Datalized

Solving Problems. With Data.

Women Who Code
“Tu primer modelo de Machine Learning”

Mayo 2018

Agenda

01 Introducción a Machine Learning

02 Conceptos teóricos

03 Preparando el entorno

04 A codear!

Introducción a Machine Learning


MACHINE LEARNING
DICTIONARY
NEURAL
PREVIOUSLY
PROBABILISTIC
INPUT
INTUITION
DATABASE
APPROACH
UNSEEN
STATISTICAL
RELATIONSHIPS
LEARN
RECOGNITION
CLASSIFIER
UNLABELED
DATUM
OVERLAP
COMPACTNESS
ACTION
ATTEMPT
CONSTRAINT
DISCOVERY
VECTOR
SPARSE
ENVIRONMENT
PERFORMANCE
LEARNER
ALGORITHM
LOGIC
DISTRIBUTION
IMAGE
SPARSELY
TIME
PROVIDED
LOCOMOTION
DEFINED
SUBSETS
GRAPHICAL
MINING
FIND
RESPECT
EXPLICITLY
APPROXIMATES
THEORY
TAXONOMY
METHOD
TREE
EXPERIENCE
SUPERVISED
CONFERENCES
EMPLOY
EXAMPLES
OBSERVATION
COEFFICIENT
TERMS
METHODS
NETWORK
INFERENCE
ARTIFICIAL
VARIABLE
TRAINING
CLASS
FEATURES
COMPUTATIONAL
IMPLAUSIBLE
BOUNDED
GRAPH
KNOWLEDGE
MACHINE
MAPS
PREDICTION
OUTPUT
UNDERLYING
CLASSIFICATION
CASES
ANALYSIS
OBSERVED
INDUCTIVE
SET
GENERALIZE
CLUSTER
UNSUPERVISED
HYPOTHESIZED
UNKNOWN
ASSUMPTIONS


SENSORIMOTOR
RESEARCHERS
APPROACH
LOGIC
SCRUFFY
COMPUTER
ALGORITHM
CONTROL
SINGULARITY
HUMANOID
FICTION
ECONOMICS
BINOMIAL
CHESS
PLAN
REASON
HIERARCHICAL
COMMONSENSE
GENERAL
TIME
ARTIFICIAL
CLASSIFICATION
CALCULUS
APPLICATION
SYNTACTIC
LEARN
CLASSIFIER
AUTOMATONS
ACTION
SPECIFIC
ENGINEERING
UNCERTAIN
OUTPUTS
INTELLIGENCE
METHODS
DIFFICULT
MATHEMATICAL
ONTOLOGY
NETWORK
CONCEPT
ARCHITECTURE
EMBODIED
SIMULATION
SWARM
EVOLUTIONARY
EARLY
AUTONOMOUSLY
DECISION
SCIENCE
PERCEPTION
RECOGNITION
NONHUMAN
BEINGS
MANIPULATION
COMPUTING
KNOWLEDGE
STATISTICAL
PROBLEM
DEVELOP
AGENT
NEURAL
CYBORGS
ARTICLE
HUMAN
CYBERNETICS
SYMBOLIC
COMPUTATION


BIG
DATA
ALSO
TOOLS
DISK
TARGET
APPLIED
SENSOR
MAY
EXAMPLES
SHARED
DEFINITION
CURRENT
MOVING
EVERY
RELATIONAL
SET
USE
LARGER
CONTINUES
SOCIAL
PRACTITIONERS
CAPTURE
WITHIN
COMPLEX
CITATION
TENS
INTERNET
BUSINESS
SETS
GENOMICS
ANALYTICS
DISTRIBUTED
MANAGEMENT
TERABYTES
CASE
ABILITY
BIOLOGICAL
PROCESSING
CAPACITY
PETABYTES
INCLUDE
SYSTEMS
TOLERABLE
HUNDREDS
NETWORKS
DATABASES
RECORDS
SEARCH
GARTNER
STORAGE
SIZE
PARALLEL
SAN
MASSIVELY
GROW
DIFFICULTY

Machine Learning

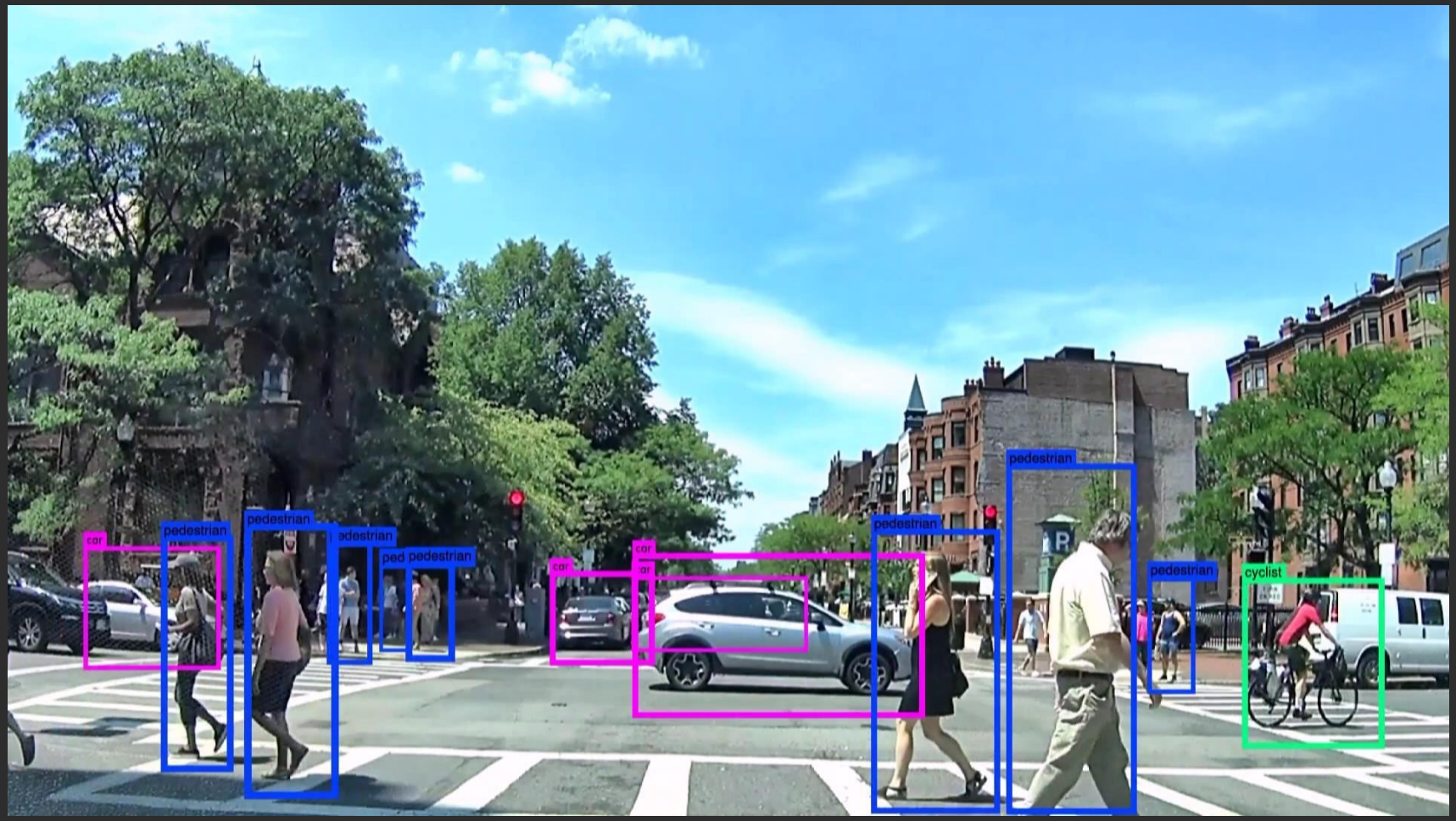
Método para desarrollar modelos predictivos a través de la extracción de patrones de un conjunto de datos

PERRO

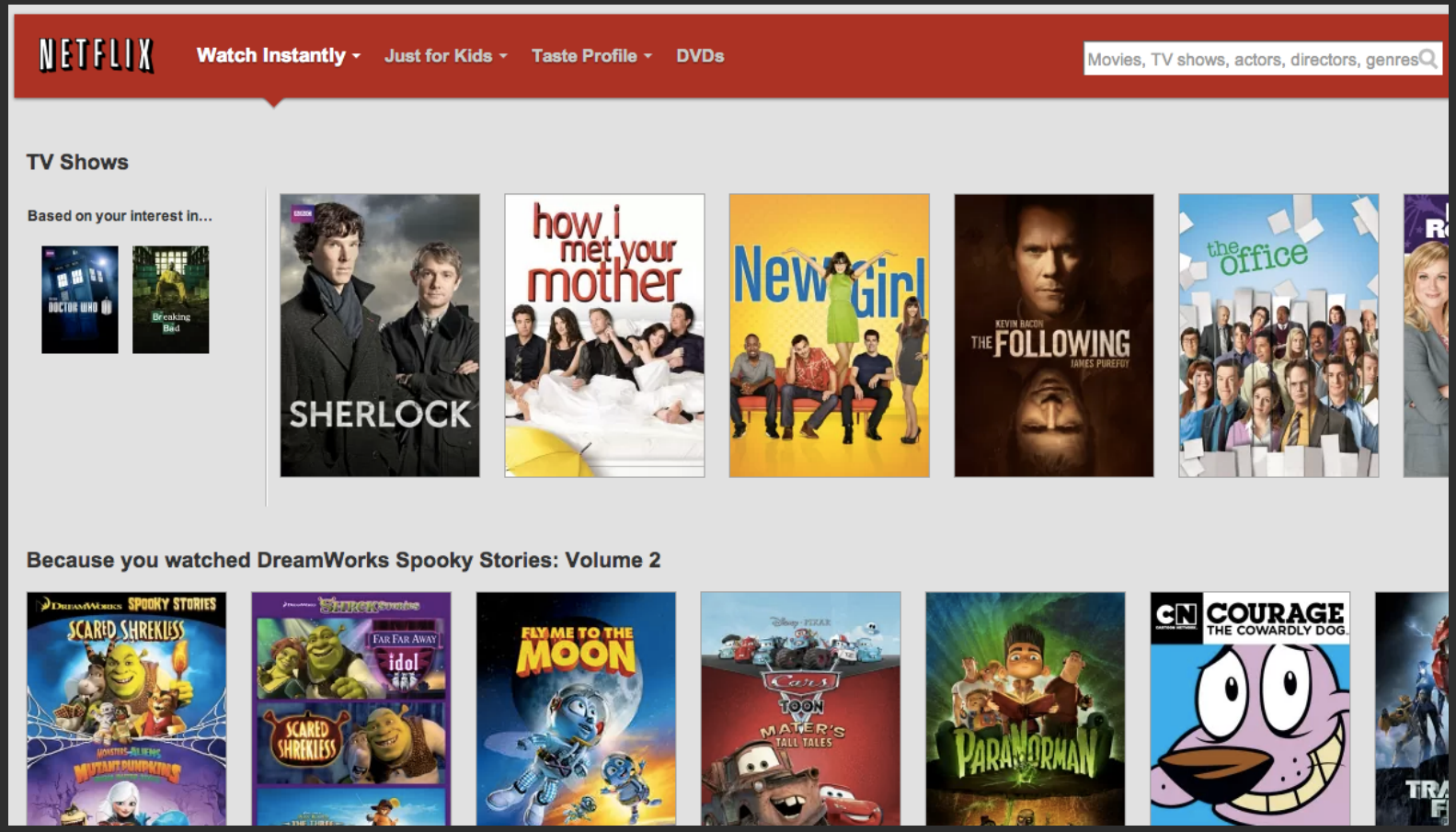
GATO

F () = ¿PERRO o GATO?

Machine Learning

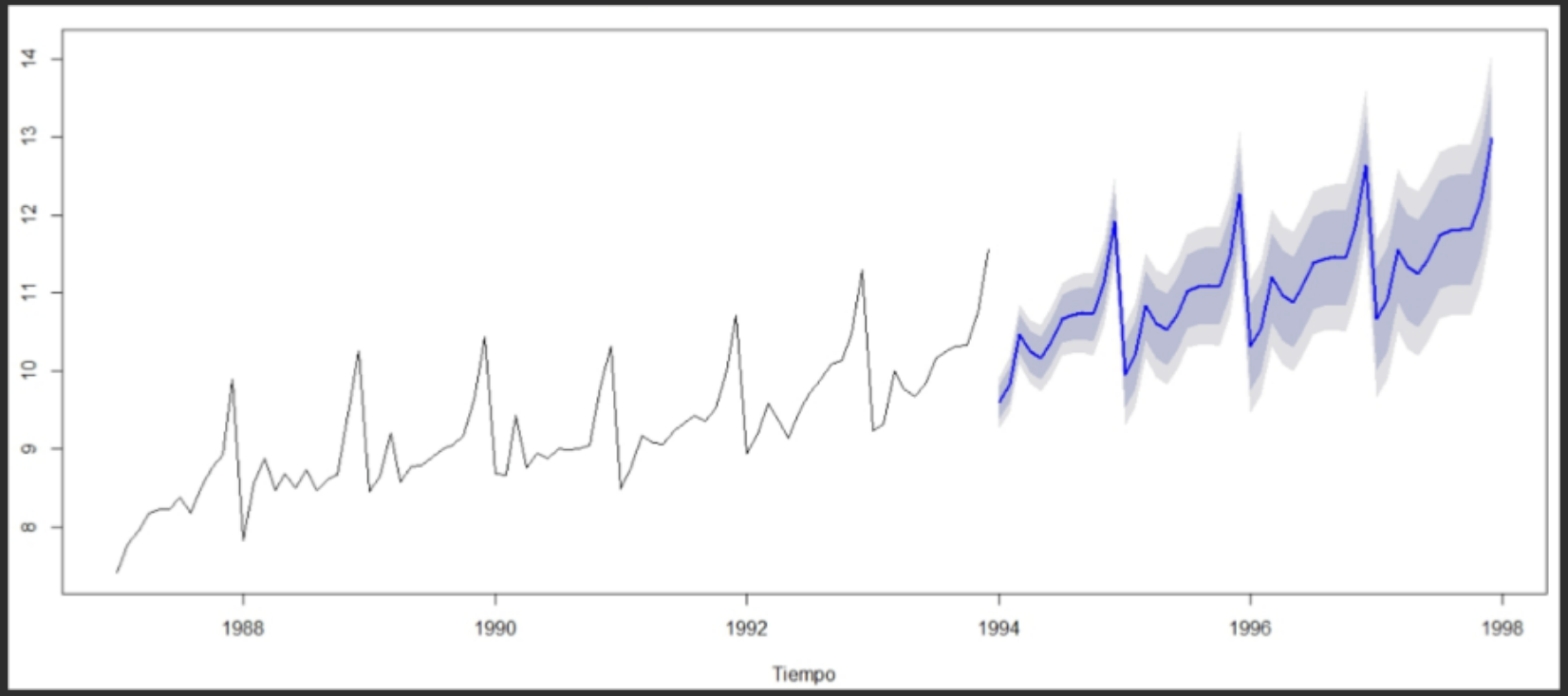


Machine Learning



Machine Learning

Machine Learning



Conceptos Teóricos

Tipos de problemas de Machine Learning

Aprendizaje Supervisado

El conjunto de datos incluye la variable objetivo que se quiere predecir.

Categorías:
Clasificación
Regresión

Algunos algoritmos:
Arboles de Decisión
Support Vector Machines
Regresión Lineal
Redes Neuronales

Aprendizaje No Supervisado

El conjunto de datos incluye la variable objetivo que se quiere predecir.

Categorías:
Segmentación
Reducción de dimensionalidad

Algunos algoritmos:
K-Means
Gaussian Mixture Models
Principal Component Analysis
Linear Discriminant Analysis

Aprendizaje Reforzado

Método para enseñarle a un *agente* a tomar la *acción* que maximicen su *recompensa* dada su interacción con el *entorno*.

Ejemplo:
Game playing (AlphaGo)

Algunos algoritmos:
Q-Learning
Temporal Difference
Deep Adversarial Networks

Tipos de problemas de Machine Learning

Aprendizaje Supervisado

El conjunto de datos incluye la variable objetivo que se quiere predecir.

Categorías:
Clasificación
Regresión

Algunos algoritmos:
Arboles de Decisión
Support Vector Machines
Regresión Lineal
Redes Neuronales

Clasificación

Predecir a que *clase* pertenece una observación.

Ejemplo: Predecir si un animal es un perro o un gato, predecir si un cliente se fugará o no, predecir el tipo de flor, etc..

Regresión

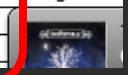
Predecir un *valor continuo*.

Ejemplo: Predecir cuantos minutos nos demoraremos en completar un pedido, predecir el precio futuro de un Bitcoin, etc..

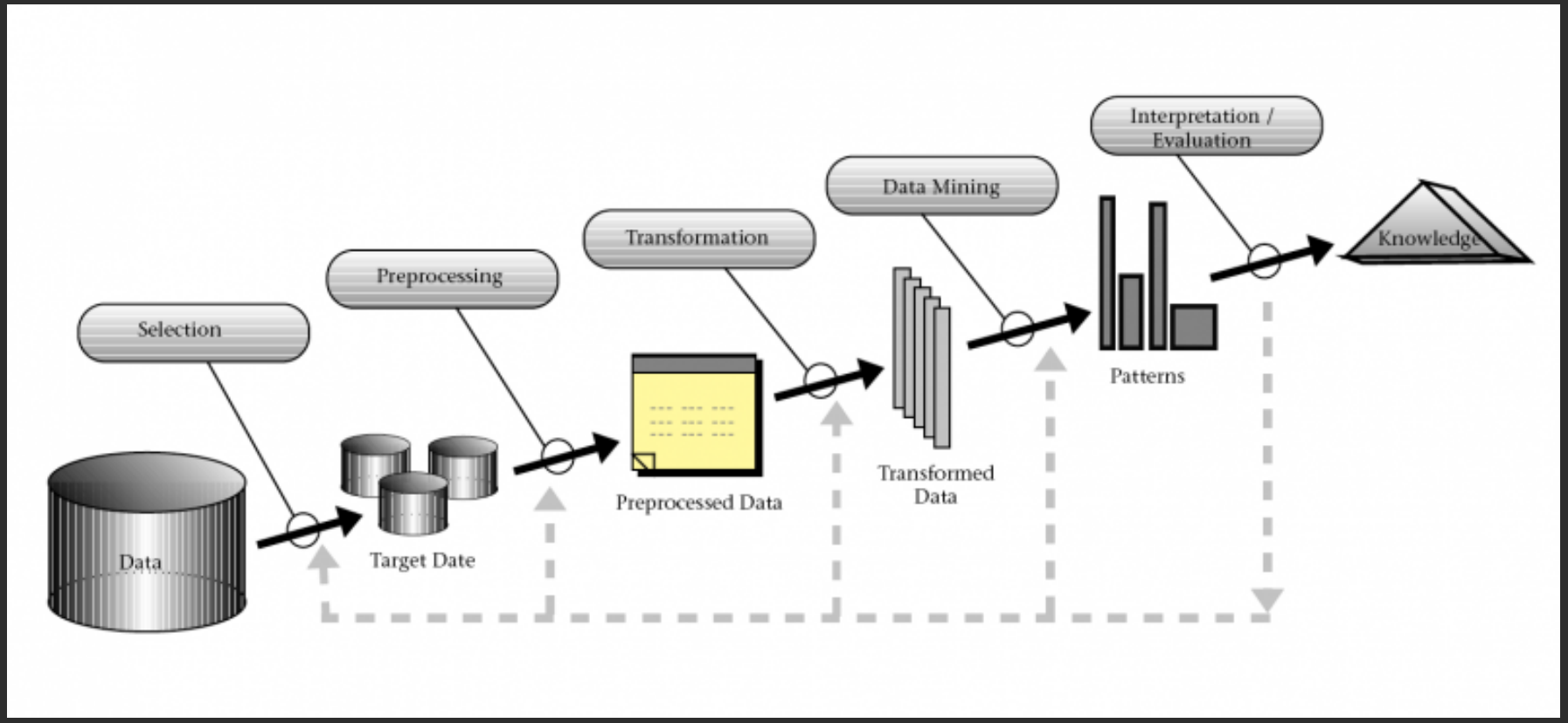
Conjunto de Datos

Conjunto de Datos

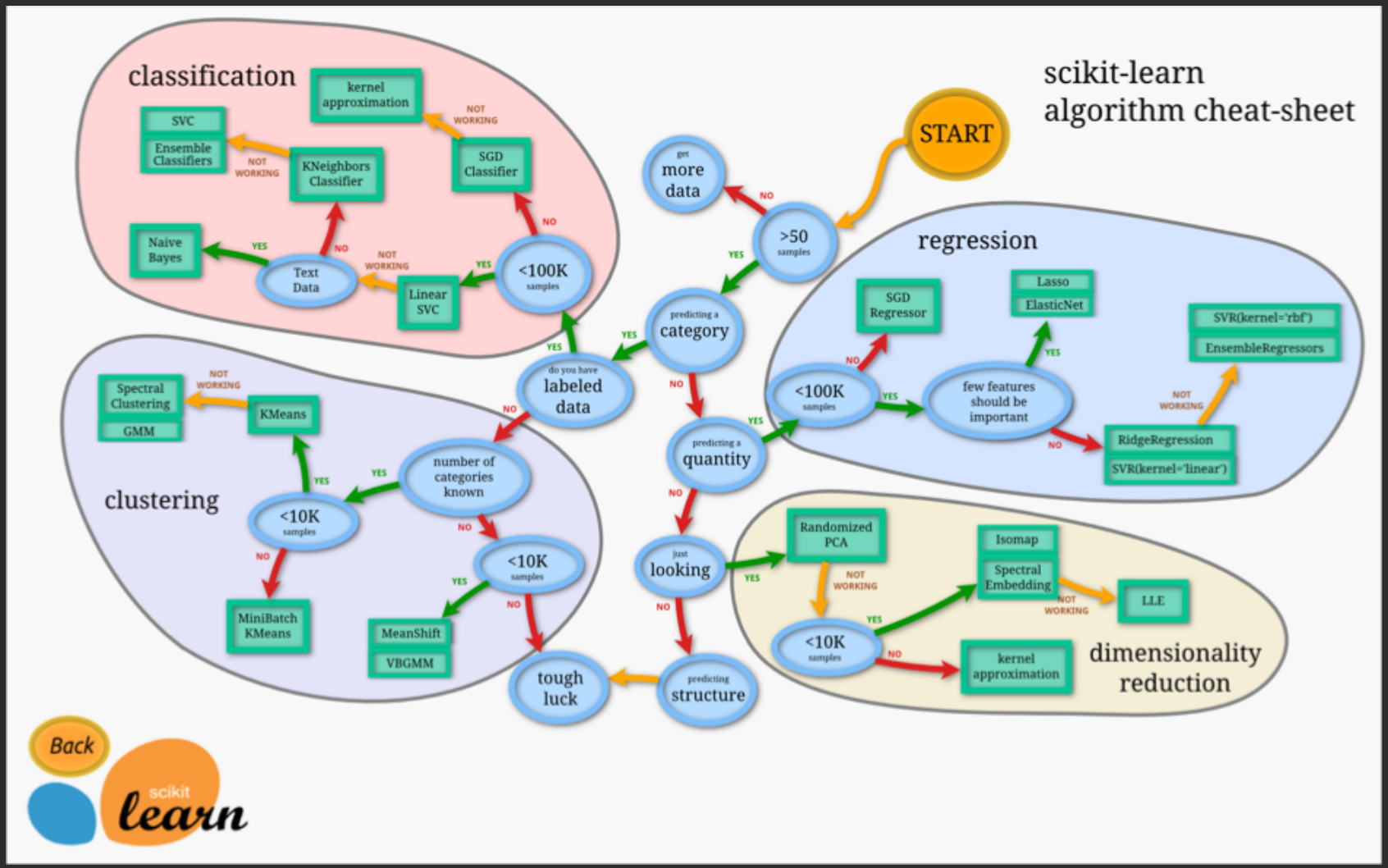
ID	GENERO	RENTA	EDAD	NIV_EDUC	E_CIVIL	COD_OFI	COD_COM	CIUDAD	COL_1	COL_2	COL_3	COL_4
1	M	1,200,000	39	TEC	CAS	42	31	COQUIMBO	3,788,129	512,102	0	412,983
2	M	1,200,000	55	TEC	CAS	16	90	SANTIAGO	0	0	0	1,527,726
3	M	800,000	47	UNV	CAS	48	87	SANTIAGO	0	800,000	0	0
4	F	700,000	47	TEC	CAS	79	148	SAN FERNANDO	1,112,326	104,780	0	1,167,713
5	F	4,730,000	32	TEC	CAS	52	88	SANTIAGO	0	750,000	0	15,614
6	F	500,000	37	TEC	CAS	139	90	SANTIAGO	0	132,982	0	0
7	M	1,700,000	56	TEC	SOL	64	314	COYHAIQUE	930,378	1,174,182	0	0
8	F	1,162,000	30	TEC	SOL	90	88	SANTIAGO	0	0	0	0
9	M	7,350,000	41	UNV	CAS	55	90	SANTIAGO	0	800,000	0	0
10	M	616,000	52	UNV	CAS	31	119	SANTIAGO	0	300,000	0	0
11	M	2,279,000	40	UNV	SEP	138	1	ARICA	4,373,067	1,000,000	0	966,131
12	M	700,000	48	UNV	CAS	74	272	VALDIVIA	0	227,199	0	1,000,000
13	F	1,000,000	28	UNV	CAS	44	173	TALCA	0	2,600,000	0	0
14	F	1,536,000	55	MED	CAS	39	84	SANTIAGO	0	347,816	0	320,337
15	M	650,000	39	TEC	SEP	76	45	PETORCA	0	388,697	0	246,770
16	M	750,000	41	MED	CAS	31	119	SANTIAGO	0	459,827	0	0
17	M	74,200	50	MED	CAS	55	90	SANTIAGO	0	500,000	1,366,018	0
18	M	700,000	33	UNV	CAS	48	89	SANTIAGO	0	686,288	0	0
19	M	744,000	48	MED	CAS	138	1	ARICA	0	103,413	0	241,818
20	F	500,000	49	MED	CAS	80	81	SANTIAGO	0	128,846	0	188,588
21	M	1,000,000	53	UNV		138	1	ARICA	609,457	351,911	0	700,000
22	F	435,000	50	UNV		131	94	SANTIAGO	0	543,111	0	0
23	M	525,000	48	UNV		148	90	SANTIAGO	0	201,292	0	0
24	M	600,000	51	UNV	SOL	90	97	SANTIAGO	0	285,070	0	130,936
25	M	547,000	48	UNV	CAS	127	93	SANTIAGO	0	892,177	0	298,710
26	M	1,044,000	49	UNV	SOL	120	116	SANTIAGO	1,438,520	232,488	0	187,326
27	F	86,000	45	MED	SOL	51	88	SANTIAGO	0	1,745,545	0	0
28	M	1,160,000	47	TEC	CAS	83	85	SANTIAGO	0	12,044	0	1,099,284
29	F	463,000	63	MED	CAS	95	316	AYSEN	0	423,805	0	289,449
30	F	6,250,000	36	MED	SOL	90	89	SANTIAGO	2,503,920	825,989	0	0
31	M	797,000	60	MED	CAS	149	108	SANTIAGO	0	300,000	0	181,933
32	M	2,349,000	53	TEC	CAS	84	85	SANTIAGO	0	0	0	0
33	F	800,000	41	MED	CAS	102	90	SANTIAGO	0	827,187	0	
34	M	893,000	52	UNV	SOL	138	1	ARICA	0	1,945,800	0	



Knowledge Discovery in Databases



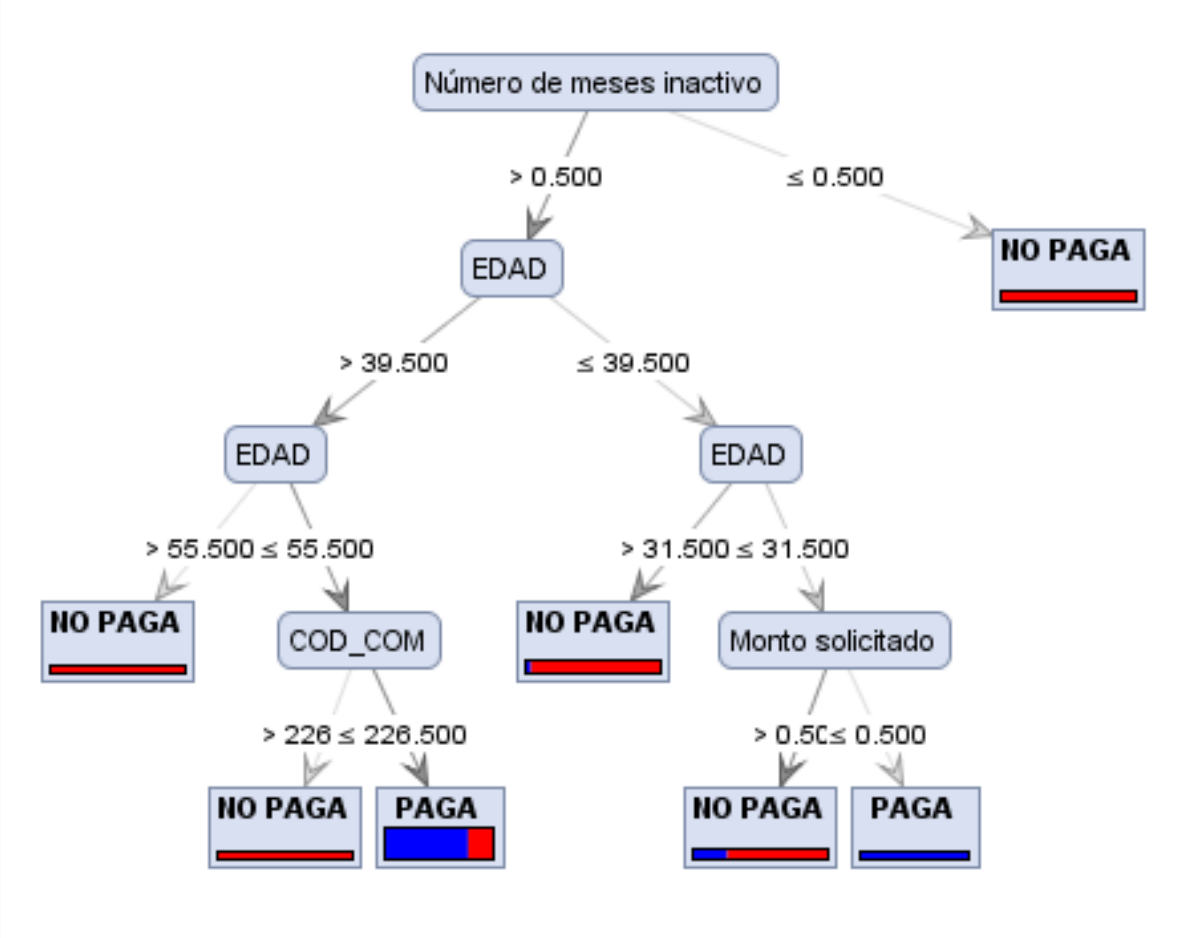
Algoritmos de Machine Learning



http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

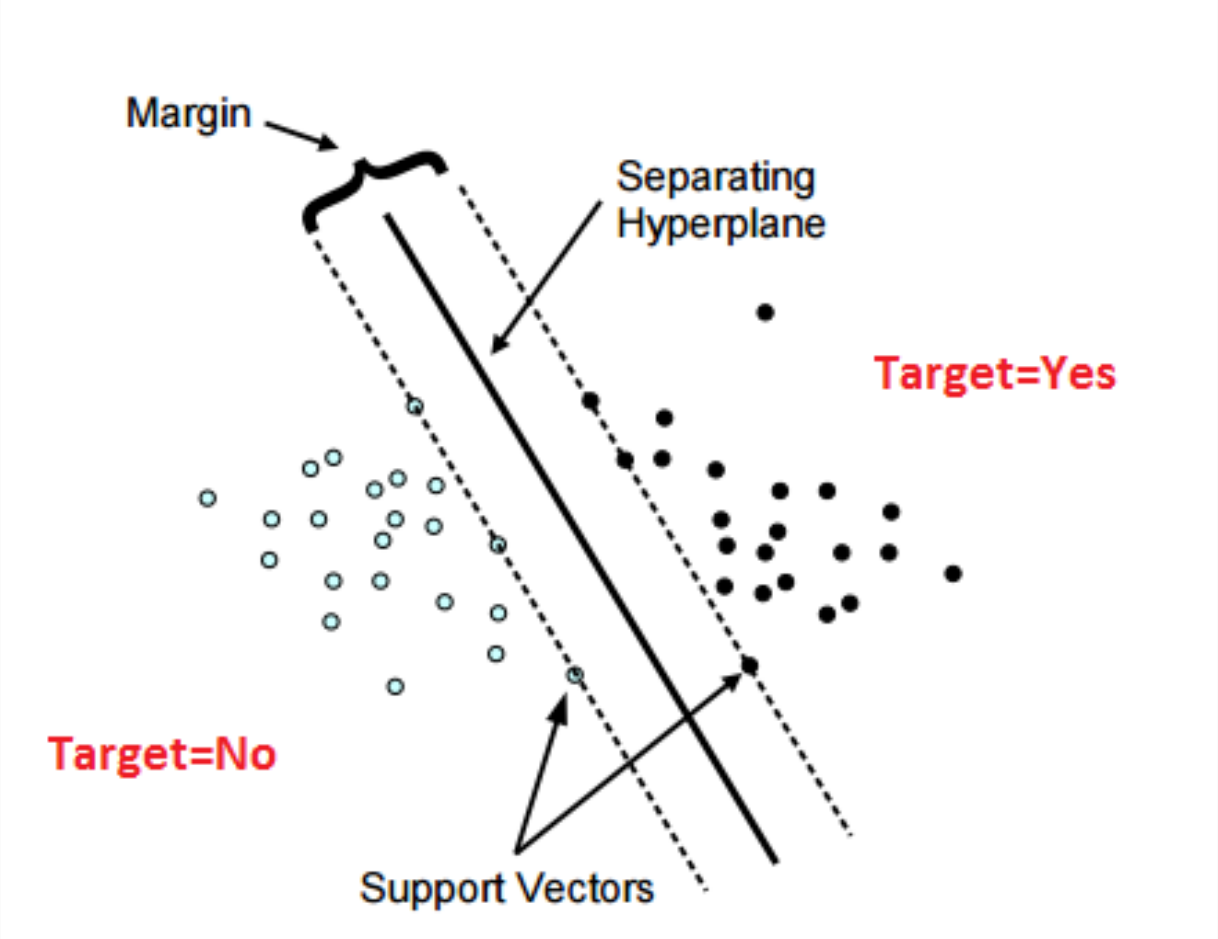
Algoritmos de Machine Learning

Arboles de decisión



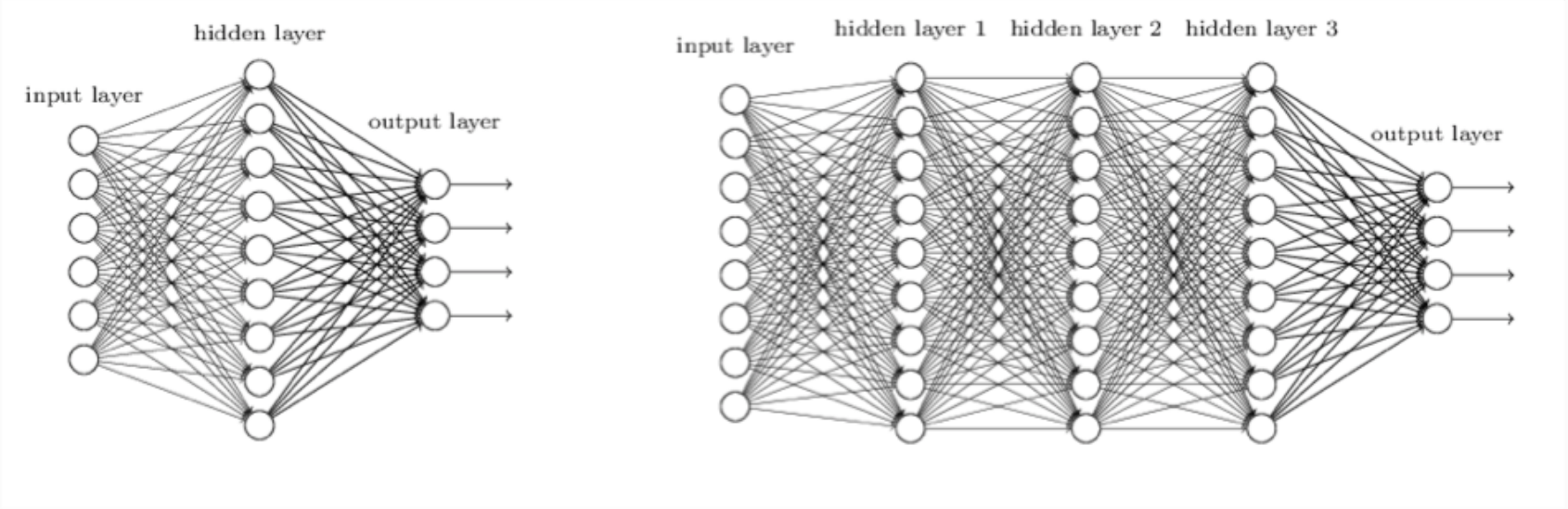
Algoritmos de Machine Learning

Support Vector Machines



Algoritmos de Machine Learning

Redes Neuronales



Preparando el entorno

Preparando el entorno

En Windows abre Anaconda Prompt

En Mac / Linux abre el Terminal

$ conda create –n wwc python=2	Crear un entorno virtual llamado wwc con python 2.7
$ source activate wwc	Activar ese entorno para la sesión actual
$ python –V	Verificar que estamos con la versión de Python correcta
$ pip install sklearn	Instalar Sklearn (librería de Machine Learning)
$ pip install pandas	Instalar Pandas (librería para manipular conjunto de datos)
$ pip install jupyter	Instalar Jupyter (librería para *documentos* con código y texto)
$pip install scipy	Instalar Scipy (librería científica que necesita sklearn)
$ pip install sklearn_pandas	Instalar sklearn_pandas (librería que facilita el preprocesamiento de conjunto de datos)

$ jupyter notebook

¡Comencemos a trabajar!

¡A codear!

¿Qué queremos hacer?

Variable Independiente

Variable Dependiente

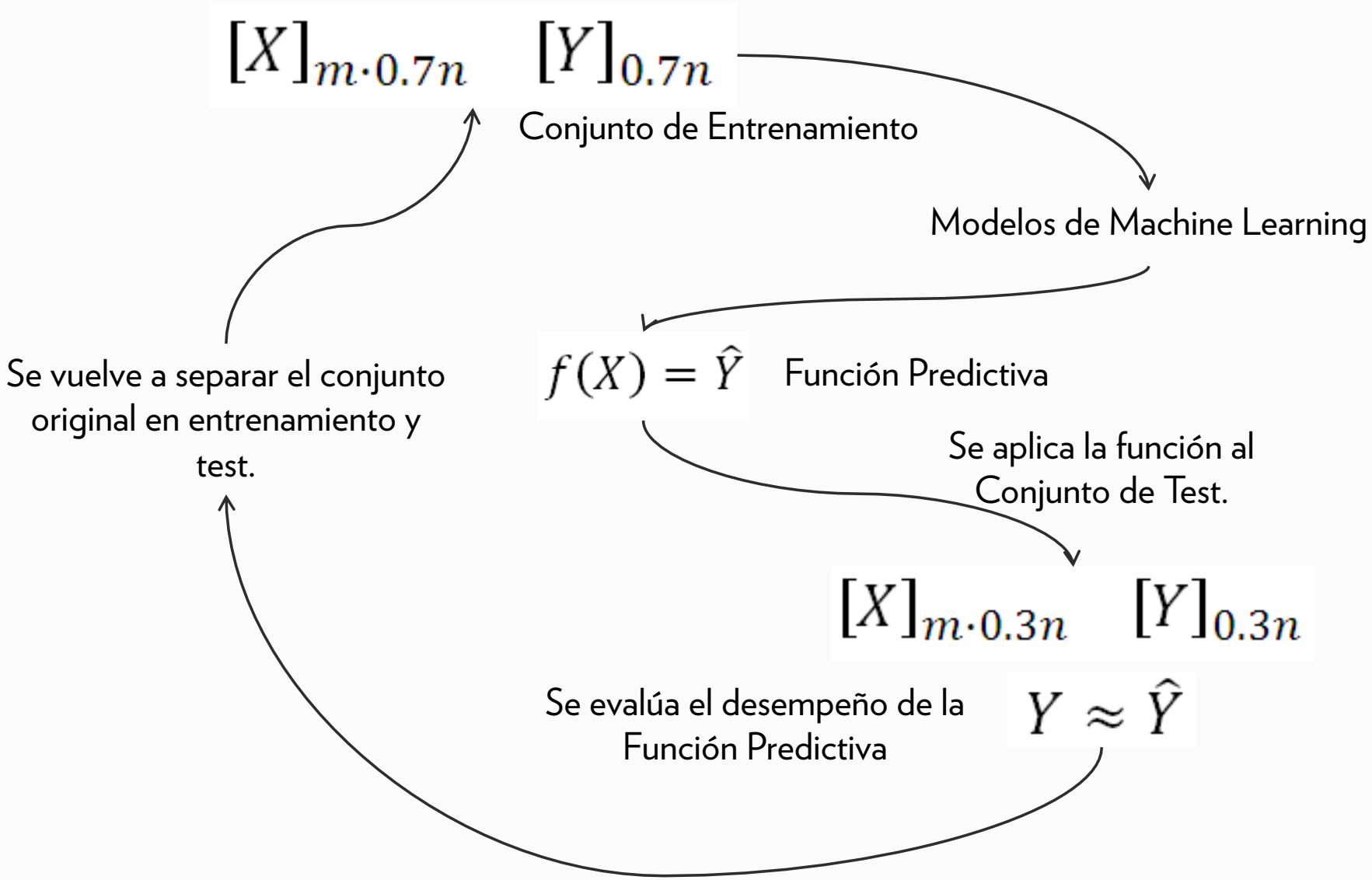
$$[X_n][Y]$$

Función predictiva

$$F(X) \rightarrow \hat{Y}$$

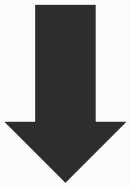
¿Cómo determinar cual es el mejor modelo?

Validación Cruzada



¿Qué hacer con variables categóricas?

day_of_week
lunes
martes
miercoles
jueves
viernes
sabado
domingo



day_of_week_lunes	day_of_week_martes	day_of_week_miercoles	day_of_week_jueves	day_of_week_viernes	day_of_week_sabado	day_of_week_domingo
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1

Datalized

Solving Problems. With Data.

Women Who Code
“Tu primer modelo de Machine Learning”

Matías Sánchez Cabrera
matias@datalized.cl
+56 9 926 99 253