EDEM
Escuela de Empresarios

EDEM Escuela de Empresarios

ANÁLISIS Y APRENDIZAJE AUTOMÁTICO

IBERIA EXPRESS

Yvonne Gala García y Jesús Prada Alonso

Curso Máster en Data Analytics - Edición 1

Fecha 08/02/2020

PROFESORES (I)

EDEM
Escuela de Empresarios

Yvonne Gala García
Responsable Advanced Analytics en Iberia Express

Estudios:

- Doctorado en Aprendizaje Automático en la Universidad Autónoma de Madrid (UAM).
- Máster Universitario en Investigación e Innovación en TIC, especialidad inteligencia computacional, UAM.
- Máster Educación Secundaria Obligatoria, especialidad matemáticas, UAM.
- Licenciatura en Matemáticas, UAM.

Experiencia Laboral:

- Responsable Advanced Analytics en Iberia Express.
- Freelance Data Scientist: Iberia Express, Piperlab.
- Profesora en EDEM y DevAcademy.
- Investigadora en el Grupo de Aprendizaje Automático de la UAM.

PROFESORES (II)

EDEM
Escuela de Empresarios

Jesús Prada Alonso
Responsable Area Machine Learning en Sigesa y Data Scientist en Iberia Express

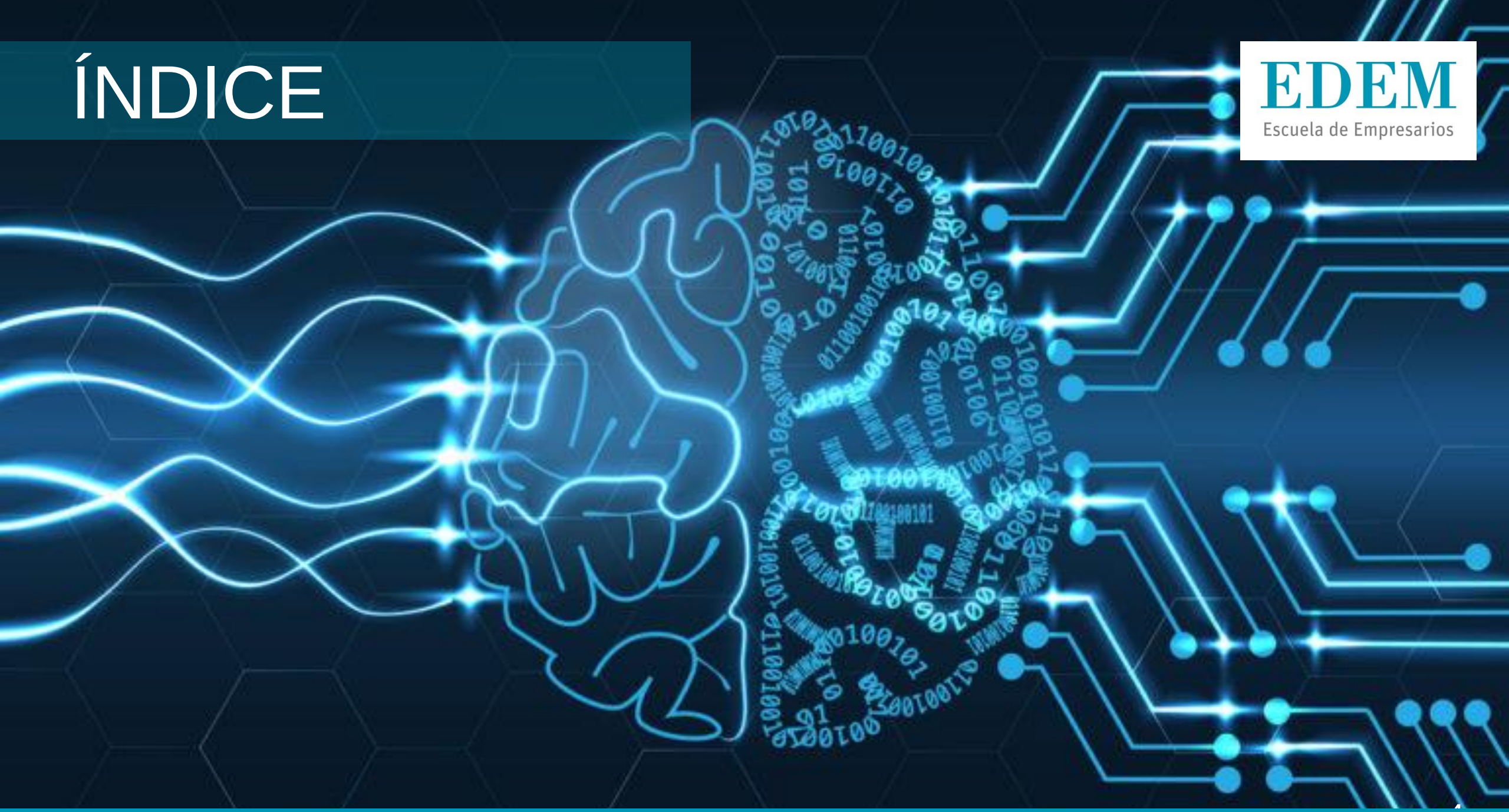
Estudios:

- Doctorado en Aprendizaje Automático en la Universidad Autónoma de Madrid (UAM).
- Doble Máster Universitario en Investigación e Innovación en TIC y en Matemáticas y Aplicaciones, UAM.
- Doble Titulación en Ingeniería Informática y Licenciatura en Matemática, UAM.

Experiencia Laboral:

- Responsable Area Machine Learning en Sigesa.
- Freelance Data Scientist: Iberia Express, M+ Vision Consortium, etc.
- Profesor en EDEM e IE School of Human Sciences and Technology, HST.
- Data Scientist en Kernel Analytics,
- Investigador en el Grupo de Aprendizaje Automático de la UAM.

ÍNDICE



EDEM
Escuela de Empresarios

ÍNDICE

EDEM
Escuela de Empresarios

BLOQUE I (27/02/2021): Introducción ML. ML en el sector turístico

TEORÍA

1. Introducción ML.
    - ¿Qué es Machine Learning?
    - Estructura proyecto ML.
    - Casos prácticos sector turístico.
2. Conceptos básicos ML.
    - Conjuntos Train/Test/Validación. Cross validation.
    - Métricas.
    - Metaparametrización.
    - Trade off bias/variance. Overfitting/underfitting.
3. Aprendizaje Supervisado Vs No supervisado.
    - Supervisado: Clasificación vs Regresión.
    - No supervisado: Clustering, Reducción de dimensionalidad.
    - Otros: Sistema de Recomendación.

ÍNDICE

EDEM
Escuela de Empresarios

BLOQUE I (27/02/2021): Introducción ML. ML en el sector turístico

PRÁCTICA

Datasets:

- Dataset scikit-learn.
- Ejemplo sistema de recomendación pequeño en sector turístico.
- Ejemplo sistema de recomendación películas.

Ejercicios:

- Introducción a scikit-learn.
- Split data set, evaluación, overfitting…
- Sistema de recomendación de vuelos.
- Sistema de recomendación de películas.

ÍNDICE

EDEM
Escuela de Empresarios

BLOQUE II (05/03/2021): Modelos ML I, regresión lineal, logística y SVM. Clasificación y Regresión.

TEORÍA

1. Regresión lineal
    - Conceptos.
    - Regularización. Lasso / Ridge.
2. Regresión logística
    - Conceptos.
    - Regularización. Lasso / Ridge.
3. SVM
    - Conceptos.
    - Clasificación.
    - Regresión.

ÍNDICE

EDEM
Escuela de Empresarios

BLOQUE II (05/03/2021): Modelos ML I, regresión lineal, logística y SVM. Clasificación y Regresión

PRÁCTICA:

Datasets:

- Dataset scikit-learn.
- Datos médicos.
- Datos aerolínea (Iberia Express).

Ejercicios:

- Script de regresión lineal, logística, SVM.
- Problema de clasificación en medicina.
- Problema de regresión en una aerolínea. Predicción de demanda.

ÍNDICE

EDEM
Escuela de Empresarios

BLOQUE III (06/03/2021): Modelos ML II, árboles de decisión, random forest, xgboost. Clasificación y Regresión.

TEORÍA

1. Árboles de decisión.
    - Conceptos.
    - Representación gráfica.
    - Hiperparámetros.
2. Random Forest.
    - Conceptos.
    - Ensembles de árboles.
    - Hiperparámetros.
3. XGBoost.
    - Conceptos.
    - Intuición Random Forest vs XGBoost.
    - Hiperparámetros.
    - Comparación. Interpretabilidad vs Accuracy

ÍNDICE

EDEM
Escuela de Empresarios

BLOQUE III (06/03/2021): Modelos ML II, árboles de decisión, random forest, xgboost. Clasificación y Regresión.

PRÁCTICA:

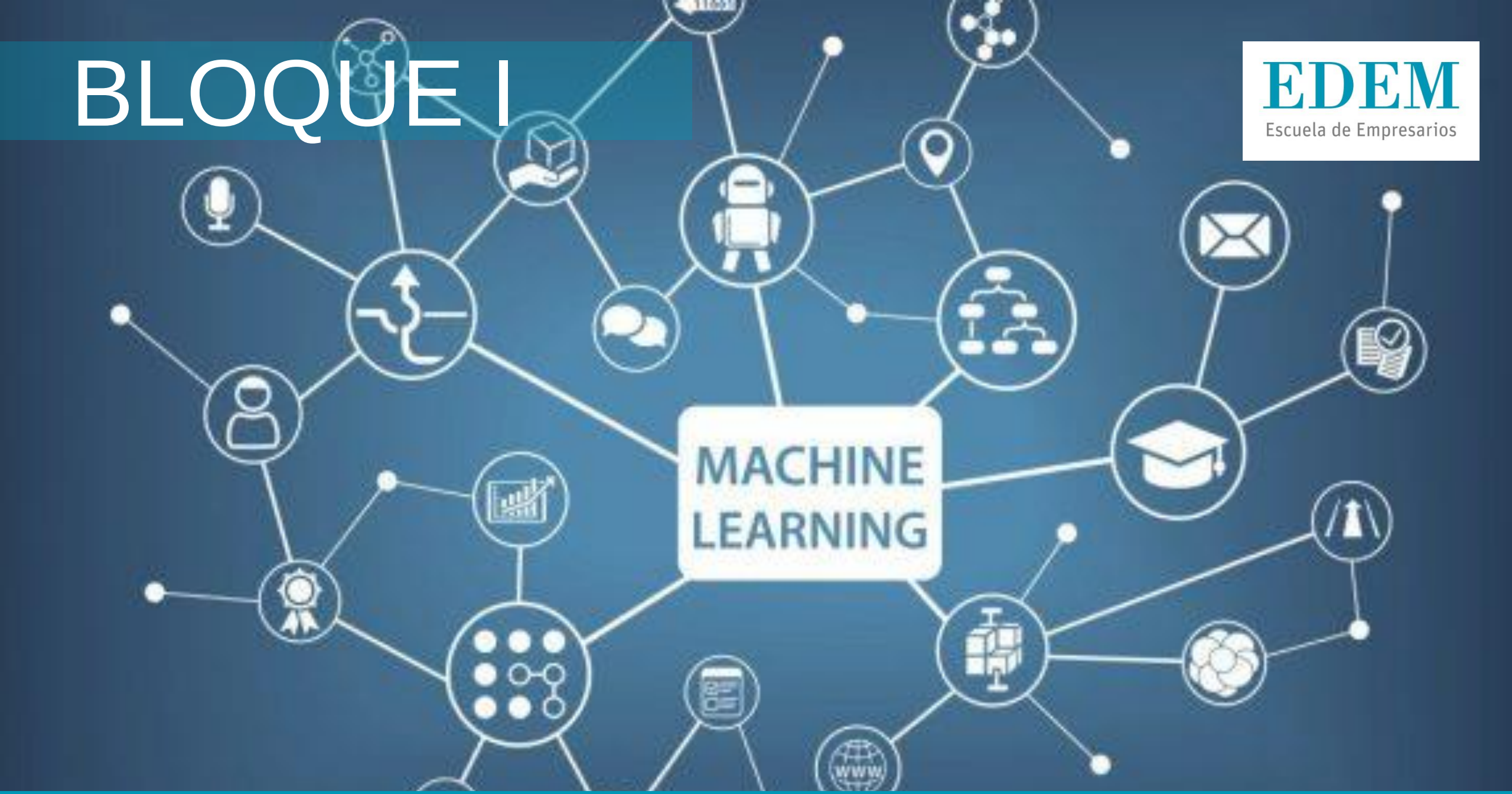
Datasets:

- Dataset scikit-learn.
- Datos médicos.
- Datos aerolínea (Iberia Express).

Ejercicios:

- Script de árboles de decisión, random forest, xgboost.
- Problema de clasificación en medicina. Comparar los 3 modelos y optimización de parámetros.
- Problema de regresión en una aerolínea. Predicción de demanda. Comparar los 3 modelos y optimización de parámetros.
- Comparación final de métricas de todos los modelos vistos.

BLOQUE I



INTRODUCCIÓN ML

EDEM
Escuela de Empresarios

Machine Learning

Introduction

What is machine learning

Machine Learning

Introduction

What is machine learning

Machine Learning

Introduction

What is machine learning

Machine Learning

Introduction

What is machine learning

¿QUÉ ES DATA SCIENCE?

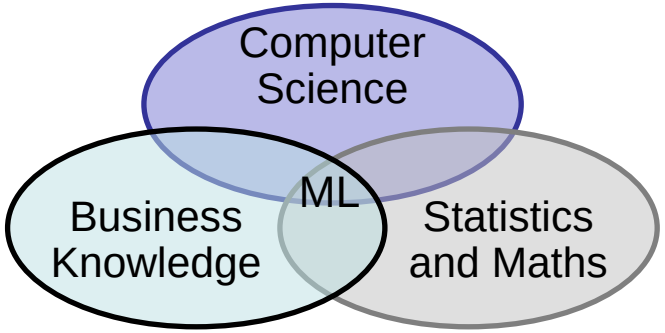
EDEM
Escuela de Empresarios

¿Qué es Data Science y Machine Learning?

Data Science es un conjunto de herramientas, técnicas y disciplinas que se enfocan en convertir grandes cantidades de datos en información útil para explicar la relación entre variables y generar modelos predictivos.

Machine Learning, ML, o **Aprendizaje Automático** es una rama de la Inteligencia Artificial cuyo objetivo es construir sistemas que aprendan automáticamente de los datos.

A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E.



ESTRUCTURA PROYECTO ML

EDEM
Escuela de Empresarios

Paso previo - Plantear un problema o hipótesis a demostrar.

Paso 1 - Recolección de datos.

Paso 2 - Preproceso.

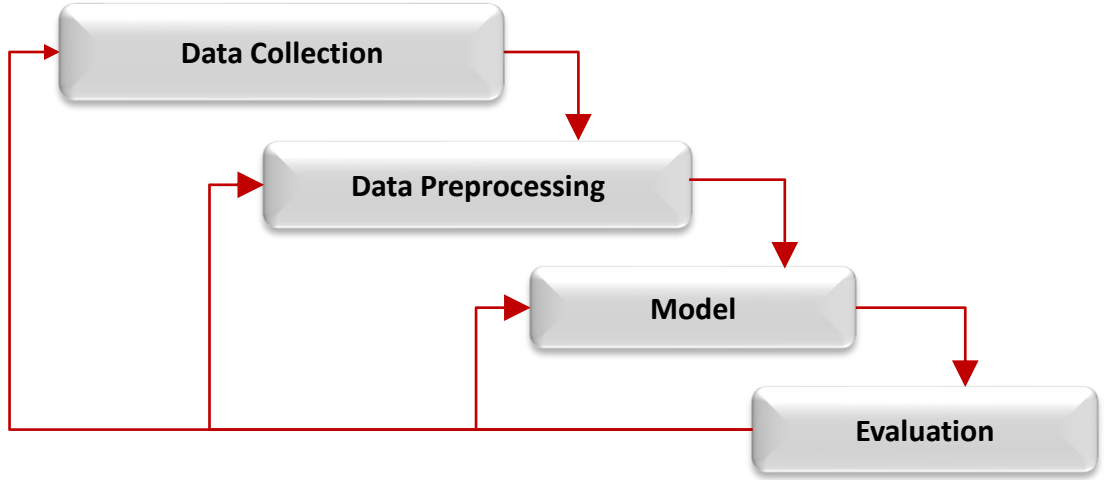
Paso 3 - Modelización.

Paso 4 - Validación.

Paso 5 - Conclusiones/informes/visualización.

Paso 6 - Puesta en producción.

Paso posterior - Mantenimiento.

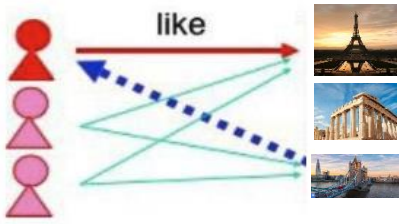


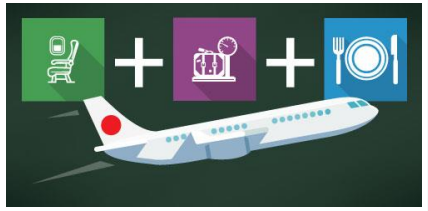
EDEM
Escuela de Empresarios

Proyectos

Recomendación

Rutas y Ancillaries





Estimaciones

Pasajeros, Maletas y No show

Optimización de precios

Billetes, Maletas y Asientos

Detección de fraude web

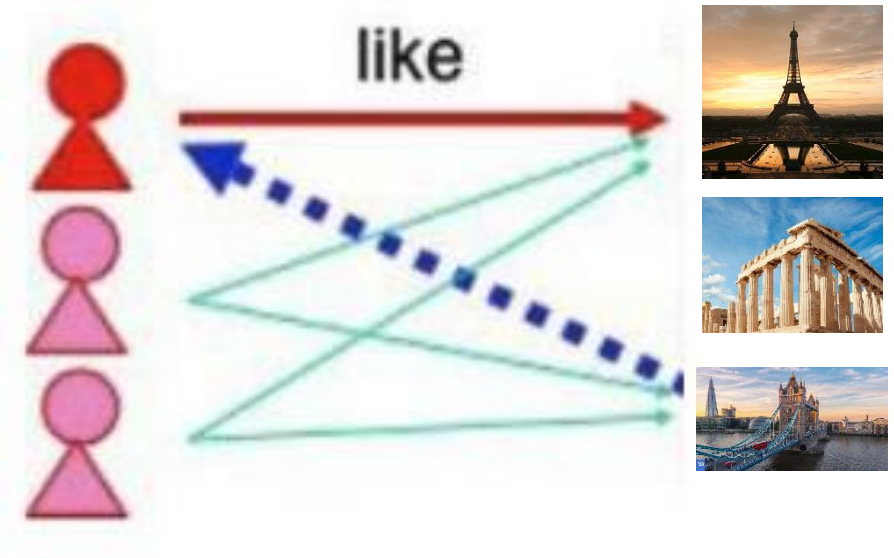
EDEM
Escuela de Empresarios

Sistema de recomendación

Elección de la ruta a recomendar a cada cliente en función de sus interacciones con Iberia Express.

Sirve para personalizar:

- Campañas de emails
- Página web:
  - Home
  - Gestión de Reserva
  - Check in



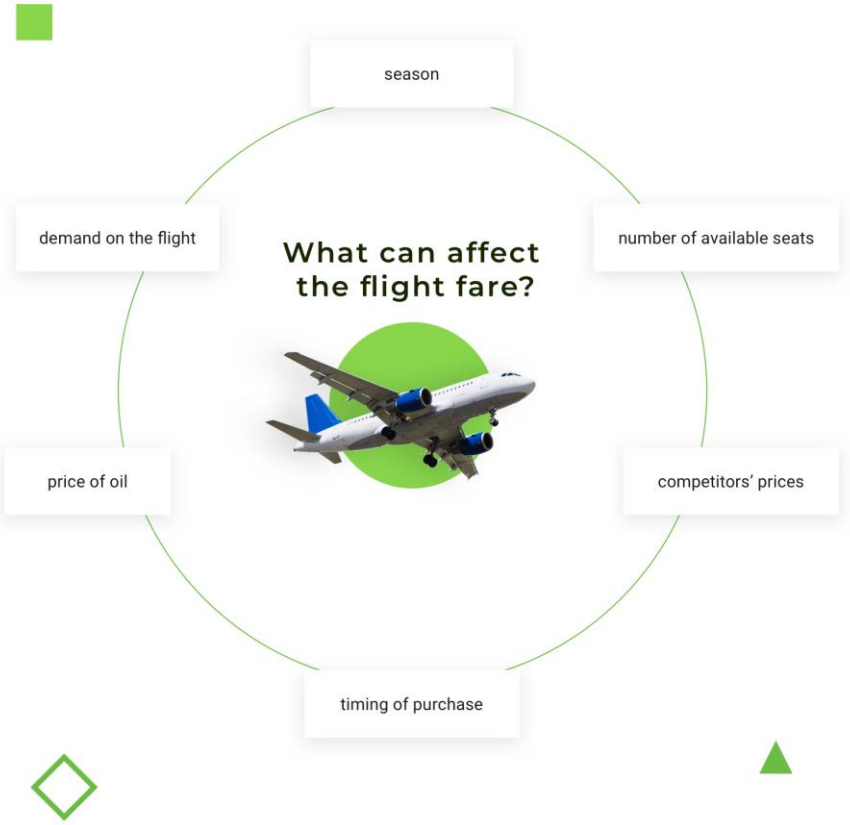
EDEM
Escuela de Empresarios

Optimización de precios

Elección del precio óptimo en billetes vendidos por la web.

Basado en:

- Características del vuelo
- Histórico de llenado del vuelo
- Momento de la compra
- Disponibilidad de asientos
- Precio de la competencia
- Búsquedas web Iberia express
- Búsquedas competencia



EDEM
Escuela de Empresarios

Overbooking

Predicción de overbooking en los aviones. Es decir, billetes vendidos por encima de la capacidad del vuelo.

Con este proyecto se optimiza el número de asientos vendidos, maximizando el revenue por vuelo y minimizando el impacto negativo sobre el cliente.

EDEM
Escuela de Empresarios

Fraude (I). Descripción

Detección de compras fraudulentas en la web.

Basado en:

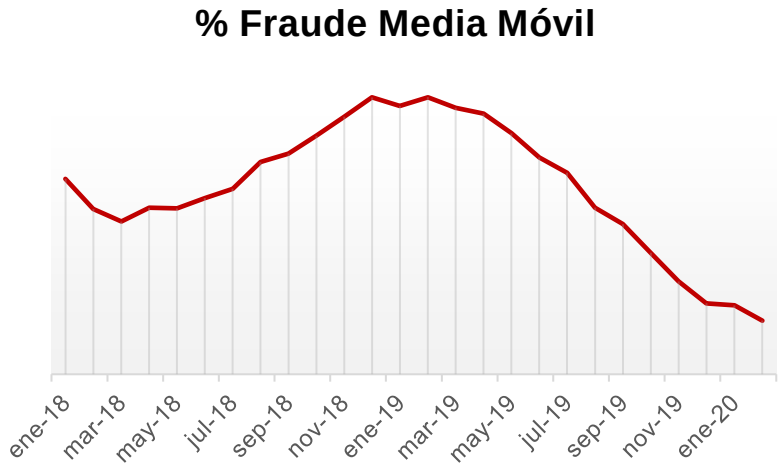
- Características del vuelo
- Características de la compra
- Características del comprador
- Histórico de fraude anterior
- Momento de la compra

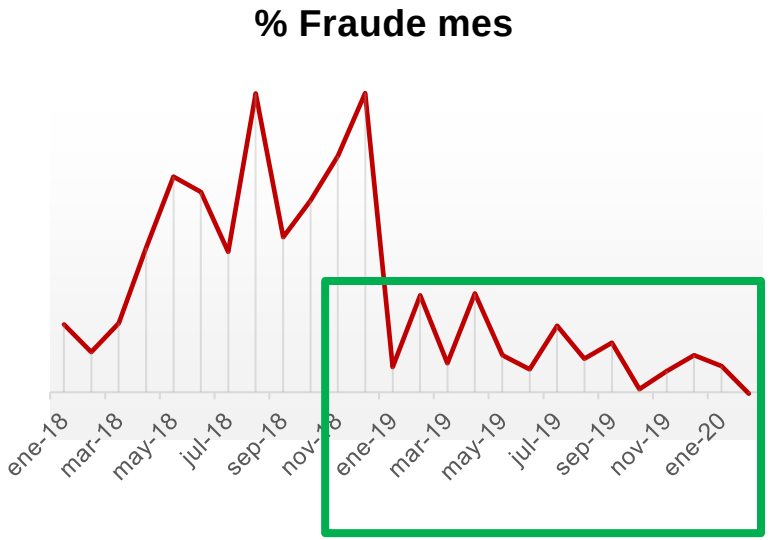
Este proyecto ha perdido importancia en el ultimo año tras la entrada en vigor de la psd2.

EDEM
Escuela de Empresarios

Fraude (II). Resultados

El modelo se puso en producción el 2019-01-01





El fraude se redujo 10 veces de noviembre de 2018 a enero de 2019

EDEM
Escuela de Empresarios

Bajada de maletas

Predicción de bajada de maletas en el embarque del avión en función del llenado y características del vuelo.

Con este proyecto se avisa a los usuarios para facturar su maleta en el mostrador de facturación evitando retrasos y haciendo mejor la experiencia del usuario.

En covid se ha dejado de hacer pues las maletas se bajan GRATIS SIEMPRE para cumplir el distanciamiento social durante el embarque.

EDEM
Escuela de Empresarios

Analítica pre y post COVID

Las aerolíneas es uno de los sectores económicos más afectados por esta crisis sanitaria.

Esto tiene una implicación directa en los datos.

- Han **cambiado los patrones de comportamiento** de los usuarios. Tanto de compra (ej: recomendaciones), como de vuelo (precios dinámicos).
- Han **cambiado las necesidades de negocio** → muchos productos no se pueden ofertar y la empresa tiene otros objetivos.
- **No tenemos un histórico** de los nuevos comportamientos de los clientes **en el entorno actual**.

Soluciones

- Adaptar los proyectos según necesidades.
- Adaptar los datos, quitando información pre covid o dándole menos importancia.

EDEM
Escuela de Empresarios

Estimación de reservas COVID

- **Objetivo:** abrir rutas que puedan tener volumen de pasajeros en los próximos meses según la evolución del COVID.
- **Datos:**
  - Incluimos solo información del periodo covid.
  - Información de la evolución de la pandemia, número de casos y vacunas.
- **Resultados:** Estimación de reservas a dos o tres meses.

Estimación no show de pasajeros COVID

- **Objetivo:** optimizar el llenado del avión cumpliendo medidas sanitarias.
- **Datos:**
  - Incluimos solo información del period covid y variables que puedan aumentar la probabilidad de no show.
- **Resultados:** predicciones de reservas y no show.

CONCEPTOS ML

EDEM
Escuela de Empresarios

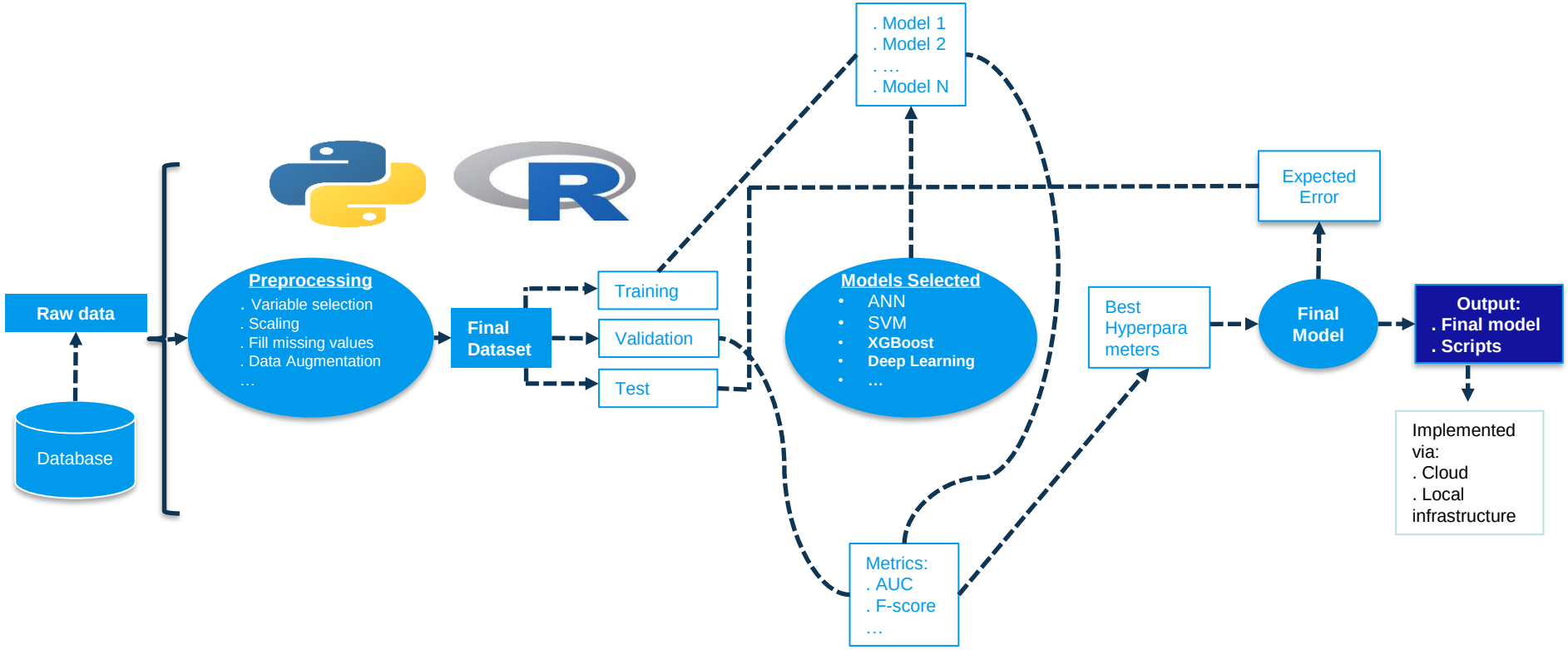
CONCEPTOS ML

EDEM
Escuela de Empresarios

- Conjuntos Train/Validación/Test. Cross-validation.
- Métricas.
- Metaparametrización.
- Trade off bias/variance. Overfitting/Underfitting.

ESQUEMA ML

EDEM
Escuela de Empresarios



EDEM
Escuela de Empresarios

Conjuntos Train/Validación/Test

Tipos de conjuntos

- Muestra de Entrenamiento (TRAINING) : Datos de los que los modelos extraen patrones.
- Muestra de Validación (VALIDATION) : Se emplea para seleccionar el mejor de los modelos entrenados cuando realizamos el ajuste de parámetros o **metamodelización**.
- Muestra de Prueba (TEST) : Proporciona el error real esperado con el modelo seleccionado.

Consideraciones de los conjuntos de train, validación y test

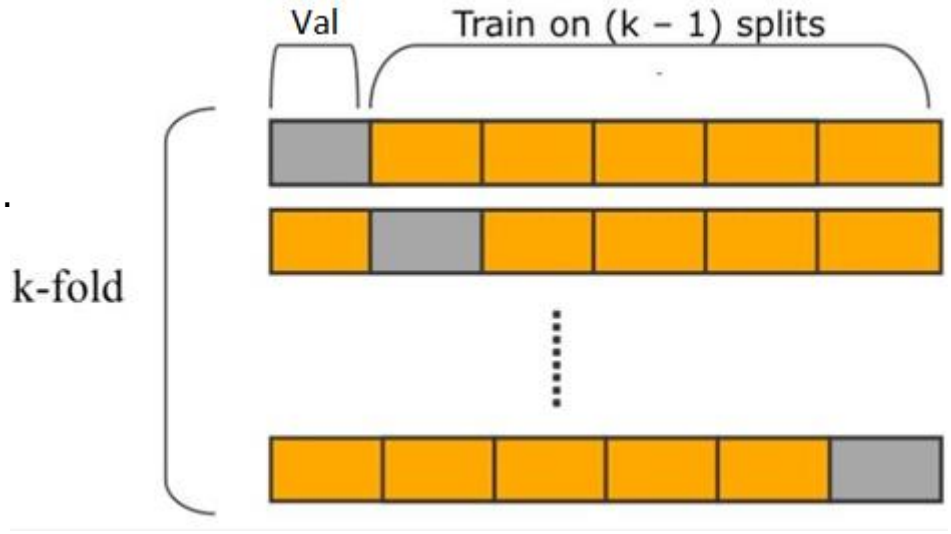
- Que sean lo **suficientemente grandes** como para generar resultados significativos desde el **punto de vista estadístico**.
- Que sean **representativos** de todo el conjunto de datos. Es decir, no elegir un conjunto de prueba con características diferentes (**sesgo**) al del conjunto de entrenamiento.
- **No existe** una solución óptima (**golden rule**) para elegir el porcentaje del total de datos asignado a cada conjunto, ya que depende del problema. Pero un estándar típico es 70/15/15.

EDEM
Escuela de Empresarios

Cross-validation

Es un método alternativo a la división en train/val/test para realizar la optimización de hiperparámetros. Permite no tener que crear un conjunto de validación, sustituyendo su funcionalidad por la siguiente metodología:

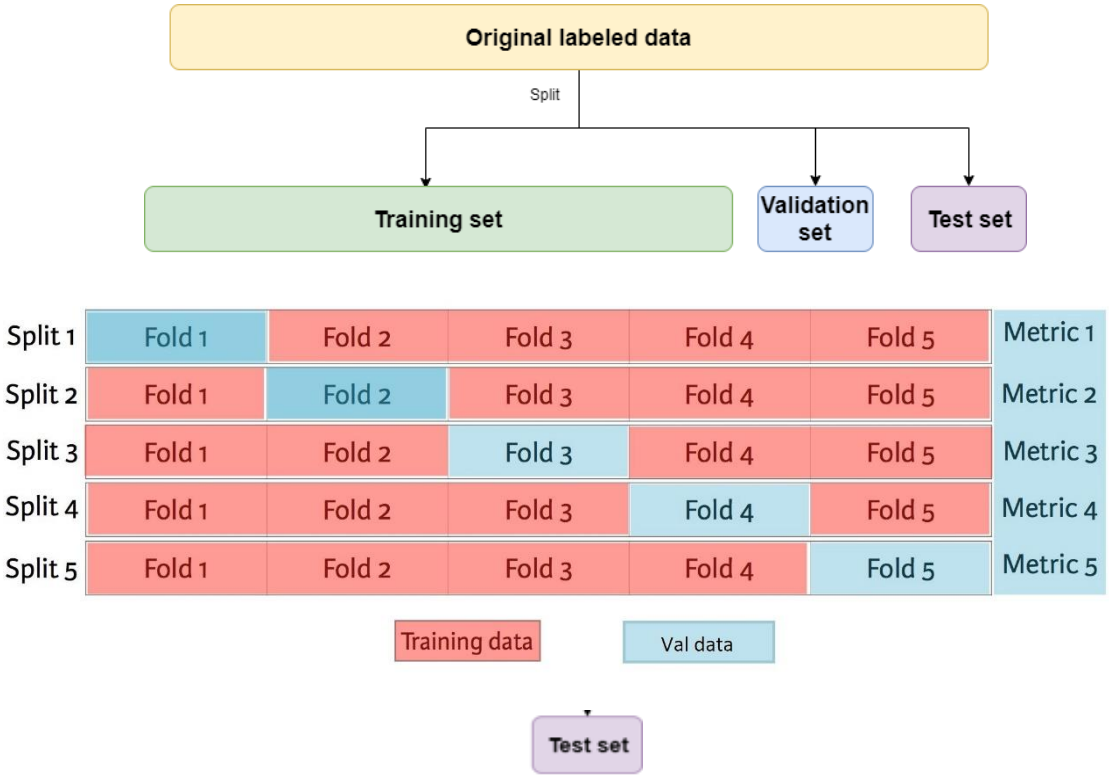
- Se hace una separación del datatest en k subconjuntos del mismo tamaño.
- Se utiliza k -1 conjuntos para entrenar y 1 para validación.
- Se repite el procedimiento k veces rotando el conjunto de validación.
- Se evalúa cada iteración con la métrica seleccionada.
- Finalmente se calcula la media de las k métricas como error de validación final.



EDEM
Escuela de Empresarios

Cross-validation VS. Validación fija

- Cross-validation puede tener efectos negativos cuando existe una dimensión temporal en el problema. ✘
- Cross-validation permite aprovechar más volumen del dataset para su uso como train en el entrenamiento de los modelos. ✓
- Cross-validation es más costoso computacionalmente. ✘
- Validación fija implica tener que estimar un porcentaje óptimo para 3 conjuntos en lugar de 2. ✓



EDEM
Escuela de Empresarios

Métricas

- Para comparar el rendimiento obtenido por cada combinación de tipo de modelos y conjunto de hiperparámetros necesitaremos de un valor numérico que nos informe de su bondad predictiva.

- Este valor numérico vendrá dado por la métrica elegida.

- La elección de esta métrica dependerá del tipo de problema, de los datos y del objetivo a solucionar.

- Ejemplo: MAE

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|, \text{ where } e_t = original_t - predict_t$$

EDEM
Escuela de Empresarios

Metaparametrización (I)

- Los modelos ML suelen incluir un conjunto de **hiperparámetros** que nos permiten controlar su comportamiento.
- De su correcta elección dependerá la bondad del modelo entrenado.
- Los hiperparámetros dependen del perfil de los datos que estamos analizando (**problem-dependent**), por lo que no es sencillo establecer un procedimiento ad-hoc para su obtención.
- Lo habitual es aplicar una técnica llamada **búsqueda en rejilla**.

EDEM
Escuela de Empresarios

Metaparametrización (II). Grid search

1. Elegimos una familia de modelos.
2. Elegimos unos hiperparámetros a optimizar, les llamaremos par1 y par2.
3. Para cada hiperparámetro, elegimos una serie de valores a probar.

par1/par2	10	100	1000
0.1	0.3	0.22	0.25
0.01	0.15	0.14	0.14
0.001	0.35	0.05	0.11

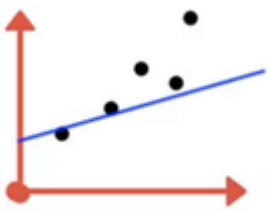
4. Entrenamos nuestro modelo sobre el conjunto de train con los diferentes hiperparámetros haciendo todas las combinaciones posibles.
5. Hacemos la predicción de los diferentes modelos sobre el conjunto de validación y calculamos el error con la métrica seleccionada.
6. Escogemos el que mejor métrica obtenga y lo aplicamos sobre el conjunto de test para ver el error final esperado de nuestro modelo.

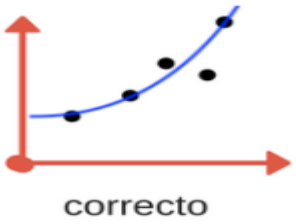
EDEM
Escuela de Empresarios

Trade off bias/variance

Bias:

- El sesgo es la diferencia entre la predicción promedio de nuestro modelo y el valor correcto que estamos tratando de predecir.
- El modelo con alto sesgo presta muy poca atención a los datos de entrenamiento y simplifica en exceso el modelo.
- Un modelo muy sesgado siempre da un error alto en los datos de train. **Underfitting**.





Variance:

- Es la variabilidad de las predicciones cuando se introducen datos que difieren entre sí.
- El modelo con alta variación se ajusta mucho a los datos de entrenamiento y no generaliza bien con datos que no ha visto antes.
- Dichos modelos funcionan muy bien con los datos de entrenamiento pero tienen altos índices de error en los datos de prueba. **Overfitting**.

EDEM
Escuela de Empresarios

Overfitting/underfitting (I)

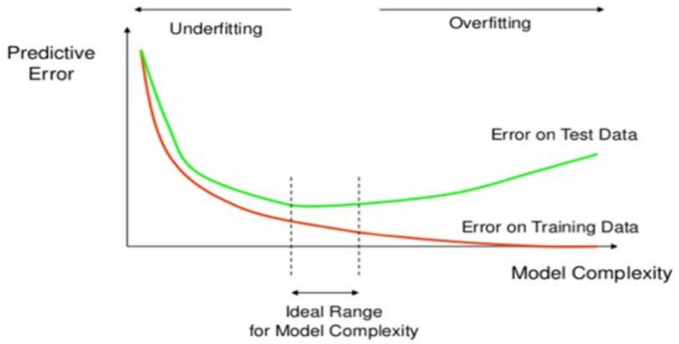
Las principales causas al obtener malos resultados en ML son el overfitting o el underfitting de los datos.

- Si nuestros datos de entrenamiento son muy pocos o nuestro modelo es demasiado sencillo no será capaz de aprender a resolver el problema → **underfitting** (High bias - Low variance).

- Cuando el algoritmo sólo se ajusta a aprender los casos particulares que le enseñamos y es incapaz de reconocer nuevos datos de entrada → **overfittting** (Low bias - High variance)

EDEM
Escuela de Empresarios

Overfitting/underfitting (II)



¿Cómo detectar el overfitting?

Si el modelo entrenado tiene en el conjunto de validación un error mucho mayor que en el conjunto de train, esto sugiere la posibilidad de un problema de overfitting.

¿Cómo detectar el underfitting?

Cuándo el error de train parece demasiado elevado, podemos tener sospechas de underfitting.

También si en el conjunto de validación sólo se acierta un tipo de clase o el único resultado que se obtiene es siempre el mismo valor, o valores similares.

EDEM
Escuela de Empresarios

Overfitting/underfitting (III)

Prevenir el overfitting

- **Cantidad mínima y variada en las muestras** tanto para entrenar el modelo como para validarlo.
- **Clases variadas y equilibradas** en cantidad: es importante que los datos de entrenamiento estén balanceados.
- **Conjunto de validación** de datos. Siempre subdividir nuestro conjunto de datos y mantener una porción del mismo «oculto» a nuestra máquina entrenada.
- Parameter Tunning o **Ajuste de Parámetros**: deberemos experimentar con distintas configuraciones hasta encontrar el equilibrio.
- A veces conviene eliminar o reducir la cantidad de características que utilizaremos para entrenar el modelo, por ejemplo cuando se tiene una cantidad excesiva de dimensiones (features), con muchas variantes distintas, sin suficientes muestras. Una herramienta útil para hacerlo es PCA.

SUPERVISADO VS NO SUPERVISADO

EDEM
Escuela de Empresarios

SUPERVISADO VS NO SUPERVISADO

EDEM
Escuela de Empresarios

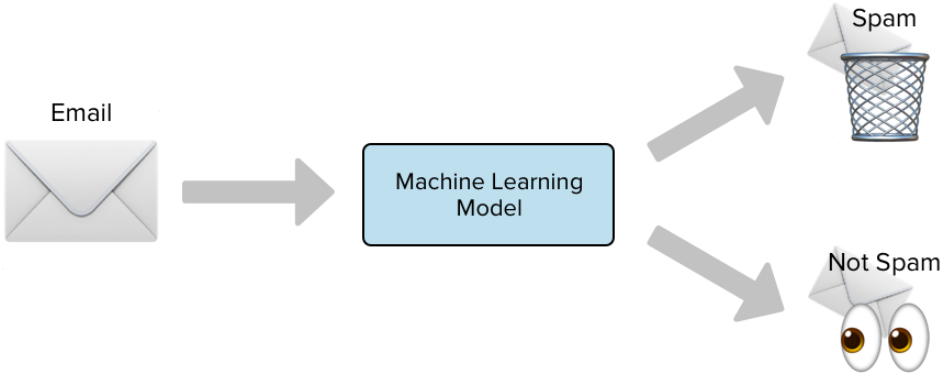
- Conceptos.
- Aprendizaje supervisado: Clasificación vs Regresión.
- Aprendizaje no supervisado: Clustering, reducción de dimensionalidad.

EDEM
Escuela de Empresarios

Aprendizaje supervisado VS no supervisado (I)

Aprendizaje supervisado:

Para entrenar el modelo se utiliza un dataset o conjunto de **muestras etiquetado (train)**. **El objetivo es predecir** la etiqueta que tendrán futuras muestras (test) que el modelo no ha visto en su entrenamiento.



Ejemplo:

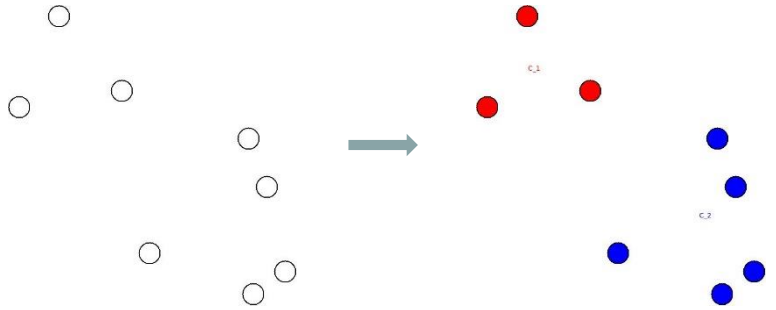
- Clasificar si un correo es spam o no.
- Predecir la producción de energía solar producida en una planta.

EDEM
Escuela de Empresarios

Aprendizaje supervisado VS no supervisado (II)

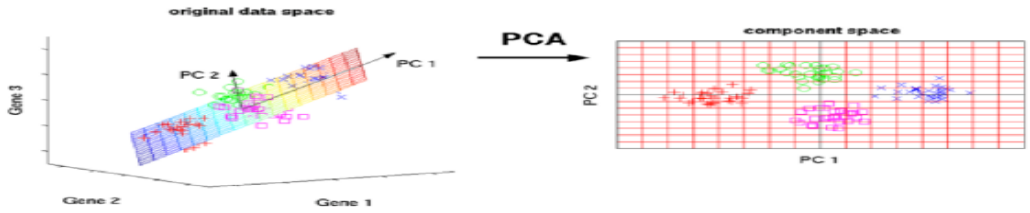
Aprendizaje no supervisado:

Para entrenar el modelo se utiliza un dataset o conjunto de **muestras sin etiquetar**. **El objetivo es encontrar patrones** en los datos para extraer conocimiento útil.



Ejemplo:

- Segmentar tus usuarios en 2 grupos.
- Reducción a 2 dimensiones.



RESUMEN

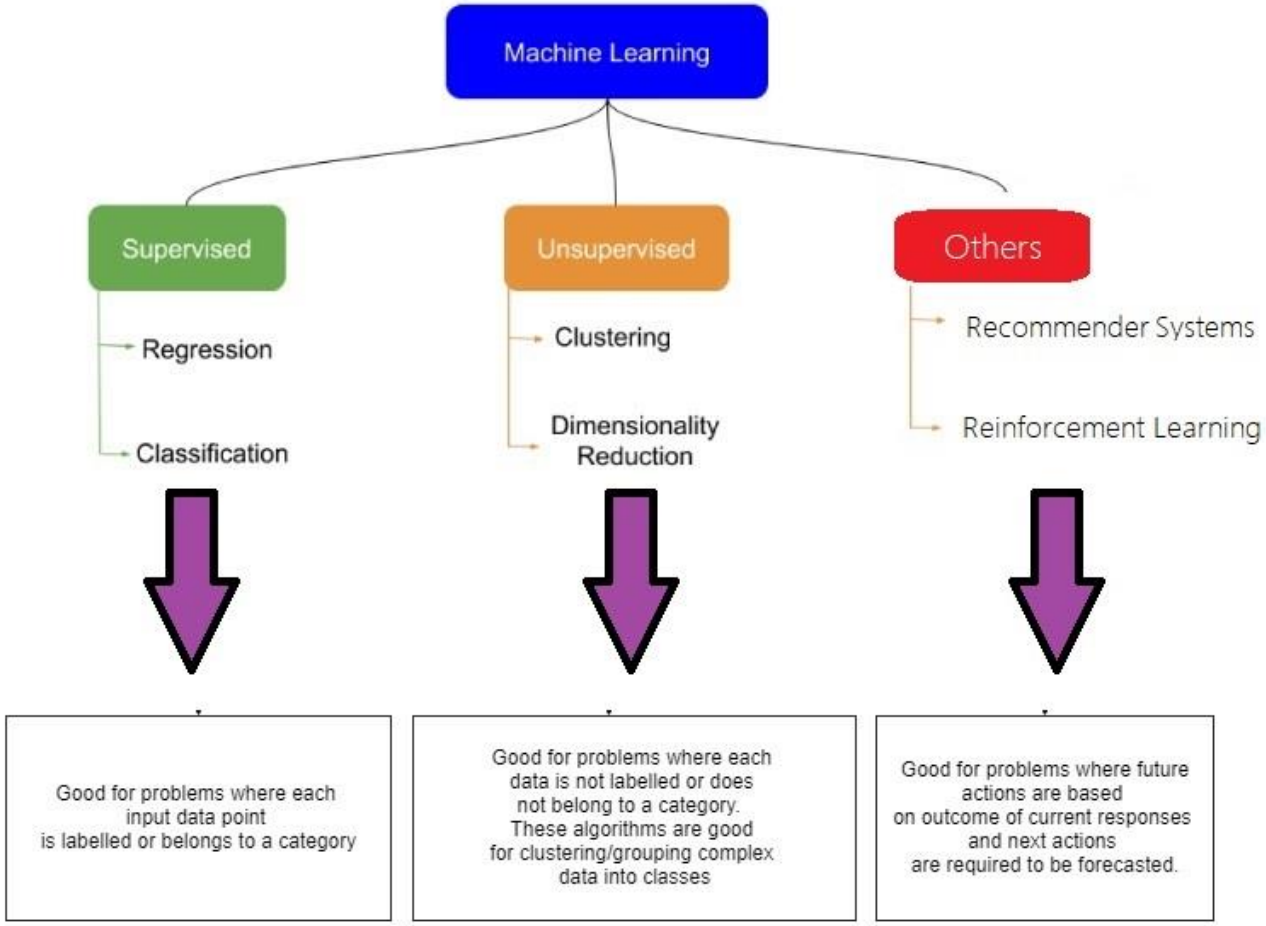
EDEM
Escuela de Empresarios

Aprendizaje Supervisado vs No Supervisado

	Supervisado	No supervisado
Etiquetas	SI	NO
Objetivo	Dar predicciones a futuro sobre el conjunto de test	Encontrar patrones en los datos o reducir dimensiones
Modelos	Regresión lineal, árboles, SVM, Redes Neuronales	Clustering, PCA
Ejemplo	Predecir si una transacción es fraudulenta	Encontrar clientes con perfiles similares

RESUMEN

EDEM
Escuela de Empresarios



EDEM
Escuela de Empresarios

Clasificación

Las etiquetas son categóricas, indicando la pertenencia de una determinada muestra a una clase en particular.

Ejemplo:

TRAIN


→ 1


→ 0


→ 1


→ 1

TEST


→ 1 ✓


→ 0 ✓


→ 1 × **¿Overfitting?**

SUPERVISADO (II)

EDEM
Escuela de Empresarios

Regresión

Las etiquetas son numéricas, indicando un valor asociado a cada muestra.

Ejemplo:

TRAIN

Superficie	Antigüedad	Ciudad	Precio
50	1	Madrid	1000€
50	1	Algete	600€
…	…	…	…
100	10	Sevilla	650

TEST

Superficie	Antigüedad	Ciudad	Precio
30	15	Madrid	?
230	5	Galicia	?
…	…	…	…
80	1	Canarias	?

EDEM
Escuela de Empresarios

Clasificación. Matriz de confusion

Nos dará un conteo de los aciertos y errores de cada una de las clases por las que estemos clasificando.

		Clasificador	
		+	-
Valor real	+	TP	FN
	-	FP	TN

- **TP – True Positives:** Es el número verdaderos positivos, es decir, de predicciones correctas para la clase +.
- **FN – False Negatives:** Es el número de falsos negativos, es decir, la predicción es negativa cuando realmente el valor tendría que ser positivo.
- **FP – False Positives**: Es el número de falsos positivos, es decir, la predicción es positiva cuando realmente el valor tendría que ser negativo
- **TN – True Negatives:** Es el número de verdaderos negativos, es decir, de predicciones correctas para la clase -.

MÉTRICAS (II)

EDEM
Escuela de Empresarios

Clasificación

- **Accuracy:** Porcentaje de aciertos del modelo.

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

- **Sensibilidad, recall, VPR:** ratio de verdaderos positivos.

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Especifidad, VNR:** ratio de verdaderos negativos.

$$\text{Especifidad} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- **Precisión**: probabilidad de que, dada una predicción positiva, la realidad sea positiva también.

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

- **F1-score**: f1-score es una medida que mezcla la precision y el recall. Mide si nuestro modelo tiene falsos positivos y falsos negativos a la vez.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

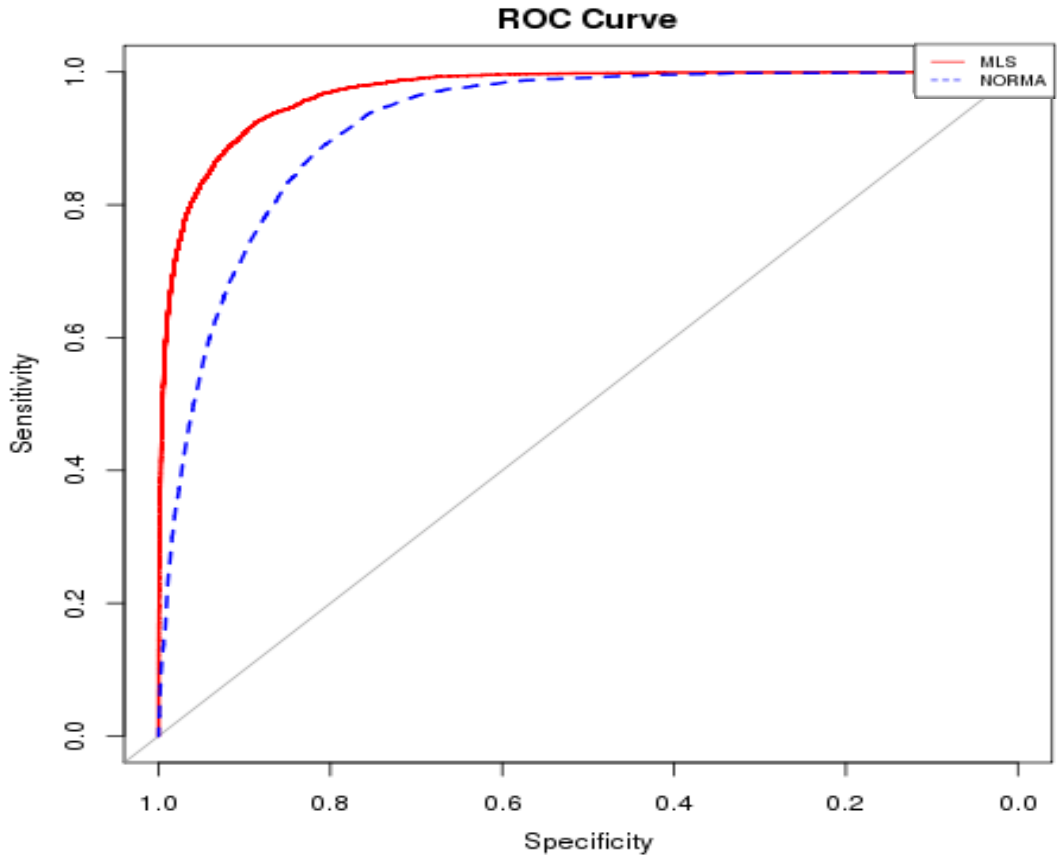
MÉTRICAS (III)

EDEM
Escuela de Empresarios

Clasificación. AUC

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Sensibilidad, recall, VPR: ratio de verdaderos positivos.



Especifidad, VNR: ratio de verdaderos negativos.

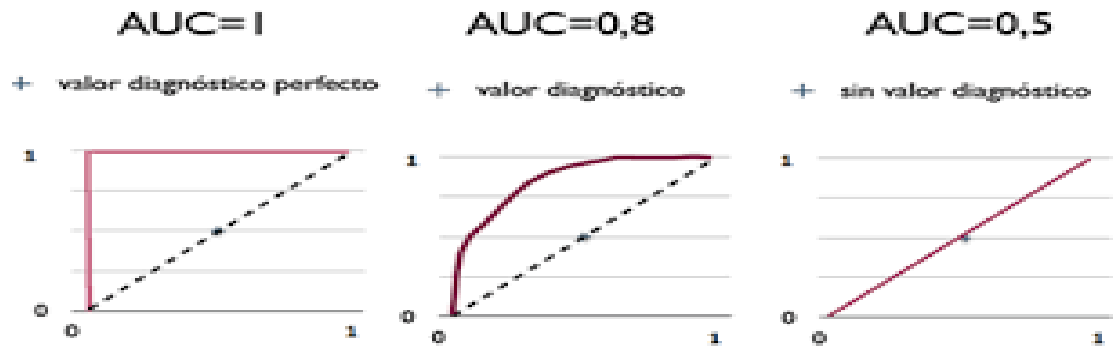
$$\text{Especifidad} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

MÉTRICAS (IV)

EDEM
Escuela de Empresarios

Clasificación. AUC

- La curva ROC se define por FPR (Ratio Falsos Positivos) y VPR (Ratio True Positive) como ejes x e y respectivamente.
- Representa los intercambios entre verdaderos positivos (beneficios) y falsos positivos (costes).
- Dado que VPR es equivalente a sensibilidad y FPR es igual a 1-especificidad, el gráfico ROC representa la sensibilidad frente a (1-especificidad).
- Cada valor umbral usado como punto de corte para distinguir entre qué es una predicción positiva y qué una negativa representa un punto en el espacio ROC.



MÉTRICAS (V)

EDEM
Escuela de Empresarios

Regresión

- **MAE o Error absoluto medio:** es la media de la diferencia absoluta entre los puntos de datos reales y el valor de predicción**.**

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

- **MSE o Error cuadrático medio:** es la media de la diferencia entre los puntos reales de datos y el valor de predicción al cuadrado. Penaliza más las diferencias mayores o extremas.

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

- **RMSE:** Raíz cuadrada del MSE. Proporciona mayor intuición que el MSE.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **MAPE o Error absoluto porcentual medio:** Permite medir error relativos a la magnitud del valor real.

$$\mathrm{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

RESUMEN

EDEM
Escuela de Empresarios

Aprendizaje Supervisado. Clasificación vs Regresión

	Clasificación	Regresión
Etiquetas	Categóricas.	Numéricas.
Ejemplo	Predecir si un email es spam (1) o no (0).	Predecir el precio de alquiler de una casa (550,632,1057…).
Métrica	AUC	MSE

NO SUPERVISADO (I)

EDEM
Escuela de Empresarios

Características

- **Objetivo:** Para entrenar el modelo se utiliza un dataset o conjunto de muestras sin etiquetar. El objetivo aquí no es predecir un target sino encontrar patrones en los datos para extraer conocimiento útil.

- **Tipos:** 2 clases principales de problemas:

    1. Clustering.

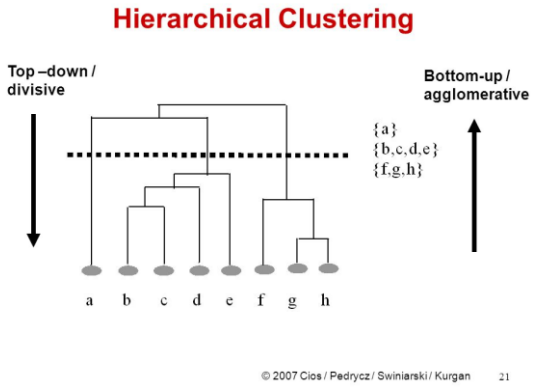
    2. Reducción de dimensionalidad.

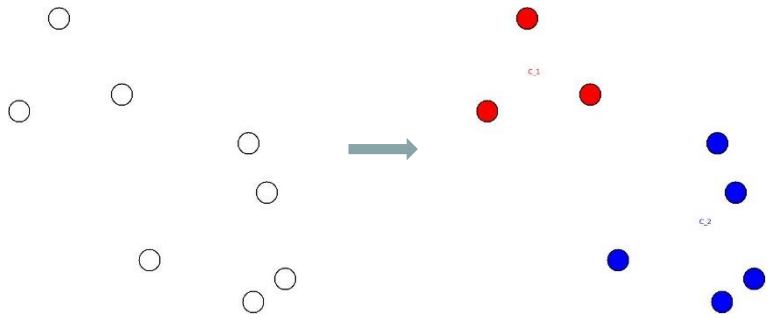
EDEM
Escuela de Empresarios

Segmentación o clustering

- **Objetivo:** Clustering es la tarea de agrupar un conjunto de objetos tales que los objetos en el mismo grupo (clúster) son más similares entre sí que a los de otros grupos.

- Métodos de agrupación:
  - **Métodos jerárquicos:** se descomponen en forma de árbol el conjunto de datos.
  - **Métodos de partición:** se crean divisiones sucesivas del conjunto de datos.





NO SUPERVISADO (III)

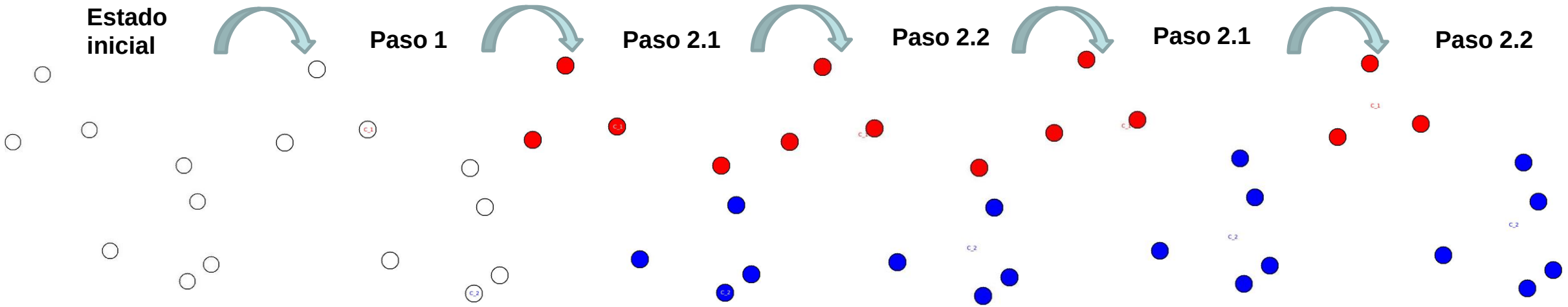
EDEM
Escuela de Empresarios

Segmentación o clustering. K- means

Ejemplo: K-means.

1. Seleccionar aleatoriamente k instancias como centros iniciales de las particiones. También puede utilizarse k puntos aleatorios del espacio de búsqueda.
2. Repetir
   1. Re/asignar instancias a la partición con el centro más próximo
   2. Recalcular el centro de cada partición (media) en función de los nuevas instancias asignadas.

Mientras haya cambios en las particiones



EDEM
Escuela de Empresarios

Segmentación o clustering. K- prototypes

- K-means solo acepta datos numéricos ya que usa la distancia euclídea. Sin embargo, hay algoritmos como k-prototypes que extienden k-means para admitir una **combinación de variables numéricas y variables categóricas**.

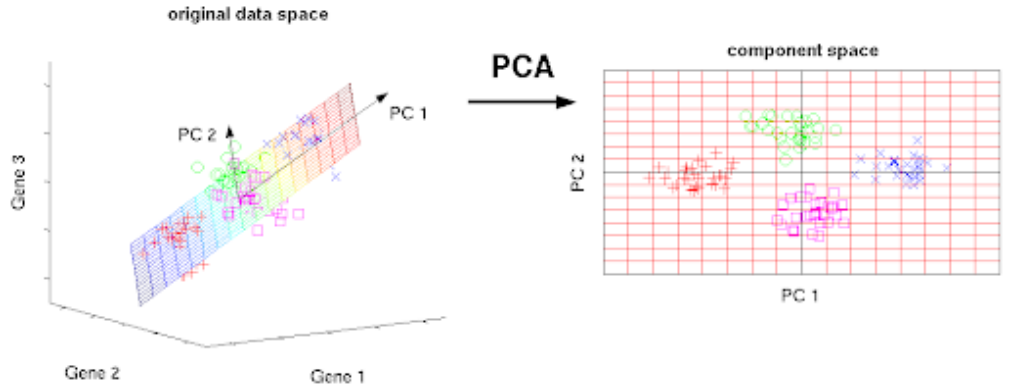
$$E = E_{num} + \lambda E_{cat}$$

- El parámetro λ sirve para **equilibrar la importancia** dada a las variables numéricas vs las categóricas.

EDEM
Escuela de Empresarios

Reducción de dimensionalidad

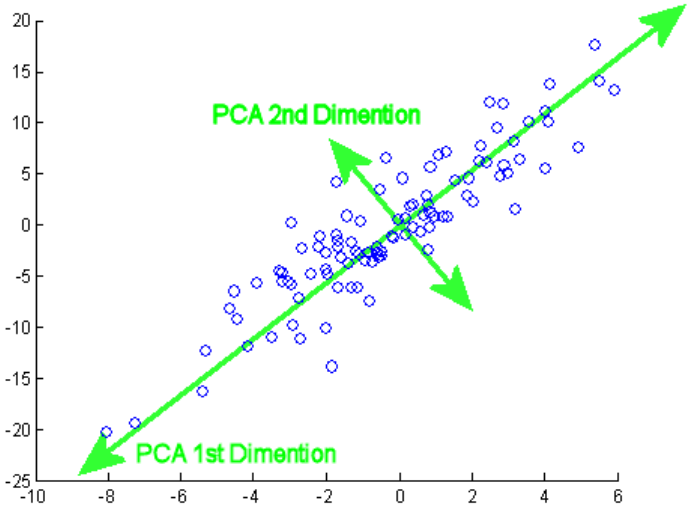
- **Objetivo:** Reducir el número de variables o columnas en un dataset. Con esto se busca:
  - Disminuir coste computacional.
  - Conseguir un nuevo dataset con menos variables irrelevantes o ruido.
  - Realizar visualizaciones.
  - …
- **Ejemplo:** PCA



EDEM
Escuela de Empresarios

Reducción de dimensionalidad. PCA

- Transforma las variables originales en un conjunto de nuevas variables, combinación de las anteriores, linealmente no correlacionadas.
- Estas variables reciben el nombre de **componentes principales**.
- Estas componentes son las que contienen la mayor información del dataset original.

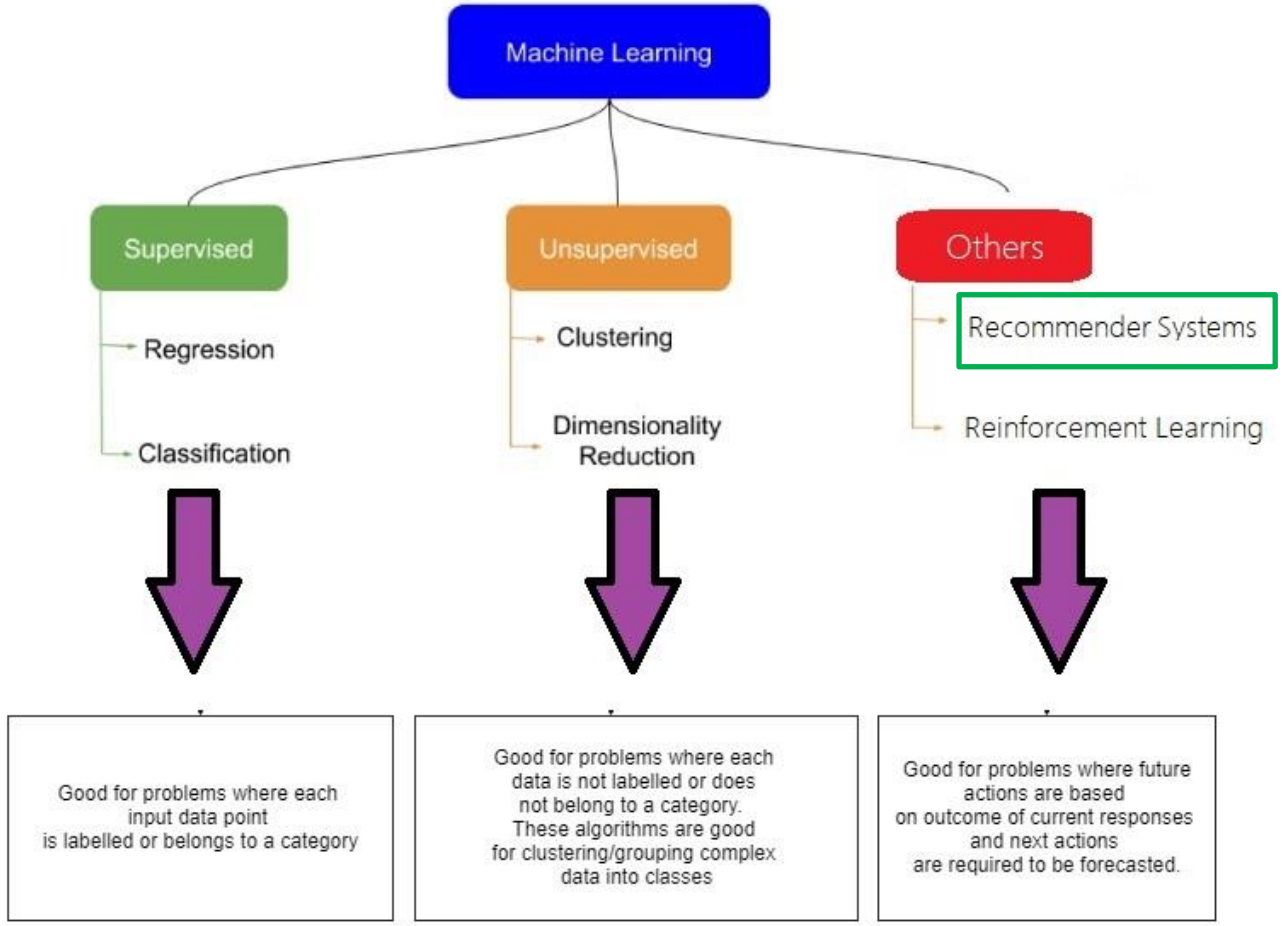


SISTEMAS DE RECOMENDACIÓN

EDEM
Escuela de Empresarios

RESUMEN

EDEM
Escuela de Empresarios



EDEM
Escuela de Empresarios

Definición

- Un **sistema de recomendación** es un algoritmo que nos permite **dar predicciones de cuál es el producto o ítem más adecuado para un usuario**.

- Los sistemas de recomendación pueden ser de varias clases según el algoritmo utilizado: basados en contenido, filtrado colaborativo, etc.

- Los sistemas de recomendación basados en algoritmos de filtrado colaborativo utilizan las valoraciones o interacciones de los usuarios sobre ciertos elementos del conjunto total para predecir interés en el resto de los elementos y recomendar los de mayor valoración predicha.

- Tipos de filtrado colaborativo:

    1. User-based Collaborative Filtering

    2. Item-based Collaborative Filtering

EDEM
Escuela de Empresarios

User Based

- **Personas con intereses similares en el pasado** es probable que tengan intereses parecidos en el futuro.
- Para predecir el interés de un usuario sobre un producto (ítem**) usamos la opinión o interacciones de usuarios similares**. Cuánto más similar sea el usuario, más tendremos en cuenta su opinión o historial de interacciones.
- Se calculan **las similitudes entre usuarios** en base a las opiniones o interacciones sobre los productos usando una determinada distancia.

$$sim(user1, user2) = \frac{\sum_j dist(puntuacion(user_1, item_j), puntuacion(user_2, item_j))}{Número\ de\ items}$$

- **Recomendaciones basadas en la actividad de usuarios similares a mí.**

- **Ejemplo:** Netflix.
  - Jesús ha visto True detective y Breaking Bad.
  - Miguel ha visto True detective, Breaking Bad y Fargo.
  - El algoritmo infiere que Jesús y Miguel son usuarios similares.
  - El sistema de recomendación **recomienda Fargo a Jesús.**

EDEM
Escuela de Empresarios

Item Based

- Si a un usuario le ha interesado en el pasado un determinado producto, es probable que en el futuro **le interesen productos similares.**
- Para predecir el interés de un usuario sobre un producto (ítem) usamos **su opinión o interacciones sobre productos similares**. Cuánto más similar sea el ítem, más tendremos el historial de puntuaciones o interacciones.
- En este caso se calculan **las similitudes entre ítems o productos** en función de las opiniones o interacciones de los usuarios.

$$sim(item_1, item_2) = \frac{\sum_j dist(puntuacion(user_j, item_1), puntuacion(user_j, item_2))}{Número\ de\ users}$$

- **Recomendaciones basadas en la actividad relacionada con productos similares a los que yo he comprado.**

- **Ejemplo:** Netflix.
  - True detective ha sido vista por Jesús, Sonia y Miguel.
  - Fargo ha sido vista por Sonia y Miguel.
  - El algoritmo infiere que True detective y Fargo son ítems similares.
  - El sistema de recomendación **recomienda Fargo a Jesús.**

EDEM
Escuela de Empresarios

Item VS User Based

Ventajas del Enfoque ítem-based

- En la mayoría de los casos disponemos de **más usuarios que ítems**, por lo que **la matriz de similitudes** final es **más escalable** al tener dimensiones más bajas. Si es una matriz de similitudes ítem-ítem, la dimensión de la matriz será NxN, donde N es el número de ítems, en vez de MxM, donde M es número de usuarios y M >> N.
- Disponemos de más interacciones por elemento en el enfoque ítem-ítem, por lo que los **resultados son más estables** y **estadísticamente significativos.**
- En muchos casos, **el inventario de productos se mantiene más estable** en el tiempo que el inventario o bolsa de usuarios.

EDEM
Escuela de Empresarios

Algoritmo Item Based

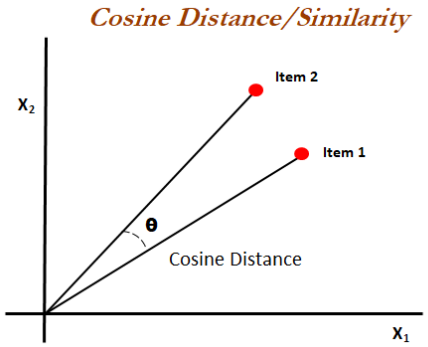
- Paso 1: Se calcula una matriz de puntuaciones o interacción usuario-ítem (LMxN).
- Paso 2: Se calcula una matriz de similitudes entre los ítems o rutas (SNxN) usando alguna distancia, por ejemplo, distancia del coseno.
- Paso 3: multiplicar matriz de interacciones por matriz de similitudes para obtener el ítem que tiene más probabilidad de interesarle a cada usuario.
- Paso 4: generar las recomendaciones (lista con n primeros elementos) mediante la ordenación de las puntuaciones para cada usuario con el resultado del paso 3.

EDEM
Escuela de Empresarios

Distancias

- Basamos el CF en similitudes ítem-item; hay que calcular una matriz de similitudes entre rutas.
- Para ello, hay multitud de medidas de similitud posibles, una de las más populares es la distancia del coseno:

$$sim_{\cos(item1,item2)} = \frac{item1 \cdot item2}{||item1|| ||item2||}$$



- También hay otras métricas de similitud como:
  - Correlación de Pearson

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

  - Distancia de Jaccard

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

EJEMPLO (I)

EDEM
Escuela de Empresarios

Supongamos que **tenemos 5 rutas** con los siguientes destinos (origen Madrid):

Ítem 1: Tenerife (TFN).

Ítem 2: Sevilla (SVQ).

Ítem 3: Santiago (SCQ).

Ítem 4: Valencia (VAL).

Ítem 5: Vigo (VGO).

7 clientes con sus interacciones:

Usuario 1:
Ha comprado a TFN y VAL.

Usuario 2:
Ha buscado en la web VAL.

Usuario 3:
Ha comprado a SCQ.
Ha buscado SCQ y VGO (2 veces).
Ha clicado en la campaña de SCQ.

Usuario 4:
Ha clicado en la campaña de TFN, SVQ y VAL.

Usuario 5:
Ha comprado a VAL, VGO, SCQ (2veces).
Ha buscado SCQ (3 veces).

Usuario 6:
Ha buscado VAL (3 veces) y SVQ.
Ha clicado en la campaña de VAL y TFN.

Usuario 7:
Ha comprado TFN.
Ha buscado TFN.
Ha clicado TFN (2 veces) y VAL.

EJEMPLO (II)

EDEM
Escuela de Empresarios

Interacciones clientes:

Usuario 1:
Ha comprado a TFN y VAL.

Usuario 2:
Ha buscado en la web VAL.

Usuario 3:
Ha comprado a SCQ.
Ha buscado SCQ y VGO (2 veces).
Ha clicado en la campaña de SCQ.

Usuario 4:
Ha clicado en la campaña de TFN, SVQ y VAL.

Usuario 5:
Ha comprado a VAL, VGO, SCQ (2veces).
Ha buscado SCQ (3 veces).

Usuario 6:
Ha buscado VAL (3 veces) y SVQ.
Ha clicado en la campaña de VAL y TFN.

Usuario 7:
Ha comprado TFN.
Ha buscado TFN.
Ha clicado TFN (2 veces) y VAL.

MATRIZ COMPRAS

	SCQ	SVQ	TFN	VAL	VGO
Usuario1	0	0	1	1	0
Usuario2	0	0	0	0	0
Usuario3	1	0	0	0	0
Usuario4	0	0	0	0	0
Usuario5	2	0	0	1	1
Usuario6	0	0	0	0	0
Usuario7	0	0	1	0	0

MATRIZ BÚSQUEDAS

	SCQ	SVQ	TFN	VAL	VGO
Usuario1	0	0	0	0	0
Usuario2	0	0	0	1	0
Usuario3	1	0	0	0	2
Usuario4	0	0	0	0	0
Usuario5	3	0	0	0	0
Usuario6	0	1	0	3	0
Usuario7	0	0	1	0	0

MATRIZ CLICS

	SCQ	SVQ	TFN	VAL	VGO
Usuario1	0	0	0	0	0
Usuario2	0	0	0	0	0
Usuario3	1	0	0	0	0
Usuario4	0	1	1	1	0
Usuario5	0	0	0	0	0
Usuario6	0	0	1	1	0
Usuario7	0	0	2	1	0

EJEMPLO (III)

EDEM
Escuela de Empresarios

MATRIZ COMPRAS

	SCQ	SVQ	TFN	VAL	VGO
Usuario1	0	0	1	1	0
Usuario2	0	0	0	0	0
Usuario3	1	0	0	0	0
Usuario4	0	0	0	0	0
Usuario5	2	0	0	1	1
Usuario6	0	0	0	0	0
Usuario7	0	0	1	0	0

MATRIZ BÚSQUEDAS

	SCQ	SVQ	TFN	VAL	VGO
Usuario1	0	0	0	0	0
Usuario2	0	0	0	1	0
Usuario3	1	0	0	0	2
Usuario4	0	0	0	0	0
Usuario5	3	0	0	0	0
Usuario6	0	1	0	3	0
Usuario7	0	0	1	0	0

MATRIZ CLICS

	SCQ	SVQ	TFN	VAL	VGO
Usuario1	0	0	0	0	0
Usuario2	0	0	0	0	0
Usuario3	1	0	0	0	0
Usuario4	0	1	1	1	0
Usuario5	0	0	0	0	0
Usuario6	0	0	1	1	0
Usuario7	0	0	2	1	0

*

MATRIZ INTERACCIÓN FINAL (I_{7x5})

	SCQ	SVQ	TFN	VAL	VGO
Usuario1	0	0	1	1	0
Usuario2	0	0	0	1	0
Usuario3	3	0	0	0	2
Usuario4	0	1	1	1	0
Usuario5	5	0	0	1	1
Usuario6	0	1	1	4	0
Usuario7	0	0	4	1	0

*Nota: Ejemplo ilustrativo simplificado, en la práctica es una combinación ponderada.

EJEMPLO (IV)

EDEM
Escuela de Empresarios

Una vez tenemos la matriz de interacción user-ítem

MATRIZ INTERACCIÓN FINAL (L_{7x5})

	SCQ	SVQ	TFN	VAL	VGO
Usuario1	0	0	1	1	0
Usuario2	0	0	0	1	0
Usuario3	3	0	0	0	2
Usuario4	0	1	1	1	0
Usuario5	5	0	0	1	1
Usuario6	0	1	1	4	0
Usuario7	0	0	4	1	0

Calculamos ahora la matriz de similitudes. Seguiremos aquí el enfoque ítem-ítem. En este caso vamos a utilizar la distancia del coseno. Si queremos obtener, por ejemplo, la similitud del ítem 1 (SCQ) con el ítem 5 (VGO) el cálculo sería.

$$sim(item_1, item_5) = \frac{(0,0,3,0,5,0,0) \cdot (0,0,2,0,1,0,0)}{||(0,0,3,0,5,0,0)||\,||(0,0,2,0,1,0,0)||} = \frac{6+5}{\sqrt{34}\sqrt{5}} = 0.84$$

S_{5x5}	SCQ	SVQ	TFN	VAL	VGO
SCQ	1	0	0	0.19	0.84
SVQ	0	1	0.32	0.77	0
TFN	0	0.32	1	0.50	0
VAL	0.19	0.77	0.50	1	0.10
VGO	0.84	0	0	0.10	1

EJEMPLO (V)

EDEM
Escuela de Empresarios

Como ejemplo, vamos a calcular la ruta que le puede resultar más interesante al cliente 4:

- Ha clicado TFN, SVQ y VAL.

Multiplicamos el vector de interés del cliente 4 extraído de la matriz de interacciones user-ítem es decir, $I_{(4,)} = (0,1,1,1,0)$, por cada una de las columnas de la matriz de similitudes S calculada.

SCQ: (0, 1, 1, 1, 0) * (1, 0, 0, 0.19, 0.84) = 0.19
SVQ: (0, 1, 1, 1, 0) * (0, 1, 0.32, 0.77, 0) = 2.09
TFN: (0, 1, 1, 1, 0) * (0, 0.32, 1, 0.50, 0) = 1.82
VAL:(0, 1, 1, 1, 0)*(0.19, 0.77, 0.5, 1, 0.1) =2.27
VGO: (0, 1, 1, 1, 0) * (0.84, 0, 0, 0.10, 1) = 0.10

S_{5x5}	SCQ	SVQ	TFN	VAL	VGO
SCQ	1	0	0	0.19	0.84
SVQ	0	1	0.32	0.77	0
TFN	0	0.32	1	0.50	0
VAL	0.19	0.77	0.50	1	0.10
VGO	0.84	0	0	0.10	1

Por tanto, en este caso le recomendaríamos Valencia al cliente 4 por haber obtenido el score de interés mayor, siendo SVQ la segunda ruta a recomendar.

Siguiendo un proceso análogo para el resto de usuarios, lo cual es equivalente a multiplicar I*S (matriz de interacciones user-ítem por matriz de similitudes ítem-ítem) obtendríamos la siguiente matriz de scores de interés:

	SCQ	SVQ	TFN	VAL	VGO
Usuario1	0,19	1,09	**1,5**	**1,5**	0,1
Usuario2	0,19	0,77	0,5	**1**	0,1
Usuario3	**4,68**	0	0	0,77	4,52
Usuario4	0,19	2,09	1,82	**2,27**	0,1
Usuario5	**6,03**	0,77	0,5	2,05	5,3
Usuario6	0,76	4,4	3,32	**5,27**	0,4
Usuario7	0,19	2,05	**4,5**	3	0,1

EDEM
Escuela de Empresarios

EDEM Escuela de Empresarios

Yvonne Gala García:
yvonne.gala@iberiaexpress.com

Jesús Prada Alonso:
TAD.jprada@gmail.com