

DIPLOMADO EN CIENCIA DE DATOS

MODELACIÓN SUPERVISADA SEMANA 1

Facultad de Estudios Superiores Acatlán

OUTLINE

Módulo 2 - Semana 1

- 1.- Objetivos y logística del curso
- 2.- Modelos supervisados en la industria
- 3.- Metodología de modelación supervisada
- 4.- Métricas de ajuste
- 5.- Optimización de Funciones de Costo

1.- Objetivos y logística del curso

OBJETIVOS Y LOGÍSTICA DEL CURSO



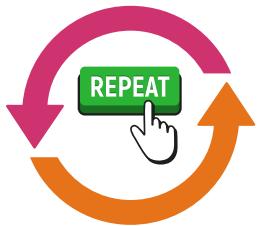
Que el alumno aprenda la teoría y práctica detrás de los modelos supervisados y sepa cómo emplearlos en la industria



OBJETIVOS

Más específicamente...

- Estructurar un proyecto de modelación supervisada
 - Preparación de datos
 - Entrenamiento del modelo
 - Evaluación del modelo



Pasar una entrevista de Ciencia de Datos

LOGÍSTICA

¿Dónde y cuándo?

- Remoto
- Jueves y Sábados (12 sesiones en 6 semanas)

Forma de evaluación

- 4 quizzes (10%)
- 1 tarea (20%)
- 1 examen Final (30%)
- 1 avance de proyecto Diplomado (40%)

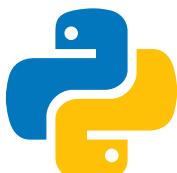
HERRAMIENTAS

Teoría

- Presentaciones

Programación

- Python
- Jupyter Notebooks



PROYECTO FINAL



Presentaciones (10%)

- 6 minutos cada uno más 1.5 minutos de preguntas
- Presentación debe contener lo más relevante de su proyecto final (a lo más 6 slides)
- Van a pasar por bloques de personas (Más información en classroom)

Reporte final (30%)

- Reporte escrito (de preferencia en Latex)
- Máximo 15 cuartillas (sin contar lo del módulo 1)
- Portada
- Índice
- Contenido de reporte final: Describir los pasos y principales resultados de desarrollar un modelo supervisado con el set de datos que eligieron
- Una sección para cada parte de la metodología CRISP-DM (Si están utilizando el mismo set de datos que en el módulo uno, presentar esos resultados en el reporte en lugar de rehacerlos)
- Ideal: Probar varios modelos supervisados y seleccionar el mejor
- Citas bibliográficas, Índice de figuras
- Subir reporte y notebook ya ejecutado con todos los resultados del proyecto (solo módulo dos)

CALENDARIO DE ENTREGAS



Feb

S	M	T	W	T	F	S
			1	2	3	4
6	7	8	9	10	11	12

1ra
semana 13 14 15 **16** 17 18 19

2da
semana 20 21 22 23 24 25 **26**

27 28

Deadline tareas

Aplicación de quizzes

Examen final

Presentaciones proyecto final

Mar

S	M	T	W	T	F	S
			1	2	3	4
6	7	8	9	10	11	12

3ra
semana 13 14 15 16 17 18 **19**

4ta
semana 20 21 22 23 24 25 **26**

5ta
semana 27 28 29 30 31

Deadline

Entrego

Deadline tareas

- 1ra: 20 marzo 23:59 PM

Quizzes

- 1ro: 26 febrero 08:15 AM
- 2do: 5 marzo 08:15 AM
- 3ro: 12 marzo 08:15 AM
- 4to: 19 marzo 08:15 AM

Examen final

- Entrego el 24 marzo 08:30 AM
- Deadline: 27 marzo 11:59 PM

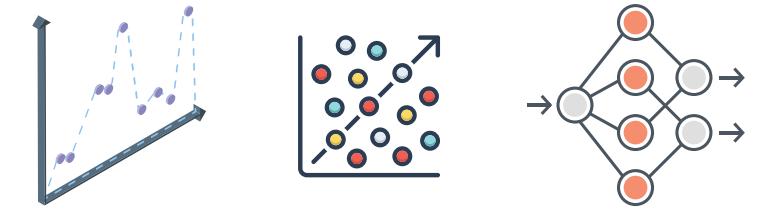
Proyecto final

- Presentaciones:** 26 marzo 2022
- Entrega reporte final:** 3 abril 2022 23:59 PM

TEMARIO



Temas que veremos a lo largo del módulo dos



Modelación supervisada

- Metodología de modelación supervisada y métricas de ajuste
- Regresión lineal
- Optimización de funciones de costo
 - Gradiente descendente
 - Gradiente estocástico descendente
- Reportes de estabilidad y desempeño
- Modelos lineales
 - Regresión Lazo
 - Regresión LARS
 - Regresión Ridge
 - Red elástica
 - Regresión Logística
 - Regresión de cresta Bayesiana
 - Regresión Bayesiana
- Análisis Discriminante
- Regresión de cresta Kernel
- Máquinas Vector Soporte
- Vecinos más cercanos
- Bayes Ingenuo
- Árboles de decisión
- Redes Neuronales
- Ensambles
- Bosque Aleatorio
 - ADABoost
 - Impulso de Árboles Gradiente
 - Clasificador Votante
 - Impulso gradiente extremo

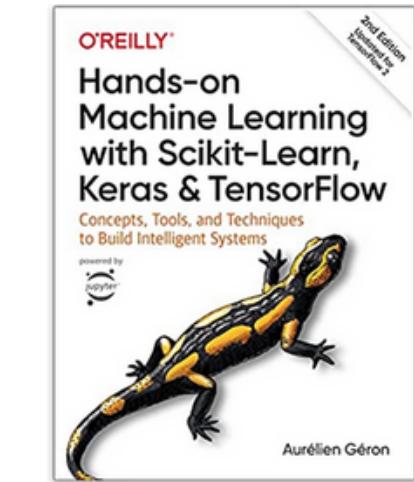
BIBLIOGRAFÍA



Libros a consultar para el módulo dos

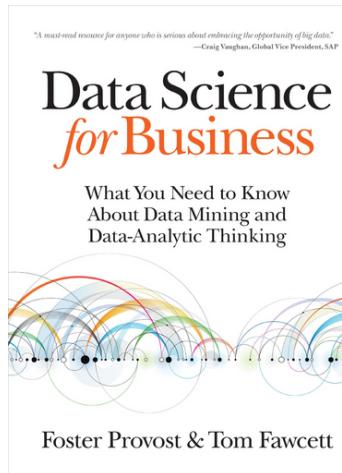
Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow

(Aurélien Géron, 2019)



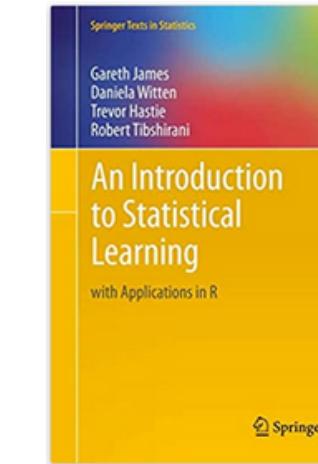
Data Science for Business

(Provost & Fawcett, 2016)



An Introduction to Statistical Learning

(James, Witten, Hastie, and Tibshirani, 2013)



Recursos en línea

(StackOverflow, TowardsDataScience, Medium, Kaggle, etc)

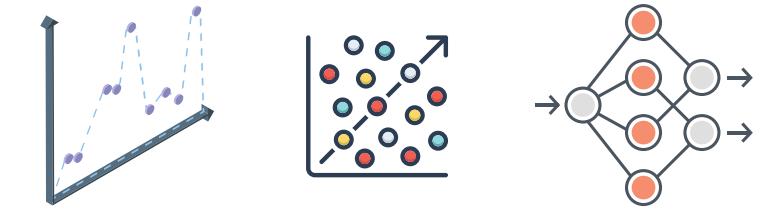


towards
data science

CURSOS RECOMENDADOS



Lista de cursos que recomiendo para ser Data Scientist



Introduction to Git

<https://www.datacamp.com/courses/introduction-to-git>

+0.5
puntos

Introduction to SQL

<https://www.datacamp.com/courses/introduction-to-sql>

+0.25
puntos

Introduction to Python

<https://www.datacamp.com/courses/intro-to-python-for-data-science>

+0.25
puntos

Data Scientist with Python - career track

<https://www.datacamp.com/tracks/data-scientist-with-python>

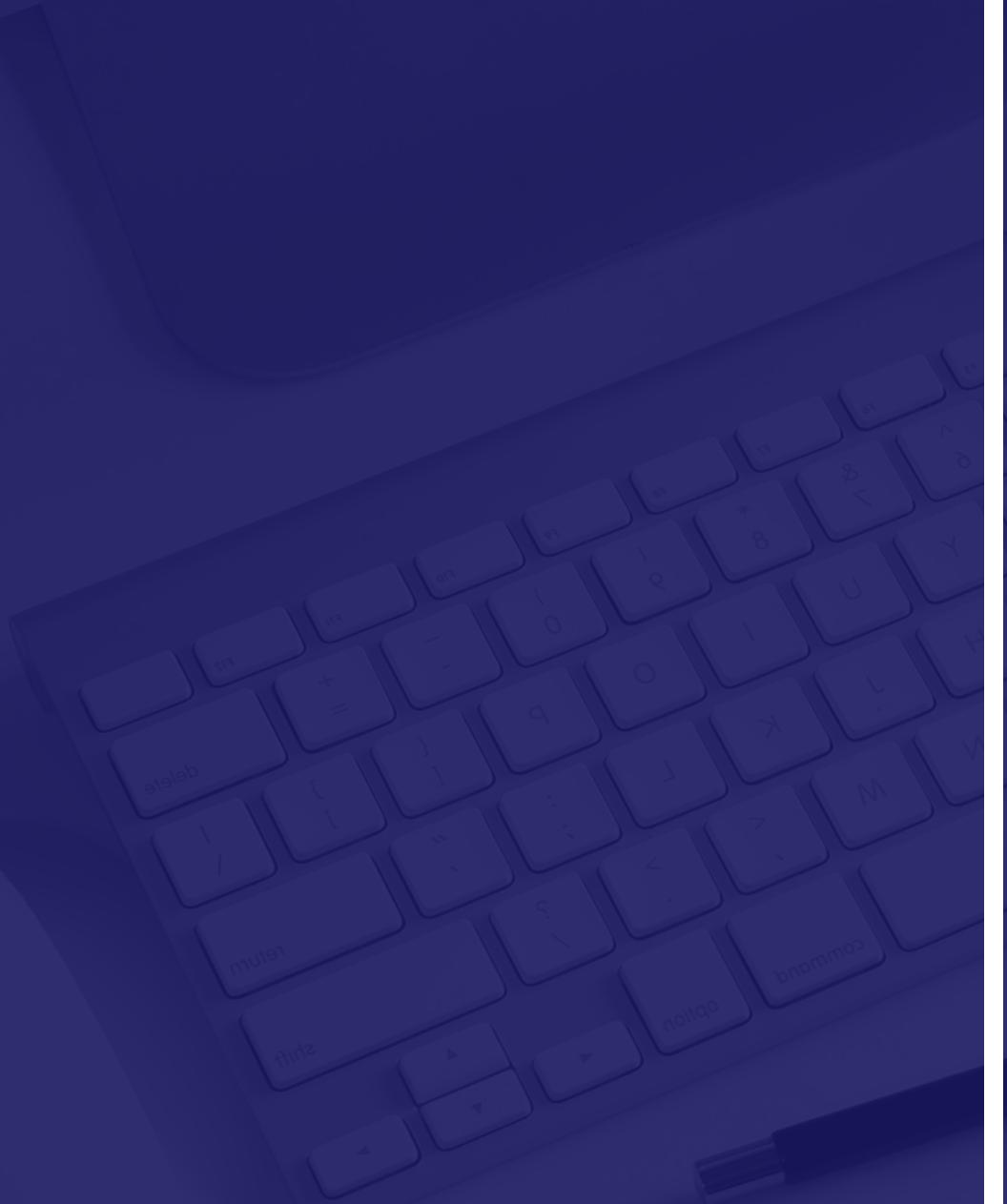


Si completan los tres cursos,
tendrán un punto extra
sobre calificación final

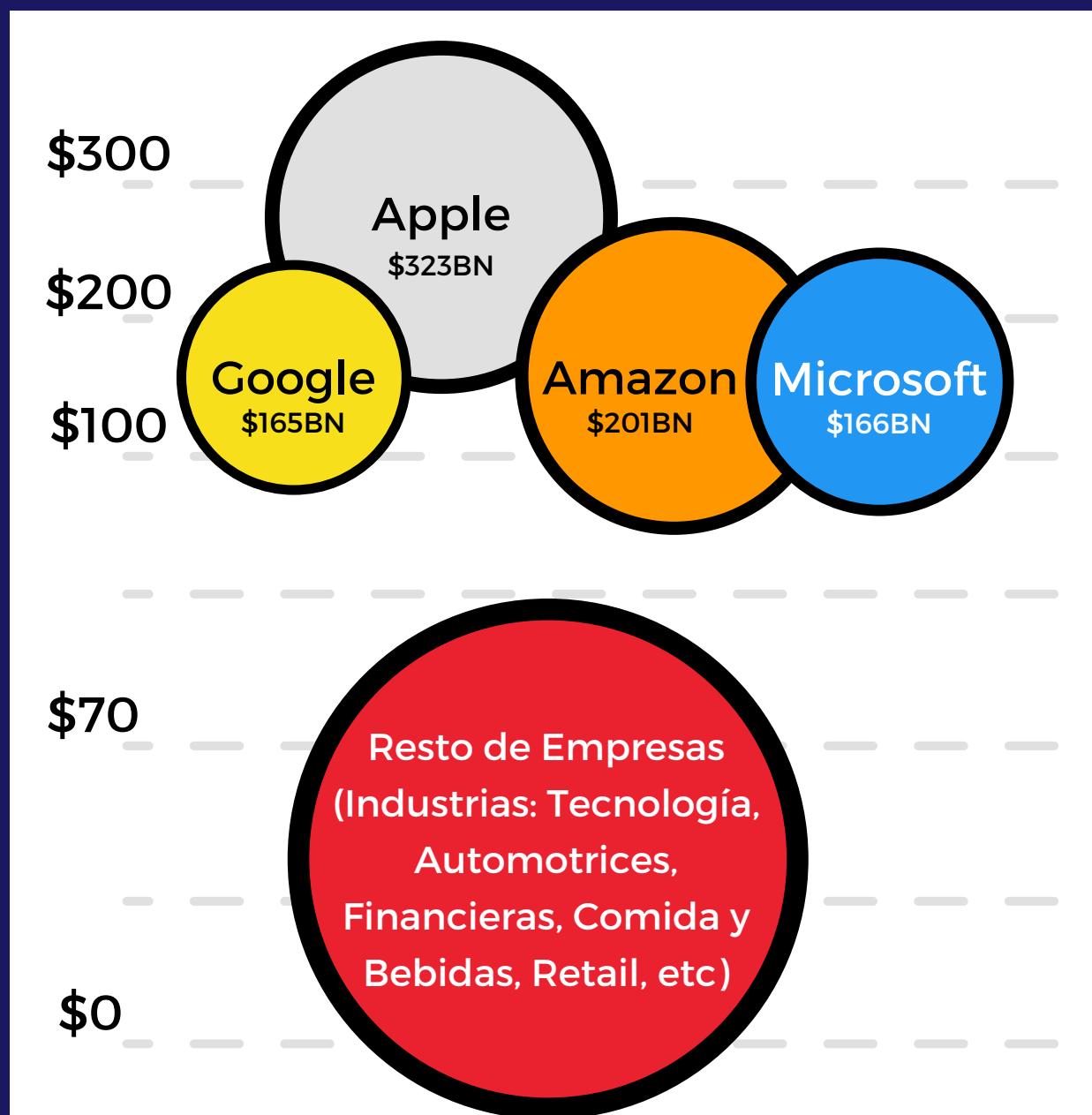
+1
punto

↓
Deberán entregar el
certificado original con su
nombre (crearé espacio en
Classroom)

2- Modelos supervisados en la industria



TOP 50 GLOBAL BRANDS MARKET CAP*



¿QUÉ HACEN LAS TOP 4 EMPRESAS?

DECISIONES DATA-DRIVEN

Decisiones basadas en análisis de datos y aprendizaje de máquina

EXPERIENCIA DIGITAL HYPER-PERSONALIZADA

Modelos para saber lo que el cliente necesita, cuándo lo necesita y de qué manera

MONETIZACIÓN DE DATOS

Productos personalizados para cada cliente que permitan una mayor monetización

INSTITUCIONES FINANCIERAS DATA-DRIVEN

Hasta hace unos años las técnicas de Aprendizaje de máquina e Inteligencia Artificial eran exclusivas de las grandes empresas tecnológicas. Otras industrias se dieron cuenta de su potencial... **incluidas las Instituciones Financieras**



FINTECHS

Nacieron Digitales

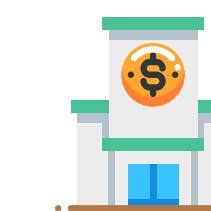
Aprendizaje de máquina y Data Science en su ADN



Revolut

N26

Experiencias hyper-personalizadas en app móvil y web



BANCA TRADICIONAL

En evolución digital

Pocos Bancos utilizan técnicas de Aprendizaje de máquina de manera automatizada



BBVA



SMS con ofertas al hacer una transacción

Experiencias hyper-personalizadas en app móvil

Asistente virtual que brinda asesoría financiera personalizada

VENTAS PERSONALIZADAS

Ofrecer la oferta ideal, en el momento adecuado, por el canal preferente...
de manera automatizada



Ventas por eventos

- Ofertas automáticas cuando se cumpla una regla de negocio
- Compras por eventos, modelo de machine learning para calificar en tiempo real los movimientos del cliente y enviar la oferta ideal por el canal preferido



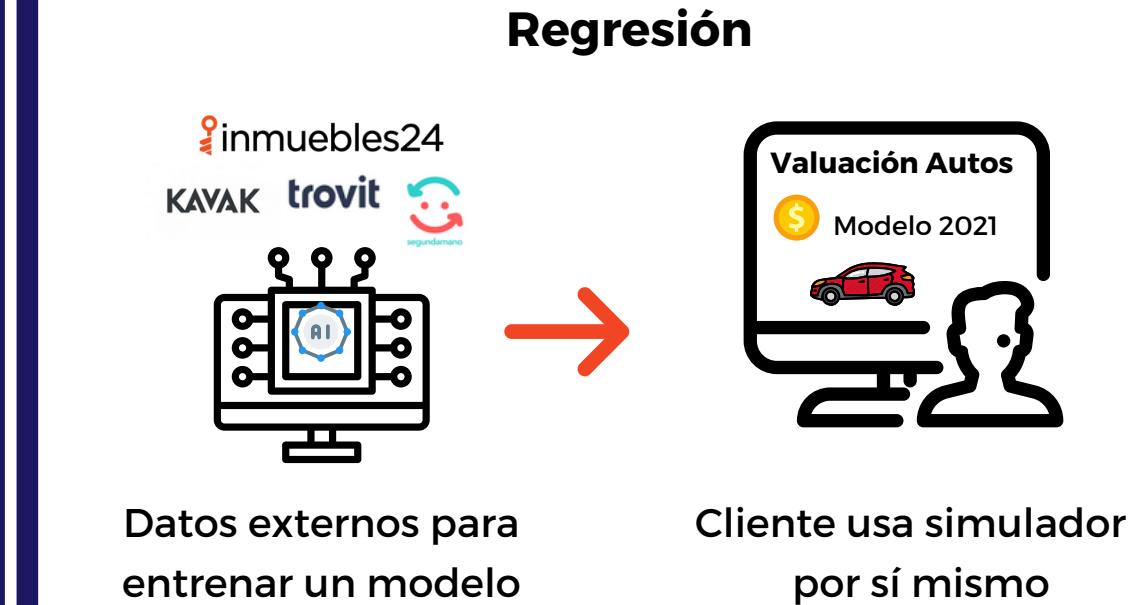
Ventas Omnichannel

- Detectar cuando un cliente ve una oferta pero no la toma. Gestión automatizada de los clientes por medio del canal óptimo para concretar la venta



Simuladores Inteligentes

- Utilizar datos de diferentes sitios de inmuebles y ventas de autos para entrenar un modelo supervisado y dar herramientas a los clientes para valuar su patrimonio



EXCELENCIA EN ATENCIÓN A CLIENTES



Escucha activa de la voz del cliente para detectar puntos de dolor y solucionarlos. Con base en lo anterior, dar la mejor atención a clientes y mejorar índice NPS

Escucha activa de la voz del cliente

- Análisis de texto y modelos supervisados para analizar comentarios y reseñas que los clientes de una empresa hacen respecto a los productos y/o servicios de la misma



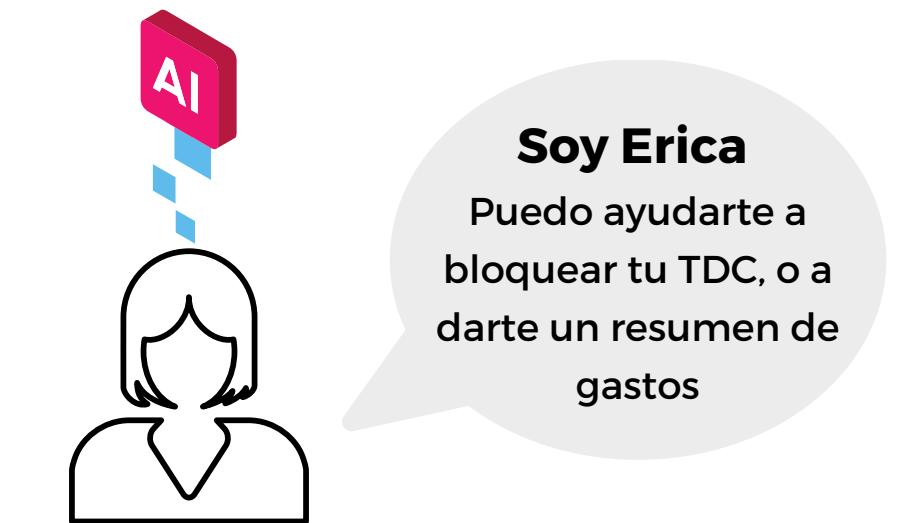
Twitter para detectar incidentes

- Escucha activa en tiempo real de los comentarios que los clientes de una empresa respecto a sus productos y/o servicios



Asistentes virtuales (muy) inteligentes

- Conocer tan bien al cliente que la empresa pueda tener una asistente virtual personalizada para cada cliente



EXPERIENCIAS PERSONALIZADAS

RESUMEN FINANCIERO INTERACTIVO

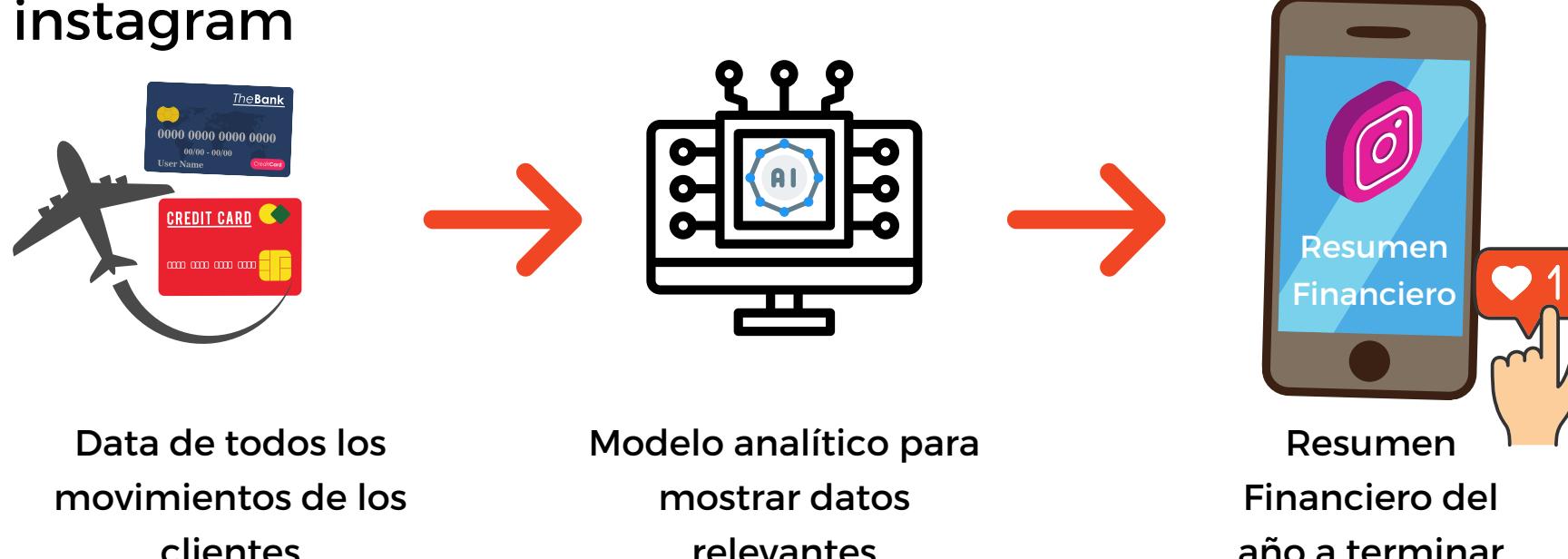
Analiza todos los movimientos en las cuentas de los clientes y muestra un resumen del año en formato de historias de Instagram



Resumen financiero de final de año

Usuario: hace compras en supermercados, en Amazon, paga en diferentes países con su tarjeta de débito

Banco: recopila toda la información financiera del usuario. Muestra datos relevantes en formato de historias de instagram

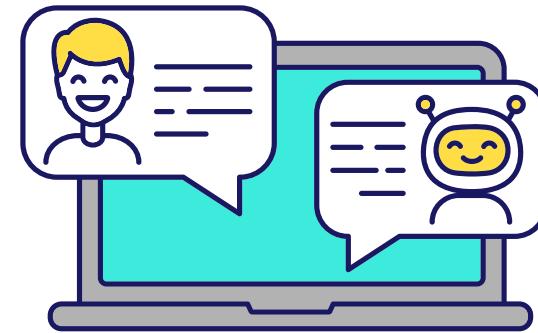


<https://www.youtube.com/watch?v=C6F0ghUMI5Q>

Revolut ya lo hace

EXCELENCIA EN ATENCIÓN A CLIENTES

ASISTENTES VIRTUALES (MUY) INTELIGENTES



El Banco conoce tan bien a sus clientes que puede brindarles un asesor financiero en forma de asistente virtual. La atención es personalizada, y funge como reemplazo de un ejecutivo

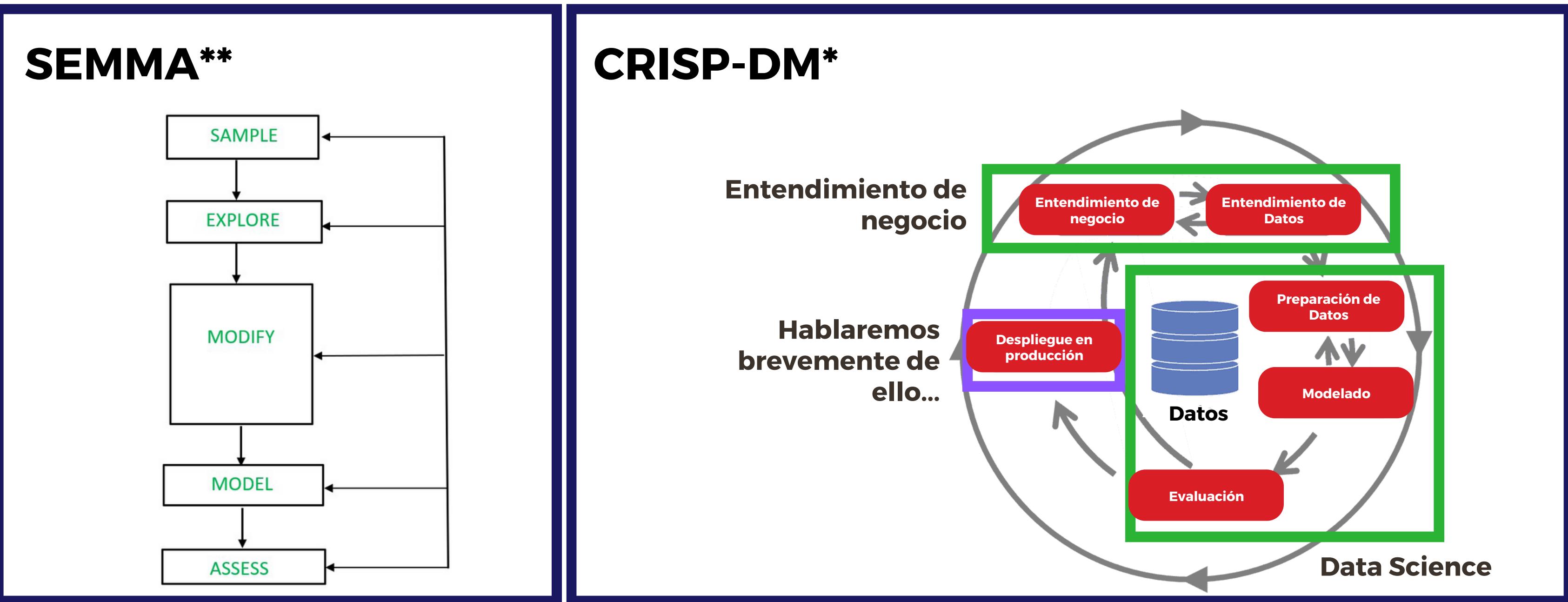
<https://www.youtube.com/watch?v=0Irg83riPzo&t=6s>

3.- Metodología de modelación supervisada y métricas de ajuste

PROYECTO DE MODELACIÓN SUPERVISADA



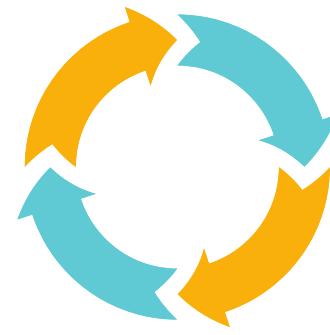
Existen diferentes procesos para llevar a cabo proyectos de modelación supervisada. CRISP-DM y SEMMA son dos procesos ampliamente usados en la industria



*What is CRISP DM?, <https://www.datascience-pm.com/crisp-dm-2/>

**SEMMA Model, <https://www.geeksforgeeks.org/semma-model/>

METODOLOGÍA CRISP-DM

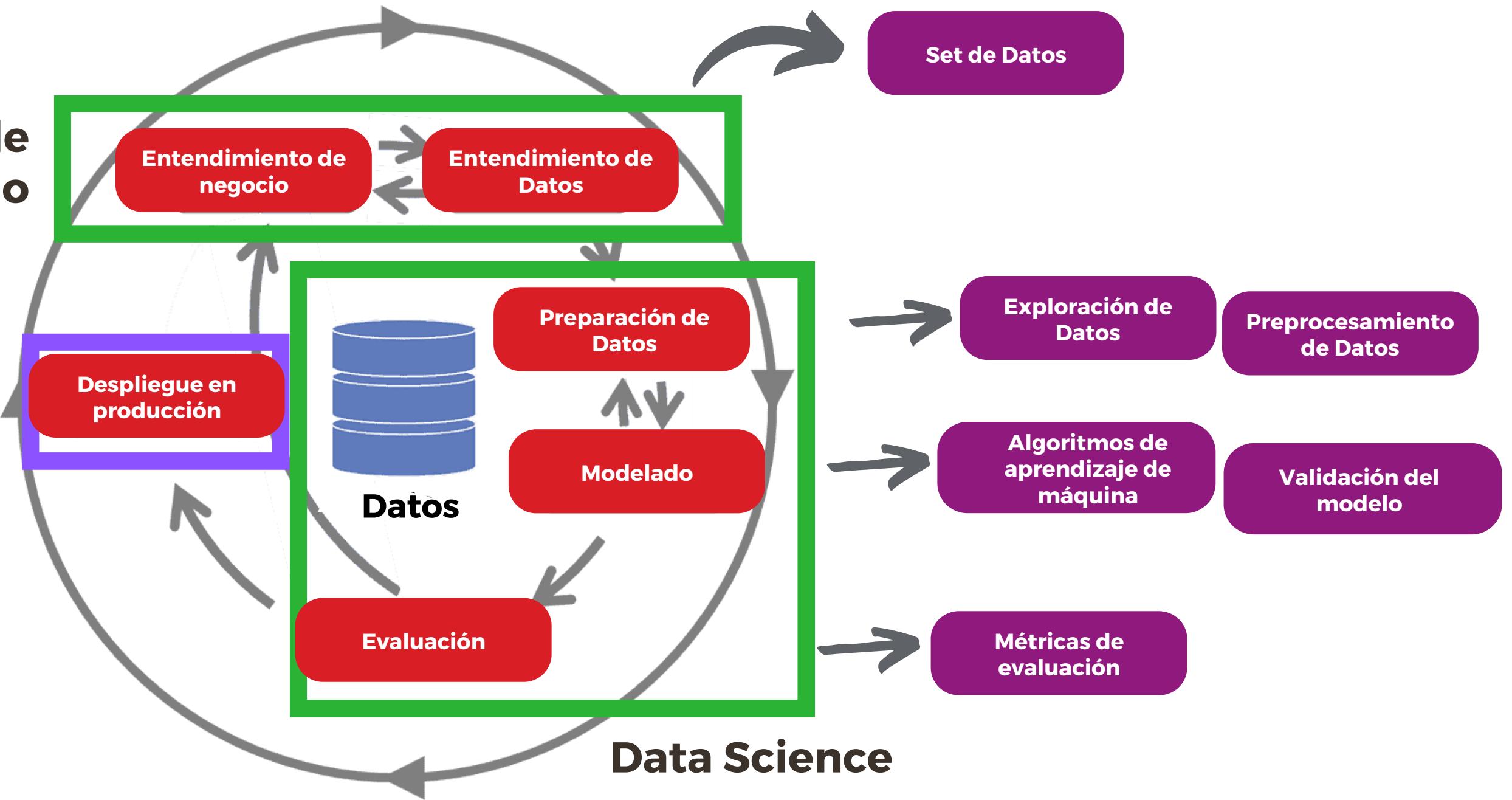


CRISP-DM: Cross Industry Standard Process for Data Mining. Proceso iterativo para llevar a cabo un proyecto de Data Mining

CRISP-DM*

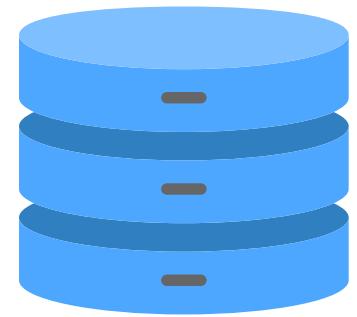
Entendimiento de negocios

Hablaremos brevemente de ello..



SET DE DATOS

DATOS PARA ENTRENAR UN MODELO



Quizás el elemento más importante a la hora de entrenar modelos de aprendizaje de máquina.

Muestra de datos

- En estadística se busca una técnica de muestreo tal que la muestra sea un buen estimador de la población completa
- En Ciencia de datos, comúnmente tenemos tantos datos que pasamos por alto algunas dificultades como lo son el **sesgo de selección**, entre otros.

Características del set de entrenamiento

Set de entrenamiento



Recolección de datos

- Encuestas
- Web scrapping
- Bases de datos de productos de alguna empresa
- Datos de sensores (IoT)

Formato de datos

- Estructurados



- No estructurados

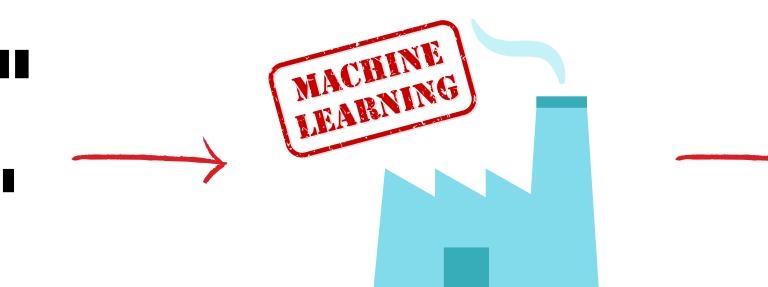


- Los algoritmos de modelación supervisada no van a funcionar bien si el set de entrenamiento es demasiado pequeño, no es representativo, presenta mucho ruido, o está contaminado con variables no relevantes para el fenómeno a predecir



Y RECUERDEN AMIGOS..

"Garbage In"
"Basura Entra"



"Garbage Out"
"Basura Sale"

SET DE DATOS

TIPOS DE VARIABLES



Las variables independientes y la variable dependiente pueden ser de los siguientes tipos

Tipos de variables

Categórica



- Número finito de valores o categorías
 - Binaria
 - Multi-Nominal
 - Ordinal

Numérica



- Continua: Tienen un número infinito no numerable de valores entre dos valores cualesquiera
- Discreta: Presentan un número infinito numerable de valores

Ejemplos

{0,1} {1ro, 2do, 3ro}
{Licenciatura, Maestría, Doctorado}

Ejemplos

Precio de una acción
Balance en tarjeta de crédito
Monto en línea de crédito

Variable dependiente - Y

Regresión

- Continúa

Clasificación

- Categórica

Ejemplos

Estatura, Peso, Precio de un barril de petróleo

Clasificación

Va a contrar o no, Sí o No, Positivo o Negativo

Clasifica las siguientes variables

Clasifica las siguientes variables (Regresión o Clasificación; y el tipo de variable)

- Tipo de cambio
- Grado de estudios
- Defraudador o no
- Categorías a predecir: {a,b,c,d,e}
- Categorías a predecir: {1,2,3,4,5}
- Gato o perro
- Precio de una casa

TAREA

SET DE DATOS

SETS DE ENTRENAMIENTO Y VALIDACIÓN



¿Cómo podríamos asegurar que un modelo va a generalizar bien a nuevos sets de datos?

Partir los datos en dos sets, **entrenamiento y validación**

Construcción del set de datos

- Fuentes de datos distintas
 - Cruzarlas para formar **X**
- Definir variable target **Y**
 - Con base en evento a predecir
- Asociar cada conjunto de variables independientes con su respectiva variable target
- Explorar set final
- Preprocesamiento (limpieza de datos)
- Partir set de datos en dos: Entrenamiento y validación

Set de entrenamiento

- Preprocesamiento (transformación de datos)
- **Entrenar modelo** de aprendizaje supervisado en este set de datos

Set de validación

- Aplicar mismas transformaciones que en el set de entrenamiento (transformación de datos)
- **Evaluar modelo** de aprendizaje supervisado en este set de datos
- Deseable que error sea pequeño. Dicho error se conoce como **estimación del error de generalización**



Si el **error de entrenamiento es pequeño** (modelo comete pocos errores en el set de entrenamiento) **pero la estimación del error de generalización es grande** (modelo comete muchos errores en el set de validación), el modelo no está generalizando bien, si no que está sobre ajustando a los datos de entrenamiento
Y RECUERDEN AMIGOS...

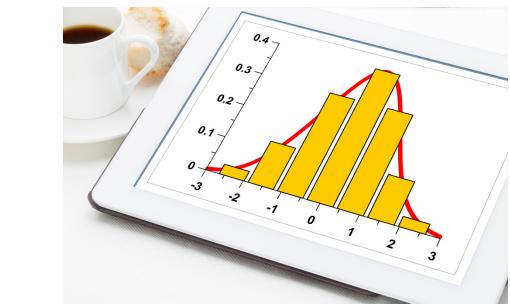
Más de esto
la siguiente
semana

EXPLORACIÓN DE DATOS

ESTADÍSTICA DESCRIPTIVA



Saber acerca de la distribución de datos y de las principales características de cada variable



Medidas de centralización

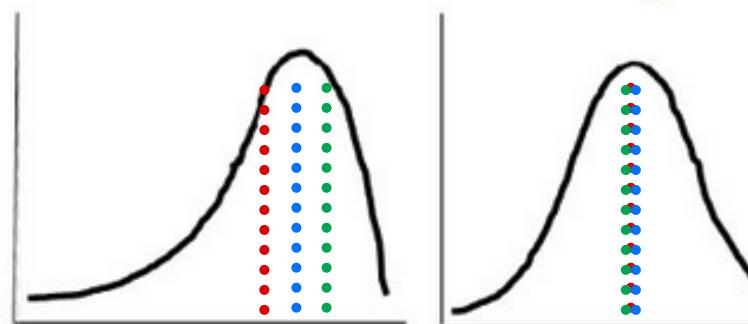
- Ayudan a determinar dónde está el centro de masa de una distribución
 - Media
 - Mediana
 - Moda

$$\bar{X} = \frac{\sum X}{N}$$

● Mediana

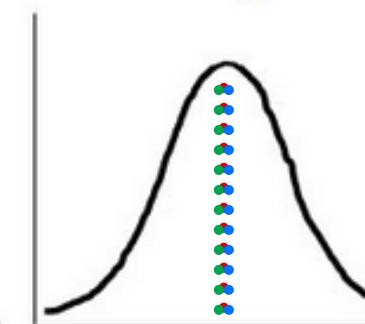
● Media

● Moda



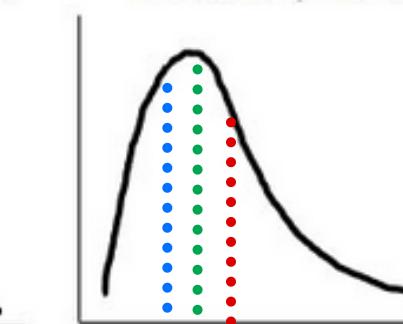
Distribución asimétrica a la izquierda

Media < Mediana



Distribución simétrica

Media = Mediana



Distribución asimétrica a la derecha

Media > Mediana

Medidas de dispersión

- Ayudan a determinar el grado de dispersión alrededor del centro de masa de una distribución
 - Varianza
 - Puede ser resumida por los cuantiles (percentiles, cuartiles)

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$$

Medidas de simetría

- Ayudan a determinar que tan simétrica es una distribución
 - Coeficiente de asimetría de Fisher
 - Coeficiente de asimetría de Pearson

$$\gamma_1 = \frac{\mu_3}{s^3}$$

EXPLORACIÓN DE DATOS

POR TIPO DE VARIABLE

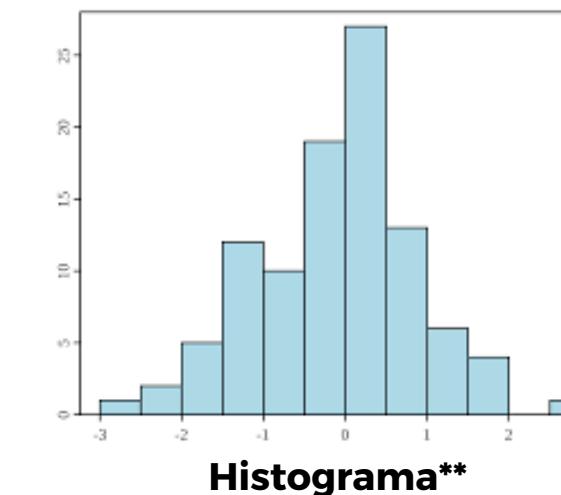
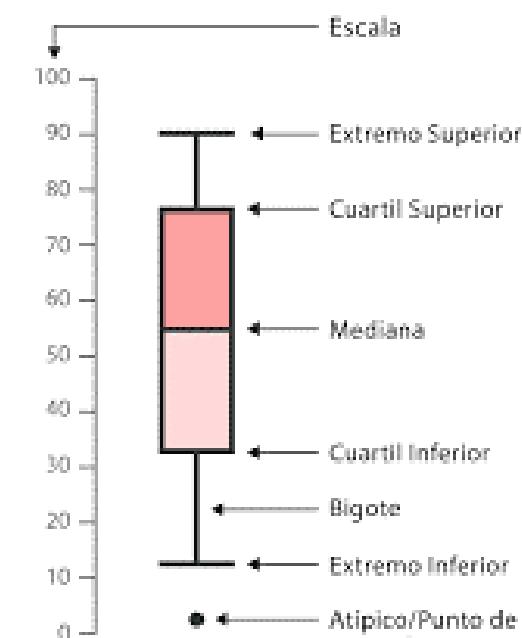


Saber acerca de la distribución de datos y de las principales características de cada variable, por tipo de variable



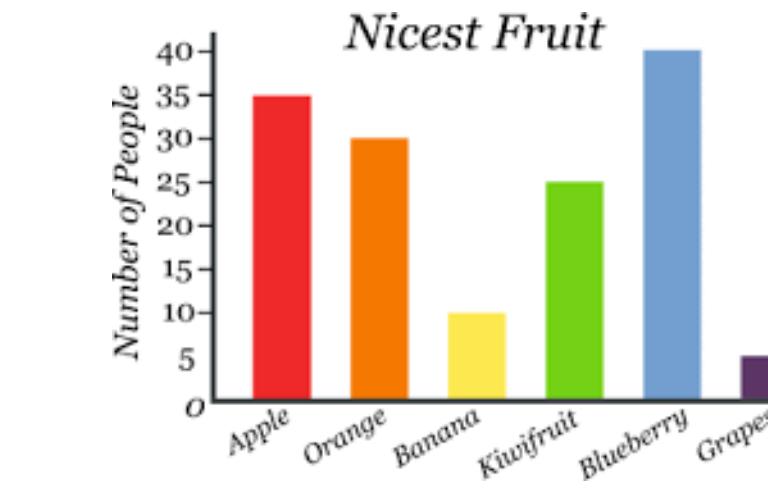
Variables continuas

- Pueden ser exploradas a través de:
 - Histograma
 - Diagrama de caja y bigotes



Variables categóricas

- Pueden ser exploradas a través de:
 - Tablas de frecuencias
 - Estratificación de variables continuas en rangos.
Ejemplo: Rangos de edad, rangos de precio, rango de conteos de descargas de una aplicación móvil
 - Visualizaciones por categoría. **Ejemplo:** Gráfica de barras**



*<https://enlamentedeachenwall.blogspot.com/2019/09/box-plot-diagrama-de-caja-y-bigote-los.html>

**<https://www.mathsisfun.com/data/bar-graphs.html>

PREPROCESAMIENTO DE DATOS

LIMPIEZA DE DATOS



Debemos limpiar el set de datos antes de entrenar modelos supervisados

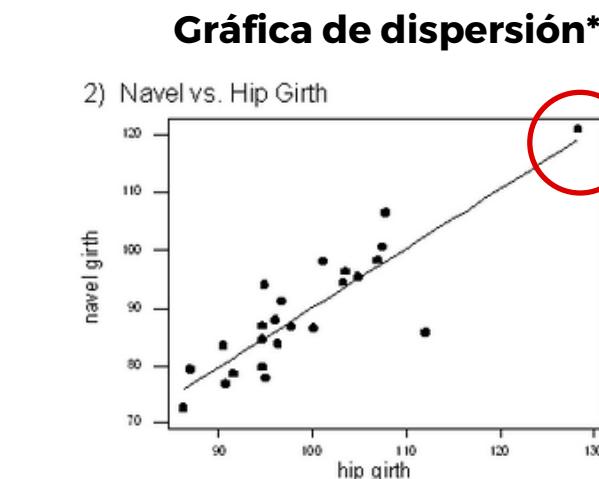
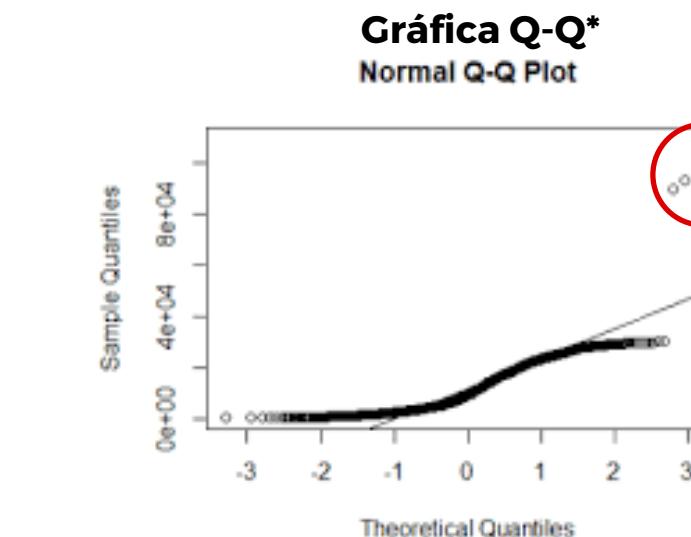


Valores ausentes

- Varias opciones, recomendado utilizar el método que da las mejores predicciones
 - Eliminar filas con valores ausentes
 - Reemplazar valor ausente con media muestral
 - Reemplazar valor ausente con media por grupos
 - Reemplazar valor ausente con predicción de otro modelo
- Para modelos de predicción, se puede evaluar cuál método da las mejores métricas de desempeño

Valores atípicos

- Existen diferentes formas de identificar valores atípicos, por ejemplo:



TAREA

¿Qué otros métodos conocen?

*<https://www.r-bloggers.com/2019/02/robust-regressions-dealing-with-outliers-in-r/>

**<https://www2.southeastern.edu/Academics/Faculty/dgurney/Math241/StatTopics/ScatAnal.htm>

PREPROCESAMIENTO DE DATOS

LIMPIEZA DE DATOS



Debemos limpiar el set de datos antes de entrenar modelos supervisados



Datos redundantes

- Checar si existen observaciones duplicadas y eliminarlas
- Correlación de variables independientes con otras variables independientes en el mismo set de datos (cuando una variable es combinación lineal de otra) puede ocasionar que los algoritmos fallen. Es necesario analizar la correlación de las variables independientes y ver si algunas resultan ser redundantes. En la industria se suelen priorizar variables con base en criterios de negocio

Error en codificación de variables

- Algunos sistemas de generación de datos codifican valores nulos con etiquetas predeterminadas (por ejemplo, 99 o 9999)
- Una variable puede ser continua, pero la codificación indica que es categórica

TAREA

Piensa en otros ejemplos donde se puede presentar un error en la codificación de variables

PREPROCESAMIENTO DE DATOS

TRANSFORMACIÓN DE DATOS



Para evitar que los algoritmos de optimización utilizados en modelos supervisados diverjan, es necesario que las variables independientes tengan escalas similares

Estandarización

- Cuantificar el fenómeno a predecir con base en el número de desviaciones estándar. Es decir, re escalar la variable independiente con base en el lugar de la distribución en que se encuentra
- Las variables estandarizadas son más fáciles de interpretar y comparar
- El resultado es una distribución con media cero y desviación estándar uno

Unidad tipificada

$$z_i = \frac{x_i - \mu}{\sigma}$$

μ Media
 σ Desviación estándar

Normalización

- Re escalar una variable a un intervalo, típicamente [-1, 1] o [0, 1]
- Método Min-max
- El resultado será una distribución que quede dentro de un intervalo de medida uno

Método Min-max*

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

PREPROCESAMIENTO DE DATOS

TRANSFORMACIÓN DE DATOS



🎯 En muchas ocasiones, es necesario hacer ingeniería de datos para crear nuevas variables independientes o para hacer que sean compatibles con los algoritmos de aprendizaje de máquina

Transformaciones numéricas

- Media
- Mediana
- Suma
- Desviación estándar
- arcoseno
- logaritmo natural

¿Cuáles otros métodos se te ocurren?

Transformaciones para variables categóricas

- One hot encoding
 - Convierte una variable categórica en diferentes variables que contienen una indicadora con base en una categoría de la variable original
 - Si hay 10 categorías, entonces habrán 10 nuevas variables

One hot encoding - variable con 3 categorías

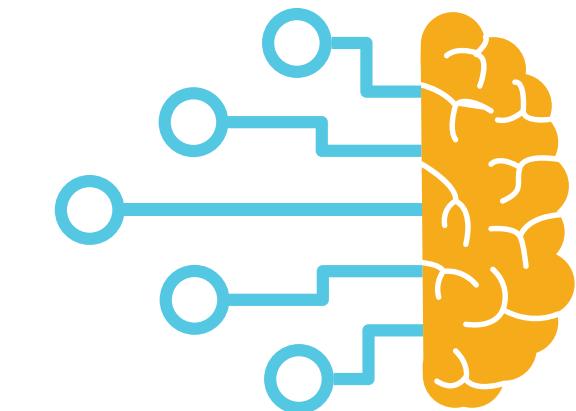
Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	0	1

ALGORITMOS DE APRENDIZAJE DE MÁQUINA

MODELOS SUPERVISADOS Y NO SUPERVISADOS



El mundo del aprendizaje automático...



Modelación Supervisada

- Tenemos una respuesta a predecir
 - Tenemos un set de datos con una respuesta dada
 - Se ha observado la respuesta en el pasado
 - Se puede etiquetar la variable respuesta
 - Queremos determinar la distribución de probabilidad $p(y|X)$
- Tipos de modelos supervisados
 - Regresión
 - Clasificación

Modelación No Supervisada

- No tenemos una respuesta a predecir
 - Tratamos de reducir la dimensión de un problema
 - Tratamos de agrupar poblaciones
 - Tratamos de encontrar una asociación latente en los datos
 - Queremos determinar la distribución de probabilidad $p(X)$
- Tipos de modelos no supervisados
 - Clustering
 - Asociación de reglas

ALGORITMOS DE APRENDIZAJE DE MÁQUINA

MODELOS SUPERVISADOS

$$Y \approx f_X$$



Cuando queremos predecir una respuesta dado un set de datos, cuando tenemos etiquetas, cuando hemos observado el problema en el pasado, etc

TIPOS DE VARIABLES

- | | |
|------------------------|--------------------------|
| Output Y | Input X |
| • Variable Target | • Atributo |
| • Variable Respuesta | • Covariable |
| • Variable Dependiente | • Variable Independiente |

¿PARA QUE USARLA?

- Interpretación: efecto de input en output
- Predicción: output de un nuevo input

REGRESIÓN

¿Cómo identificarla?

- Variable **target numérica**
- Ejemplos: # contratos, # productos vendidos, precio de un inmueble o un auto, precio de una acción tras IPO de empresa, rating de una película, etc



CLASIFICACIÓN

¿Cómo identificarla?

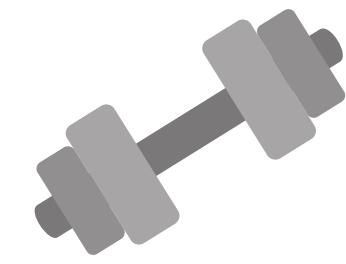
- Variable **target categórica** (binaria, nominal, multi nominal)
- Ejemplos: Venta o no venta; Respuesta o no respuesta; Sí o no; Negativo, positivo o neutral, entre otros

0101
1001
0110 YES
NO



ALGORITMOS DE APRENDIZAJE DE MÁQUINA

ENTRENAMIENTO DE UN MODELO SUPERVISADO



¿Qué significa entrenar un modelo?

- Dado un set de datos (\mathbf{x}_n, y_n) , deseamos encontrar los parámetros $\mathbf{W} = (w_0, w_1, \dots, w_D)$ óptimos. Tal que $y_n \approx f(\mathbf{x}_n)$
- Para determinar dichos parámetros, es necesario utilizar un algoritmo de optimización



4.- Métricas de ajuste



MÉTRICAS DE AJUSTE

ERROR DE REGRESIÓN Y CLASIFICACIÓN



🎯 Para evaluar el desempeño de un modelo supervisado necesitamos utilizar las siguientes métricas de ajuste

Regresión

- Medir la distancia entre una observación y la predicción hecha por el modelo de regresión
 - Error cuadrático medio (Mean Squared Error - MSE)

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

- Error absoluto medio (Mean Absolute Error - MAE)

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

- Coeficiente de determinación (R^2)



Objetivo: Minimizar el error... pues eso significa que nuestra predicción es muy parecida a lo que estamos tratando de predecir

Clasificación

- Para targets categóricas, no existe una fórmula para medir distancias, pero si podemos determinar cuántas observaciones clasificamos correctamente

Matriz de confusión - Target binaria

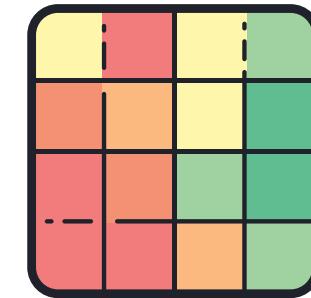
Valor real

		Positivo (1)	Negativo (0)
Predicción	Positivo (1)	Verdaderos Positivos (VP)	Falsos Positivos (FP)
	Negativo (0)	Falsos Negativos (FN)	Verdaderos Negativos (VN)

Caso positivo: etiqueta que estamos tratando de predecir. No necesariamente un evento positivo

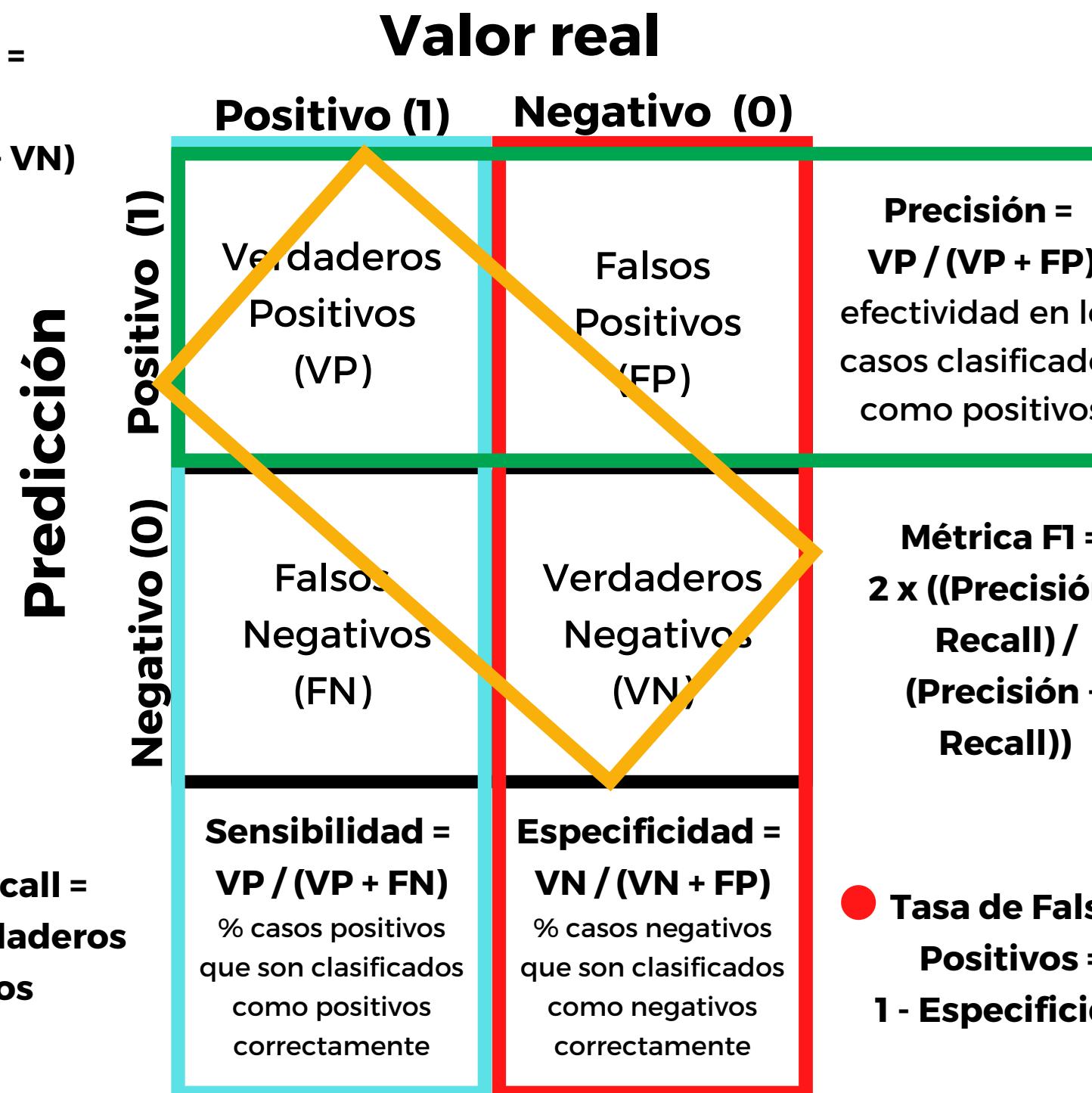
MÉTRICAS DE AJUSTE

MÉTRICAS PARA CLASIFICACIÓN BINARIA



Muchas métricas salen de la matriz de confusión de un ejercicio de clasificación binario

Efectividad =
$$\frac{VP + VN}{(VP + FP + FN + VN)}$$



Tasa de Recall =
Tasa de Verdaderos Positivos

Precisión =
$$\frac{VP}{VP + FP}$$

efectividad en los casos clasificados como positivos

Métrica F1 =
$$\frac{2 \times ((\text{Precisión} \times \text{Recall}) / (\text{Precisión} + \text{Recall}))}{(\text{Precisión} + \text{Recall})}$$

Tasa de Falsos Positivos =
$$1 - \text{Especificidad}$$

Warning

Para targets desequilibradas, no es recomendable usar únicamente la métrica de efectividad para evaluar un modelo

TAREA ¿Por qué?

Hechos

- Métrica F1 es la media armónica de las métricas precisión and recall. Dicha métrica favorece clasificadores que tienen precisión y recall similares
- Precisión/Recall trade-off: Incrementar la métrica de precisión conlleva a reducir el recall y vice versa

TAREA

Piensa en un caso de negocio donde valga la pena priorizar la métrica de precisión sobre el recall

MÉTRICAS DE AJUSTE

EJEMPLOS PARA REGRESIONES Y CLASIFICADORES



🎯 Calcula las siguientes métricas de ajuste para el ejercicio de regresión y para el de clasificación

Regresión

TAREA

- Calcula el error cuadrático medio de la siguiente predicción
 $y(\text{predicción}) = (43.6, 44.4, 45.2, 46, 46.8)$
 $y(\text{valor actual}) = (41, 45, 49, 47, 44)$

Clasificación

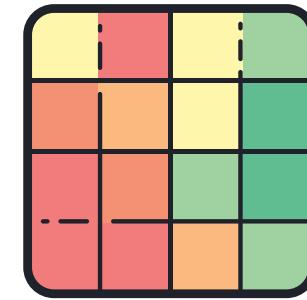
TAREA

		Valor real	
		Positivo (1)	Negativo (0)
Predicción	Positivo (1)	27	4
	Negativo (0)	12	37

- Calcula: todas las métricas que aparecen en el slide anterior

MÉTRICAS DE AJUSTE

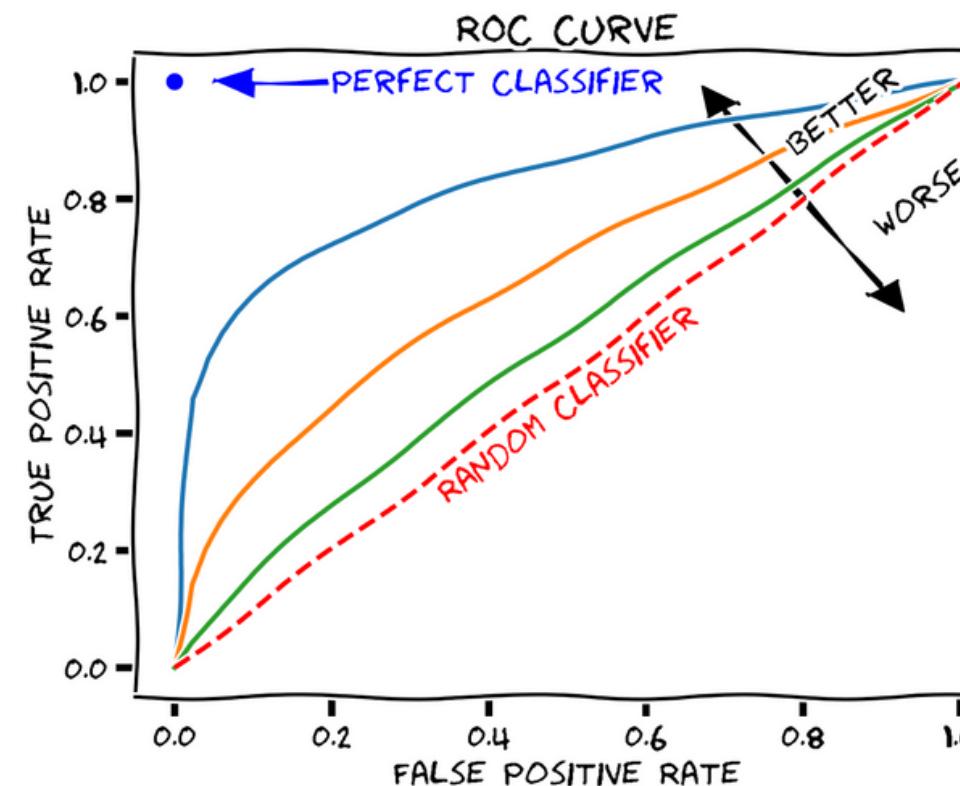
CLASIFICACIÓN - GRÁFICAS Y MÉTRICAS IMPORTANTES



Las siguientes gráficas ofrecen una manera visual para evaluar modelos de clasificación

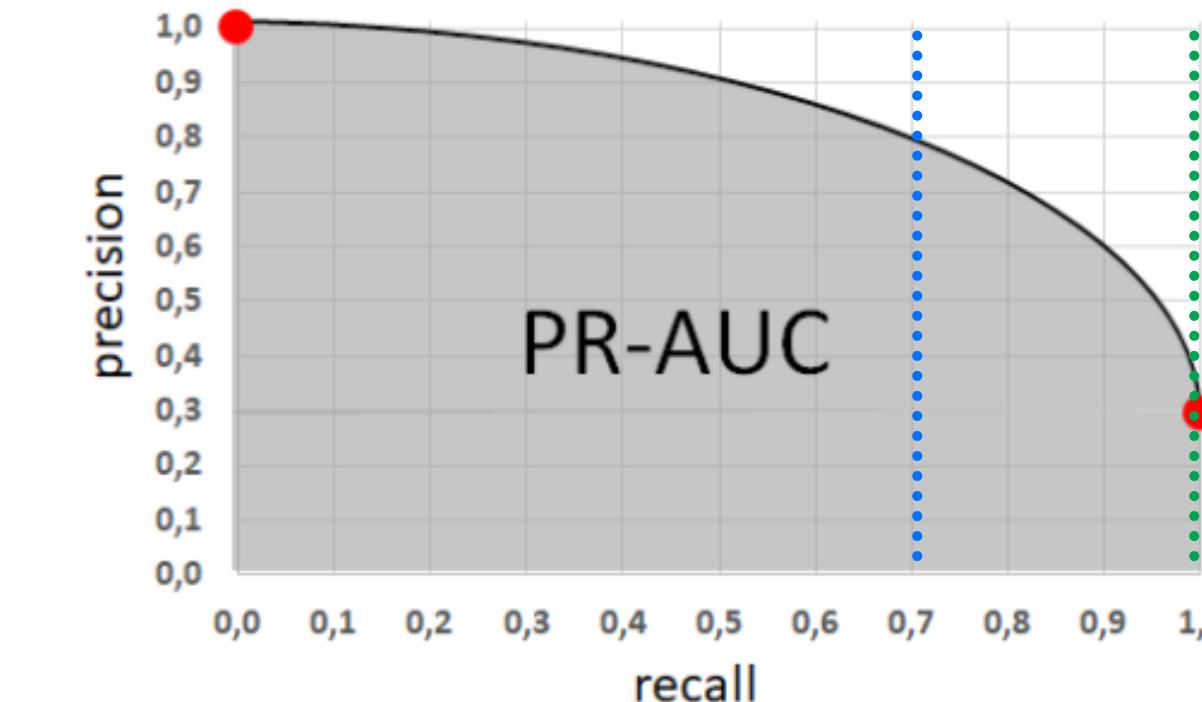
Curva ROC* y métrica ROC AUC

- Receiver operating characteristic, gráfica de la tasa de verdaderos positivos versus tasa de falsos positivos.
- Métrica ROC AUC es el área bajo la curva ROC
- Clasificador perfecto tendría métrica ROC AUC igual a 1. Un clasificador que predice el azar tendría una métrica de 0.5



Curva Precisión - Recall**

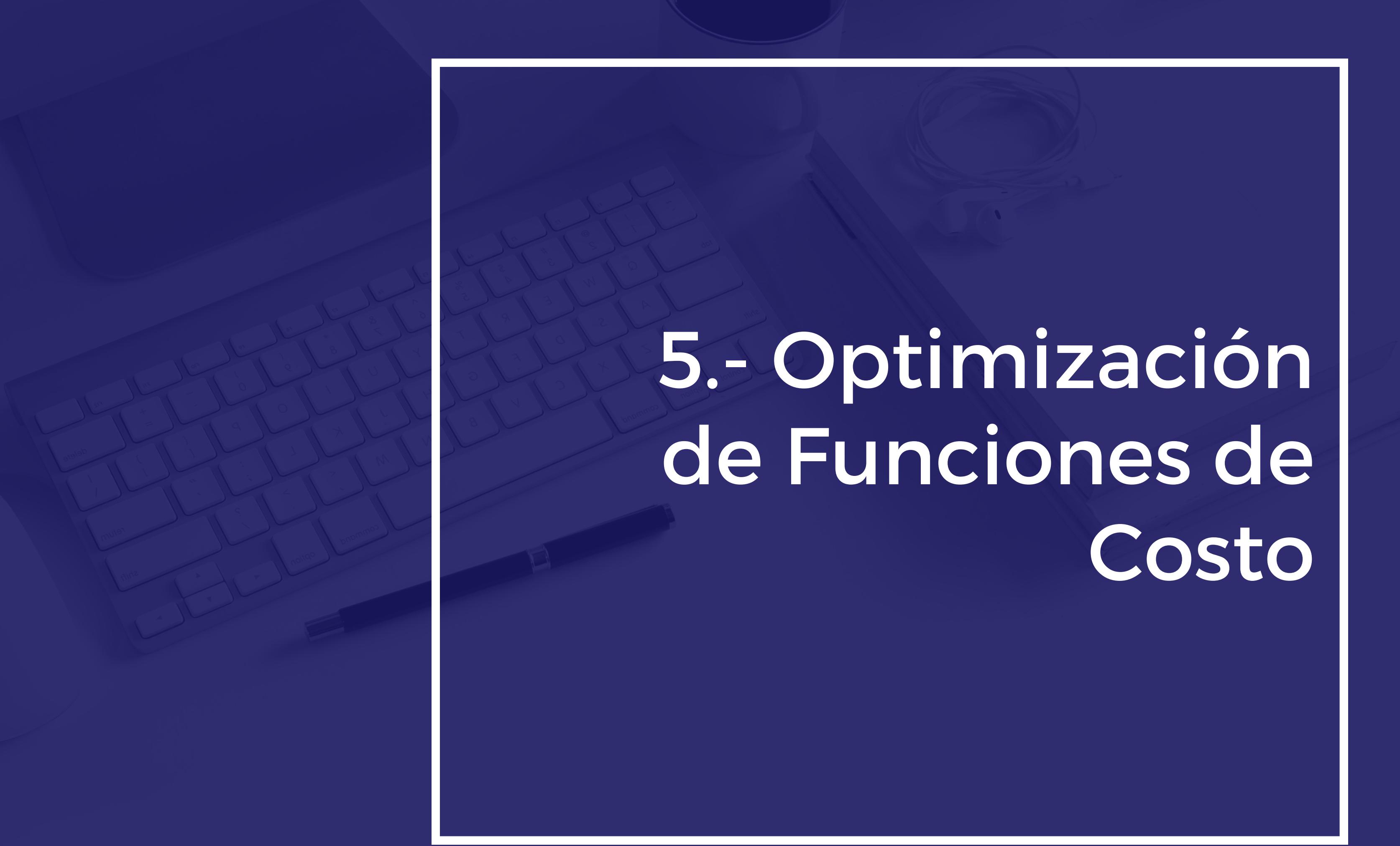
- Buena manera de visualizar el trade-off entre las métricas de precisión y recall
- Con una métrica de 70% de recall lograremos tener un 80% de precisión
- Con una métrica de 100% de recall lograremos tener un 30% de precisión



*<https://towardsdatascience.com/precision-recall-curves-for-imbalanced-and-healthcare-related-data-sets-e3bc76575d1e>

**<https://towardsdatascience.com/gaining-an-intuitive-understanding-of-precision-and-recall-3b9df37804a7>

5.- Optimización de Funciones de Costo



OPTIMIZACIÓN DE FUNCIONES DE COSTO

ENTRENAMIENTO DE UN MODELO



¿Qué significa entrenar un modelo?

Entrenamiento de un modelo supervisado

- Dado un data set (\mathbf{x}_n, y_n) , deseamos encontrar los parámetros $\mathbf{W} = (w_0, w_1, \dots, w_D)$ óptimos. Tal que $y_n \approx f(\mathbf{x}_n)$
- Para determinar dichos parámetros, es necesario encontrar una función de costo $\mathcal{L}(\mathbf{W})$ y encontrar el punto óptimo $\mathbf{W} = (w_0, w_1, \dots, w_D)$, donde la función alcanza un mínimo global, i.e

distancia entre
predicción y valor real

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) \quad \text{tal que } \mathbf{W} \in R^D$$

- Existen varias maneras de dar solución al problema de optimización planteado anteriormente
- Deseable: encontrar el mínimo global para la función de costo

OPTIMIZACIÓN DE FUNCIONES DE COSTO

FUNCIONES DE COSTO Y PROPIEDADES



 Las funciones de costo juegan un papel importante en el entrenamiento de modelos supervisados

Funciones de Costo

- También llamada función de pérdida, es utilizada para entrenar parámetros que expliquen un set de datos
- Cuantifica que tan bueno es el desempeño del modelo. En otras palabras, cuantifica qué tan costosos son los errores cometidos por el modelo

Dos propiedades deseadas

- Cuando la variable target es numérica, se desea lo siguiente:
 - Función de costo sea simétrica alrededor del número cero
 - Función de costo penalice errores grandes de manera similar a los errores muy grandes

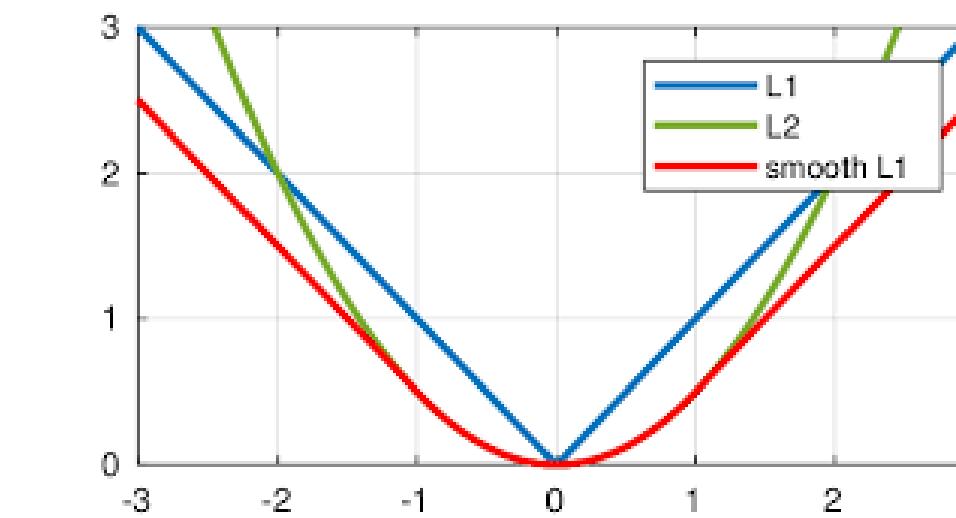
Funciones de costo populares

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

- ✓ Simétrica alrededor del cero
- ✗ No robusta cuando el data set contiene outliers

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

- ✓ Simétrica alrededor del cero
- ✓ Robusta cuando el data set contiene outliers



OPTIMIZACIÓN DE FUNCIONES DE COSTO

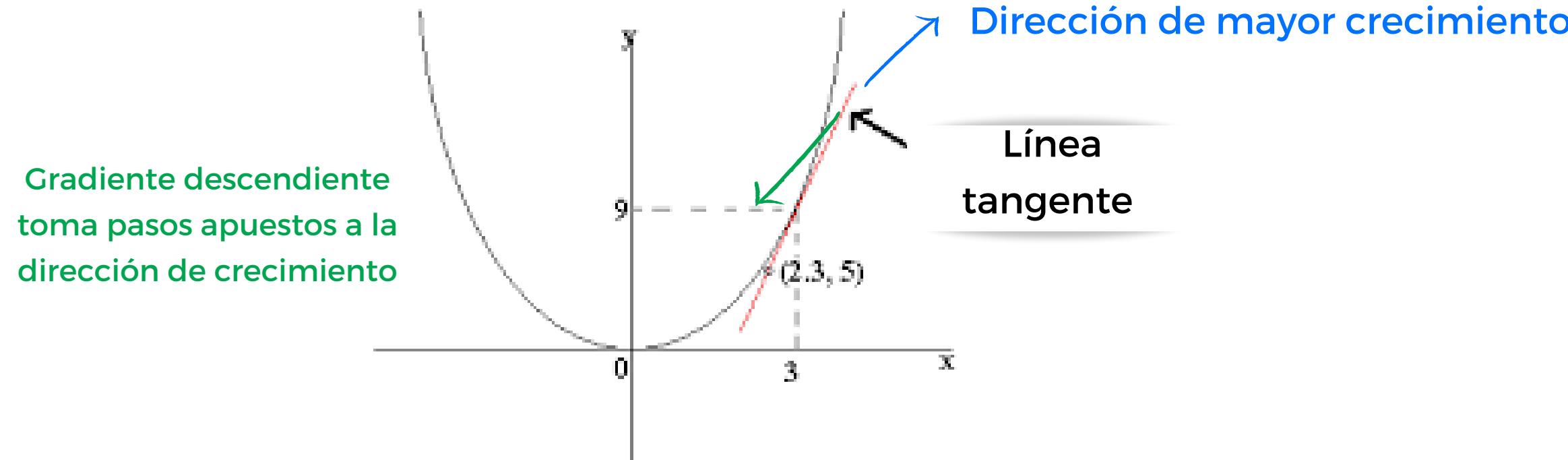
MÉTODOS DE OPTIMIZACIÓN - GRADIENTE DESCENDENTE



🎯 Algoritmo de optimización para encontrar parámetros óptimos tal que función de costo toque su punto mínimo al ser evaluada en esos parámetros

Optimización suave

- El gradiente (derivada) existe para una determinada función de costo
- Cuando el gradiente es evaluado en un punto, tenemos una pendiente tangente a la función en ese punto.
Dicha tangente apunta en la dirección del mayor incremento de la función
- Para minimizar una función de costo, damos de manera iterativa pasos en sentido contrario a la dirección del gradiente



ALGORITMOS DE APRENDIZAJE DE MÁQUINA

ENTRENAMIENTO DE UN MODELO SUPERVISADO



¿Qué significa entrenar un modelo?

Definición

- Un modelo de regresión lineal asume una relación lineal entre la variable dependiente y las independientes
- Target numérica
- Un set de datos está compuesto por el par (\mathbf{x}_n, y_n)
- Regresión lineal simple

$$y_n \approx f(\mathbf{x}_n) = w_0 + w_1 x_{n1}$$

$$\mathbf{W} = (w_0, w_1) \quad \text{Parámetros del modelo}$$

- Regresión lineal múltiple

$$\begin{aligned} y_n &\approx f(\mathbf{x}_n) = w_0 + w_1 x_{n1} + \dots + w_D x_{nD} \\ &= \mathbf{x}_n^T \mathbf{W} \end{aligned}$$

REGRESIÓN LINEAL

CORRELACIÓN \neq CAUSALIDAD



🎯 La regresión lineal encuentra la correlación lineal entre inputs y output, más no la relación causal del fenómeno. Interpreta tus resultados con cuidado

Correlación Espuria

¿Qué ven de raro en la siguiente imagen?

