Hello Hello

#HackemCON2020

Angelica Landazabal

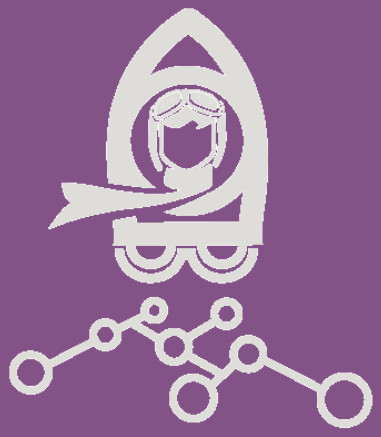
- Computer Science Engineering and a technology lover. (Karateka too)
- I like the development of web applications, marketing and, of course, data analysis ;)
- It motivates me to continue learning about various IT topics, personal / professional growth and helping others so they are also motivated to learn.
- "Keep going" is my motto :)

Angelica.landazabal@hackem.org

ALandazabal

@ALandazabal15

django girls Colombia



PionerasDev

Hackem
Cybersecurity Research Group

Platzi


SciPy
LATAM·2019
7ª CONFERENCIA
LATINOAMERICANA DE
PYTHON CIENTÍFICO
OCT 8 - 10, 2019 • BOGOTÁ, COLOMBIA • CONF.SCIPYLA.ORG

Lindsey Heagy

Travis Oliphant

Damián Avila

https://www.youtube.com/channel/UClb88lwUvlFikmhTzVGsVGA/featured

"Modelos de Clasificación con Python"

¿Qué es clasificar?

Objetivos:

- Entender qué son los modelos de clasificación y sus tipos.
- Conocer el modelo Naïve Bayes.
- Conocer el modelo árboles de decisión.
- Conocer el modelo random Forest.

¿Qué necesitamos?

colab

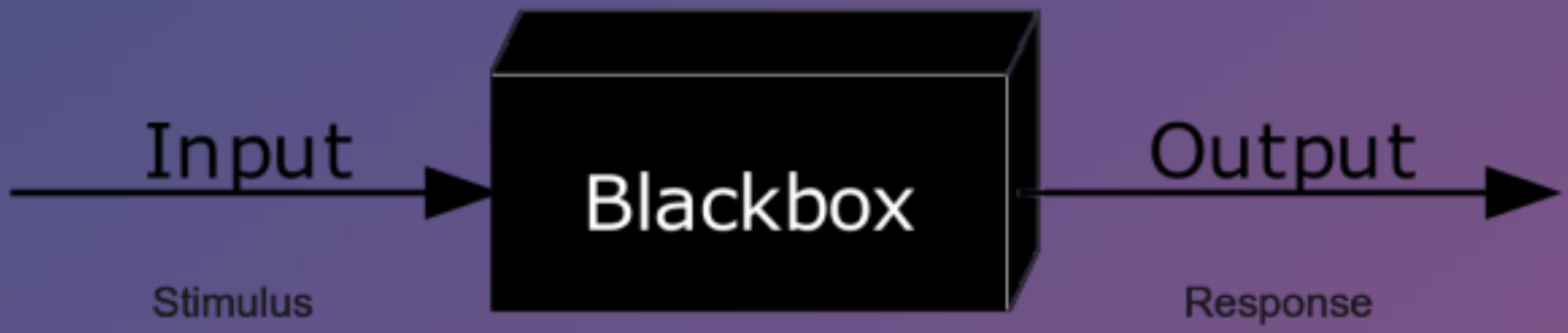
https://github.com/ALandazabal/

¿Qué es clasificar?

Es la tarea de asignar objetos a una categoría predefinida.

Es un problema penetrante que engloba muchas aplicaciones diversas. Un ejemplo incluye detectar mensajes de spam en emails basados sobre la cabecera y el contenido del mensaje.

Modelo de clasificación



- **Modelo descriptivo: Un modelo de clasificación puede servirse como una herramienta explicativa para distinguir entre objetos de diferentes clases.**

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

The vertebrate data set (Tan et al.)

- **Modelo predictivo: Un modelo de clasificación puede ser usado para predecir las etiquetas de las clases de registros desconocidos.**

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
gila monster	cold-blooded	scales	no	no	no	yes	yes	?

Characteristics of a Gila monster.(Tan et al.)

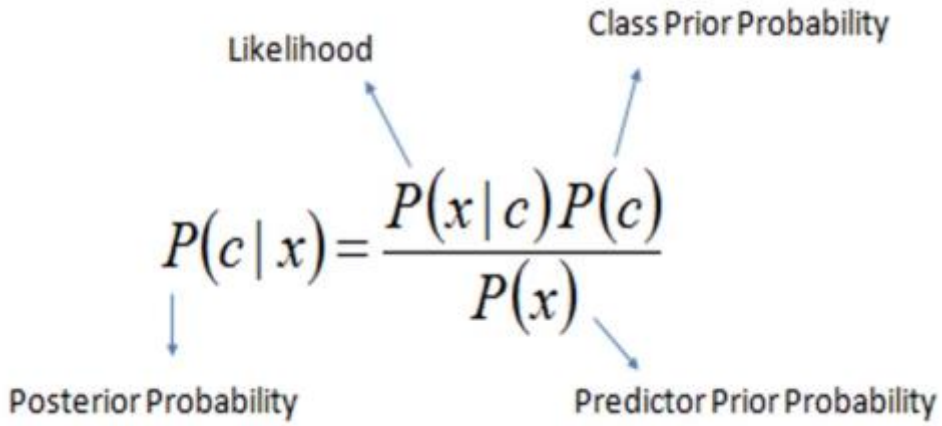
Técnicas de clasificación

,

Naive Bayes

"Un clasificador Naive Bayes estima la probabilidad de la clase condicional asumiendo que los atributos son condicionalmente independientes".

Tan et al (2006)



,

Naive Bayes

Total	Spam		Ham	
100	70		30	
gratis	55	0,78	5	0,16
productos	40	0,57	10	0,33
…	…	…	…	…
gratis y productos	31	0,44	1,5	0,05

$$P(Spam \mid \text{gratis y productos}) = \frac{31}{31 + 1{,}5} = 0{,}95 \Rightarrow 95\%$$

$$P(Ham \mid \text{gratis y productos}) = \frac{1{,}5}{31 + 1{,}5} = 0{,}05 \Rightarrow 5\%$$

,

Naive Bayes

- **Bolsa de palabras:** Es una representación matemática de prueba de un conjunto de datos. Este contiene todas las palabras únicas de un documento y la frecuencia de ocurrencia de cada uno.

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1

,

Árbol de decisión

Es una estructura jerárquica que consiste en una serie de nodos hasta llegar al borde.

El árbol posee 3 tipos de nodos:

- Nodo raíz: Es el nodo superior siendo el único que no posee entradas y puede tener cero o más salidas hacía otros nodos.
- Nodos internos: cada nodo posee exactamente una entrada de información y dos o más salidas hacía otros nodos internos u hojas.
- Hojas o nodos terminales: cada uno tiene exactamente una entrada y ninguna salida.

,

Árbol de decisión

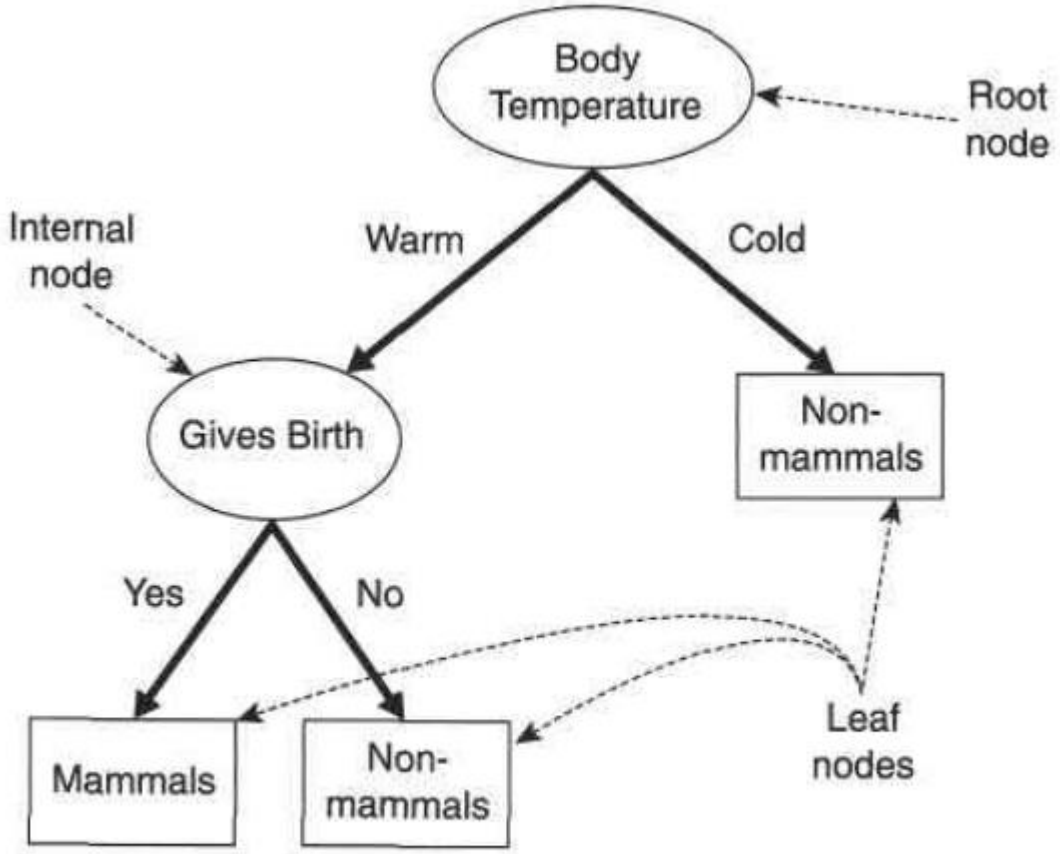
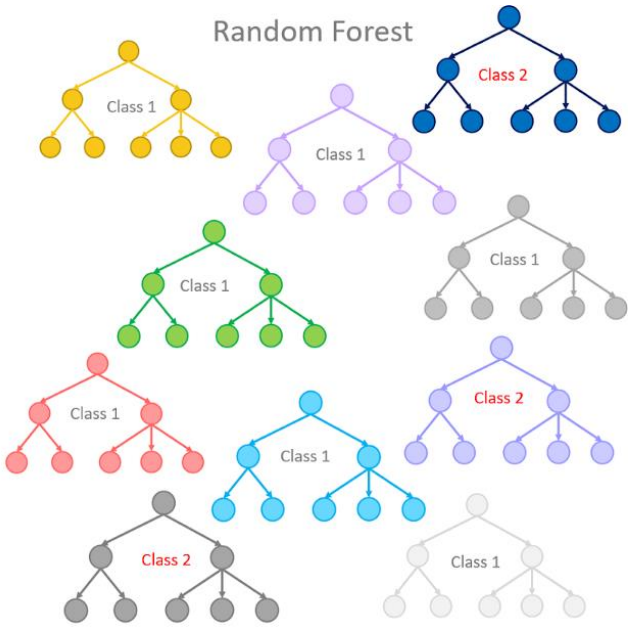


Figure 4.4. A decision tree for the mammal classification problem.

,

Random Forest

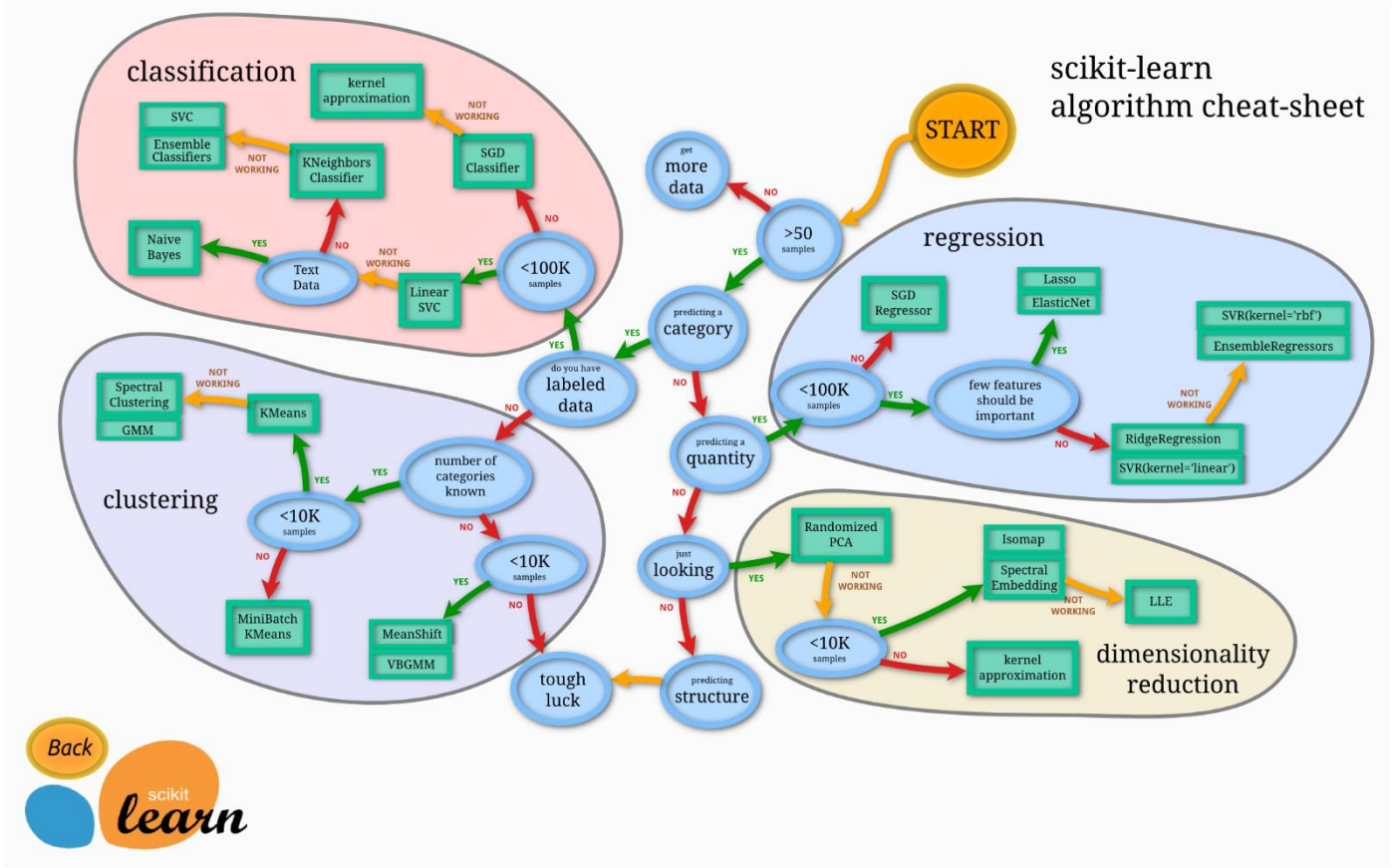
Es considerado un ensamble de diversos árboles de decisión donde cada árbol es generado en base a los valores de un conjunto independiente de vectores aleatorios.



Otras técnicas de clasificación

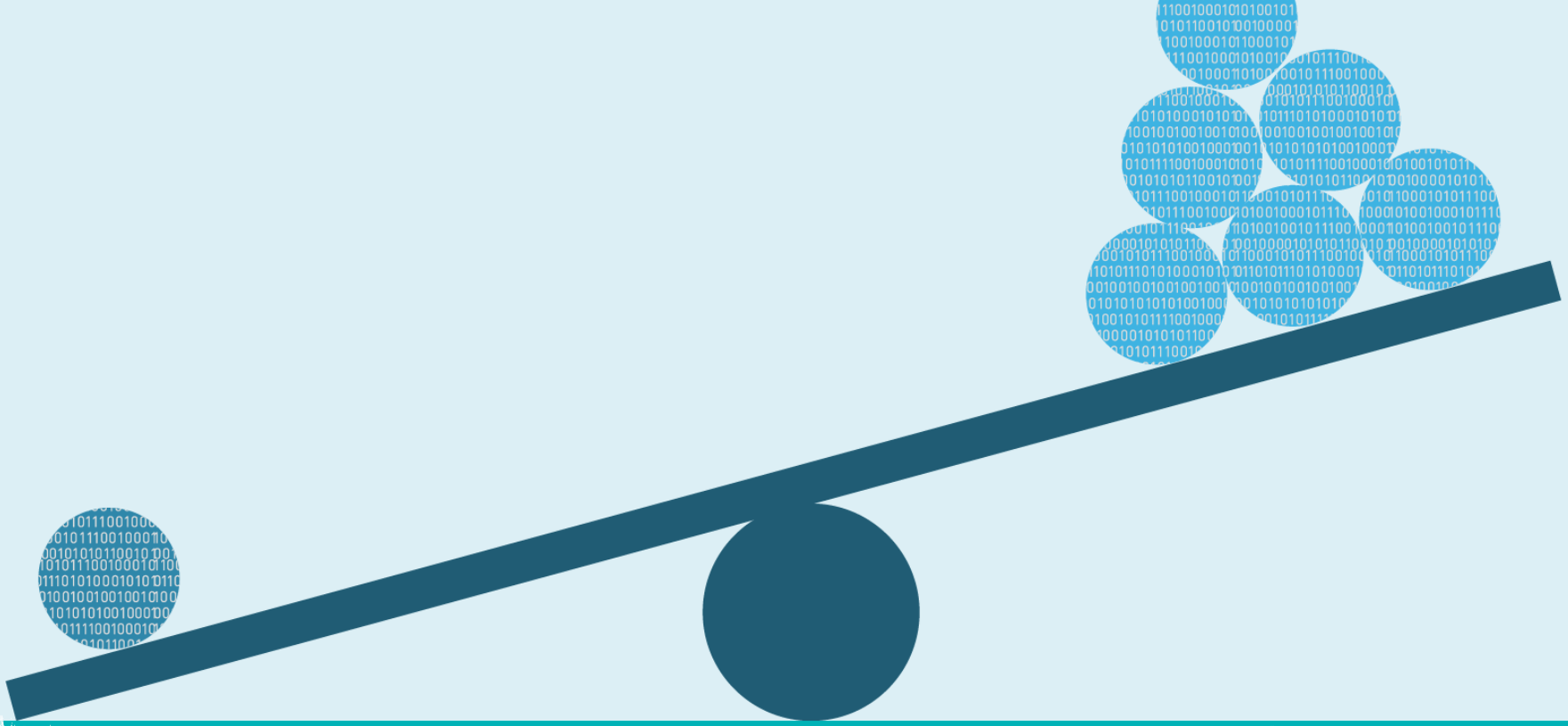
- Clasificadores lineales.
- Regresión logística.
- Vecinos más cercanos.
- Support Vector Machines.
- Stochastic Gradient Descent.
- Neural network models (Supervised).

Otras técnicas de clasificación



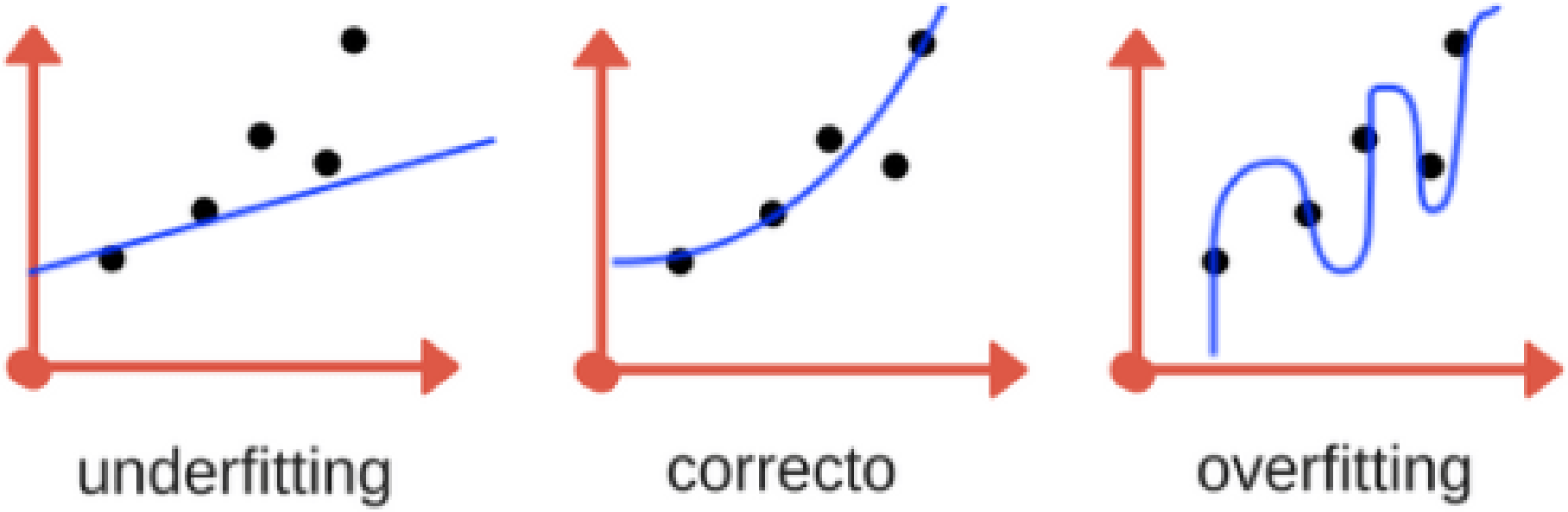
Objetivos:

- Entender qué son los modelos de clasificación y sus tipos. ✓
- Conocer el modelo Naïve Bayes. ✓
- Conocer el modelo árboles de decisión. ✓
- Conocer el modelo Random Forest. ✓



A tener en cuenta

- Desbalanceo: Esto es uno de los problemas comunes cuando trabajamos con clasificación de datos donde solemos encontrar que en nuestro conjunto de datos de entrenamiento contamos con que alguna de las clases cuenta con muy pocas muestras.



A tener en cuenta

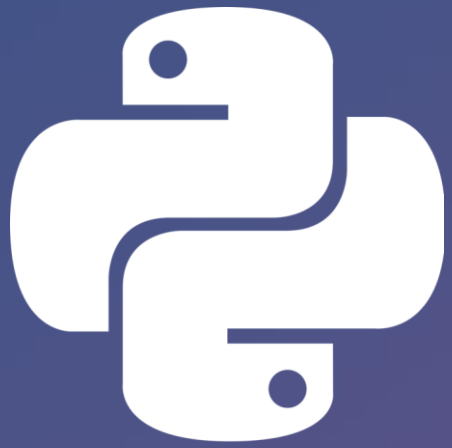
- Overfitting: se da cuando el modelo es demasiado complejo y lo que hace es ajustarse demasiado cerca a los datos de entrenamiento.
- Underfitting: Ocurre porque el modelo aún tiene que aprender la verdadera estructura de los datos.

Referencias

- Tan, P., Steinbach, M. y Kumar, V. (2006). Introduction to Data Mining. Boston, Estados Unidos de América: PEARSON
- Raschka, S. (2015). Python Machine Learning. Birmingham, Reino Unido: Packt Publishing
- Richert, W. y Coelho L. (2013). Building Machine Learning Systems with Python. Birmingham, Reino Unido: Packt Publishing
- Parra F. (2019) Estadística y Machine Learning con R. Recuperado de https://bookdown.org/content/2274/portada.html
- Bagnato J. (s/f) Aprende Machine Learning antes de que sea demasiado tarde. Recuperado de https://www.aprendemachinelearning.com/
- Barrios J. (2019) La matriz de confusión y sus métricas. Recuperado de: https://www.juanbarrios.com/matriz-de-confusion-y-sus-metricas/

Referencias

- Matriz de confusión: https://es.wikipedia.org/wiki/Matriz_de_confusi%C3%B3n#:~:text=En%20el%20campo%20de%20la,se%20emplea%20en%20aprendizaje%20supervisado.
- Curva ROC: https://es.wikipedia.org/wiki/Curva_ROC
- Credit Card Fraud Detection: https://www.kaggle.com/mlg-ulb/creditcardfraud/data
- Scikit-learn: https://scikit-learn.org/stable/index.html
- Curso Machine Learning Aplicado con Python (Platzi): https://platzi.com/clases/scikit/
- Curso de Introducción al Pensamiento probabilístico (Platzi): https://platzi.com/clases/probabilistica/



“Las personas mienten, los datos no”
PlatziConf 2019


angelica.landazabal@hackem.org


ALandazabal


@ALandazabal15