Clase 2 – Aprendizaje Automático

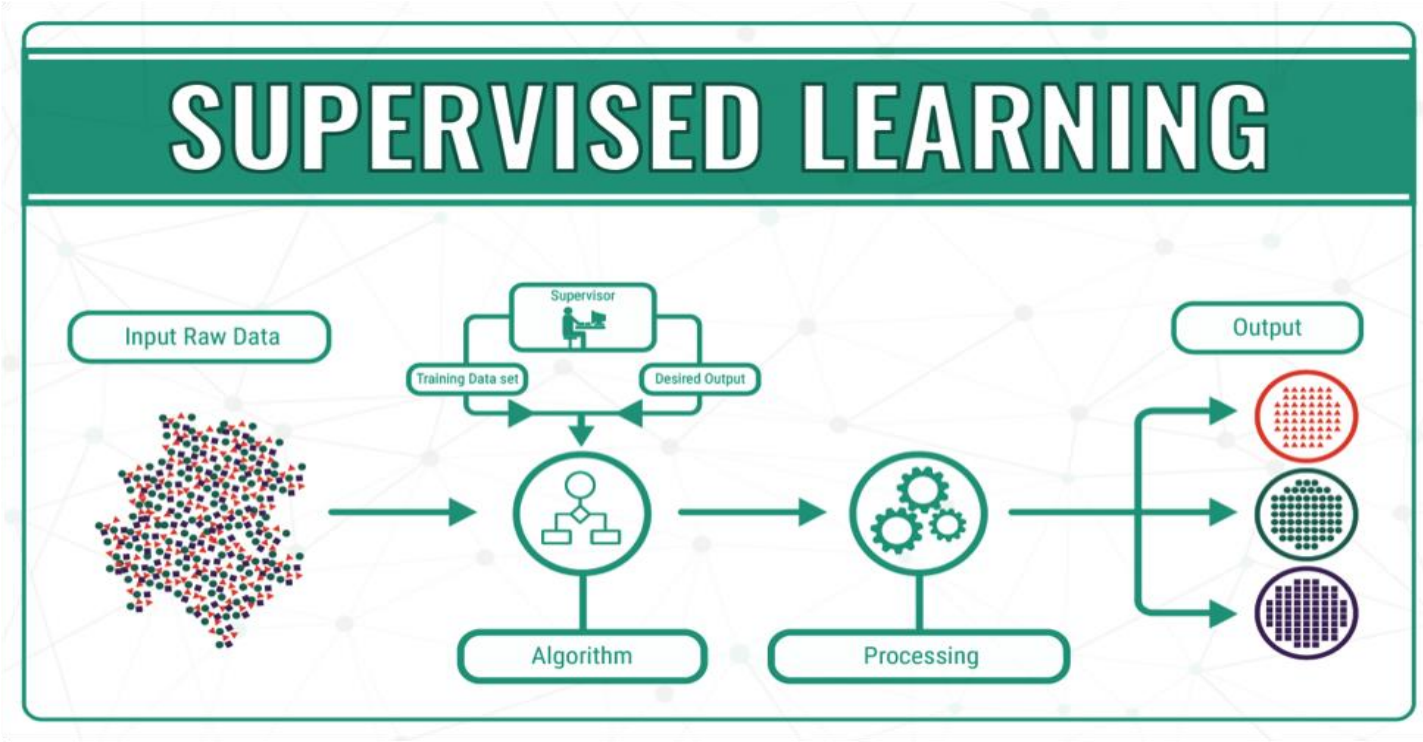
Aprendizaje Supervisado

Universidad Nacional de Luján

Objetivos de la clase

- **Reconocer** los conceptos básicos asociados a Machine Learning
- **Identificar** las principales métricas para evaluar la performance de un modelo de regresión
- Reconocer las etapas de Entrenamiento, Test y validación

Aprendizaje Supervisado

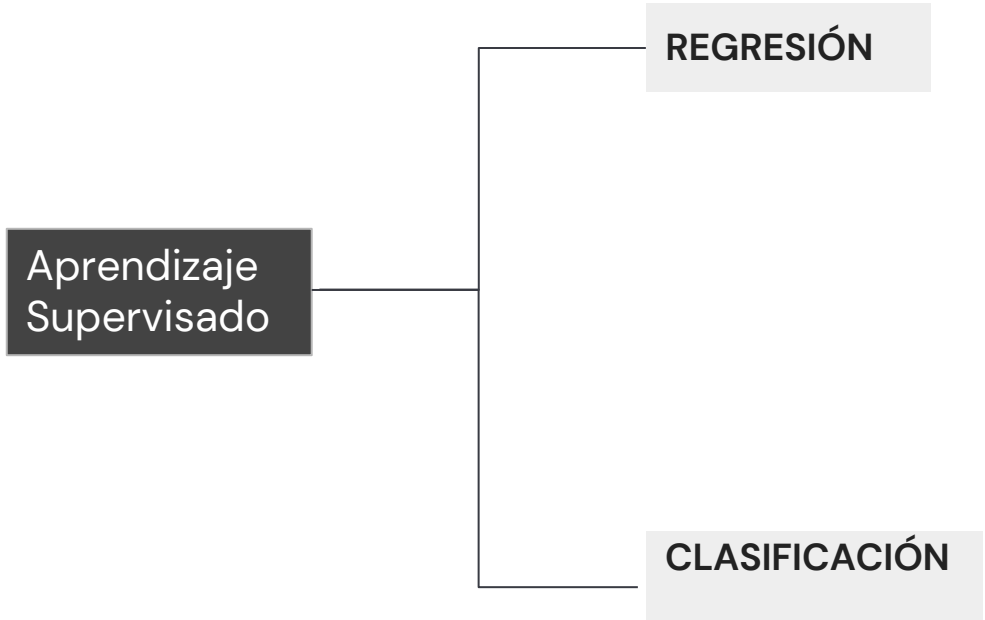


Para pensar....

¿Cuál es el objetivo principal del aprendizaje supervisado?

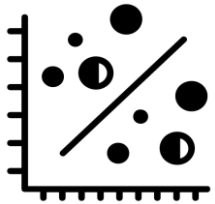
Predecir las respuestas que habrá en el futuro, gracias al entrenamiento del algoritmo con datos conocidos del pasado (datos históricos).

Mapa de conceptos



Algoritmos de Regresión

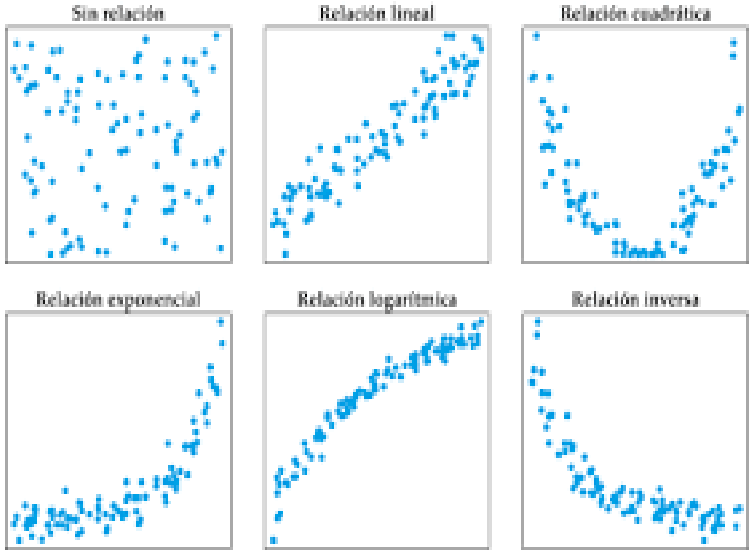
Problemas de Regresión



•Tenemos dos grandes tipos: problemas lineales y no lineales:

•**Problemas lineales:** son aquellos donde los coeficientes que acompañan a las variables del modelo son lineales.

Problemas no lineales: son todos aquellos en donde no se cumple el supuesto del modelo lineal, por ejemplo, una serie de Fourier o de crecimiento Weibull.



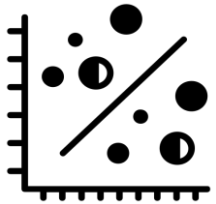
Precio de una casa

105.000 U$S

93.000 U$S

??? U$S

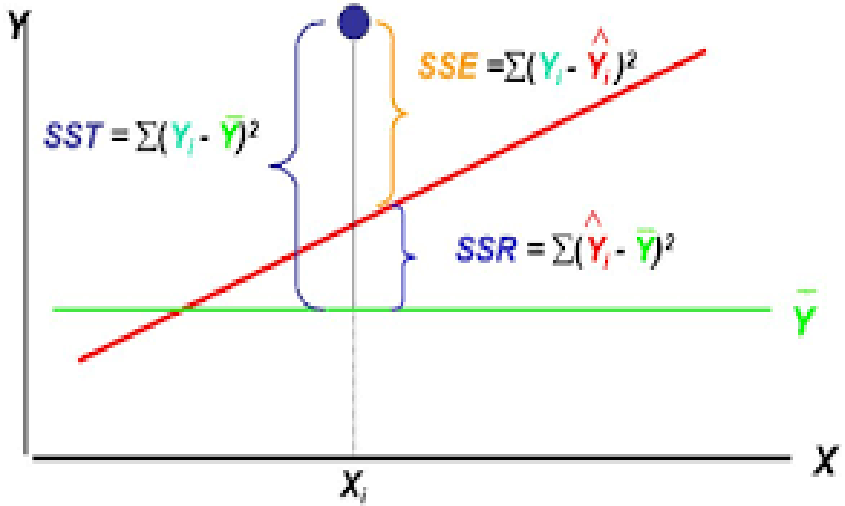
Regresión Lineal



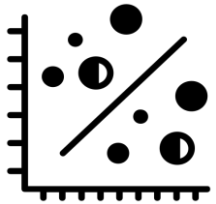
La estimación de α y β se hace por medio del método de **mínimos cuadrados,** donde se busca minimizar la suma de cuadrados de los errores dada por:

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

test set; predicted vaue; actual value



Regresión Lineal Simple y Múltiple



Linear Regression: Single Variable

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

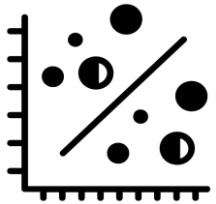
Predicted output — Coefficients — Input — Error

Linear Regression: Multiple Variables

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

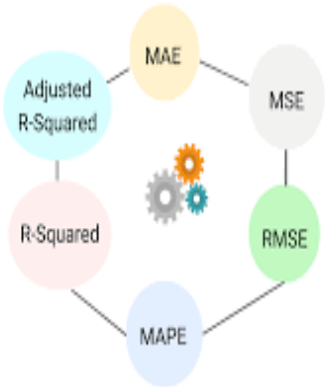
Métricas para algoritmos de regresión

Métricas RL

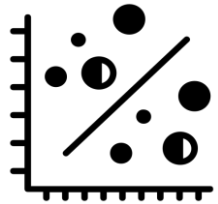


En la actualidad hay muchas formas para estimar el rendimiento y evaluar el ajuste del modelo de regresión, las más importantes son:

- **Error Cuadrático Medio** (RMSE, por sus siglas en inglés, Root Mean Squared Error).
- **Error Absoluto Medio** (MAE, Mean Absolute Error).
- **R-Cuadrado**
- **MAPE**



Métricas RL



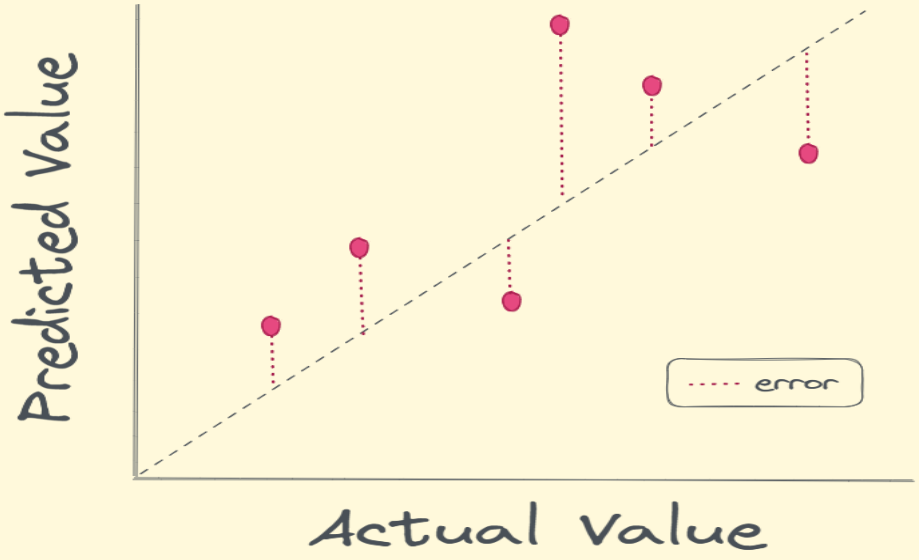
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

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

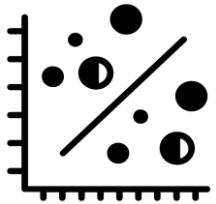
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$



Resumen



1. Las métricas son vitales para cualquier modelo de aprendizaje automático.

2. El error cuadrático medio (**MSE**) mide el error cuadrático medio de nuestras predicciones. Su problema es que, dado que los valores se elevan al cuadrado, la unidad de medida cambia.

3. Para salvar esta deficiencia, analizamos otra métrica llamada **RMSE**, que revierte el valor a su unidad de medida original tomando una raíz cuadrada.

4. El **MAPE** se puede usar para comparar dos modelos de diferentes escalas.

5. El **R2** aumenta a medida que se agregan variables independientes al modelo, el R2_ajustado resuelve en parte el problema con el R2.

6. El **MAE** tiene la ventaja de que, dado que se toma el valor absoluto, todos los errores se ponderarán en la misma escala lineal.

Entrenamiento y Validación

Entrenamiento

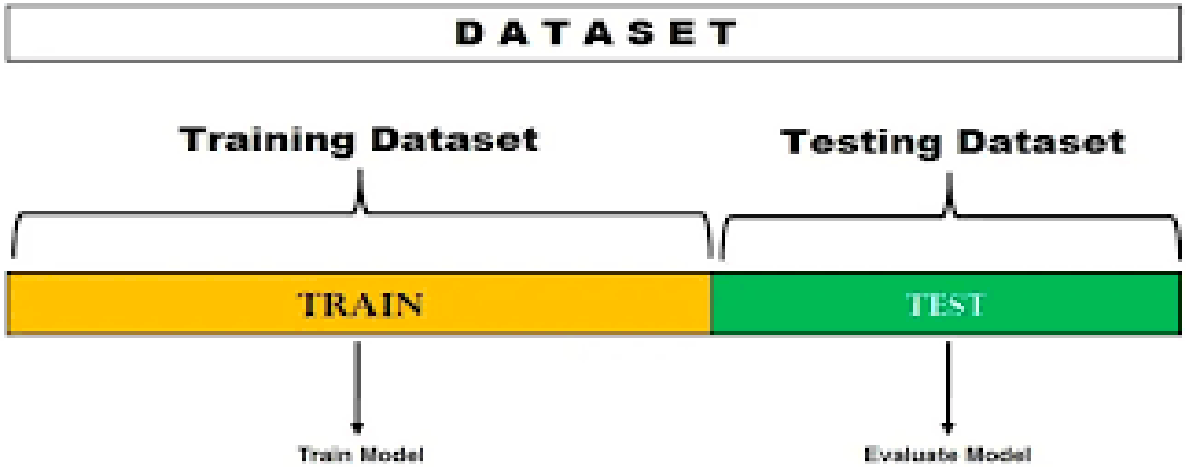
•Proceso en el que se **detectan los patrones de un conjunto de datos**, es decir, es el **corazón del machine learning**.
Cuando identificamos los patrones, se pueden hacer **predicciones** con **nuevos datos** que se incorporen al sistema.

Validación

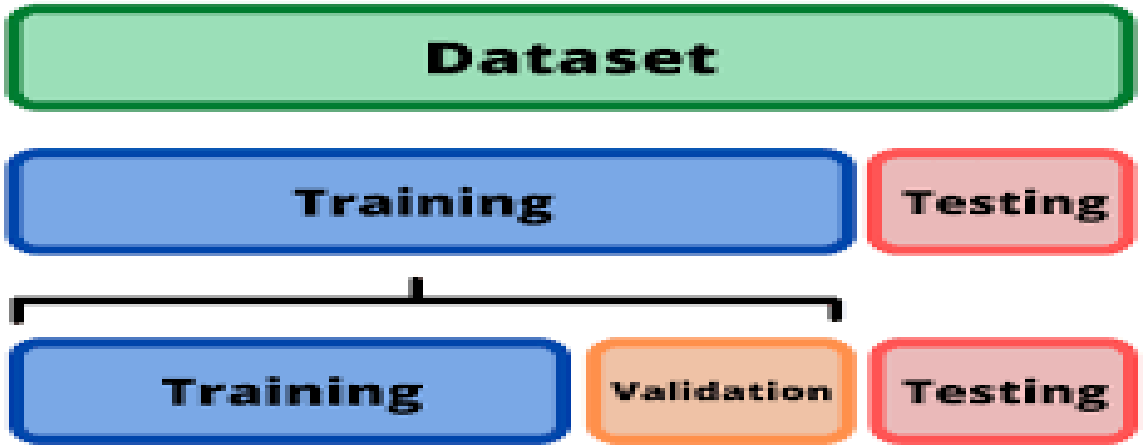
•**Proceso de evaluar un modelo entrenado sobre un conjunto de datos de prueba.** Esto proporciona la capacidad de **generalización** de un modelo de ML.

•Para poder evaluarlo correctamente, hay que realizar "split de datos" es decir, separar nuestro dataset original en **"Datos de Entrenamiento",** que serán usados justamente para entrenar a nuestro modelo y en **"Datos de Test o de Testing"** que serán aquellos datos que utilizaremos para evaluar la performance de nuestro modelo.

Porcentaje de training y test



Porcentaje de training y test



Break

¡5 minutos y volvemos!

¿Preguntas?

Glosario

Aprendizaje Supervisado: subcategoría del aprendizaje automático y la inteligencia artificial que cuenta con datos etiquetados (históricos) para aprender de comportamiento de una variable particular.

Problemas de regresión: son aquellos donde la variable respuesta es una variable continua (e.g, predicción de ventas)

Regresión: es un método para determinar la influencia de variables independientes en una variable dependiente.

Training: fracción de datos (usualmente 70-80%) que se utiliza para entrenar algoritmos de Machine Learning supervisado con el fin de entender patrones y tendencias.

Validación: fracción de datos (usualmente 20-30%) que se utiliza para validar algoritmos de Machine Learning supervisado con el fin de identificar si el modelo aprendió correctamente.

Resumen
de la clase hoy

- Aprendizaje Supervisado
- Algortimo de Regresión
- Métricas para regresión


Muchas
Gracias!