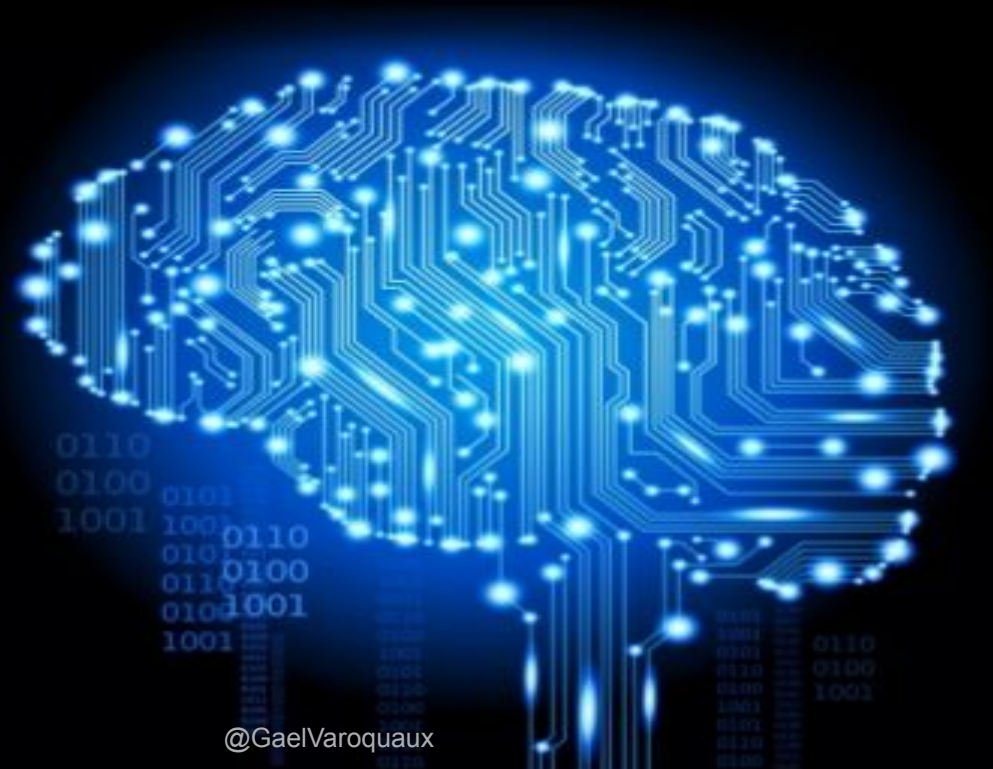
TECHIDEAS
ENGINEER YOUR DREAMS

¿Qué es el Aprendizaje Automático?

Machine Learning

Machine learning in a nutshell

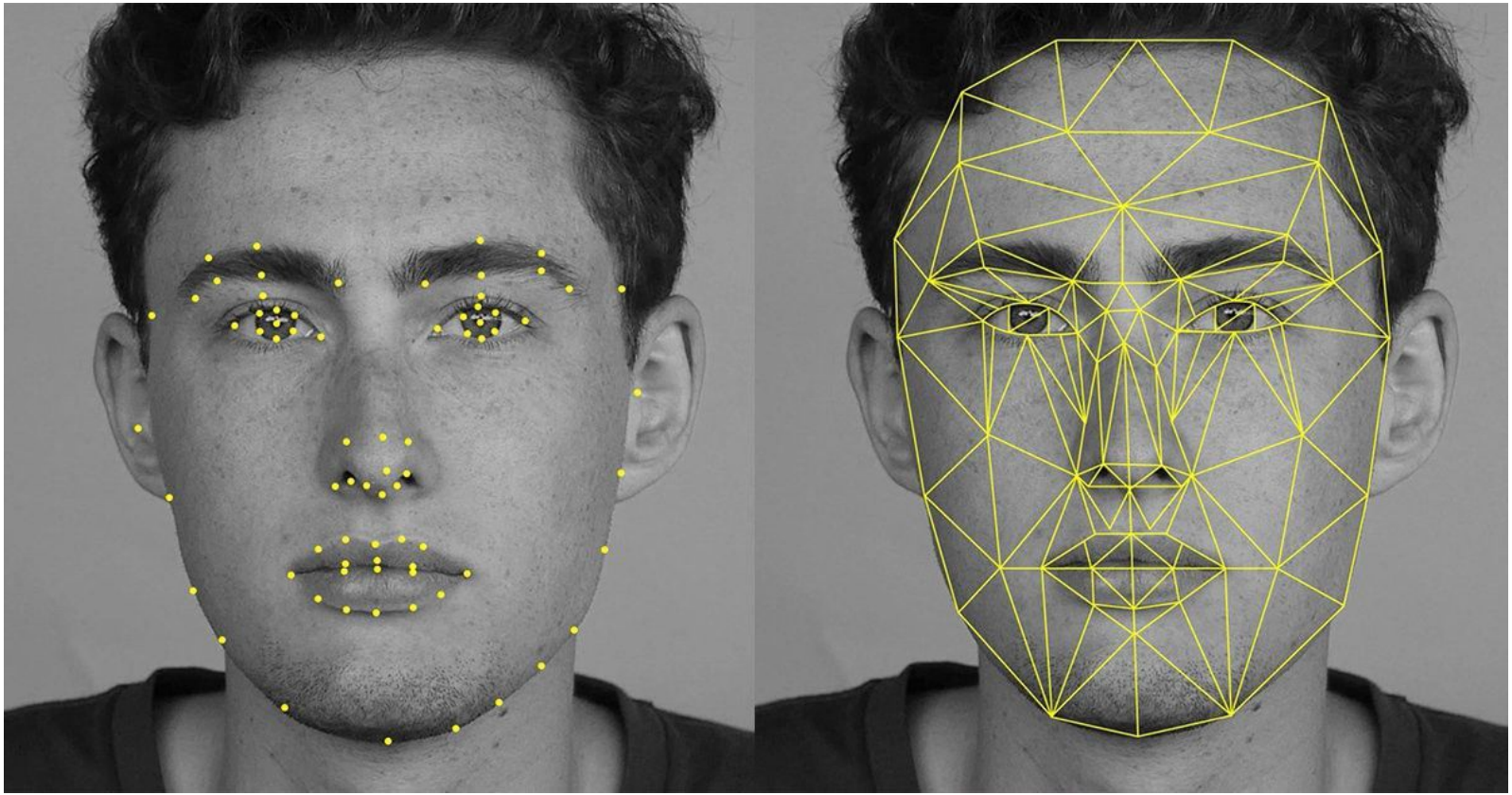
Machine learning is about making prediction from data



TECHIDEAS
ENGINEER YOUR DREAMS

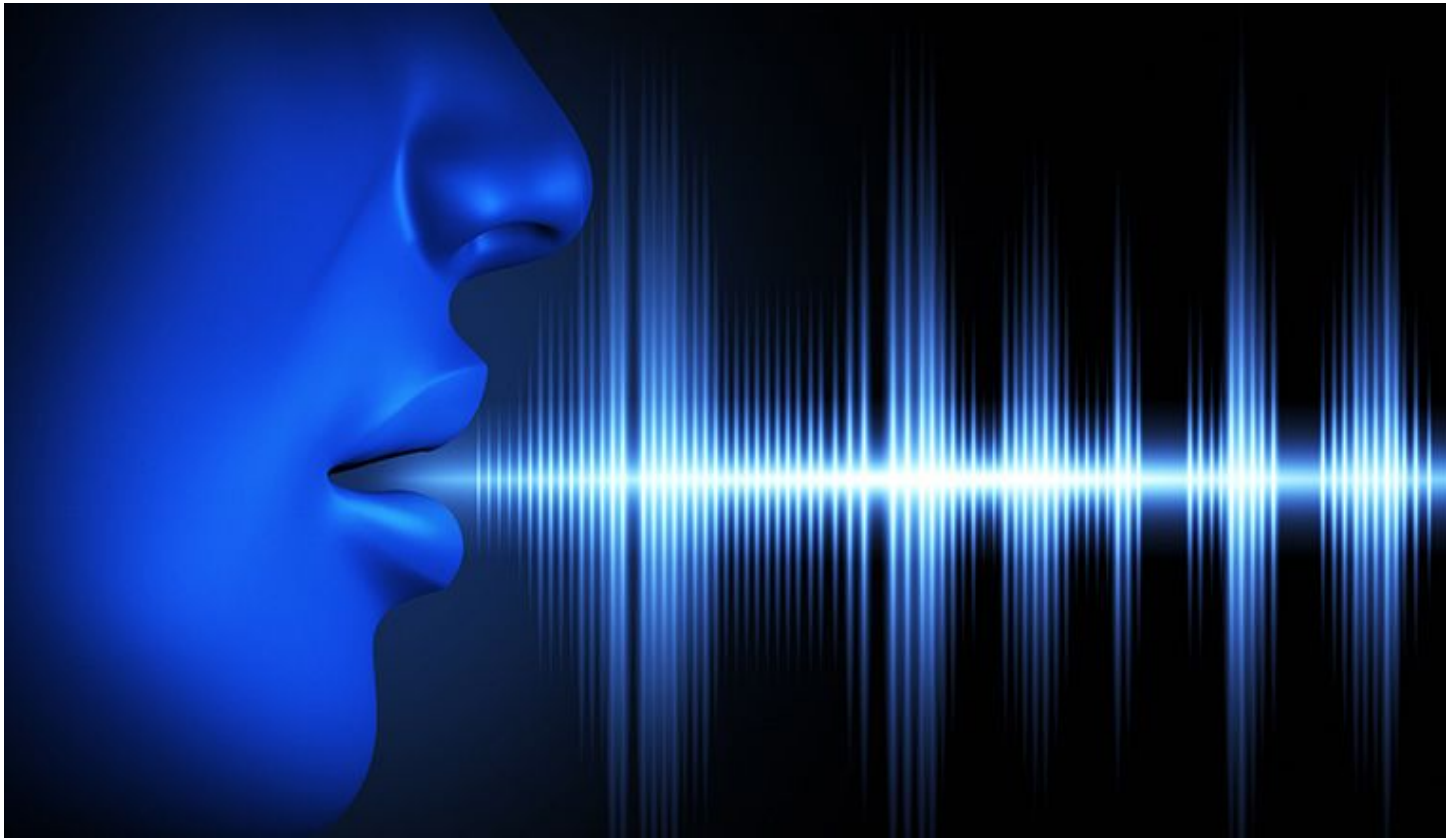
¿Cuáles son sus Aplicaciones?

Reconocimiento de Imágenes

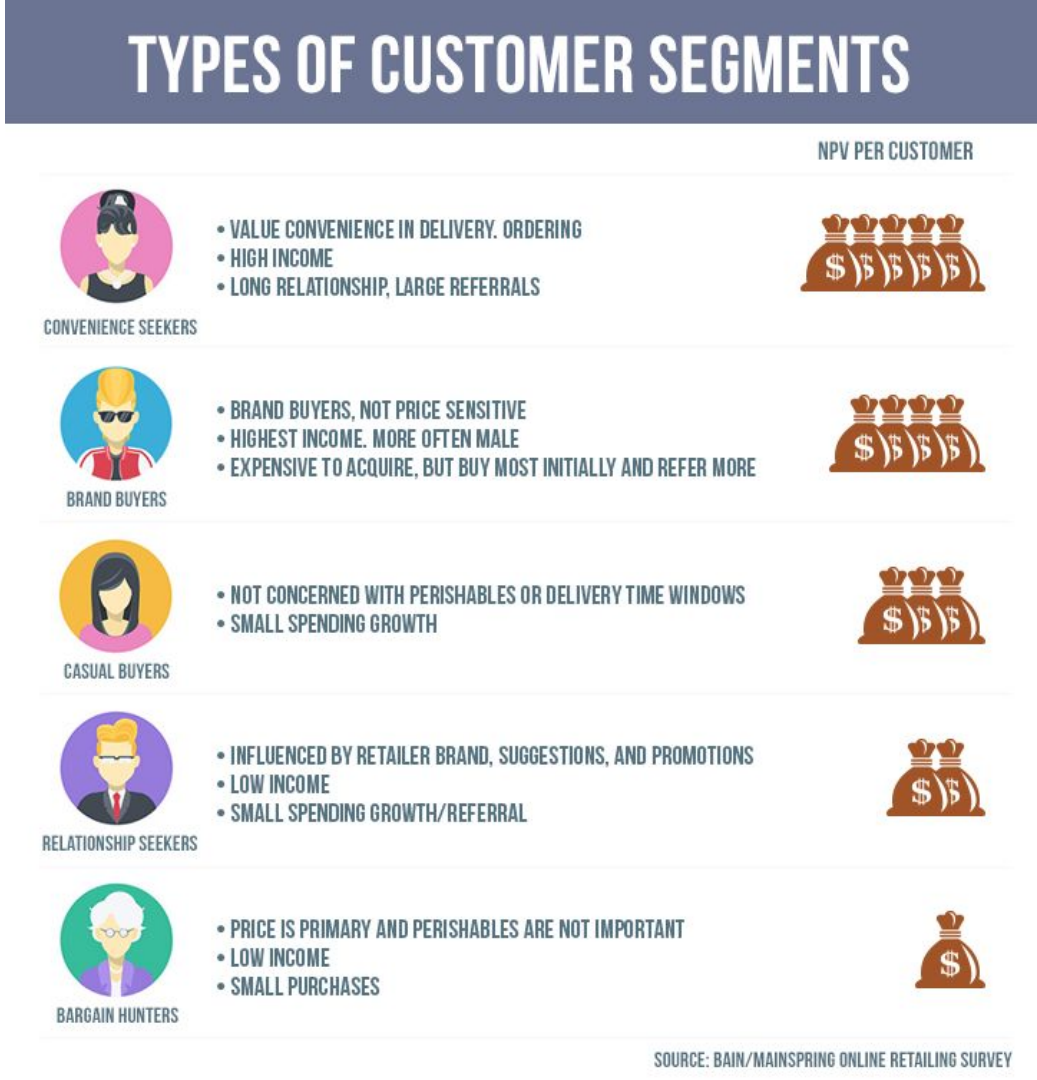


Sistema de Recomendación

Reconocimiento de Voz



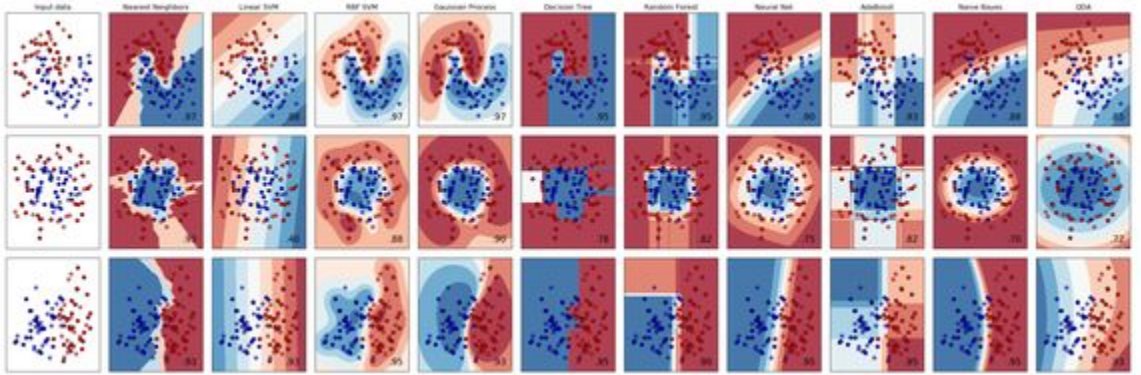
Segmentación de Clientes



TECHIDEAS
ENGINEER YOUR DREAMS

¿Qué es *Scikit-Learn*?

Definición



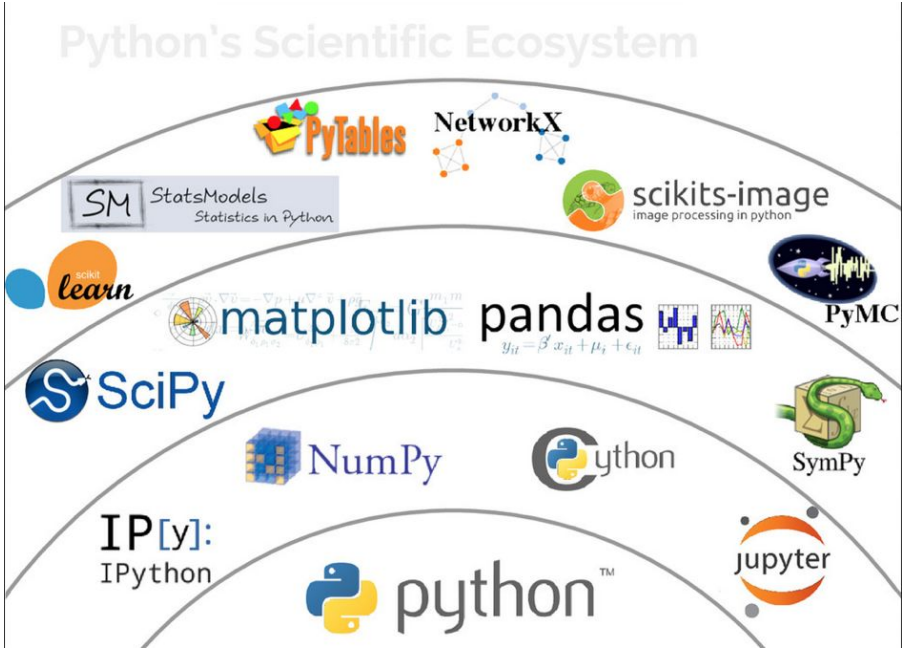
Scikit-learn es una librería de Python para dar soporte a los algoritmos de aprendizaje automático *(Machine Learning)*.

Scikit-learn soporta algoritmos de clasificación, regresión y agrupamiento.

Scikit-learn está basado en NumPy, SciPy y matplotlib

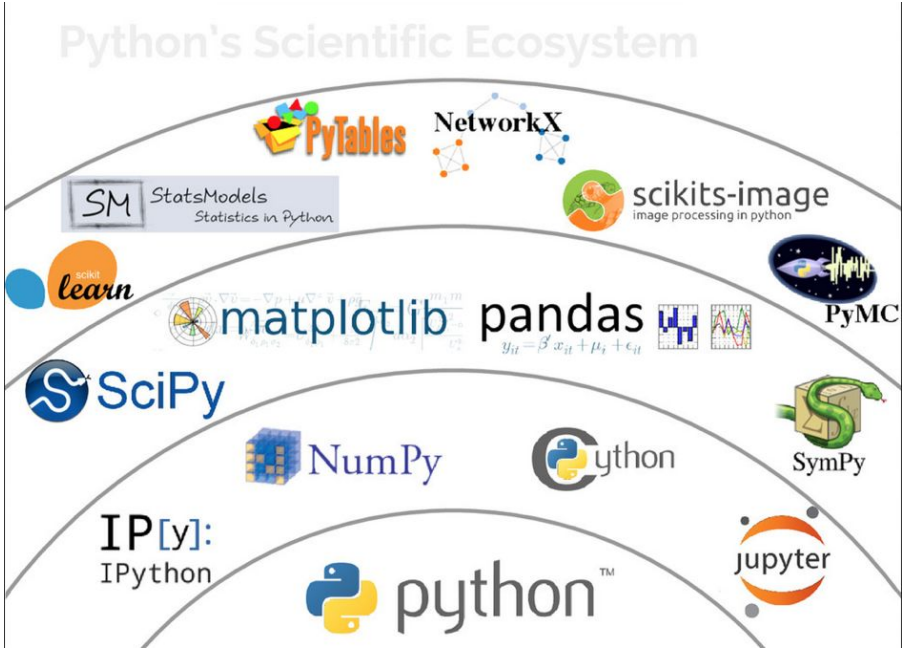
Ecosistema Científico para Python

Scikit-Learn es un conjunto de herramientas para facilitar el aprendizaje basado en los datos.



Ecosistema Científico para Python

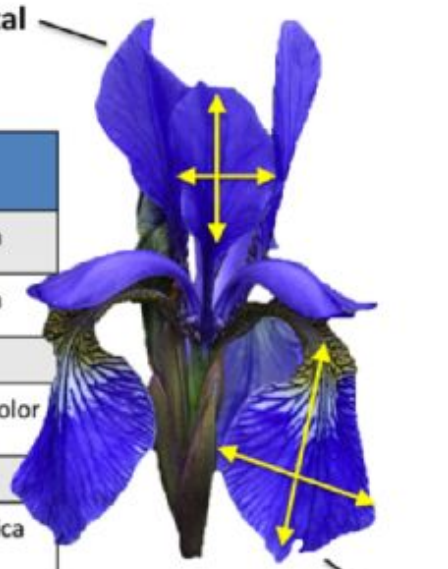
Scikit-Learn es un conjunto de herramientas para facilitar el aprendizaje basado en los datos.



scikit-learn Data Structure

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

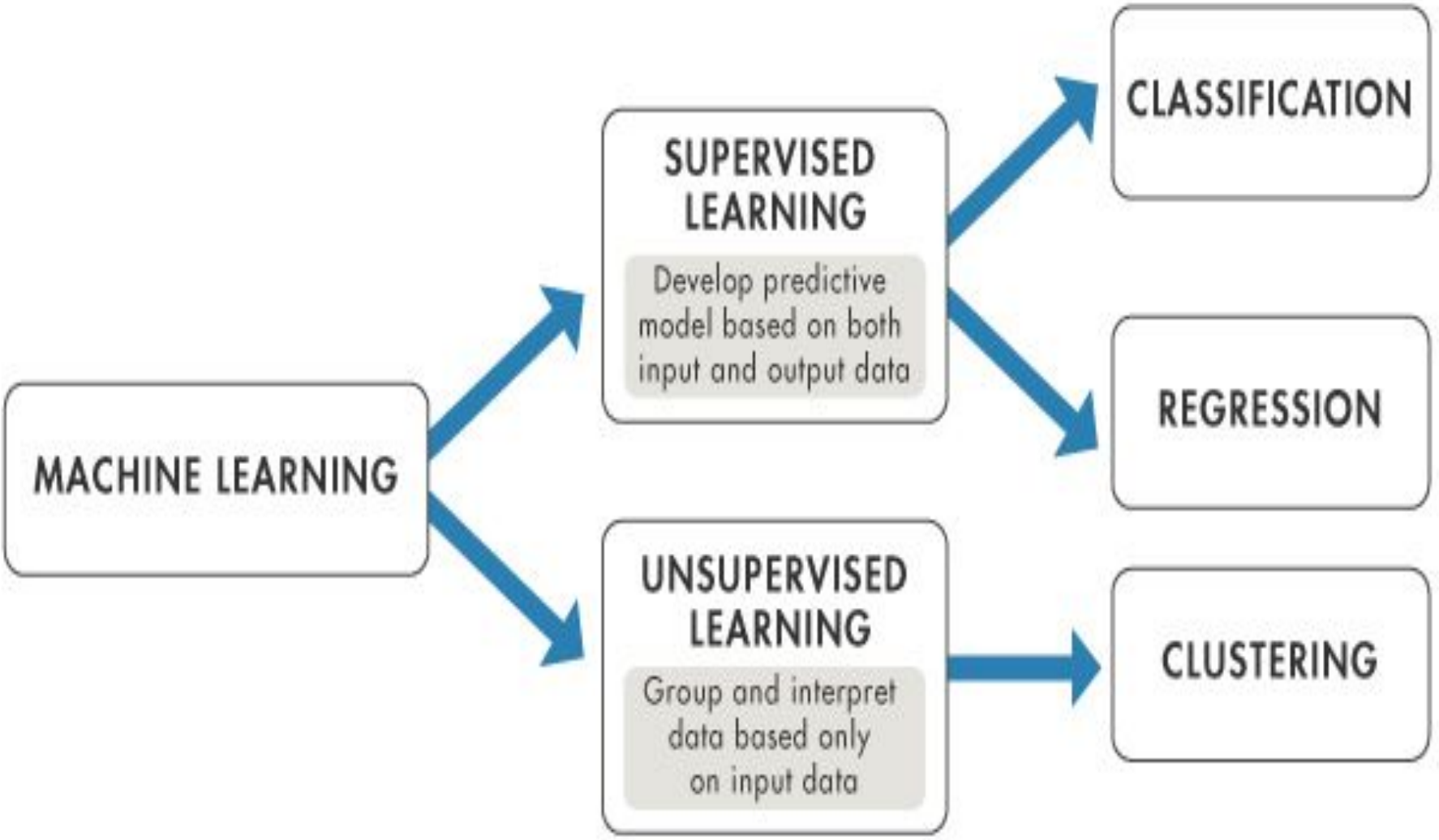


Features
(attributes, measurements, dimensions)

Class labels
(targets)

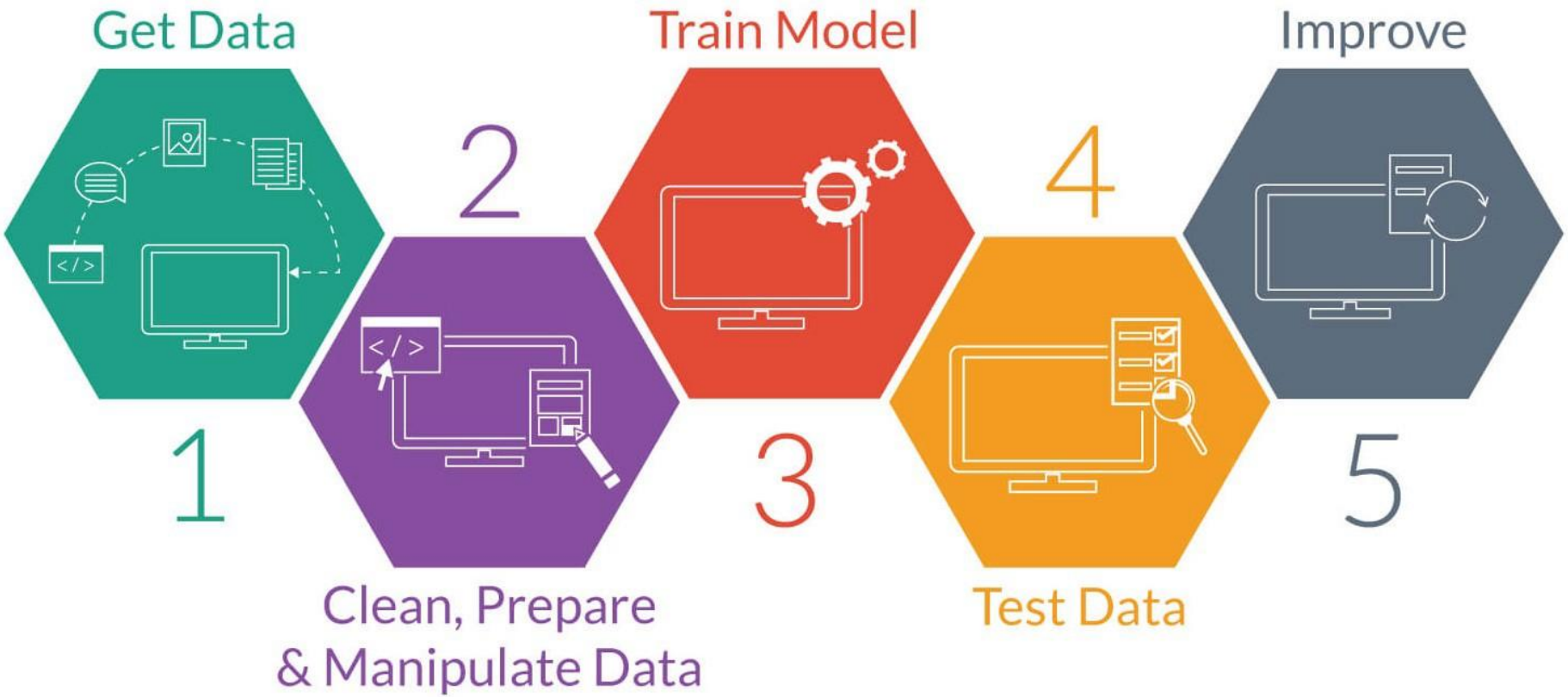
TECHIDEAS
ENGINEER YOUR DREAMS

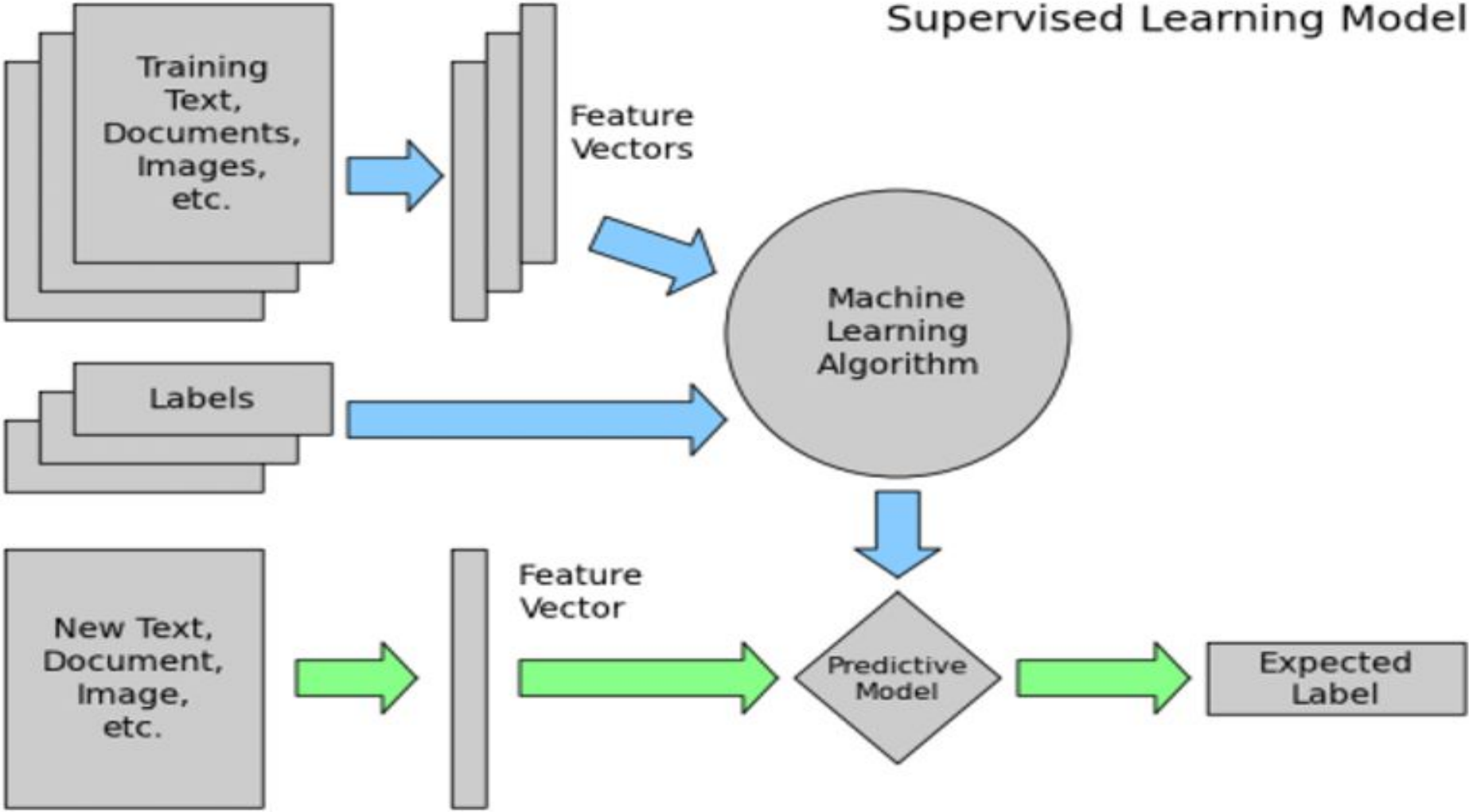
Características del Aprendizaje Automático

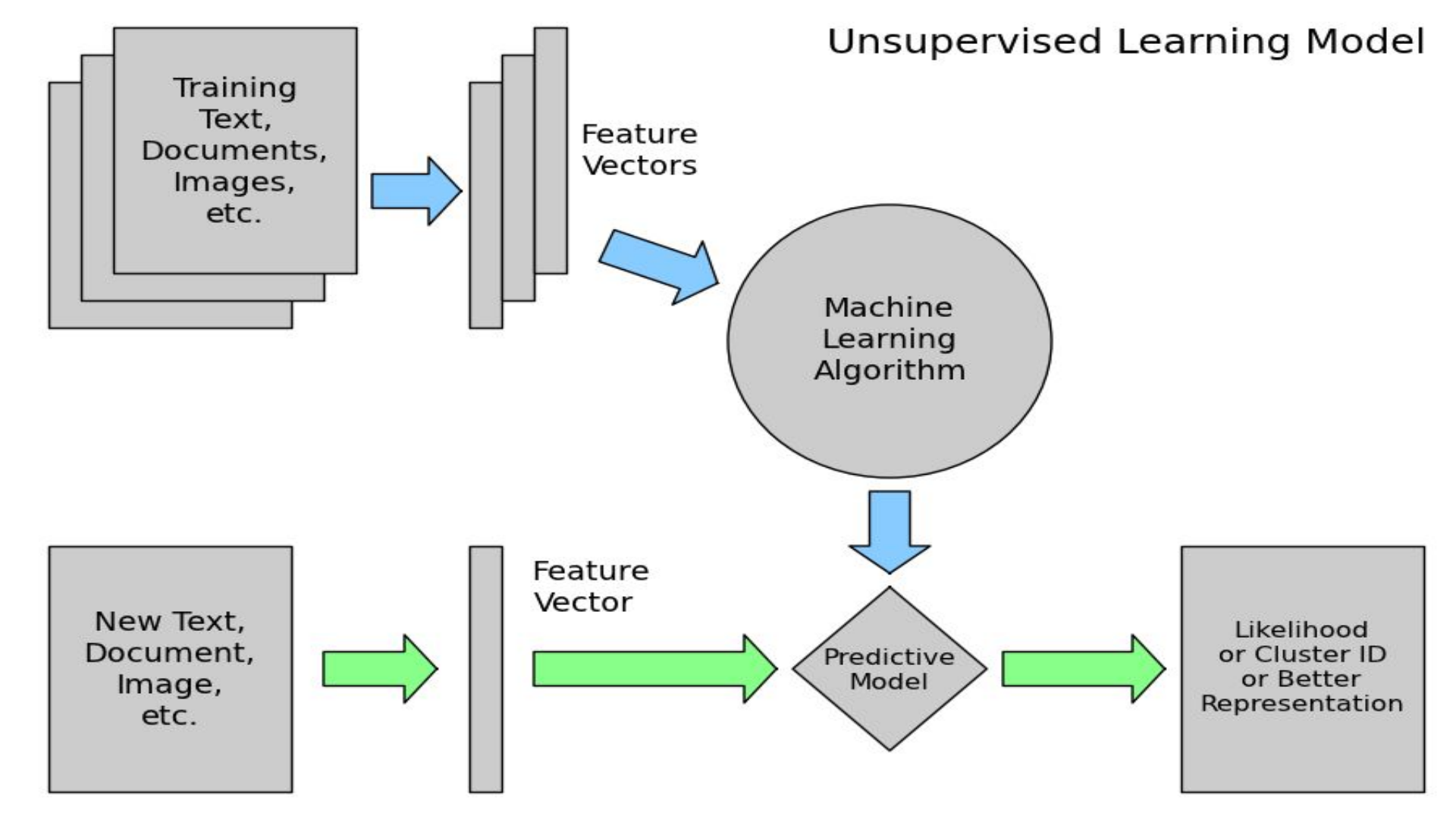

MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

TECHIDEAS
ENGINEER YOUR DREAMS

Procesos del Aprendizaje Automático


Get Data
Train Model
Improve
2
4
1
3
5
Clean, Prepare
& Manipulate Data
Test Data

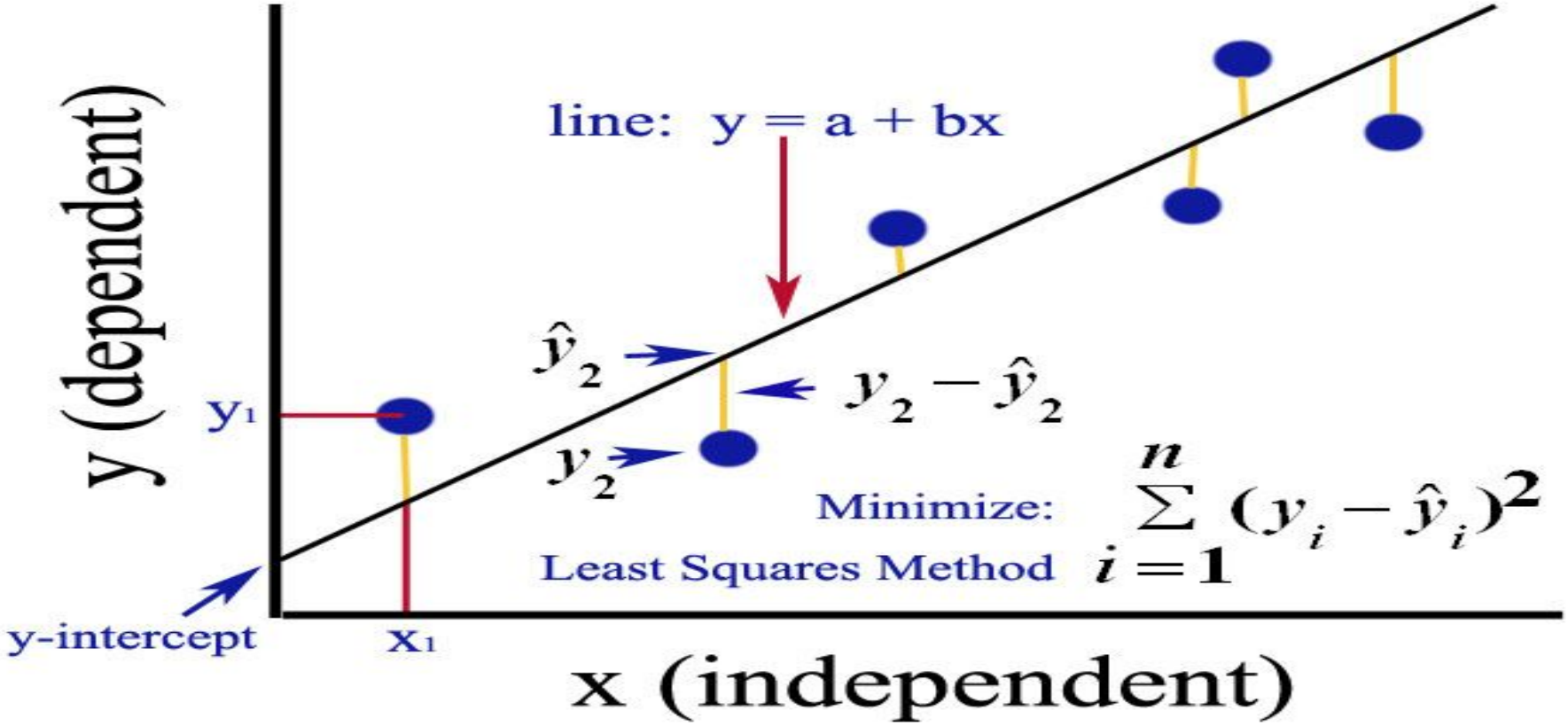

Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label


Unsupervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Likelihood or Cluster ID or Better Representation

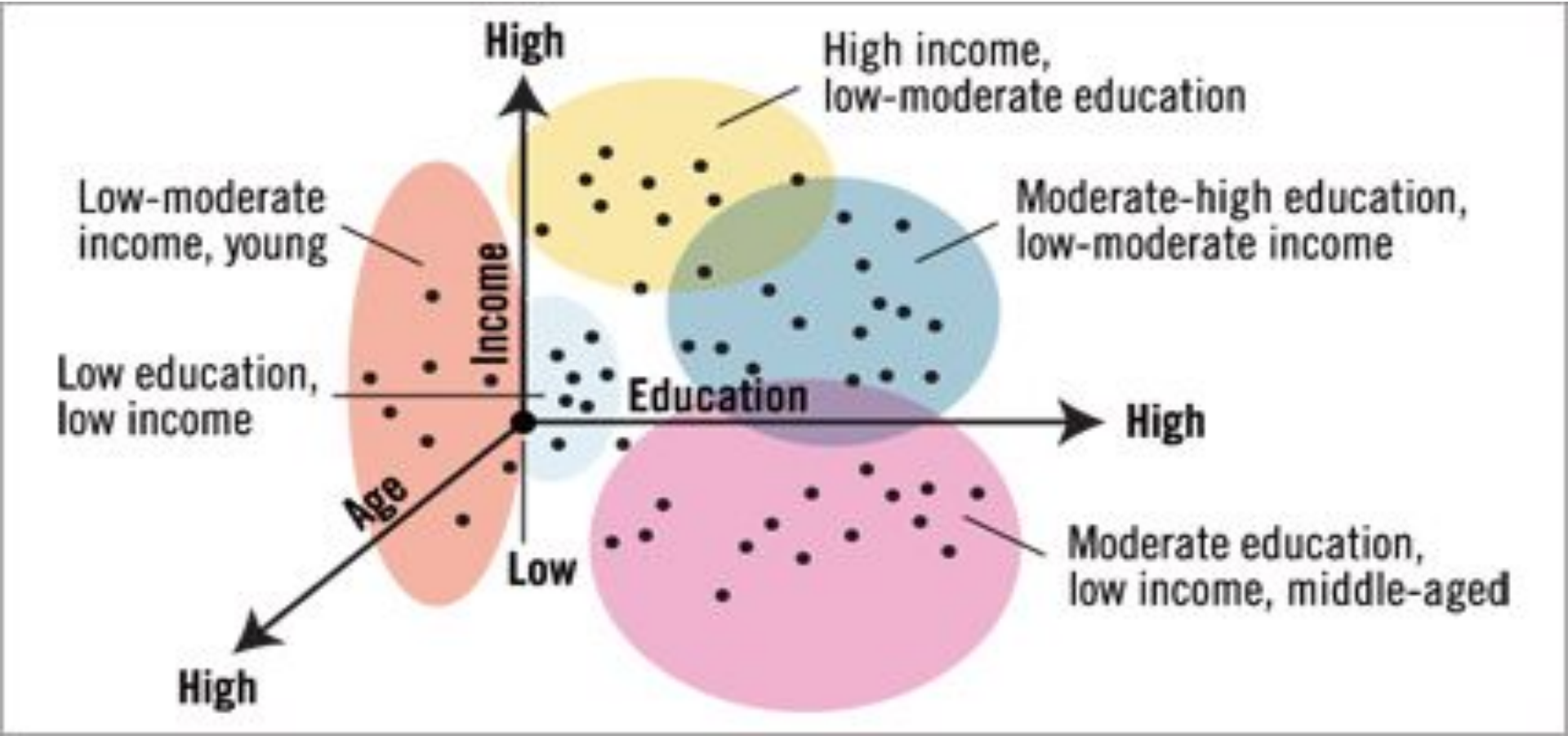
Modelos Predictivos

Clasificación

Regresión


line: y = a + bx
y (dependent)
y₁
$\hat{y}_2$
$y_2 - \hat{y}_2$
y_2
Minimize: $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Least Squares Method
y-intercept
x₁
x (independent)

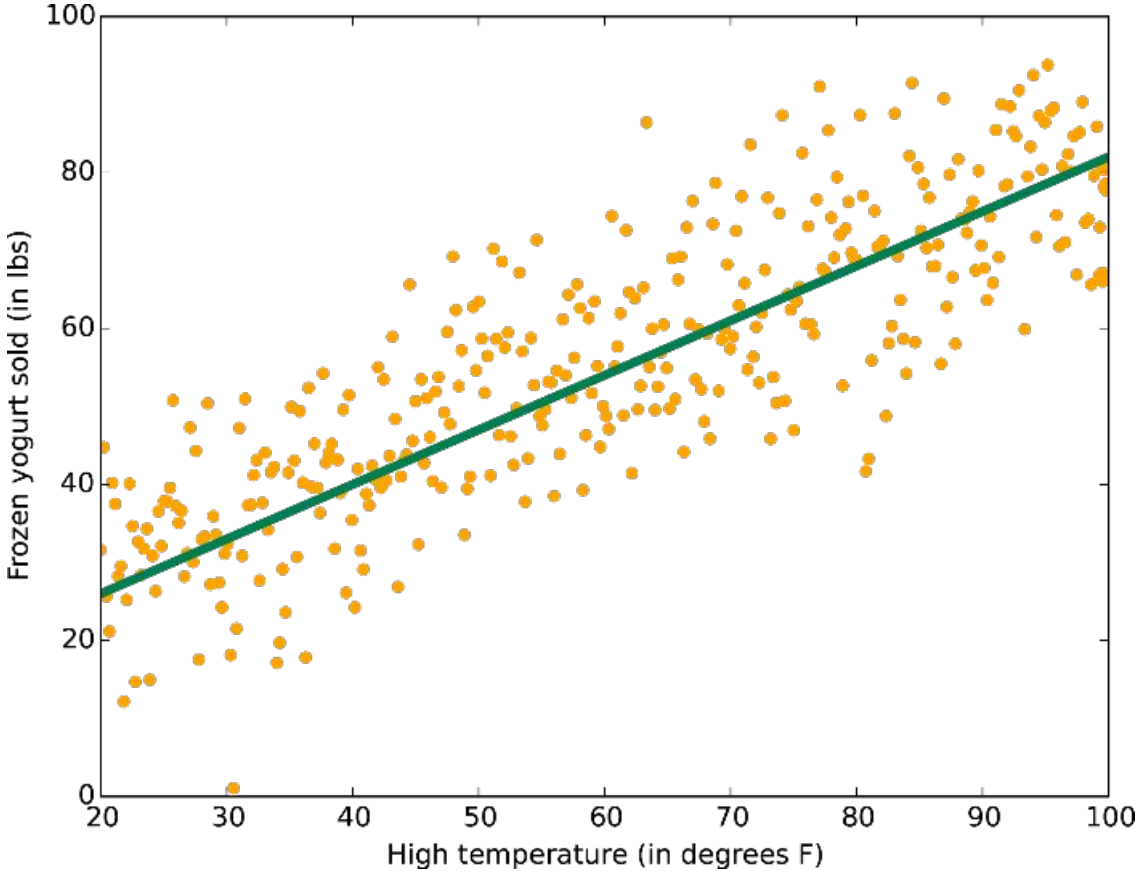
Agrupación


High
High income,
low-moderate education
Low-moderate
income, young
Moderate-high education,
low-moderate income
Income
Low education,
low income
Education
High
Age
Low
Moderate education,
low income, middle-aged
High

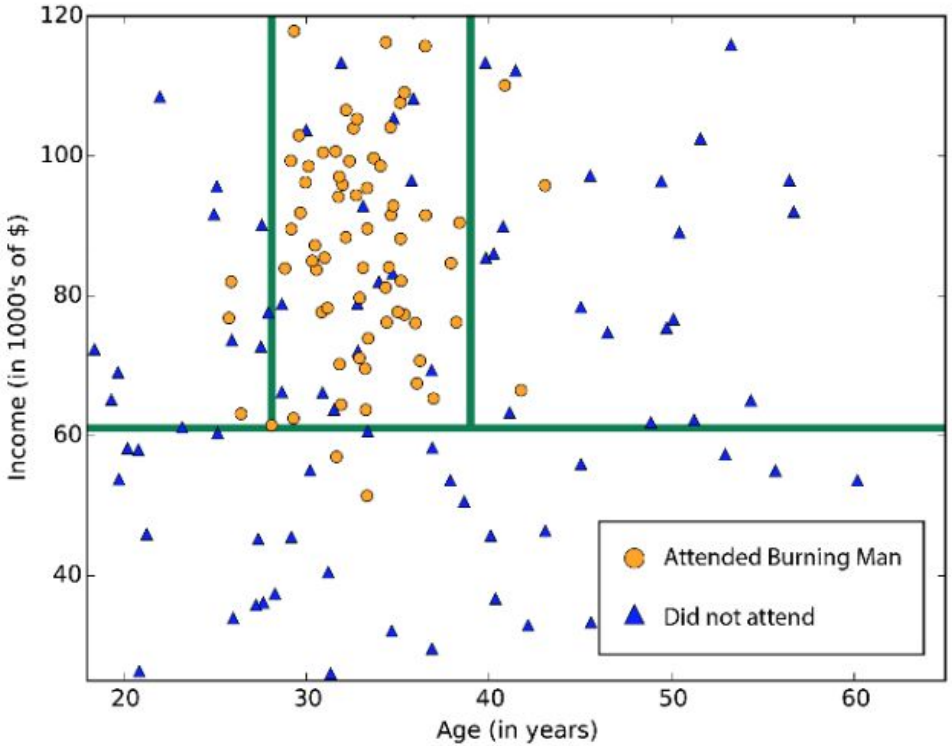
TECHIDEAS
ENGINEER YOUR DREAMS

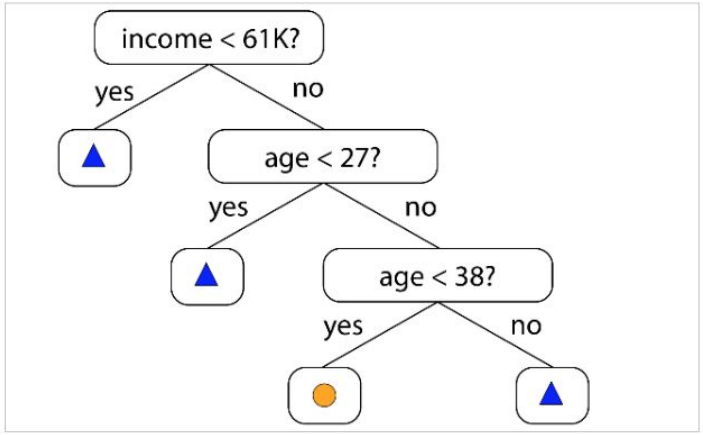
Algoritmos de Aprendizaje

Regresión Lineal

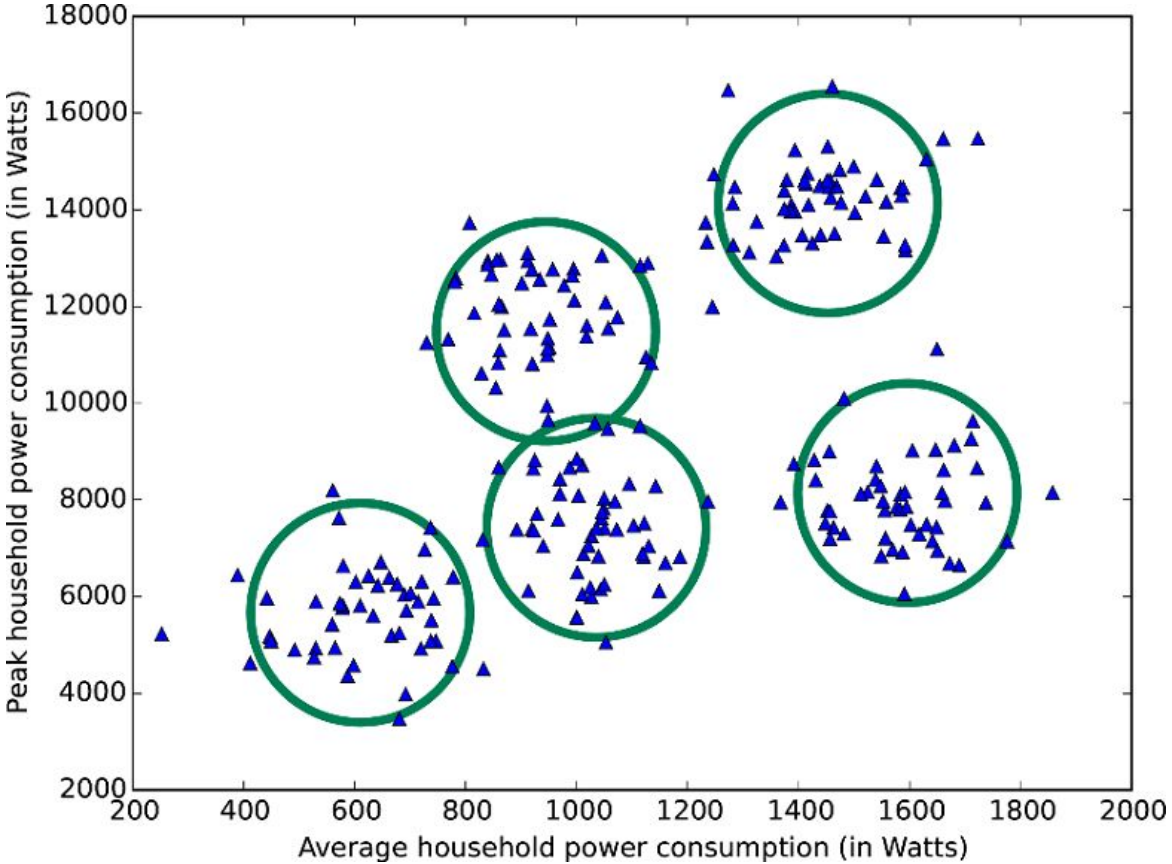


Clasificación - Árboles de Decisión

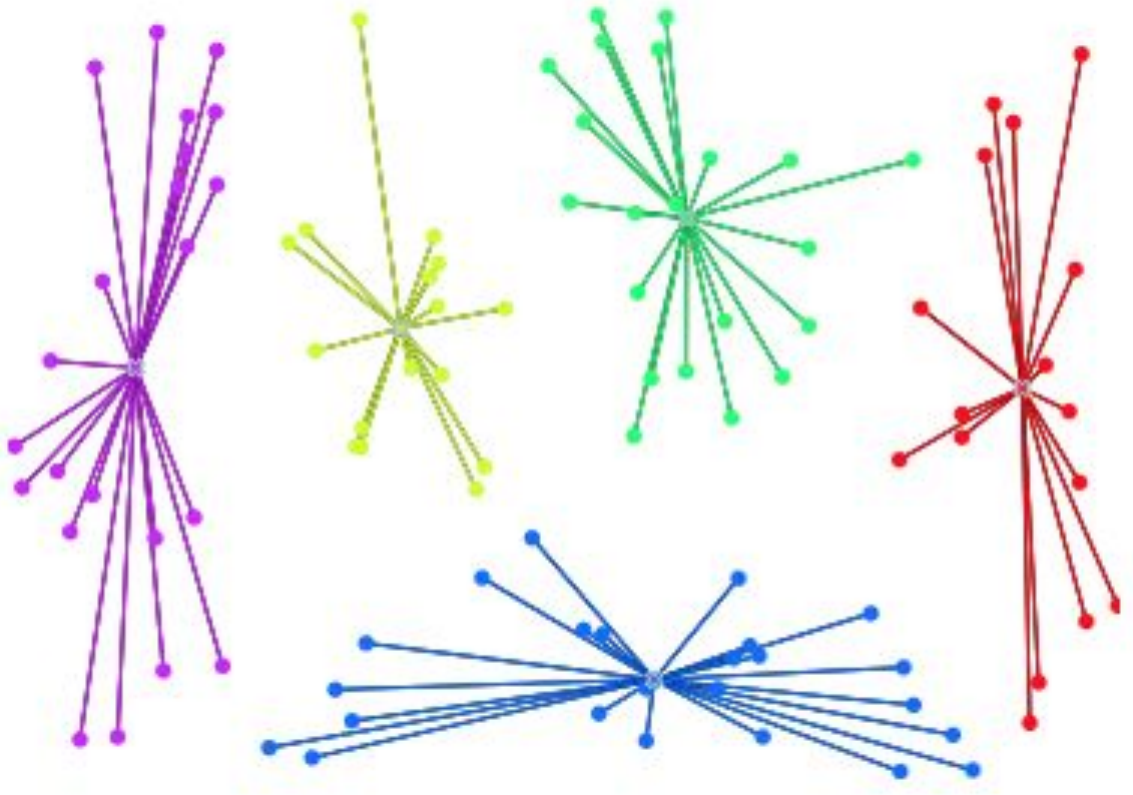




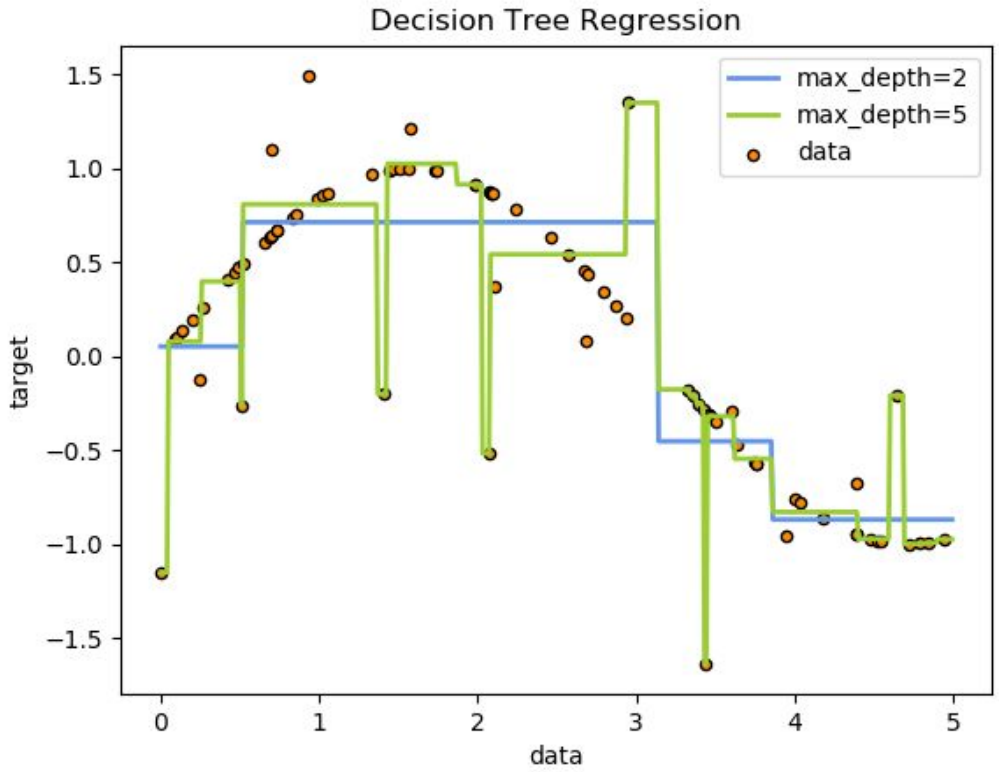
Agrupación: K-Means

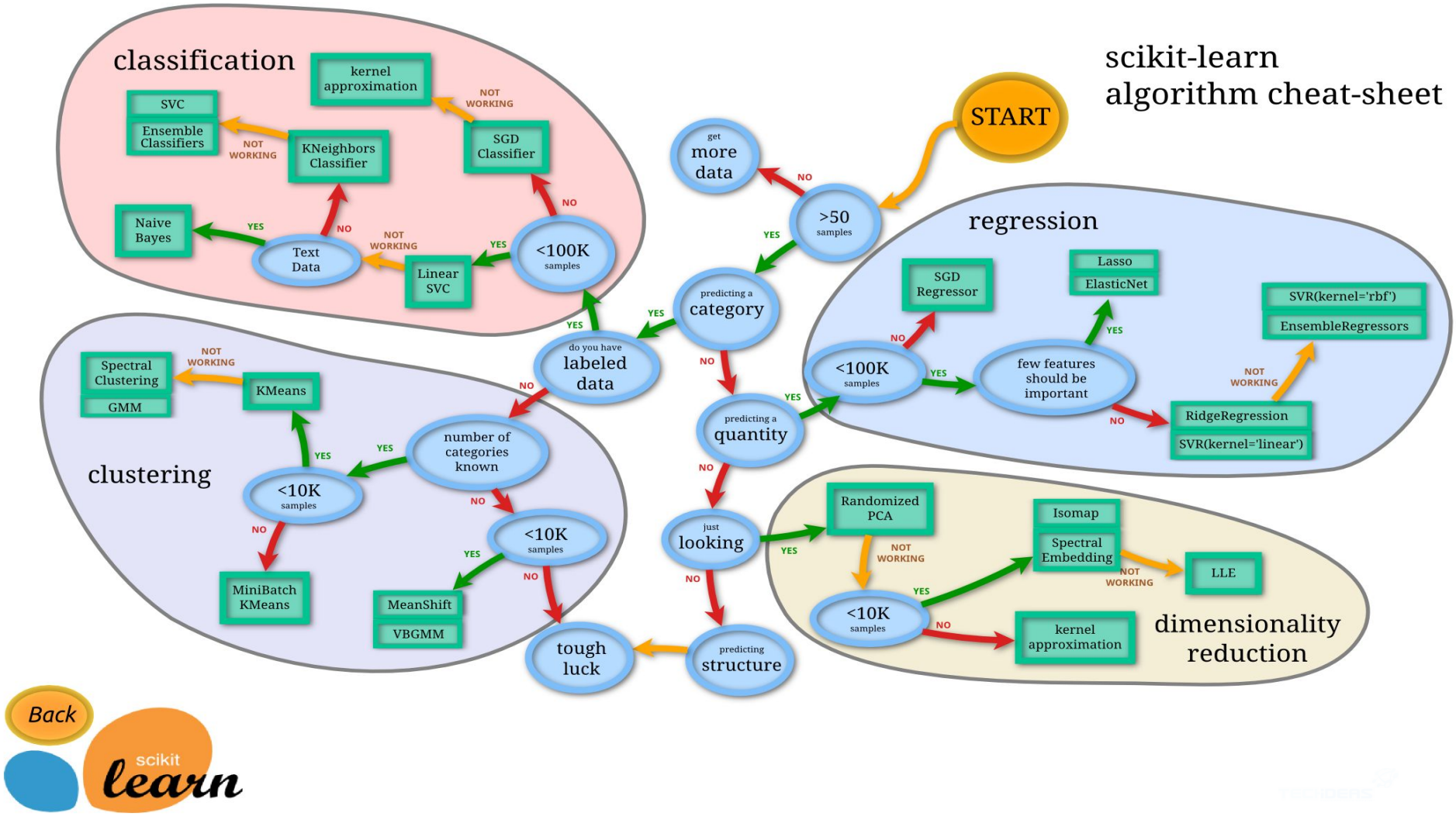


Agrupación: K-Means Visualization



Regresión basada en Arboles de Decisión




scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
NO
do you have labeled data
YES
NO
predicting a quantity
YES
NO
just looking
YES
NO
predicting structure
tough luck
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
YES
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
regression
SGD Regressor
NO
<100K samples
YES
few features should be important
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
dimensionality reduction
Randomized PCA
YES
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
Back
scikit learn

Thank you

TECHIDEAS
ENGINEER YOUR DREAMS

Pablo Hernández
CEO & Co-Founder
@phermar