Introducción al aprendizaje de máquina

Aprendizaje Automático

Juan David Martínez
jdmartinev@eafit.edu.co

Objetivos de aprendizaje

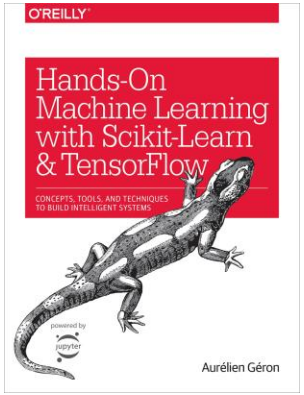
Aprender principios del aprendizaje automático, técnicas convencionales de preprocesamiento de datos, algoritmos básicos de ML y evaluación de modelos.

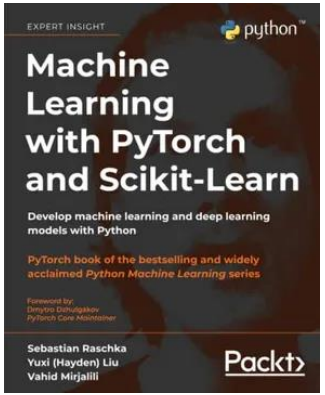
Familiarizarse con métodos tradicionales de ML y librerías de preprocesamiento de datos, incluyendo métodos para detectar y corregir el sobre-entrenamiento

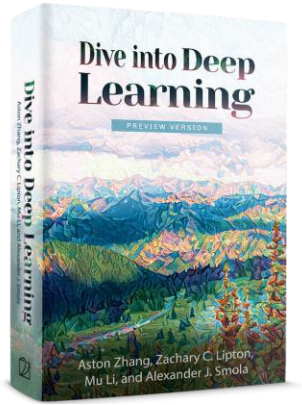
Agenda

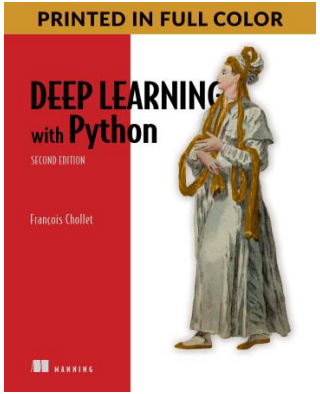
- Introducción a inteligencia artificial y aprendizaje de máquina
- Tipos de aprendizaje de máquina
- Evaluación de modelos
  - Entrenamiento, validación, prueba
  - Sobreentrenamiento
- Análisis exploratorio de datos
- K vecinos más cercanos

Recursos para aprender y practicar









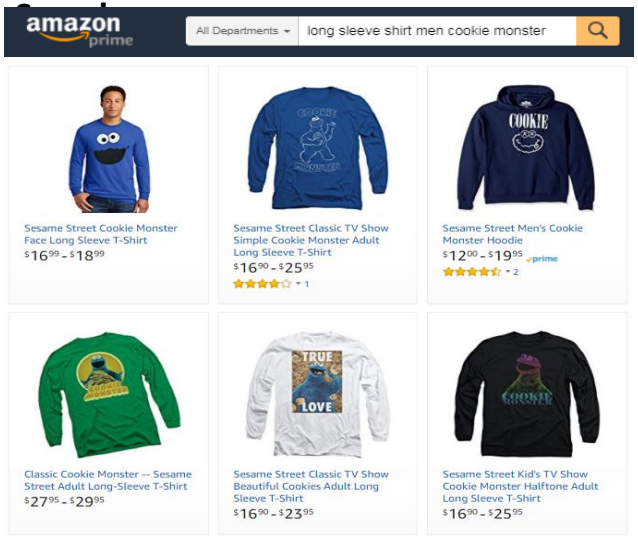
Amazon SageMaker

colab

kaggle

Aplicaciones de ML

Business/ML	Description	Example
Ranking	Helping users find the most relevant thing	Ranking algorithm within Amazon
Recommendation	Giving users the item they may be most interested in	
Classification	Figuring out what kind of thing something is	
Regression	Predicting a numerical value of a thing	
Clustering	Putting similar things together	
Anomaly Detection	Finding uncommon things	



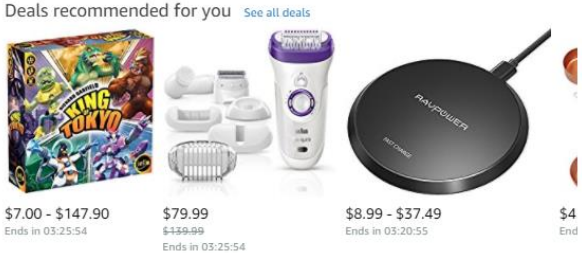
https://github.com/UBC-CS/cpsc330/blob/master/lectures/01_intro.ipynb

Aplicaciones de ML

Business/ML	Description
Ranking	Helping users find the most relevant thing
Recommendation	Giving users the item they may be most interested in
Classification	Figuring out what kind of thing something is
Regression	Predicting a numerical value of a thing
Clustering	Putting similar things together
Anomaly Detection	Finding uncommon things

Example

Recommendations across the website



Amazon's Choice

Aplicaciones de ML

Business/ML	Description	Example
Ranking	Helping users find the most relevant thing	Product classification for our catalog
Recommendation	Giving users the item they may be most interested in	
Classification	Figuring out what kind of thing something is	
Regression	Predicting a numerical value of a thing	
Clustering	Putting similar things together	
Anomaly Detection	Finding uncommon things	

High-Low Dress

Straight Dress

Striped Skirt

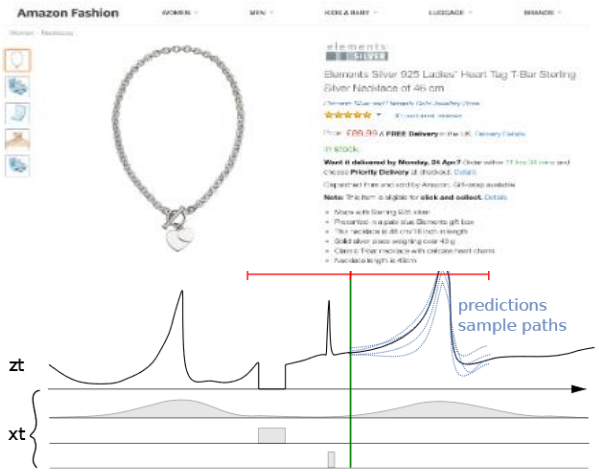
Graphic Shirt

Aplicaciones de ML

Business/ML	Description
Ranking	Helping users find the most relevant thing
Recommendation	Giving users the item they may be most interested in
Classification	Figuring out what kind of thing something is
Regression	Predicting a numerical value of a thing
Clustering	Putting similar things together
Anomaly Detection	Finding uncommon things

Example

Predicting sales for specific ASINs



Seasonality | Out of stock | Promotions

Aplicaciones de ML

Business/ML	Description
Ranking	Helping users find the most relevant thing
Recommendation	Giving users the item they may be most interested in
Classification	Figuring out what kind of thing something is
Regression	Predicting a numerical value of a thing
Clustering	Putting similar things together
Anomaly Detection	Finding uncommon things

Example

Aplicaciones de ML

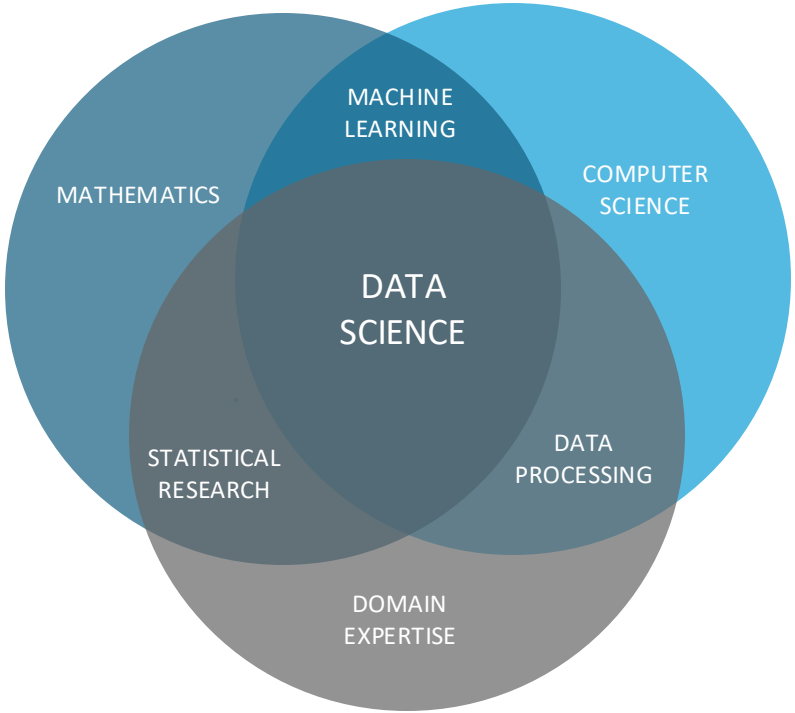
Business/ML	Description
Ranking	Helping users find the most relevant thing
Recommendation	Giving users the item they may be most interested in
Classification	Figuring out what kind of thing something is
Regression	Predicting a numerical value of a thing
Clustering	Putting similar things together
Anomaly Detection	Finding uncommon things

Example

Fruit freshness

¿Qué es ciencia de datos?

Wikipedia:
“a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights** from structured and unstructured **data**.”

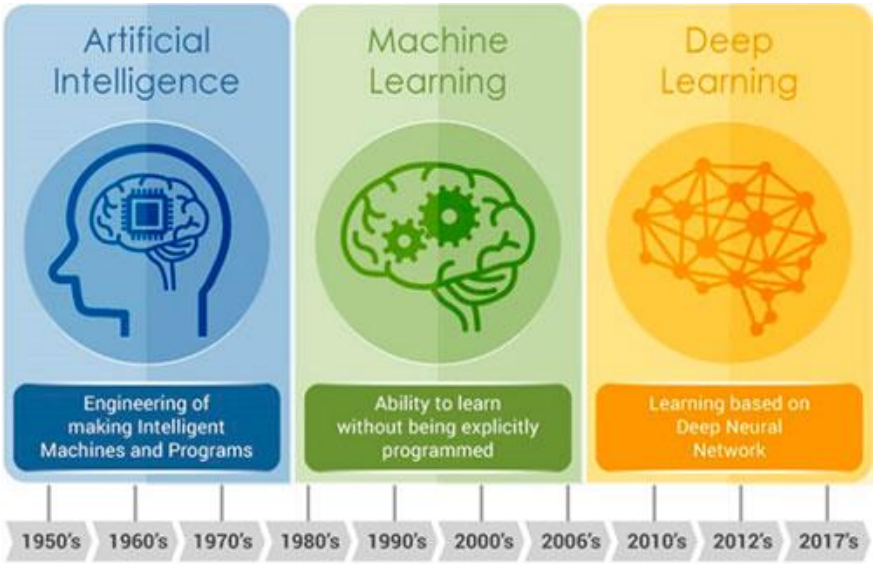


¿Qué es aprendizaje de máquina?

"Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy"

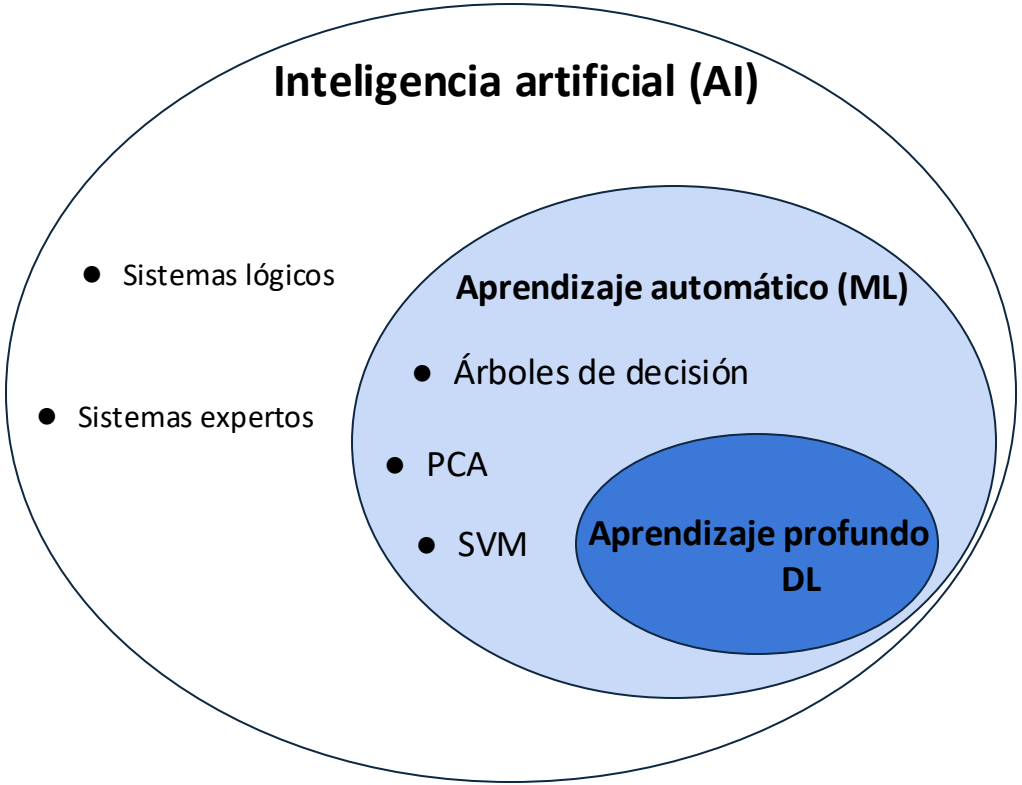
Los algoritmos de ML se caracterizan por su habilidad para **aprender de los datos** sin estar explícitamente programados.

¡ML se utiliza comúnmente para hacer predicciones!



https://otech.uaeh.edu.mx/noti/index.php/machine-learning/los-conceptos-de-machine-learning-y-deep-learning-en-la-industria/

AI vs ML vs DL



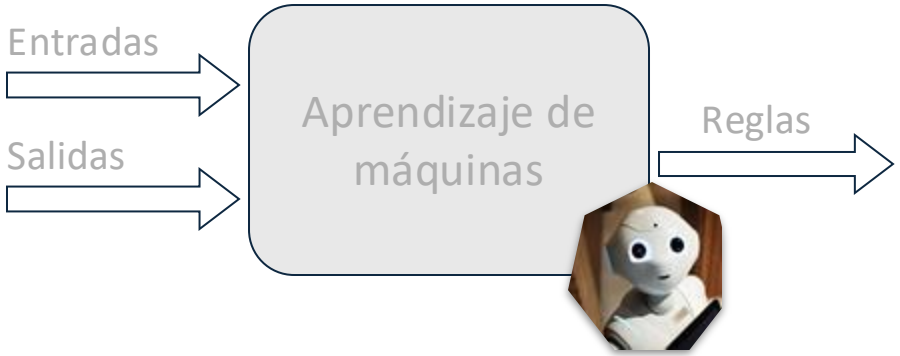
Inteligencia artificial: estudia los comportamientos inteligentes en dispositivos.

Machine learning: estudia los algoritmos que mejoran su rendimiento incorporando nuevos datos a un modelo existente.

Deep learning: usa algoritmos de redes neuronales artificiales para extraer características cada vez más complejas a partir de los datos de entrada, buscando mejorar su rendimiento

ML vs programación tradicional

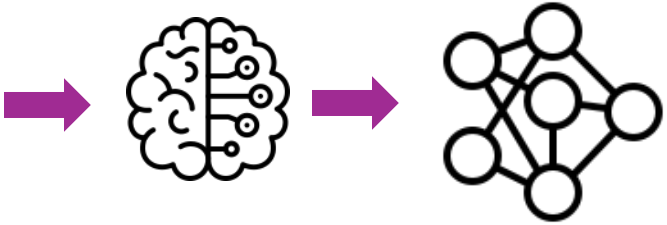
Se definen las reglas y el programador las implementa.



El algoritmo aprende las reglas en lugar de ser implementadas por un programador.

Tipos de ML: Aprendizaje supervisado

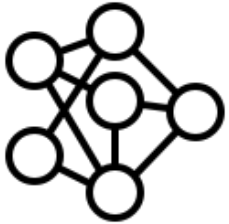
Training



Output: Model that maps images to a labels

Deployment

Input: Image



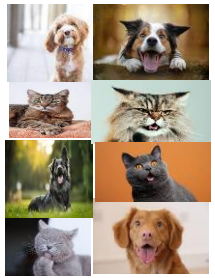
$P(dog) = 0{,}99$

$P(cat) = 0{,}01$

Output: Probability for each label

Tipos de ML: Aprendizaje no supervisado

Training



Input: Images (without labels)

Output: Model that finds similar groups (clusters)

Deployment

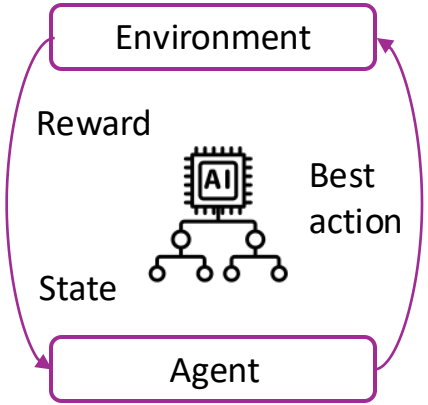
$Cluster\ 1$

Input: Image

Output: group

Tipos de ML: Aprendizaje por refuerzo

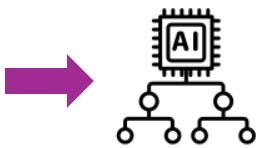
Training



Input: Images (without labels)

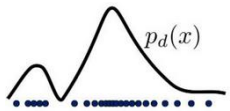
Output: best action

Deployment

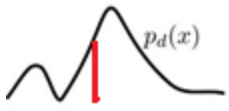


Tipos de ML: Modelos generativos

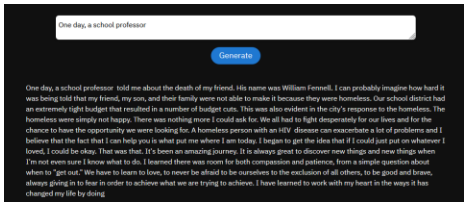
Training



Deployment

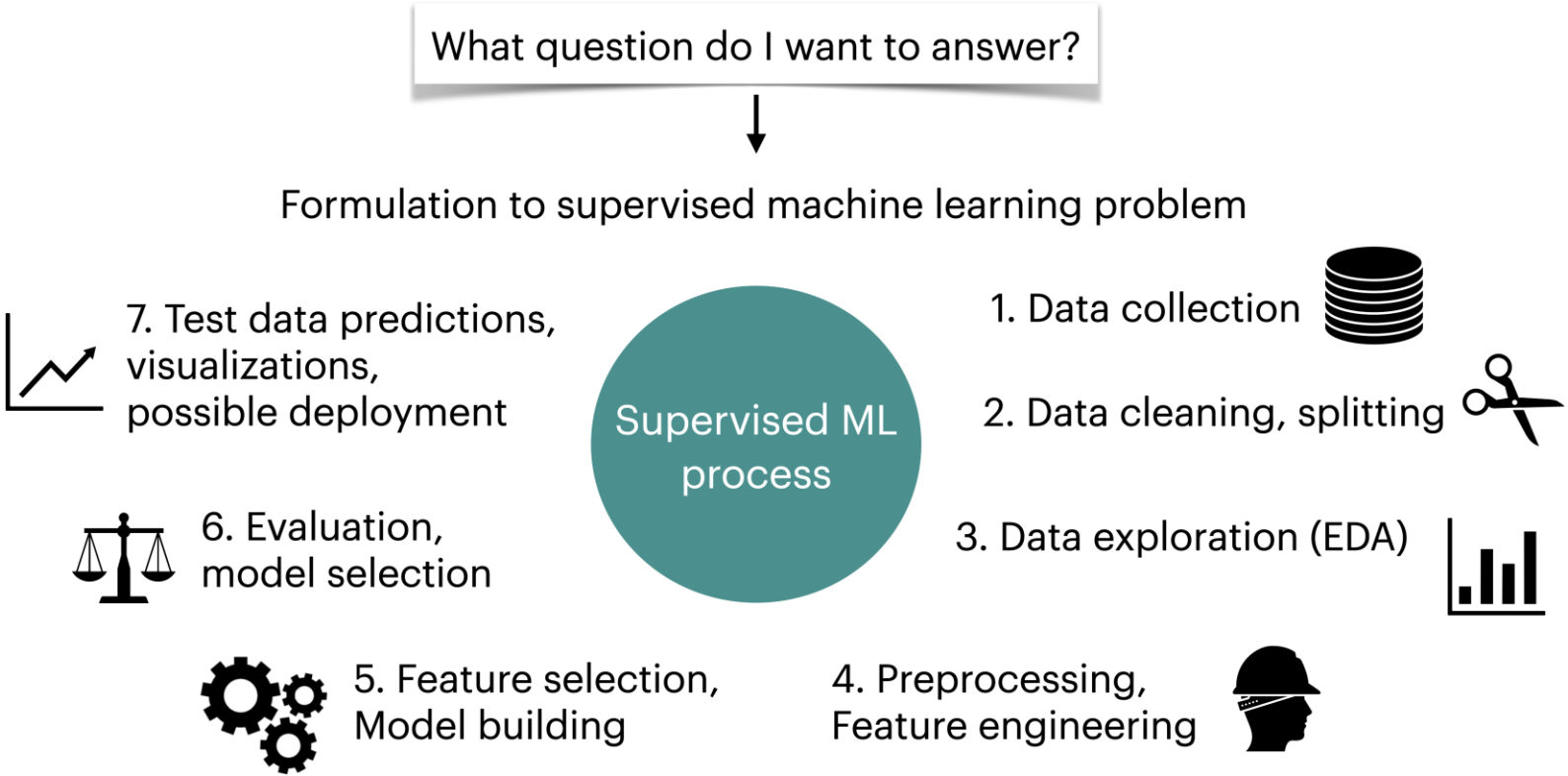


https://thispersondoesnotexist.com/



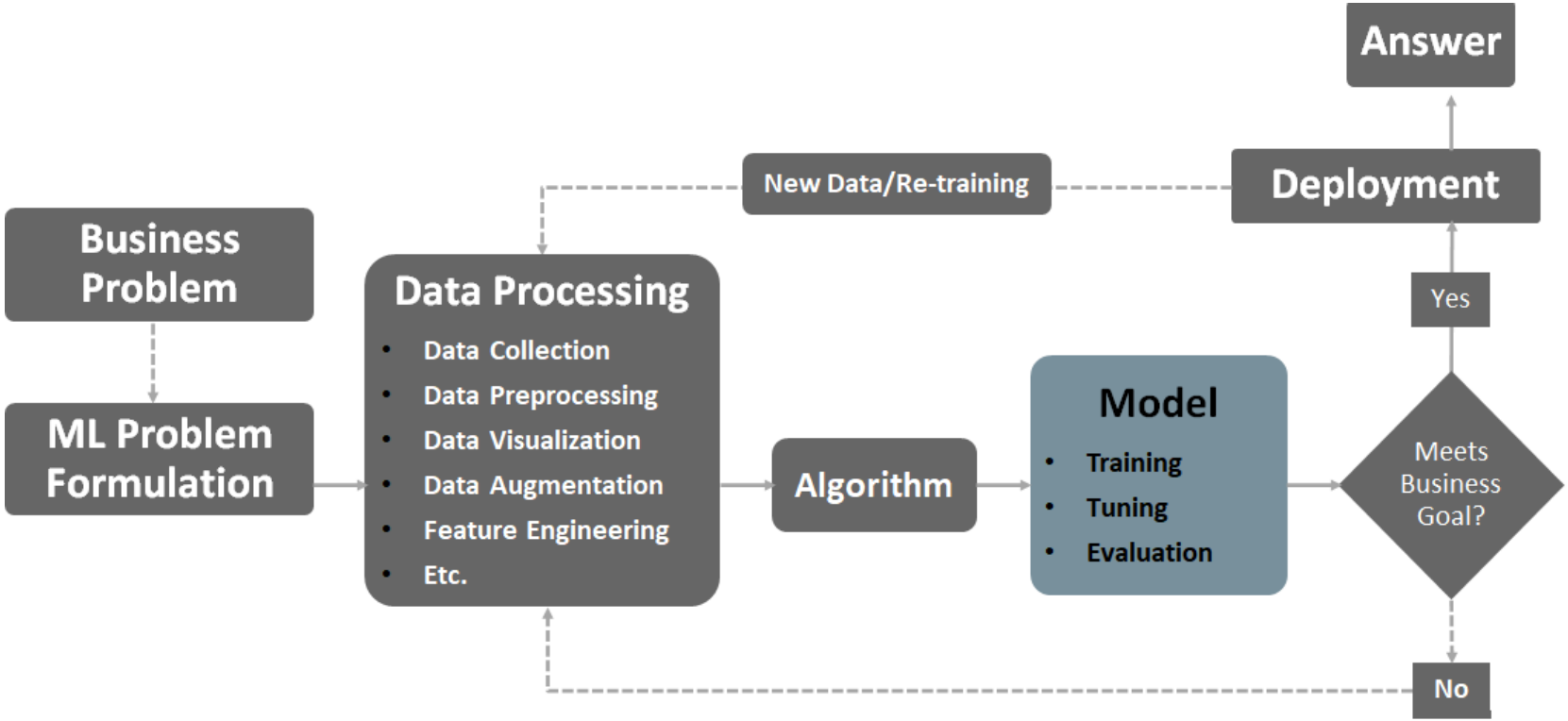
https://hyperwriteai.com

Pasos para entrenar un modelo de ML

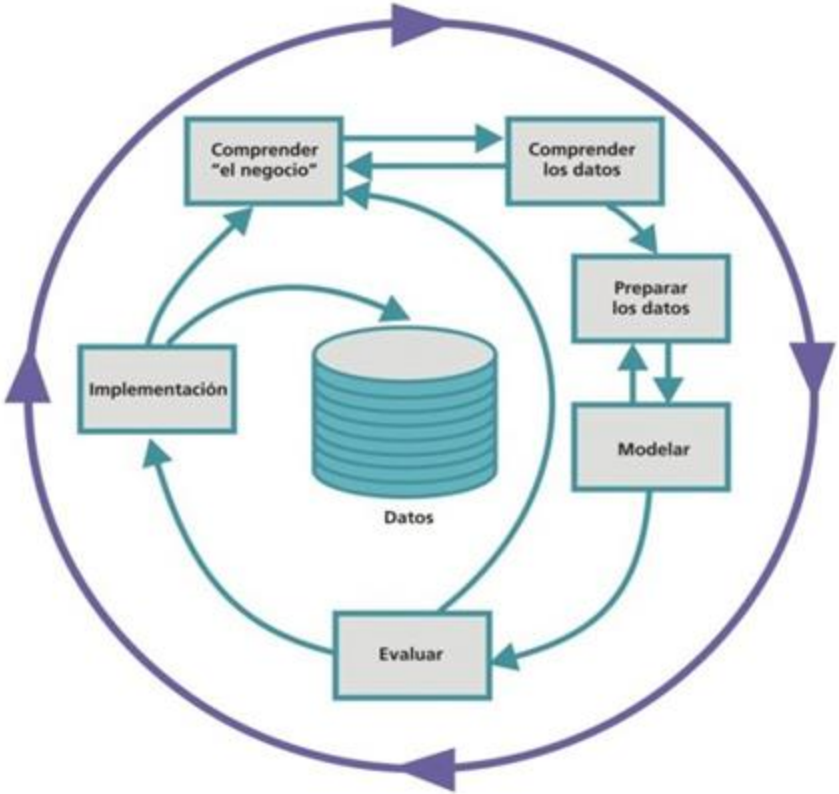


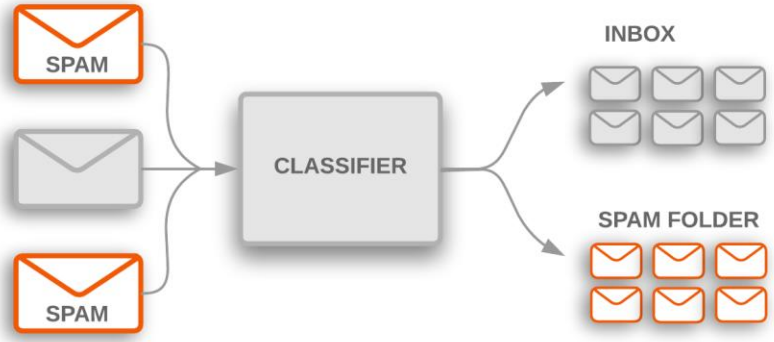
https://github.com/UBC-CS/cpsc330/blob/master/lectures/01_intro.ipynb

Cilco de vida de un sistema de ML

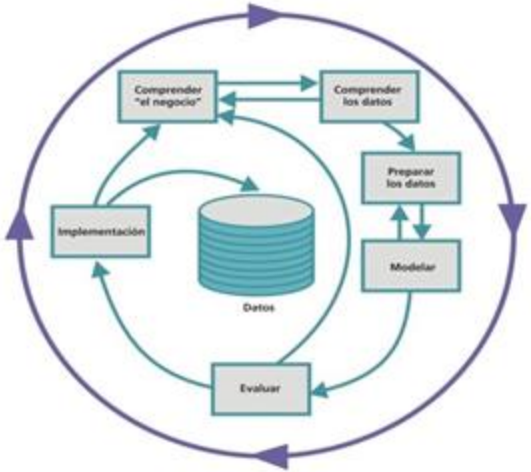


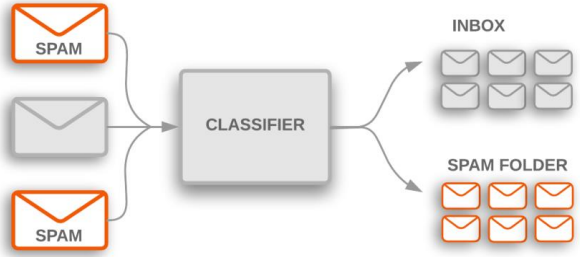
Cross Industry Standard Process for Data Mining (CRISP-DM)





CRISP-DM – Comprender el negocio





Problema:

- Los clientes se quejan del spam
- Analizar qué tan grave es el problema
- ¿Hay una solución simple?
- ¿Podemos solucionarlo con ML?

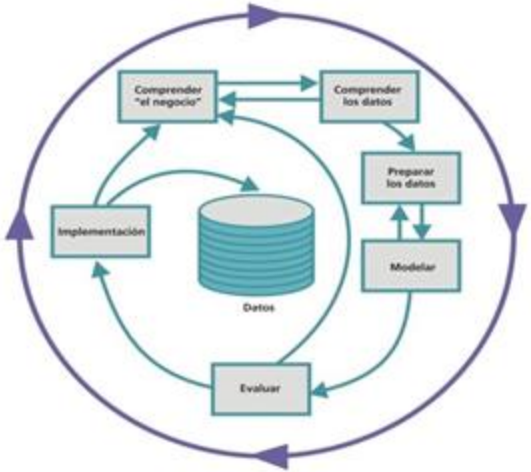
Meta:

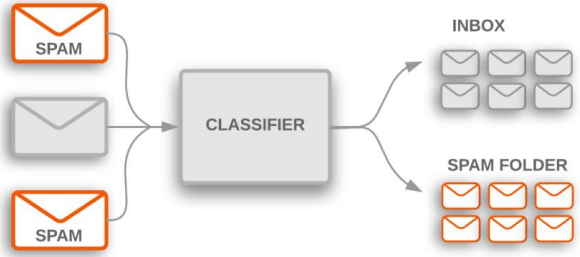
- Reducir la cantidad de mensajes spam
- Reducir la cantidad de quejas recibidas por los usuarios

Meta medible:

- Reducir la cantidad de spam en un 50%

CRISP-DM – Comprender los datos





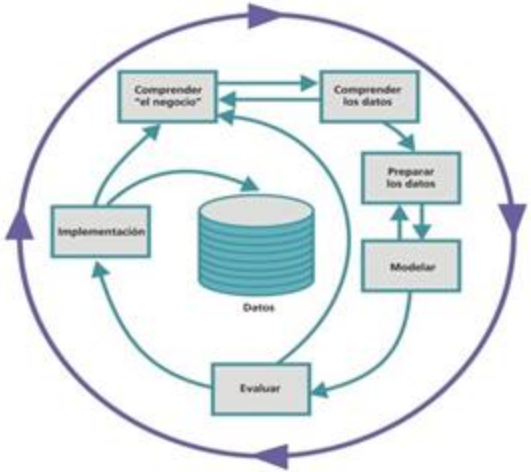
Identificar las fuentes de datos

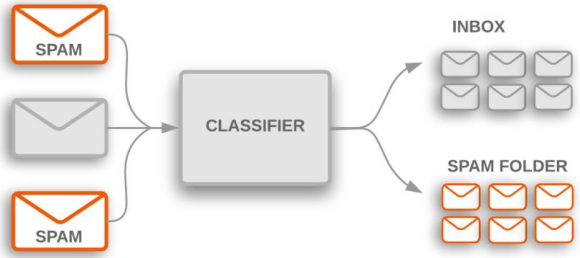
- Botón de spam
- ¿Los datos son lo suficientemente buenos?
- ¿Los datos son confiables?
- ¿Los almacenamos de forma correcta?
- ¿La base de datos es lo suficientemente grande?
- ¿Necesitamos recolectar más datos?

Preparar los datos

- Transformar los datos para que se puedan entregar a un algoritmo de ML
- Limpieza, imputación de datos faltantes, transformación de variables
- Feature engineering

CRISP-DM – Modelar





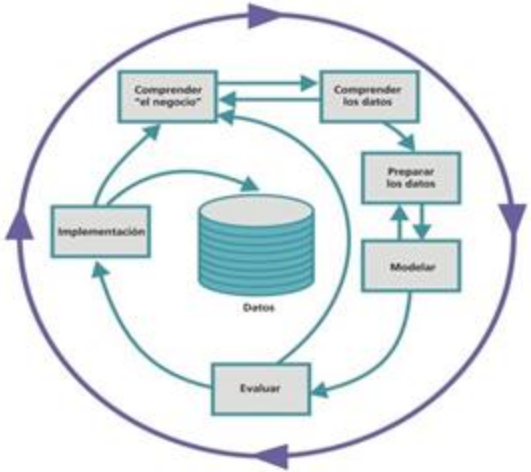
Entrenamiento del modelo

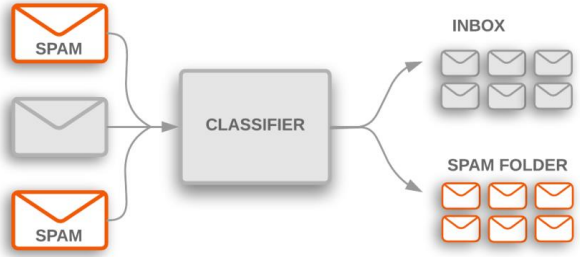
- Intentar diferentes modelos que sean adecuados para los datos con los que está trabajando
- Seleccionar el mejor
- Si el modelo no es lo suficientemente bueno, devolverse al paso anterior

Preparar los datos

- Incluir más datos y/o más características

CRISP-DM – Evaluar





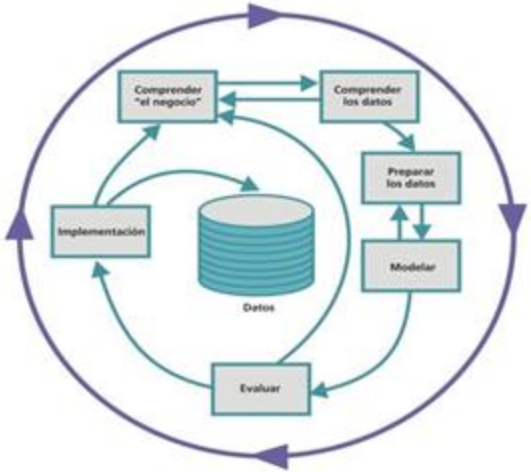
Evaluación del modelo

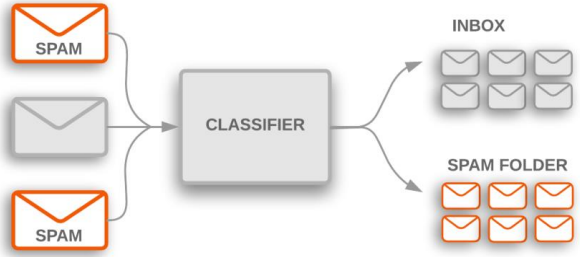
- ¿El modelo es lo suficientemente bueno?
- ¿Alcanzamos la meta propuesta?

Análisis retrospectivo

- Re-evaluar la meta: reducir la cantidad de spam en un 30%
- Dejar de trabajar en el proyecto ☹

CRISP-DM – Desplegar

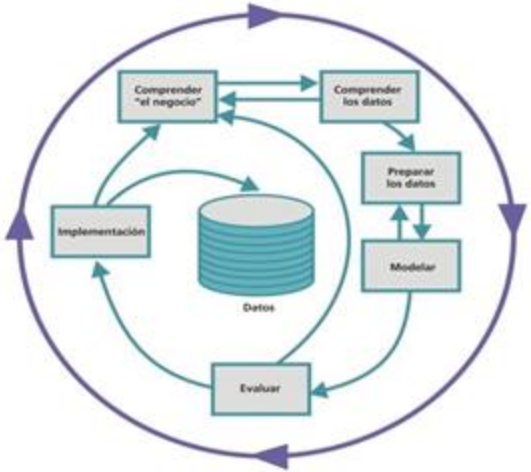


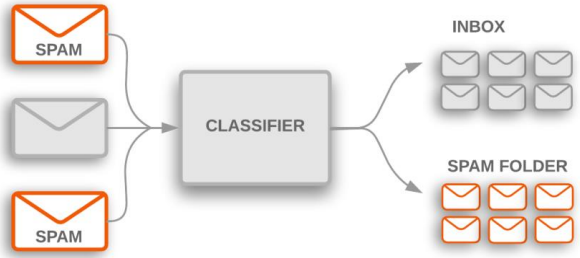


Despliegue del modelo

- Poner el modelo disponible para todos los usuarios
- Monitorear el rendimiento del modelo en el tiempo (MLOps)
- Asegurar la calidad u mantenibilidad del proyecto
- Re-entrenar cuando sea necesario

CRISP-DM – Iterar





Iterar

- ¡Los proyectos con ML requieren varias iteraciones!

CRISP-DM – Resumen

- **Entendimiento del problema:** Definir una meta medible. Preguntarse, ¿necesitamos ML?
- **Entendimiento de los datos:** ¿Tenemos los datos, son lo suficientemente buenos?
- **Preparación de los datos:** Transformar los datos en una tabla para entregar a un algoritmo de ML
- **Modelado:** Entrenar y seleccionar el mejor modelo
- **Evaluación:** Evaluar si se cumplió la meta
- **Despliegue:** Poner disponible el modelo para todos los usuarios
- **Iterar:** Empezar simple, aprender de la retroalimentación, mejorar

Notación matemática

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Vector de características

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix} \in \Re^{d \times 1}$$

$$y^{(i)} \in \Re$$

$$y^{(i)} \in [0, 1, \cdots, C]$$

Matriz de características

$$\mathbf{X} = \begin{bmatrix} \cdots & \mathbf{x}^{(1)\top} & \cdots \\ \cdots & \mathbf{x}^{(2)\top} & \cdots \\ & \vdots & \\ \cdots & \mathbf{x}^{(n)\top} & \cdots \end{bmatrix} \in \Re^{n \times d} \quad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

Ejemplo – Food delivery

Problema:

- Predecir si una orden de comida llegará a tiempo

Datos:

- Información de los últimos 50 pedidos

Solución:

- Clasificación, problema bi-clase 1 tarde

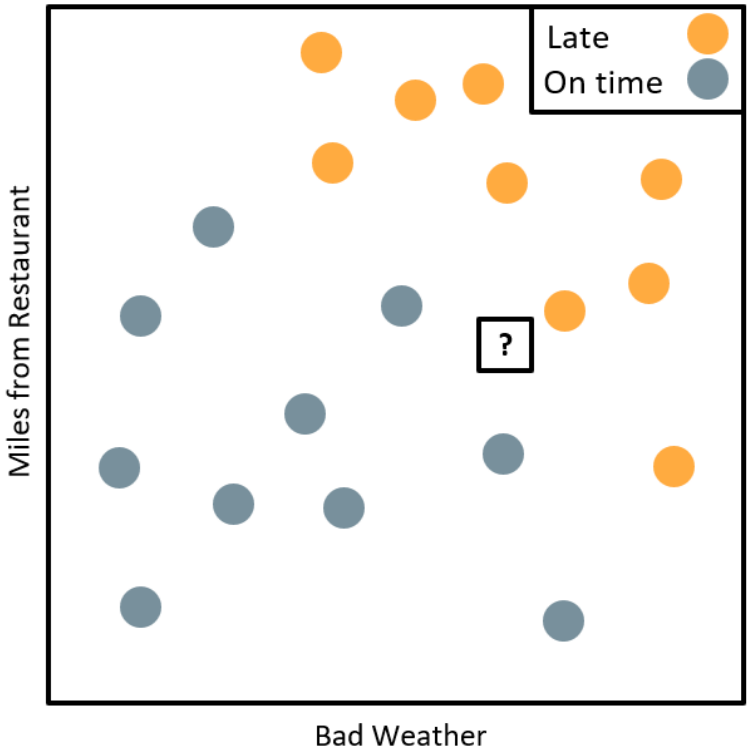
  0 a tiempo

BadWeather	RushHour	MilesFromRestaurant	UrbanAddress	Late
10	1	5	1	0
78	0	7	0	1
14	1	2	1	0
58	1	4.2	1	1
82	0	7.8	0	0
…	…	…	…	…

Ejemplo – Food delivery

K-Vecinos más cercanos (kNN):

- Predice la etiqueta de una nueva muestra basado en la etiqueta de las muestras más cercanas
- A qué clase pertenece ? ?
- Calcule la distancia desde k hasta todos las muestras
- Halle los k veciones más cercanos (escojamos k=3)
- Escoja la clase mayoritaria



Métricas de evaluación - regresión

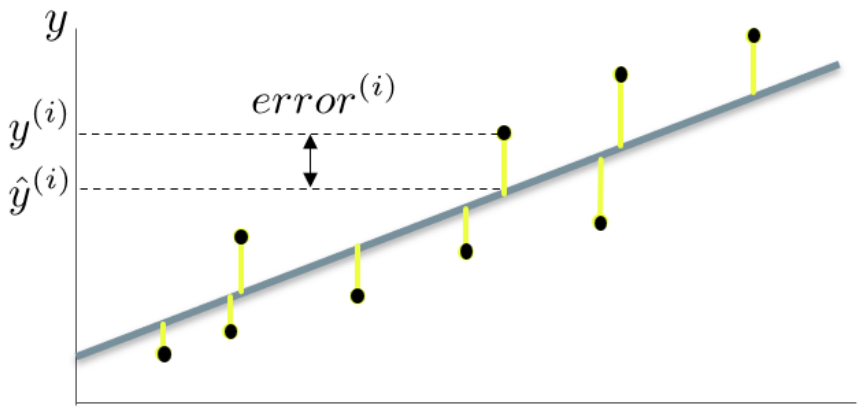
Metrics	Equations
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=0}^{n}(y^{(i)} - \hat{y}^{(i)})^2$
Root Mean Squared Error (RMSE)	$RMS = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(y^{(i)} - \hat{y}^{(i)})^2}$
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum_{i=0}^{n}\lvert y^{(i)} - \hat{y}^{(i)}\rvert$
R Squared (R^2)	$R^2 = 1 - \frac{\sum_{i=0}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=0}^{n}(y^{(i)} - \bar{y})^2}$

$y^{(i)}$: Data values

$\hat{y}^{(i)}$: Predicted values

$\bar{y}$: Mean value of data values, $\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y^{(i)}$

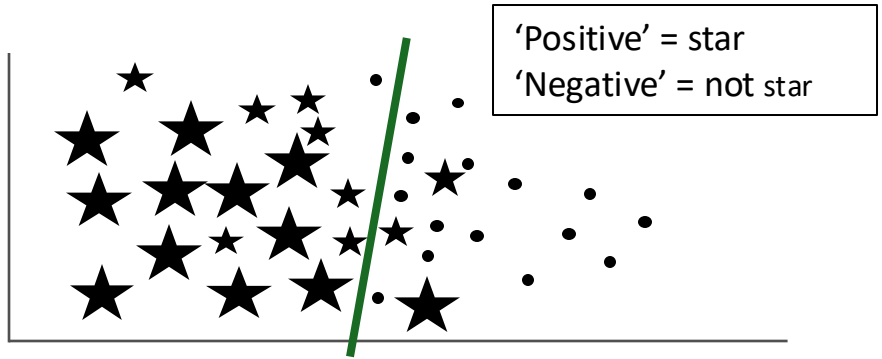
n : Number of data records



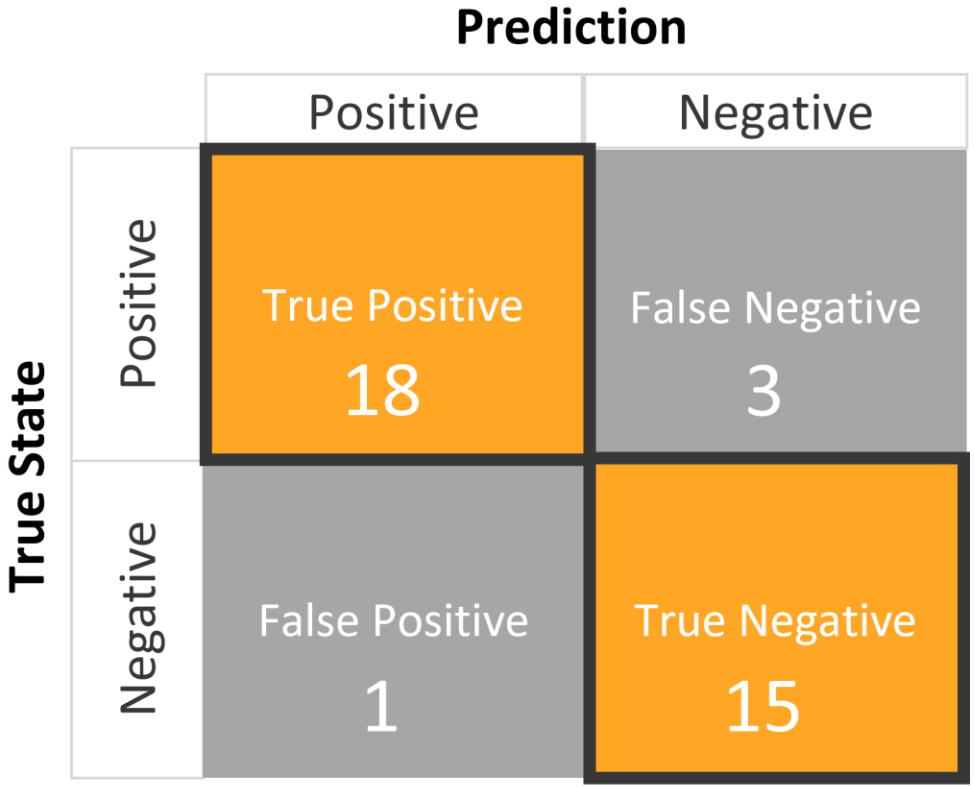
Métricas de evaluación - clasificación

True State \ Prediction	Positive	Negative
Positive	True Positive 18	False Negative 3
Negative	False Positive 1	True Negative 15

True Positive (TP): Predicción Positiva cuando la muestra es Positiva.
False Positive (FP): Predicción positiva cuando la muestra es Negativa.
False Negative (FN): Predicción Negativa cuando la muestra es Negativa
True Negative (TN): Predicción Negativa cuando la muestra es Negativa



Métricas de evaluación - clasificación



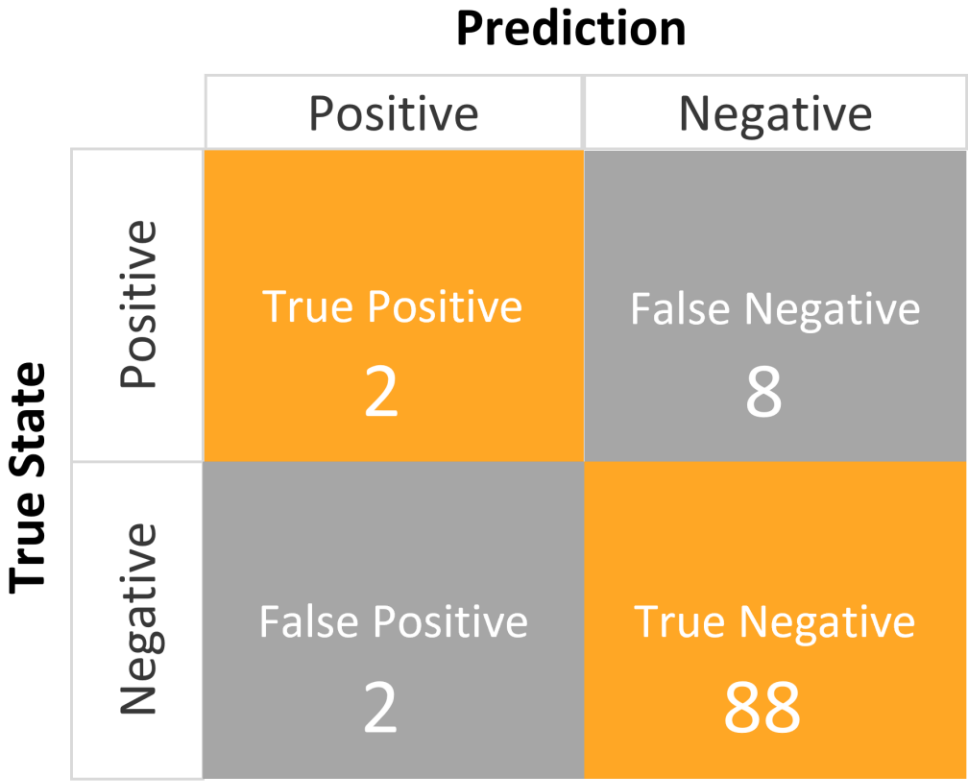
Accuracy*: Porcentaje de muestras clasificadas correctamente

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Accuracy = \frac{18 + 15}{18 + 1 + 3 + 15} = 0.89$$

$$^{*}(bad)\ 0 \leq Accuracy \leq 1\ (good)$$

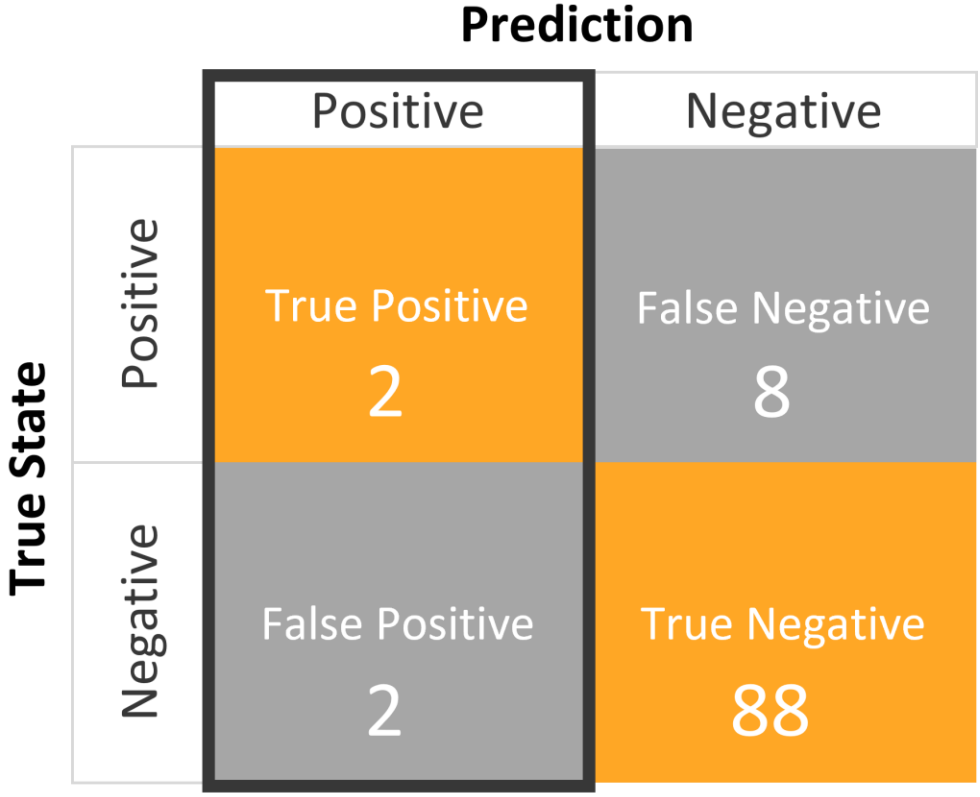
Métricas de evaluación - clasificación



Paradoja de Accuracy alta: Accuracy es engañosa cuando los datos están desbalanceados - pocos TP, la clase rara, y muchos TN, la clase dominante. **Accuracy alto aún cuando no se predice bien la clase rara.**

$$Accuracy = \frac{2 + 88}{2 + 2 + 8 + 88} = 0.90$$

Métricas de evaluación - clasificación



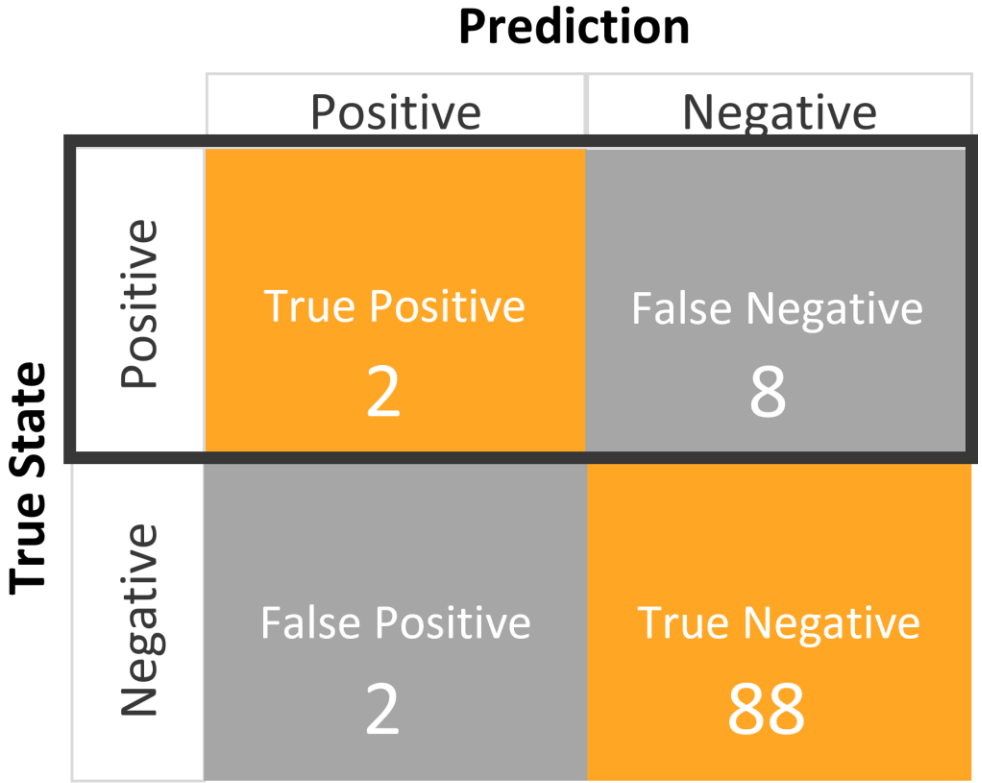
Precision*: Es la relación entre los positivos identificados correctamente (verdaderos positivos) y todos los positivos identificados.

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

$$Precision = \frac{2}{2 + 2} = 0.50$$

$^{*}(bad)\, 0 \leq Precision \leq 1\, (good)$

Métricas de evaluación - clasificación



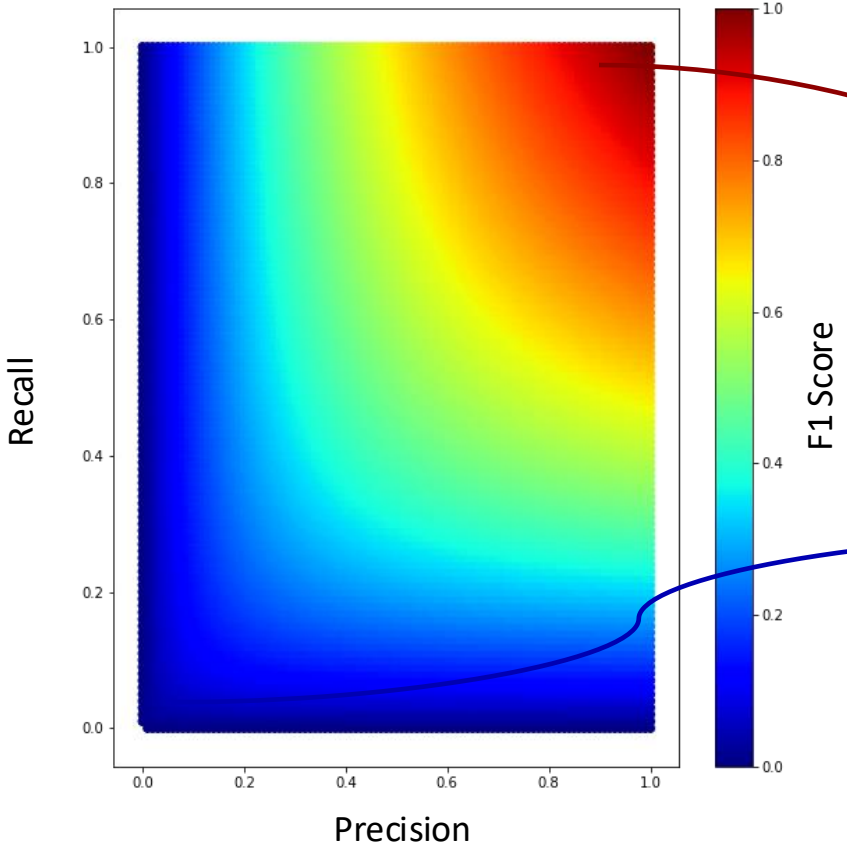
Recall*: Es la relación entre los positivos identificados correctamente (verdaderos positivos) y todos los positivos. Mide la habilidad del modelo para predecir una muestra positiva.

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

$$Recall = \frac{2}{2 + 8} = 0.20$$

$^{*}(bad)\ 0 \leq Recall \leq 1\ (good)$

Métricas de evaluación - clasificación



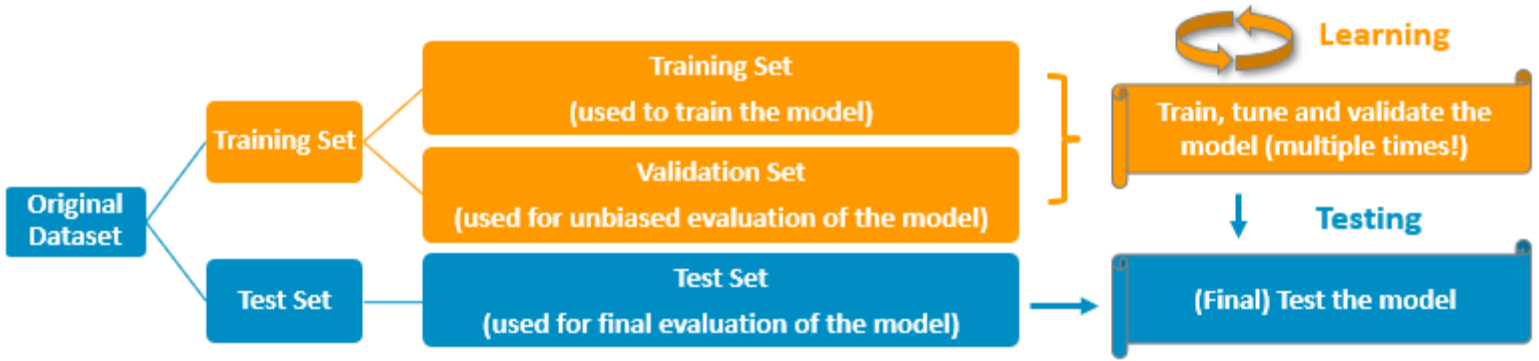
F1 score*:Una métrica combinada, la media harmonica de Presicion y Recall.

$$F1\,Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

$$^{*}(bad)\ 0 \leq F1\,Score \leq 1\ (good)$$

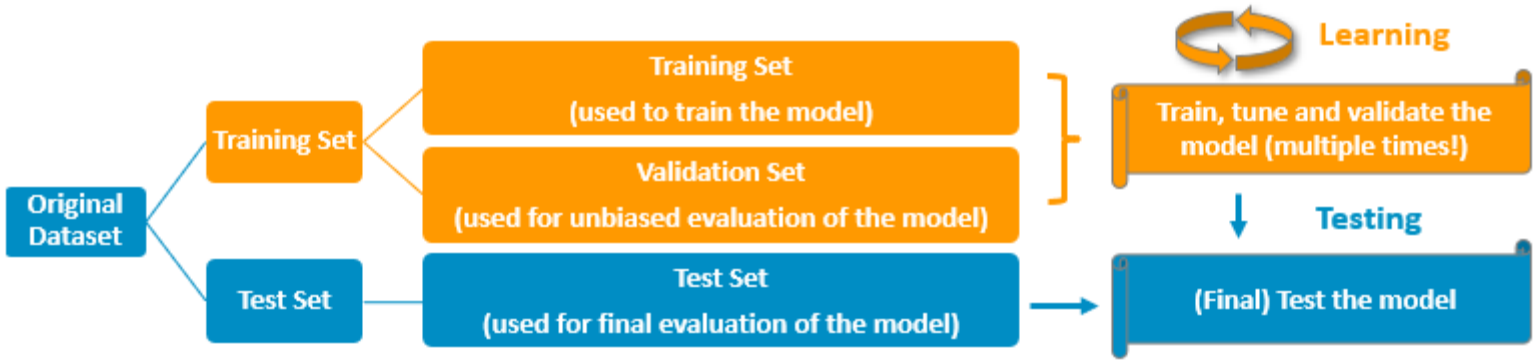
Bajo cuando Precision o Recall son bajos
Alto cuando Precision y Recall son altos

Conjuntos de entrenamiento, validación y prueba



El **conjunto de test no se utiliza para el aprendizaje (entrenamiento),** solo se usa para asegurar que el modelo **generaliza** bien en **nuevos datos "desconocidos".**

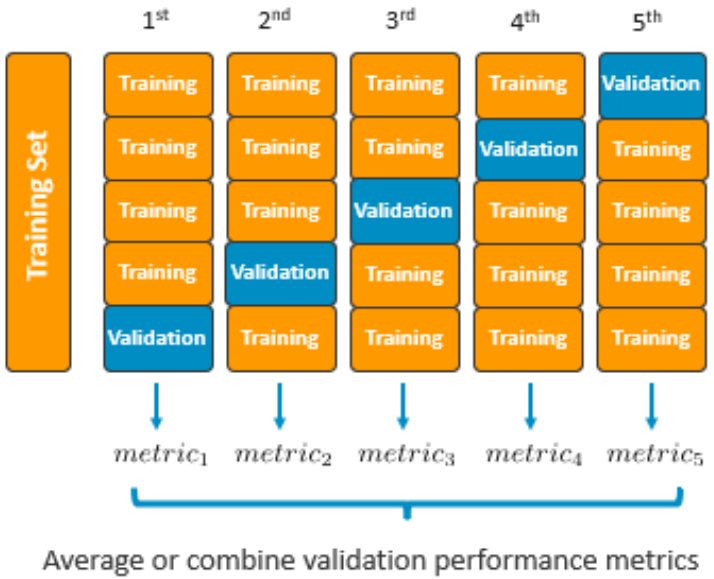
Conjuntos de entrenamiento, validación y prueba



	bad_weather	is_rush_hour	mile_distance	urban_address	late
0	0.0	1.0	5.00	1.0	0.0
1	1.0	0.0	7.00	0.0	1.0
2	0.0	1.0	2.00	1.0	0.0
3	1.0	1.0	4.20	1.0	0.0
4	0.0	0.0	7.80	0.0	1.0
5	1.0	0.0	3.90	1.0	0.0
6	0.0	1.0	4.00	1.0	0.0
7	1.0	1.0	2.00	0.0	0.0
8	0.0	0.0	3.50	0.0	1.0
9	1.0	0.0	2.60	1.0	0.0
10	0.0	0.0	4.10	0.0	1.0

Generalmente se “barajan” los datos antes de hacer las particiones para evitar sesgos en los datos resultantes

Conjuntos de entrenamiento, validación y prueba



K-fold cross-validation: Es una técnica de validación para ver que tan bien generaliza un modelo a un conjunto de validación independiente.

Se utilizan **K muestras reservadas** para validar el modelo, cada vez entrenando con las muestras restantes:

- Dividir el conjunto de entrenamiento en K grupos (folds).
- Repetir K veces el siguiente procedimiento:
  - Reservar el K^{th} fold para **validación**
  - Entrenar el modelo en los folds restantes
  - Calcular el desempeño en el fold de validación
- Combinar la métrica de rendimiento calculada

Sobre y sub entrenamiento

Subentrenamiento: El modelo no es lo suficientemente bueno para describir las relaciones entre los datos de entrada **X** y la variable de salida **y.**

- El modelo es muy simple para capturar patrones importantes en los datos
- El modelo tendrá un rendimiento bajo en los conjuntos de entrenamiento y validación

Sobre y sub entrenamiento

Sobreentrenamiento: El modelo memoriza o imita los datos de entrenamiento, y falla generalizando para datos desconocidos (datos de test).

- El modelo es demasiado complejo.
- El modelo aprende patrones de ruido y no de relaciones entre variables.
- Tendrá un rendimiento muy bueno en los datos de entrenamiento pero muy malo en los datos de validación o prueba.

Sobre y sub entrenamiento

Buen ajuste: El modelo captura las relaciones generales entre los datos de entrada **X** y la varianble de salida **y**.

- El modelo no es ni muy simple ni muy complejo.
- El modelo descifra las relaciones subyacentes de los datos de entrenamiento y no el ruido.
- El modelo tendrá un buen rendimiento en los conjuntos de entrenamiento, validación y prueba.