

Formación y experiencia



Nicolás Balparda

INSTRUCTOR

- *Estrategia tecnológica*
- *Arquitectura de datos y sistemas*
- *Big Data*
- *Análisis de datos*

FORMACIÓN

- Ingeniero en sistemas
- MsC en Gestión de Datos e Innovación tecnológica
- MsC en Big Data y Business Analytics (en curso)

Presentación

- a. Nombre
- b. ¿Qué rol ocupa en su trabajo?
- c. ¿Participaste en los cursos de las semanas pasadas?

Temario del curso

1

Introducción

Terminología
Beneficios
Casos de uso

2

Preparación de datos

Construcción del dataset
Transformación de datos

3

Tipos de aprendizaje

Supervisado
No supervisado
Por refuerzo

4

Cómo entrenar un modelo

5

Cómo validar un modelo

Medidas más utilizadas
Overfitting vs Underfitting

6

Cómo desplegar un modelo

7

Demostración y ejercicio práctico en JupyterHub

Introducción

A decorative graphic consisting of multiple thin, overlapping, wavy lines in a light gray color, creating a sense of motion and depth. The lines flow from the left side of the slide, under the title, and continue across the bottom of the slide.

Dinámica

Consigna

- ¿Qué es el machine learning?
- ¿Han tenido algún ejemplo cercano dónde hayan visto la aplicación de ML?

15 minutos

Dinámica

¿Qué es Machine Learning?

"El campo de estudio que brinda a las computadoras la capacidad de aprender sin ser programadas explícitamente" Arthur Samuel.

"Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y medida de desempeño P , si su desempeño en las tareas T , medido por P , mejora con la experiencia E " Tom Mitchell.



A decorative graphic consisting of multiple thin, overlapping, wavy lines in a light gray color, creating a sense of motion and depth. The lines flow from the left side of the slide towards the right, curving and undulating across the middle section.

Introducción

Terminología

Terminología

○ Etiquetas

Valores que queremos predecir.

○ Atributos

Las variables de entrada. Pueden ser 1 como miles.

○ Algoritmo

Conjunto de instrucciones ordenadas que ofrecen una solución a un problema.

○ Modelo

Diseño del algoritmo de Machine Learning, se visualiza como una función que se aplica a los datos nuevos de entrada, donde la salida es el resultado final.

A decorative graphic consisting of multiple thin, overlapping wavy lines in a light gray color, creating a sense of motion and depth. The lines flow from the left side of the slide towards the right, with some lines curving upwards and others downwards, creating a complex, organic pattern.

Introducción

Beneficios

Beneficios

Tendencias y patrones se identifican con facilidad

Toma de decisiones en tiempo real

Amplias aplicaciones

Predicciones valiosas para tomar mejores decisiones

Automatización

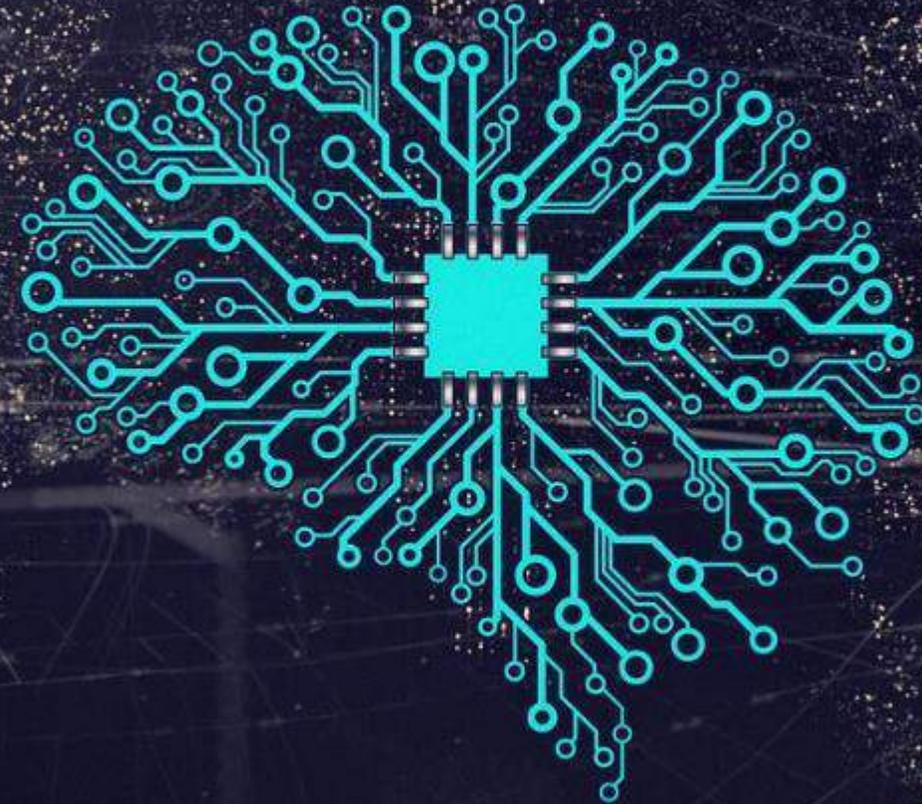
Permite adaptarse sin intervención humana

Analizar grandes volúmenes de datos

Beneficios

Extraer la información correcta y tomar una decisión a partir de los datos es difícil.

El aprendizaje automático aprovecha los algoritmos de ML y aprende de los datos existentes. Los resultados ayudan a las empresas a tomar la decisión correcta. Permite a las organizaciones transformar datos en conocimiento e inteligencia procesable. La información puede integrarse en los procesos cotidianos de la institución.



Introducción

Casos de uso

Algunos de los principales casos de uso

DetECCIÓN FACIAL

Las aplicaciones del reconocimiento facial forman parte de las tecnologías de ML más populares que ya están funcionando a nuestro alrededor. Desde etiquetar a personas en redes sociales a desbloqueo de móviles, y por supuesto medidas de seguridad en aeropuertos o en lugares públicos.



GMAIL

Para mantener al usuario protegido de virus y de recibir correos sospechosos o fraudulentos, la plataforma de email de Google integra el Machine Learning para evitar el correo no deseado (o spam) en la bandeja de entrada. El sistema «entiende y aprende» de los ejemplos pasados, para tomar decisiones en el futuro basándose en ellos. No solo esto, con la función Smart Reply, puede responder correos de forma similar a como tú lo harías.



Algunos de los principales casos de uso

Vehículos autónomos

Se está utilizando el ML en los coches para que estos se manejen solos, reconociendo la ruta, tomando en cuenta los coches y el entorno que le rodea, cumpliendo las leyes de tránsito, respetando la velocidad y manteniéndose por la vía de circulación. Con el ML aprenden de sus errores y también del comportamiento de los otros conductores. Empresas como Tesla usan tecnología Deep Learning para el reconocimiento de los patrones de conducción y para predecir accidentes segundos antes de que se produzcan.



Diagnósticos médicos

Gracias a la recolección de datos y el Machine Learning se pueden detectar con mucha anticipación enfermedades como el Cáncer de mama al conocer los síntomas de los pacientes.



Algunos de los principales casos de uso

Chatbots

Los chatbots utilizan ML cuando necesitan reconocer las intenciones de las frases que intercambian con los usuarios en un diálogo. Pero también ponen en funcionamiento su ML cuando, después de reconocer determinadas intenciones por parte de los usuarios, responden con entidades. Todo esto lo hacen sobre el diálogo, que es la estructura propia de la conversación.



Detección de fraudes

Empresas como PayPal y monedas electrónicas como BITCOIN ya utilizan el ML para combatir el blanqueo de dinero y los fraudes. Las empresas utilizan la tecnología para reconocer millones de transacciones y detectar cuáles son las fraudulentas, quién es el vendedor y el comprador.



Preparación de datos

A decorative graphic consisting of multiple thin, overlapping, wavy lines in a light gray color, creating a sense of motion and depth across the lower half of the slide.

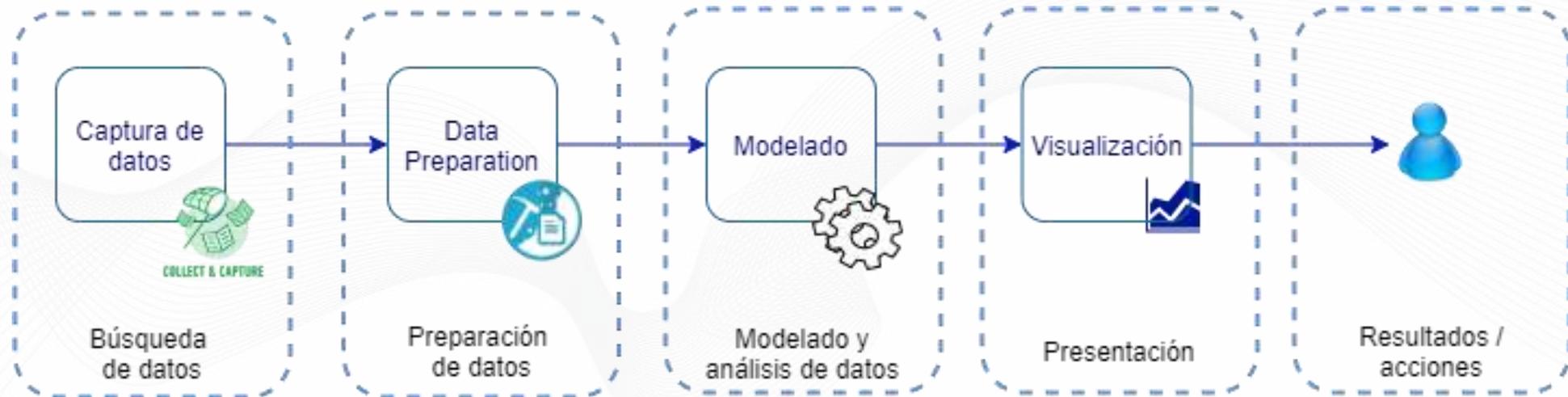


¡Datos y más datos!

¿Demasiados datos o
demasiada información?
Importancia de la
preparación de los datos

Proceso de análisis de datos

El siguiente proceso refleja los pasos recomendados para llevar a cabo un proceso de análisis de datos sobre una arquitectura de analítica avanzada.



En esta sección haremos foco en la etapa de Preparación de Datos

Dinámica

Consigna

- ¿Por qué entienden que es importante la preparación de datos?

10 minutos

Dinámica

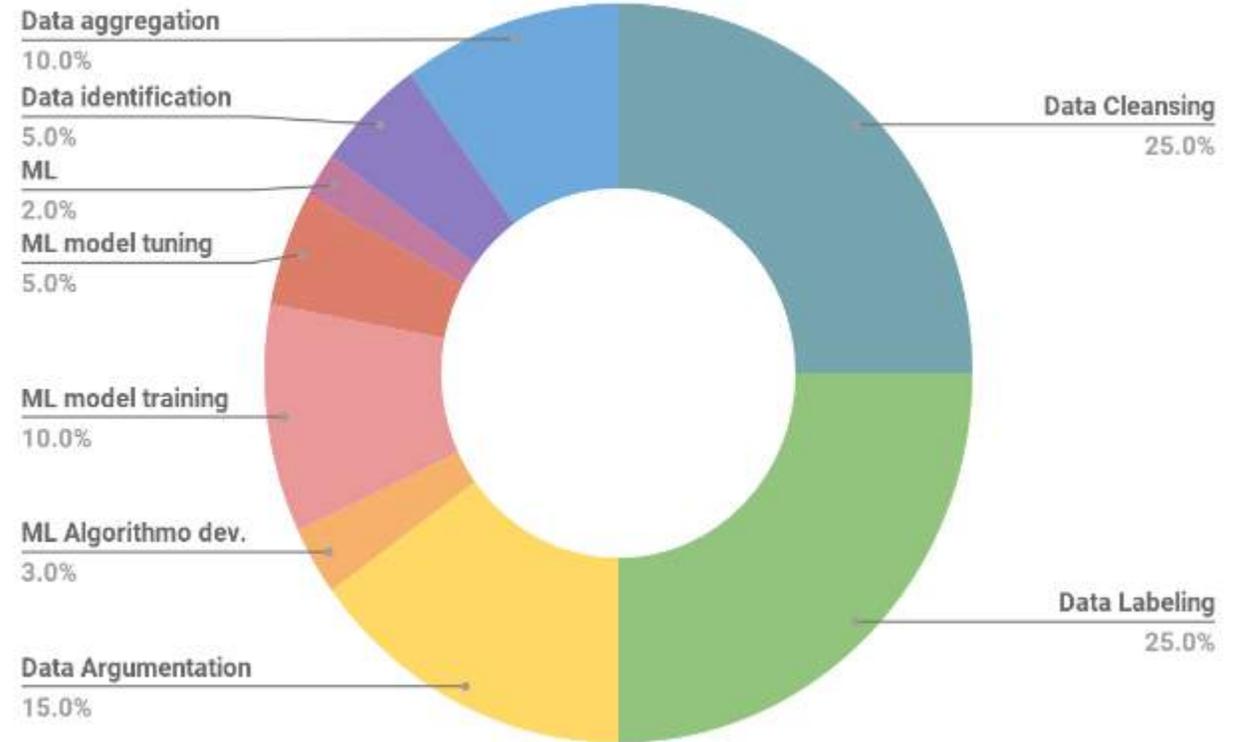
¿Por qué es importante la preparación de datos?

Los datos son el alma de los proyectos de ML.

El proceso de preparación de datos es uno de los principales desafíos que afectan a la mayoría de los proyectos.

Las tareas de preparación de datos requieren más del 80% del tiempo dedicado.

La calidad y el tamaño del conjunto de datos son mucho más importantes que el algoritmo que utilices.



Proceso de preparación de datos

Identificamos dos grandes etapas en el proceso de preparación de datos sobre una arquitectura de analítica avanzada.

Construcción del dataset

- Recolección de los datos en crudo.
- Identificación de atributos y etiquetar fuentes.
- Seleccionar una estrategia de muestreo.
- Dividir los datos.



Transformación de datos

- Detección y tratamiento de outliers.
- Escalación de características.
- Datos faltantes.
- Feature Engineering.

The background features a series of overlapping, wavy lines in a light gray color, creating a sense of motion and depth. These lines are composed of many thin, parallel strokes that form a fluid, organic shape across the slide.

Preparación de datos

Construcción del dataset

Construcción del dataset

- Los modelos simples con datasets grandes generalmente son mejores que los modelos sofisticados con datasets pequeños.
- La calidad también importa: no sirve de nada tener muchos datos si su calidad es inadecuada
- Siempre considere qué datos están disponibles para su modelo en el momento de la predicción.
- A veces es necesario unir varias fuentes de datos.
- Si hay demasiados datos se debe seleccionar un subconjunto de ejemplos para el entrenamiento.
- Dividir los datos en conjuntos de entrenamiento, validación y de prueba.



A decorative background element consisting of multiple overlapping, wavy lines in a light gray color, creating a sense of motion and depth. The lines are most prominent in the lower half of the slide, framing the text.

Preparación de datos

Transformación de datos

Detección y tratamiento de outliers

Los outliers son valores atípicos:

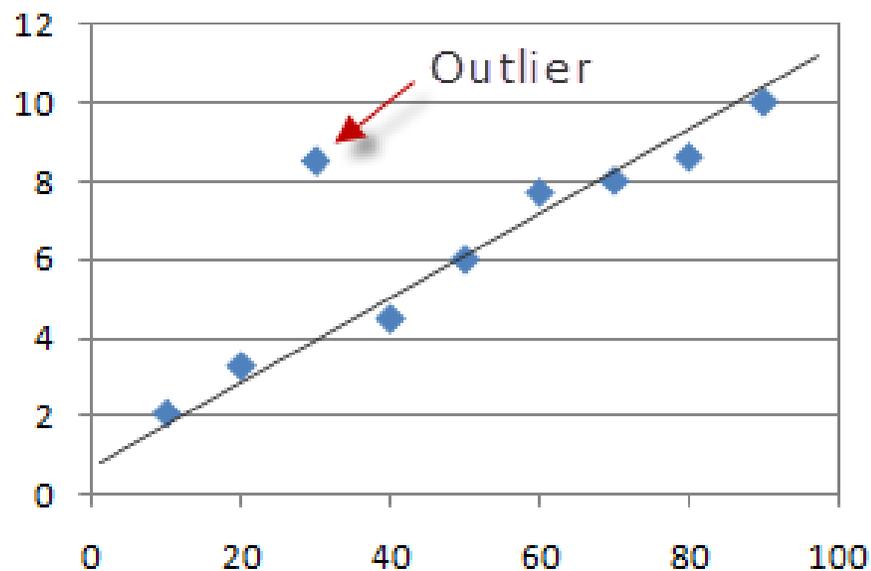
- Valores extremos altos y bajos genuinos en el conjunto de datos.
- Introducidos por error humano o mecánico.
- Introducidos para reemplazar valores faltantes.

Detección:

- Análisis de valores extremos.
- Metodo Z-score.
- K Means clustering.
- Visualizando los datos.

Tratamiento:

- Imputación media / mediana o aleatoria.
- Trimming.
- Top, Bottom and Zero Coding.
- Discretización de variables.



Escalado de Características

La mayoría de los algoritmos de ML tienen mejor rendimiento cuando tratan con características que están en la misma escala.

Normalización:

Se refiere a reescalar las características en un rango $[0,1]$, que es un caso especial de la escalación "min-max".

Estandarización:

Consiste en centrar las columnas de características con respecto a media 0 con desviación estándar 1 de forma que las columnas de características tengan los mismos parámetros que una distribución normal estándar.

Datos faltantes

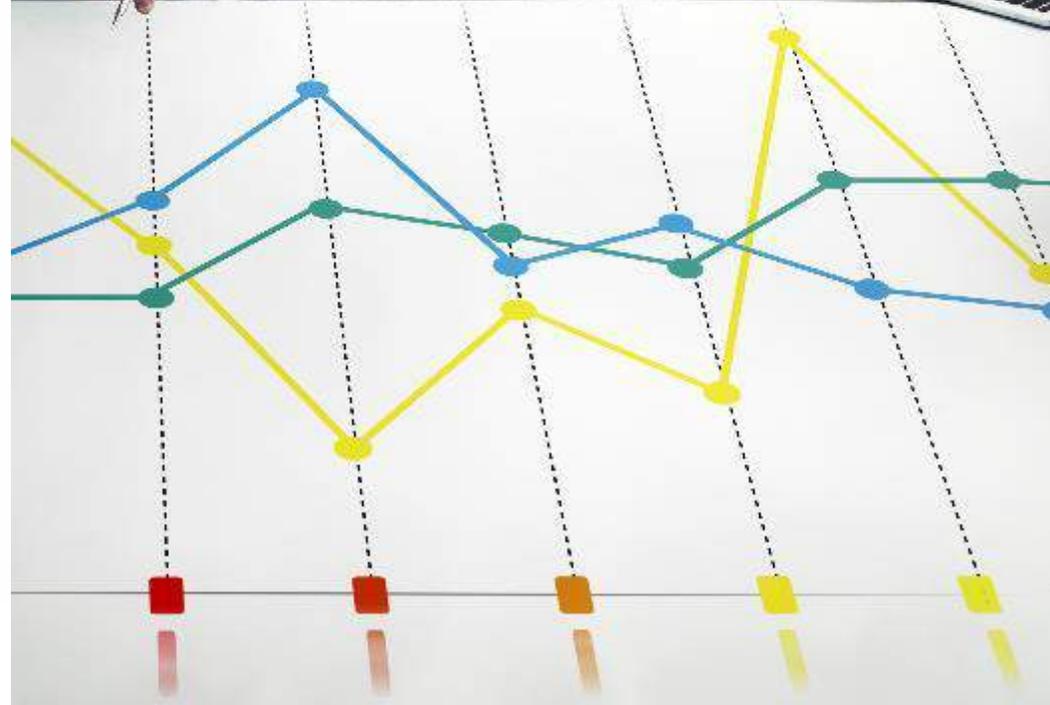
La mayoría de algoritmos no pueden manejar esos valores faltantes.

Eliminación

Eliminar las muestras o características con campos vacíos (corremos el riesgo de borrar información relevante o demasiadas muestras).

Inferencia

Inferir su valor, ésta generalmente es la mejor alternativa a seguir.



Feature Engineering

Es el proceso de determinar qué características podrían ser útiles para entrenar un modelo y crear esas características mediante la transformación de datos.

Transformar datos numéricos

- Normalizar
- Bucketing
- Logaritmo - $\log(x)$
- Reciprocal - $1 / x$
- Raíz cuadrada - \sqrt{x}
- Exponencial - $\exp(x)$
- Yeo-Johnson
- Box-Cox

Transformar datos categóricos

- Vocabulario.
- Hashing.
- Híbrido de Hashing y Vocabulario

Automated Feature Engineering

	Date_Time_Combined	Status
0	2018-02-14 20:40	Delayed
1	2018-02-15 10:30	On Time
2	2018-02-14 07:40	On Time
3	2018-02-15 18:10	Delayed
4	2018-02-14 10:20	On Time

	Hour_Of_Day	Status
0	20	Delayed
1	10	On Time
2	7	On Time
3	18	Delayed
4	10	On Time

Tipos de aprendizaje

A decorative graphic consisting of multiple thin, overlapping, wavy lines in a light gray color, creating a sense of motion and depth across the lower half of the slide.



¿Qué pasaría si pudieras...

...hacer que las
máquinas
aprendan de
sus errores?

Tipos de aprendizaje

Supervisado

A decorative background element consisting of multiple overlapping, wavy lines in a light gray color, creating a sense of motion and depth. The lines are most prominent in the lower half of the slide, framing the text.

Aprendizaje supervisado

Se refiere a entrena un modelo utilizando muestras de datos etiquetados.

El algoritmo genera una función que relaciona las variables de entrada con la salida deseada.

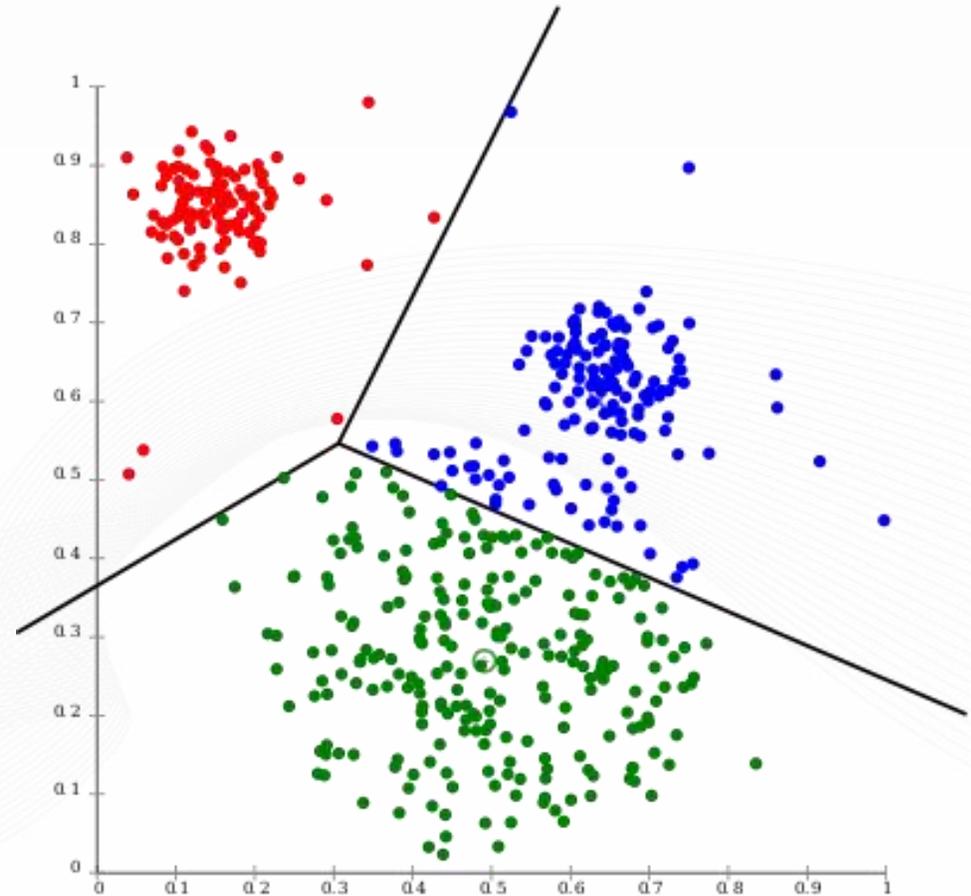
Dos tipos de problemas:

- Clasificación.
- Regresión.



Clasificación

Un modelo de clasificación predice valores discretos. Los datos se separan en un cierto número de clases.



Algoritmos de Clasificación

Logistic Regression

A pesar de su nombre no es un algoritmo para problemas de regresión sino que es un método para problemas de clasificación, en los que se obtienen un valor binario entre 0 y 1.

Decision Trees

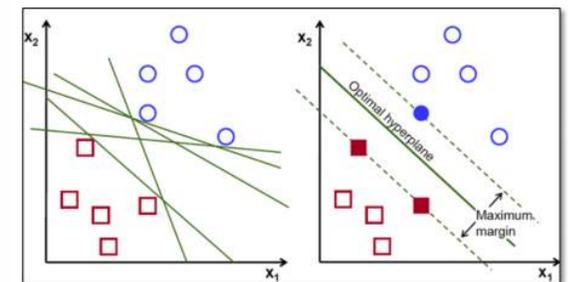
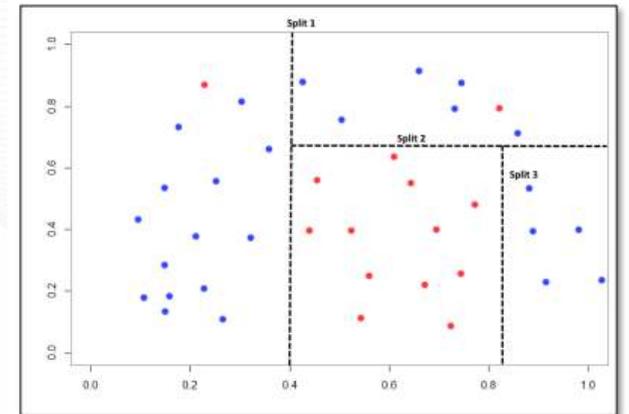
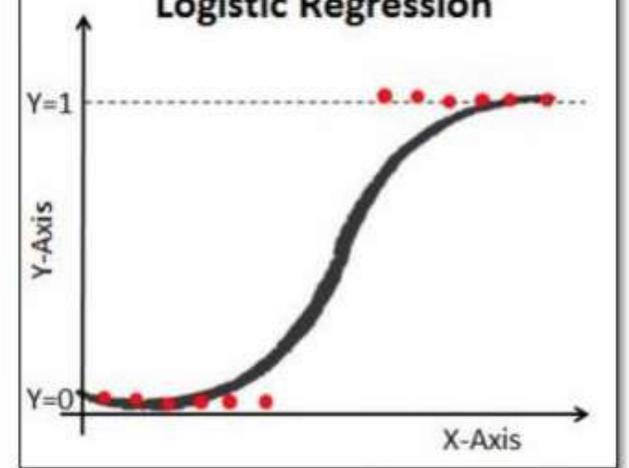
Cada nodo interno denota una prueba en un atributo, cada rama representa el resultado de una prueba, y cada hoja nodo tiene una etiqueta. El nodo superior en un árbol es el nodo raíz (algoritmo también aplicado en regresión).

Random Forest

Es un método que combina una cantidad grande de árboles de decisión independientes probados sobre conjuntos de datos aleatorios con igual distribución (algoritmo también aplicado en regresión).

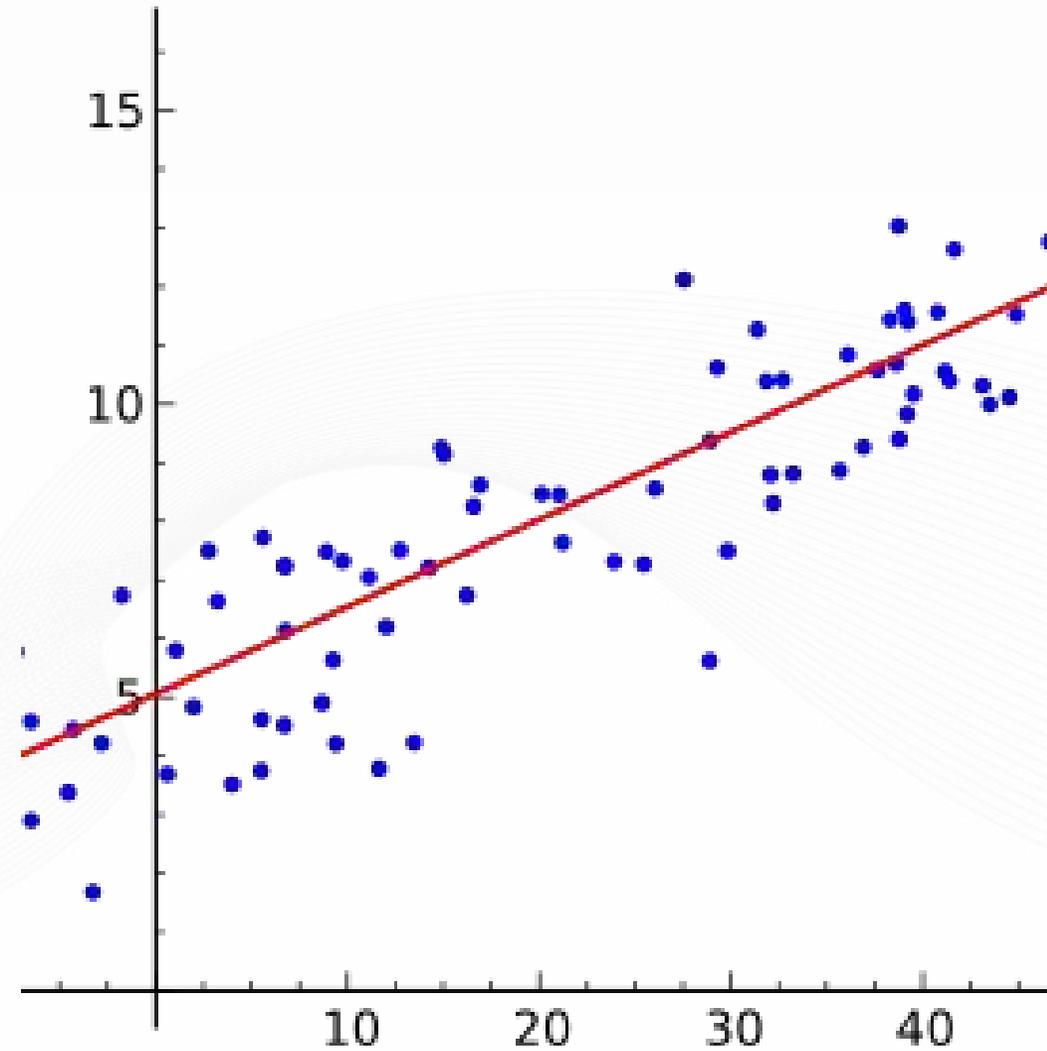
Support vector Machines (SVMs)

El objetivo es encontrar un hiperplano en un espacio N-dimensional que clasifica claramente los puntos de datos (algoritmo también aplicado en regresión).



Regresión

Un modelo de regresión predice valores continuos.



Algoritmos de Regresión

Linear Regression

Da como resultado un valor continuo a partir de una combinación lineal de atributos de entrada.

Regression Trees

Árboles de decisión, donde la variable de destino puede tomar valores continuos.

Non-Linear Regression

Es una combinación no lineal de los parámetros del modelo y depende de una o más variables independientes.

Bayesian Linear Regression

Es un enfoque de regresión lineal en la que se lleva a cabo el análisis estadístico en el contexto de la inferencia bayesiana.

Polynomial Regression

Es un caso especial de regresión lineal donde ajustamos una ecuación polinómica en los datos con una relación curvilínea entre la variable objetivo y las variables independientes.

The background of the slide features a series of overlapping, wavy lines in a light gray color, creating a sense of motion and depth. These lines are most prominent on the left and right sides, framing the central text.

Tipos de aprendizaje

No supervisado

Aprendizaje no supervisado

Se refiere a entrenar un modelos utilizando muestras de datos sin estar etiquetados.

El algoritmo aprende de los datos de entrada, sin disponer la información sobre las salidas.

El algoritmo es capaz de reconocer patrones que refieren a las salidas.

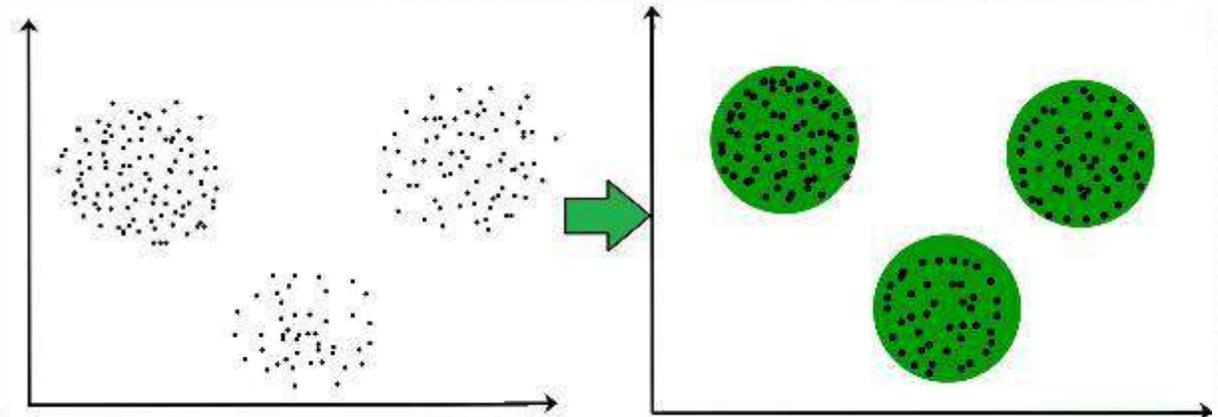
Tipos de problemas:

- Clustering.
- Asociación de reglas.
- Detección de anomalías.
- Reducción de dimensiones.



Clustering

Se utilizan para agrupar datos existentes de los que desconocemos sus características en común o queremos descubrirlas. Estos métodos intentan crear puntos centrales y jerarquías para diferenciar grupos y descubrir características comunes por cercanía.



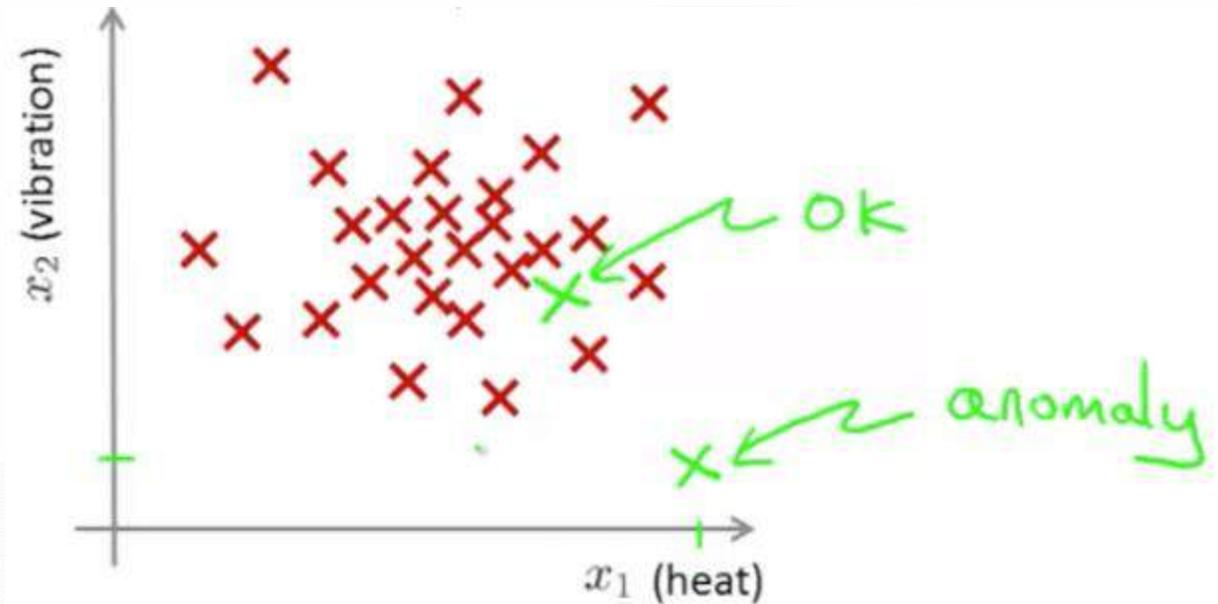
Asociación de reglas

Se utilizan para encontrar asociaciones y relaciones entre grandes conjuntos de elementos de datos. Estas reglas muestran con qué frecuencia se produce un conjunto de elementos en una transacción.



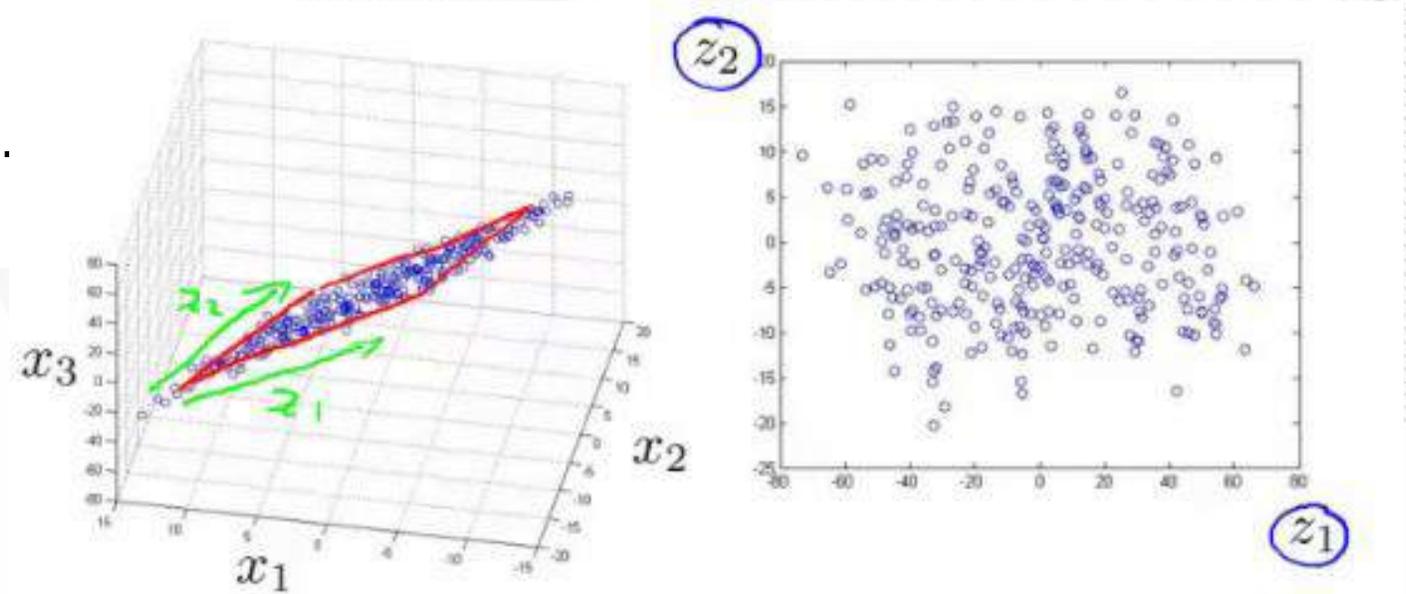
Detección de anomalías

La detección de anomalías se refiere a la identificación de elementos o eventos que no se ajustan a un patrón esperado u otros elementos en un conjunto de datos que generalmente no son detectables por un humano.



Reducción de dimensiones

Son técnicas para poder reducir la dimensión de los datos. Esto se puede utilizar para evitar sobreajustes o para poder interpretar los datos, ya sea mediante factores o gráficamente.



The background features a series of overlapping, wavy lines in a light gray color, creating a sense of motion and depth. These lines are composed of many thin, parallel lines that curve and flow across the slide.

Tipos de aprendizaje

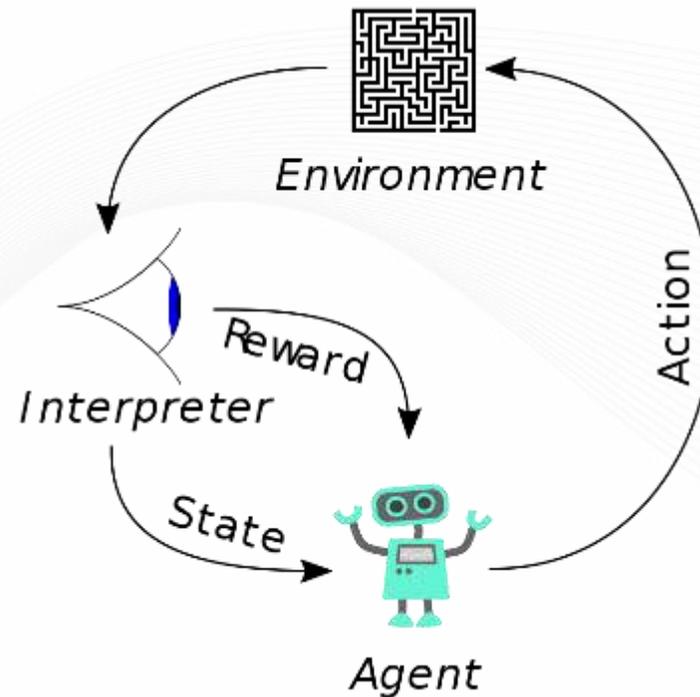
Por refuerzo

Definición

El aprendizaje por refuerzo o Reinforcement Learning, es otro tipo de aprendizaje de ML donde el objetivo es el desarrollo de un sistema (que recibe el nombre de **agente**) que se desea que mejore su eficiencia realizando cierta tarea basándose en la **interacción con su entorno**.

Para ello, el agente recibe **recompensas** (rewards en inglés) que le permiten adaptar su comportamiento.

La principal diferencia entre las técnicas clásicas y algoritmos de aprendizaje por refuerzo es que este último no es necesario el conocimiento de los procesos de decisión (MDP), y justamente se utilizan en procesos de decisión donde los **métodos exactos se convierten en no viables**.



Algoritmos más utilizados

Criterio de optimalidad

Bajo ciertas condiciones de regularidad adicional está entonces bien definida la expectativa de la recompensa total para cualquier política y cualquier distribución inicial sobre los estados. En este sentido, una política se refiere a una asignación de cierta distribución de probabilidad sobre las acciones a todas las historias posibles.

Fuerza bruta

El enfoque por fuerza bruta implica las dos etapas siguientes:

1. Para cada política posible, muestrear resultados.
2. Elija la política con el mayor retorno esperado.

El principal problema es que el número de políticas puede ser extremadamente grande, o incluso infinito.

Acercamiento al valor de la función

Las funciones de valor intentan encontrar una política que maximice el retorno al mantener un conjunto de estimaciones de los rendimientos esperados por alguna política (por lo general, ya sea la "corriente" o la óptima):

1. Método Montecarlo.
2. Métodos de diferencias temporales.

Búsqueda política directa

Se basa en la búsqueda en algún subconjunto del espacio de la política, en cuyo caso el problema se convierte en una instancia de optimización estocástica. Los dos enfoques disponibles se basan en el gradiente y métodos de gradiente libre.

Cómo entrenar un modelo





¿Cómo entrenar un modelo de ML?

El proceso de entrenamiento de un modelo de ML consiste en proporcionar datos de entrenamiento de los cuales aprende el algoritmo de ML.

El término modelo de ML se refiere al artefacto de modelo que se crea en el proceso de entrenamiento.

Separaremos el conjunto de datos inicial en 2: conjunto de entrenamiento (train) y conjunto de evaluación(test).

Solo se utiliza el conjunto de entrenamiento para entrenar el modelo.

Una vez que se entrenó el modelo debemos evaluar qué tan bueno es.

¿Cómo validar un modelo?

The background of the slide features a series of overlapping, wavy, light gray lines that create a sense of motion and depth, resembling a stylized wave or a series of overlapping curves.



¿Cómo validar un modelo?

Medidas más utilizadas

Medidas más utilizadas en modelos de Regresión

Error medio absoluto

Esta métrica de regresión es el valor medio de la diferencia absoluta entre el valor real y el valor predicho.

Error cuadrático medio

Calcula el valor medio de la diferencia al cuadrado entre el valor real y el predicho para todos los puntos de datos.

Raíz del error cuadrático medio

Es la raíz cuadrada del error cuadrático medio. Es fácil de interpretar en comparación con el error cuadrático medio y utiliza valores absolutos más pequeños, lo que es útil para los cálculos informáticos.

Medidas más utilizadas en Modelos de Clasificación

Matriz de confusión

Esta matriz se utiliza para evaluar la precisión de un clasificador (TP, FN, FP, TN).

Métrica de exactitud

Indica el número de elementos clasificados correctamente en comparación con el número total de artículos.

Métrica de exhaustividad

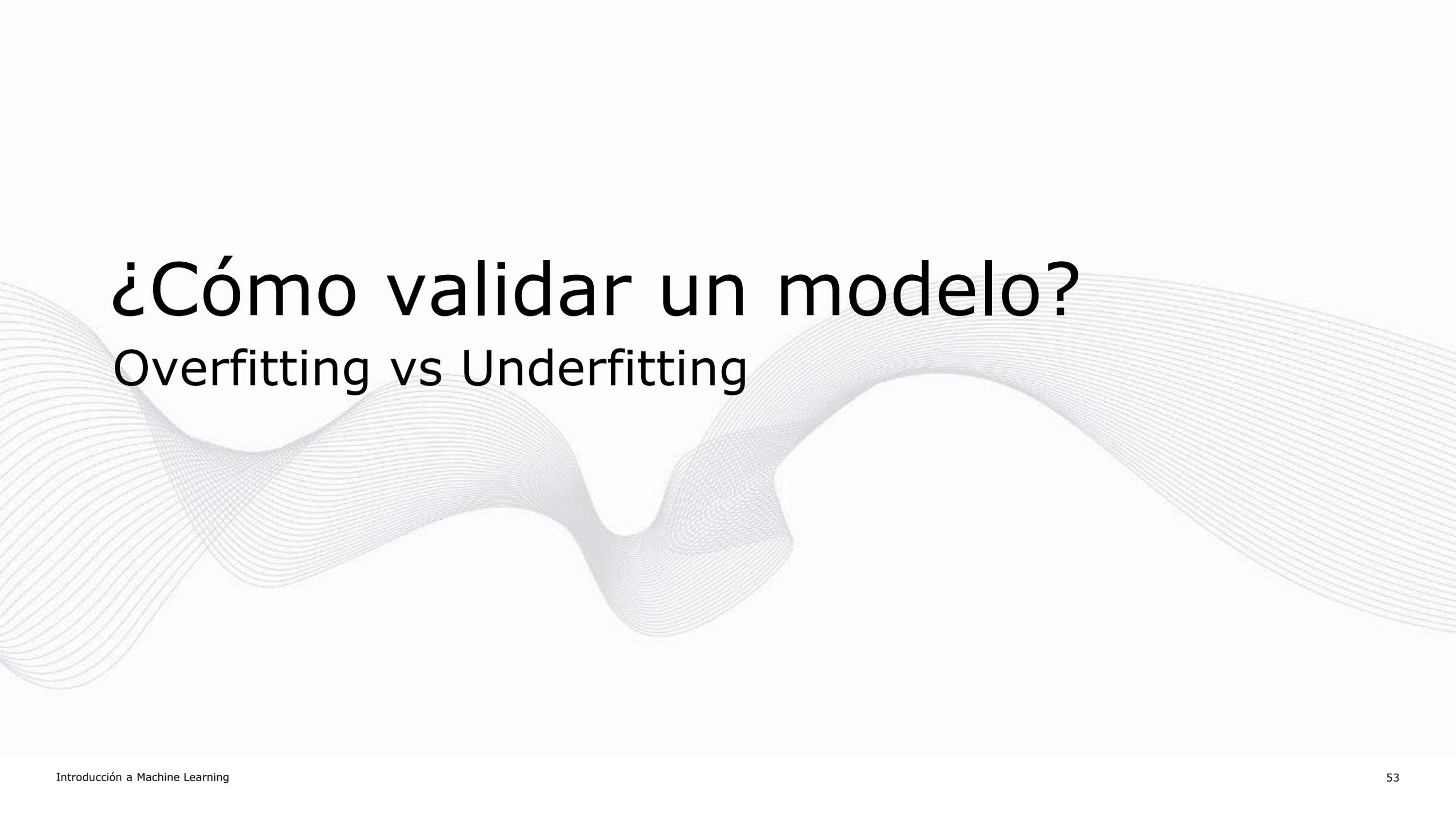
Muestra la cantidad de verdaderos positivos que el modelo ha clasificado en función del número total de valores positivos.

Métrica de precisión

Esta métrica representa el número de verdaderos positivos que son realmente positivos en comparación con el número total de valores positivos predichos.

Puntuación F1

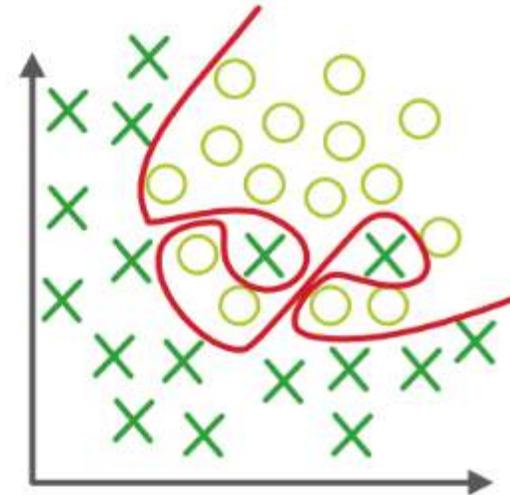
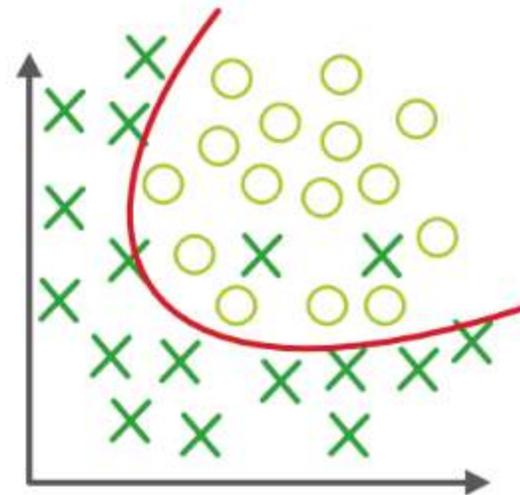
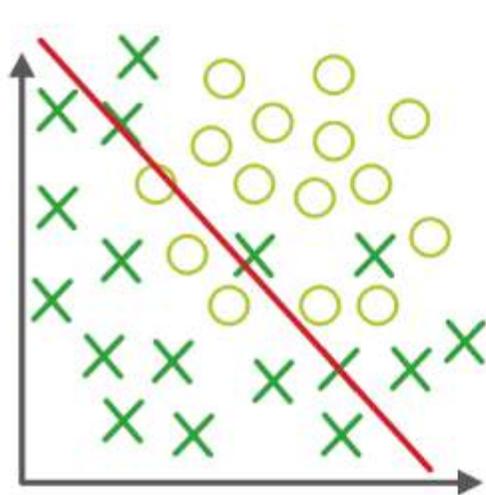
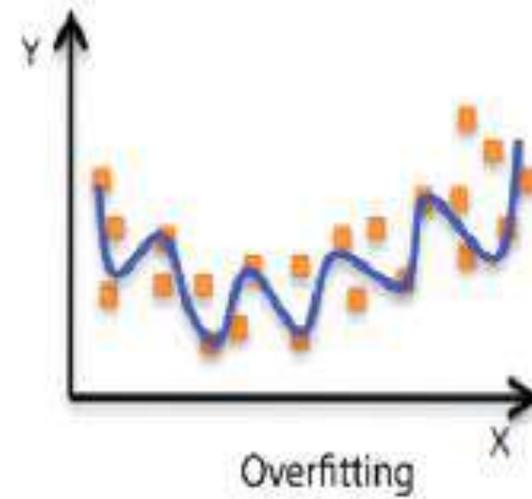
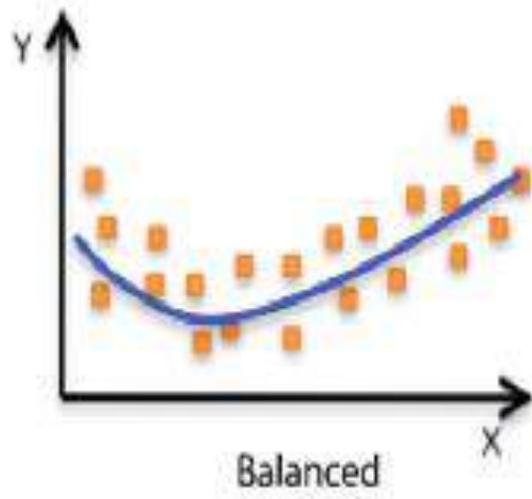
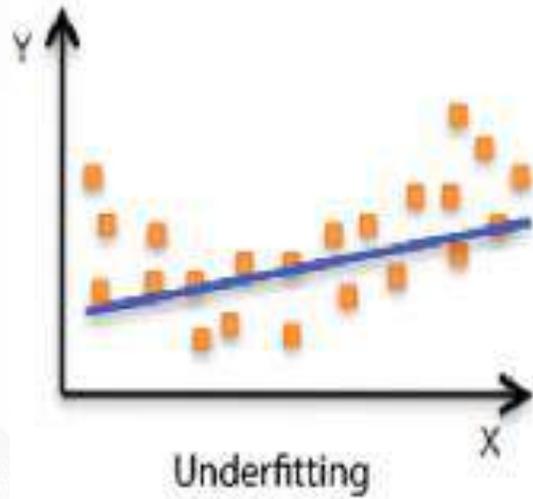
Es la combinación de las métricas de precisión y exhaustividad y sirve de compromiso entre ellas. La mejor puntuación F1 es igual a 1 y la peor a 0.



¿Cómo validar un modelo?

Overfitting vs Underfitting

Overfitting vs Underfitting



Overfitting vs Underfitting

Overfitting sucede cuando se entrena demasiado o se entrena con datos extraños, el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo.

Underfitting sucede cuando se entrena poco o con datos de solo un tipo, lo que resulta en una baja generalización y predicciones poco confiables.

Técnicas para evitar el sobreajuste

- **Oversampling**
Cuando se hace la muestra se lo fuerza a que tenga datos de clase para asegurarse de que el modelo se va a entrenar con esa clase.
- **Stratified sampling**
Hacer la muestra de manera que contenga la misma proporción sobre los datos para determinadas columnas.
- **Early stopping**
Para que el modelo no se entrene con los detalles de los datos del training set, lo detiene antes.
- **Train-validation-test sets**
Se utiliza un validation set para corregir errores en las predicciones producidos durante el training. El test set se utiliza para testear el modelo al final.
- **Cross-validation**
Utiliza train-validation-test, alternando los datos con los que se entrena.
- **Regularization**
Penaliza los parámetros para reducir "la importancia".
- **Pruning**
Se "poda" el árbol en el entrenamiento. Solamente es aplicable a CART.

¿Cómo desplegar un modelo?

The background of the slide features a series of overlapping, wavy, light gray lines that create a sense of motion and depth, resembling a stylized wave or a series of concentric, slightly offset curves.

¿Cómo desplegar un modelo?

	Patrón 1 (REST API)	Patrón 2 (BBDD compartida)	Patrón 3 (Streaming)
Formación	Batch	Batch	Streaming
Predicción	Sobre la marcha	Batch	Streaming
Entrega de resultados de predicción	Vía REST API	A través de la BBDD compartida	Streaming vía cola de mensajes
Latencia para predicción	Media	Alta	Muy baja
Dificultad de gestión del sistema	Media	Fácil	Muy difícil



¿Qué usan las grandes compañías?

Que las grandes compañías hayan creado sus propias plataformas para el despliegue de soluciones de ML indica que es un tema complejo.

Algunos ejemplos:

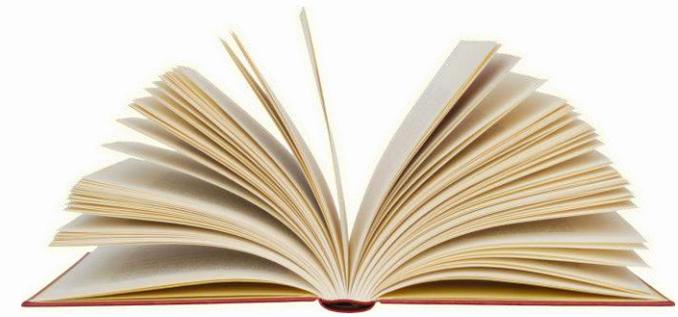
- La plataforma de Uber se llama Michelangelo.
- Facebook tiene FBLearnner Flow
- Google tiene TFX.
- Databricks creó MLFlow.

Demostración y ejercicio práctico en JupyterHub



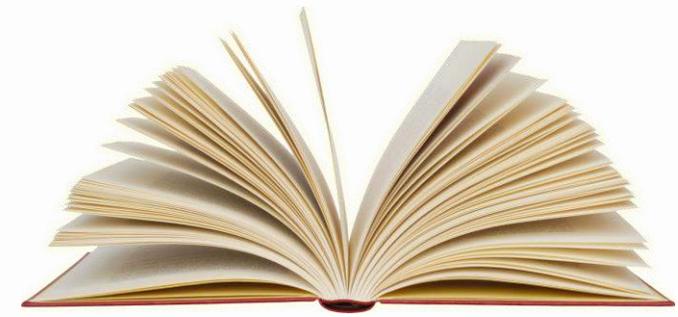
Muchas gracias.

Bibliografía y sitios relacionados



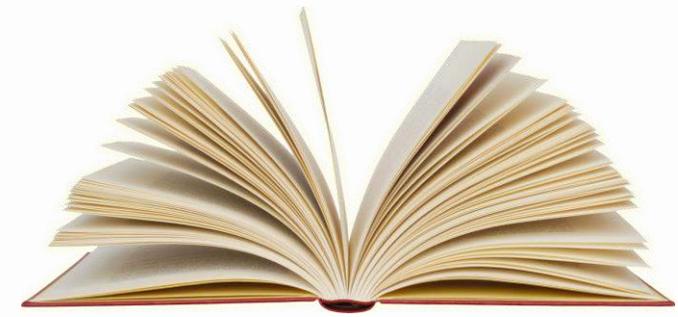
- 1) [Sternkopf H, Mueller R \(2018\). Doing Good with Data: Development of a Maturity Model for Data Literacy in Non-governmental Organizations. *Proceedings of the 51st Hawaii International Conference on System Sciences*.](#)
- 2) [Ridsdale C, Rothwell J, Smith M, Ali-Hassan, Bliemel M, Irvine D, Kelley D, Matwin S, Wuetherick B. Strategies and Best Practices for Data Literacy Education. *Dalhousie University*.](#)
- 3) [Grillenberger A, Romeike R \(2018\). Developing a Theoretically Founded Data Literacy Competency Model. *WIPSCCE*.](#)
- 4) [Data to the people \(2018\). Databilities. *Sitio web Wix*.](#)
- 5) [Bonikowska A, Sanmartin C, Frenette M \(2019\). Data Literacy: What It Is and How to Measure It in the Public Service. *Sitio web Statics Canada*.](#)
- 6) [*Sitio web Data Literacy Project*.](#)
- 7) [Department of the Prime Minister and Cabinet \(2016\). Data skills and capability in the Australian Public Service. *Sitio web Australian Government*.](#)
- 8) [*Sitio web Open Data Institute*.](#)

Bibliografía y sitios relacionados



- 9) [AGESIC, Presidencia de la República Oriental del Uruguay \(2019\). Plan de Gobierno Digital 2020: Transformación con equidad. Sitio web Presidencia de la República Oriental del Uruguay.](#)
- 10) DAMA International (2019). DAMA – DMBOK: Data Management Body of Knowledge. *Technics Publications*.
- 11) [AGESIC \(2019\). Uruguay: Política de Datos para la Transformación Digital. Sitio web Presidencia de la República Oriental del Uruguay.](#)
- 12) [Azevedo A, Santos M.F. \(2008\). KDD, SEMMA and CRISP-DM: A PARALLEL OVERVIEW.](#)
- 12) [IBM Analytics. Foundation Methodology for Data Science IBM.](#)
- 13) [AGESIC \(2019\). Framework de Análisis de Datos. Sitio web Presidencia de la República Oriental del Uruguay.](#)
- 14) [AGESIC \(2019\). Marco de referencia para la gestión de calidad de datos. Sitio web Presidencia de la República Oriental del Uruguay.](#)
- 15) [Política y estrategia de datos para la transformación digital \(2018\). AGESIC. Sitio web Presidencia de la República Oriental del Uruguay.](#)
- 16) [AGESIC \(2019\). Estrategia de Inteligencia Artificial para el Gobierno Digital. Sitio web Presidencia de la República Oriental del Uruguay.](#)

Bibliografía y sitios relacionados



- 17) [Google Developers. Curso intensivo de aprendizaje automático](#)
- 18) [Google Developers. Data Preparation and Feature Engineering in ML](#)
- 19) [Google Developer. Glosario sobre aprendizaje automático](#)
- 20) [Kaggle. Intro to Machine Learning](#)
- 21) [Kaggle. Intermediate Machine Learning](#)
- 22) Data Engineering, Preparation, and Labeling for AI 2019 [CGR-DE100]
- 23) [Amazon Machine Learning](#)
- 24) [Kaggle - Titanic: Machine Learning from Disaster](#)