



SECRETARÍA DE  
INNOVACIÓN

# CIENCIA DE DATOS



# Agenda

## Sesión 16/18

### Qué es el modelado de datos

- ¿Qué es modelado de datos?
- La regresión lineal
- scikit-learn
- Sobre el Aprendizaje Automático

# ¿Qué es modelado de datos?

En términos matemáticos, se habla de “modelar” a la creación de un modelo, una reconstrucción simplificada (¡simplificada en extremo!) de cómo funciona un proceso observado en el mundo real. En un modelo de datos, siempre tenemos al menos:

- Una variable resultante, siempre una sola, también llamada variable “dependiente”
- Una o más variables predictoras, también llamadas “explicativas”

# ¿Qué es modelado de datos?

El modelado de datos puede ser utilizado para dos propósitos:

**1. Predecir el valor** de una variable resultante en base a valores conocidos de las variables predictoras. Aquí no interesa tanto entender cómo es que las variables interactúan entre sí, o por qué lo hacen.

Mientras las predicciones sean acertadas, o se acerquen lo suficiente, el modelo cumple su cometido. Los modelos predictivos se emplean en una enorme variedad de aplicaciones: inversión en bolsa, prevención de fraude, publicidad online, fijación de primas en seguros de riesgo, etc.

# ¿Qué es modelado de datos?

**2. Explicar** la relación entre una variable dependiente y todas las demás (las explicativas), buscando determinar si la relación es significativa. Los modelos explicativos son los que se favorecen en investigación académica, ya que ayudan a entender el fenómeno modelado.

Existen muchísimas técnicas para modelar datos, algunas de ellas simples como la **regresión lineal**, y otras mucho más complejas, como las **redes neuronales**.

# Regresión lineal simple

**La regresión lineal**, fácil de explicar y muy fácil de resolver con la ayuda de una computadora, es el caballito de batalla del modelado estadístico.

A pesar de que no es adecuada para ciertos tipo de datos, y de que existen métodos más modernos que explotan con intensidad el potencial de las computadoras, la regresión lineal sigue siendo la herramienta más común. Un poco por costumbre, y otro porque es el método más fácil de interpretar, lo que favorece entender y comunicar sus resultados.

# Regresión lineal simple

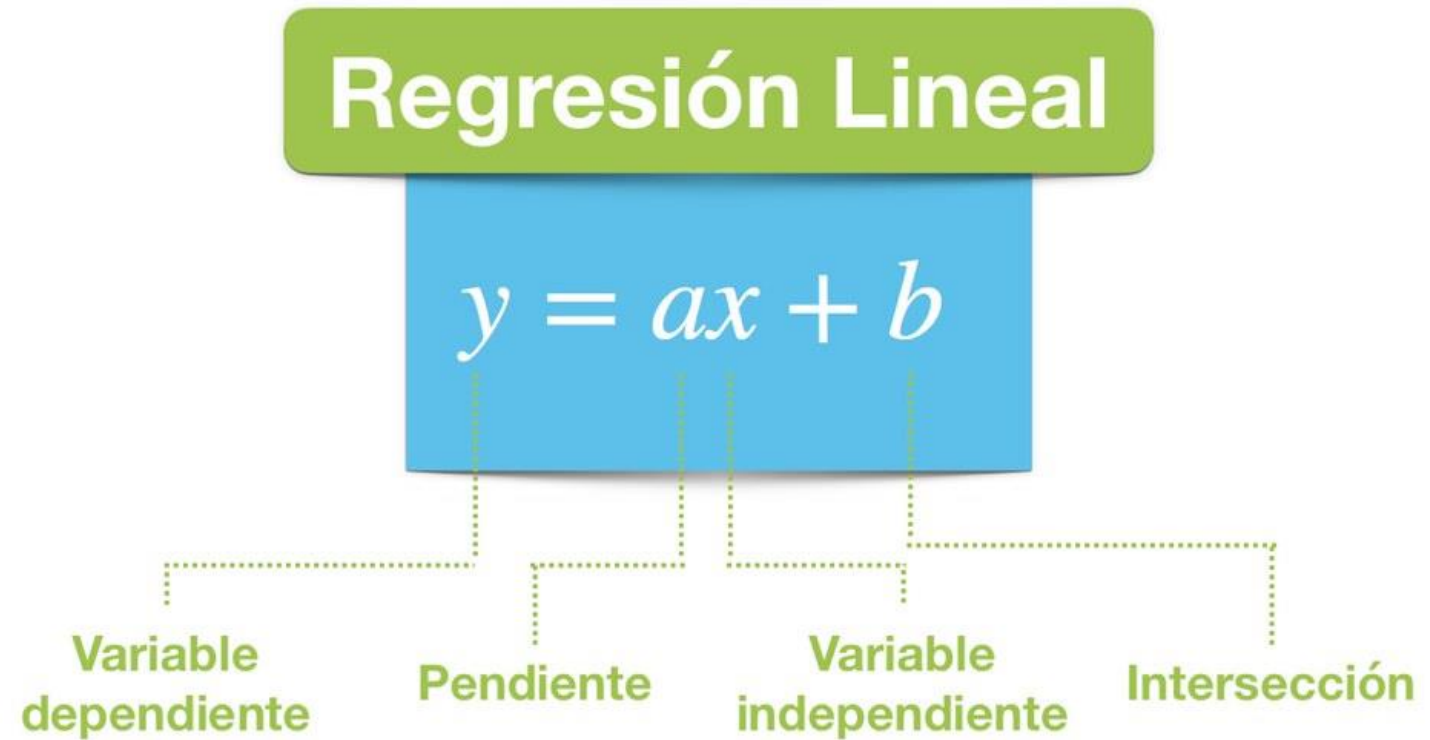
La **regresión lineal** es un método estadístico que trata de **modelar la relación** entre una variable continua y una o más variables independientes mediante el ajuste de una ecuación lineal.

Se llama **regresión lineal simple** cuando solo hay una variable independiente y **regresión lineal múltiple** cuando hay más de una.

Dependiendo del contexto, a la variable modelada se le conoce como variable dependiente o variable respuesta, y a las variables independientes como regresores, predictores o features.

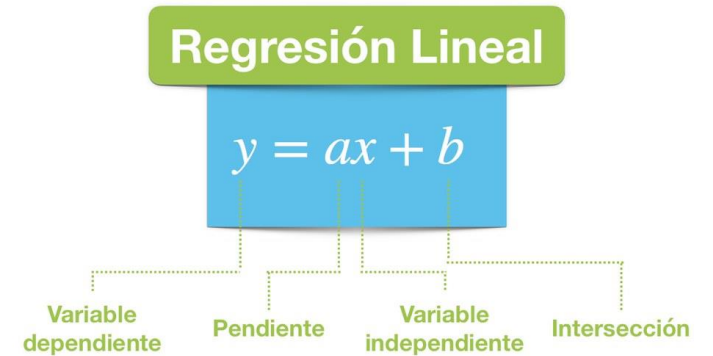


# Regresión lineal simple



Esta es la ecuación de Regresión Lineal Simple. Se llama simple porque solo hay una variable independiente involucrada, que vendría siendo “x”.

# Regresión lineal simple



Donde:

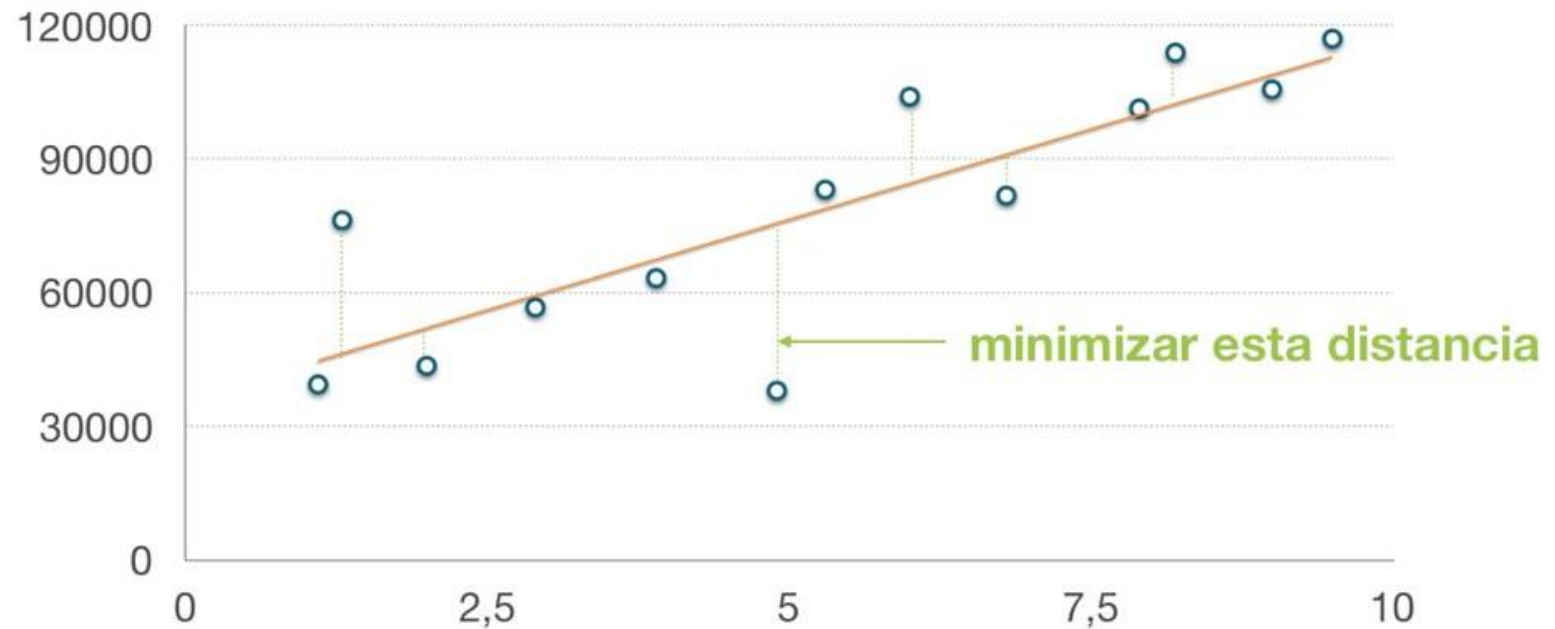
**y** – es la variable dependiente o la variable a predecir.

**x** – es la variable independiente o la variable que usamos para hacer una predicción.

**a** – es la pendiente o el valor que debe ser determinado, se le conoce como coeficiente y es una especie de magnitud de cambio que pasa por y cuando x cambia.

**b** – es la constante que debe ser determinada, se le conoce como intercepto porque cuando x es igual a 0, entonces  $y = b$ .

# Regresión lineal simple



El objetivo con Regresión Lineal Simple es minimizar la distancia vertical entre todos los datos y nuestra línea, por lo tanto, para determinar la mejor línea, debemos minimizar la distancia entre todos los puntos y la distancia de nuestra línea. Existen muchos métodos para cumplir con este objetivo, pero todos estos métodos tienen un solo objetivo que es el de minimizar la distancia.

# Regresión lineal simple

Donde:

**y** – es la variable dependiente o la variable a predecir.

**x** – es la variable independiente o la variable que usamos para hacer una predicción.

**a** – es la pendiente o el valor que debe ser determinado, se le conoce como coeficiente y es una especie de magnitud de cambio que pasa por y cuando x cambia.

**b** – es la constante que debe ser determinada, se le conoce como intercepto porque cuando x es igual a 0, entonces  $y = b$ .

# Regresión lineal simple

Una forma en que el modelo de regresión encuentre la mejor línea de ajustes es utilizando el criterio de mínimos cuadrados para reducir el error.

El error es una parte inevitable del proceso de predicción, no importa cuán poderoso sea el algoritmo que elijamos, siempre habrá un error irreducible. Sabemos que no podemos eliminar por completo el error, pero aún podemos intentar reducirlo al nivel más bajo. Justamente es en este momento en que se usa la técnica conocida como mínimos cuadrados.

La técnica de mínimos cuadrado intenta reducir la suma de los errores al cuadrado, buscando el mejor valor posible de los coeficientes de regresión.

# Regresión lineal simple

Los mínimos cuadrados no es la única técnica para usar en Regresión Lineal pero se selecciona debido:

Utiliza un error cuadrado que tiene buenas propiedades matemáticas, por lo que es más fácil diferenciar y calcular el descenso del gradiente.

Es fácil de analizar y computacionalmente más rápido, es decir, puede aplicarse rápidamente a conjuntos de datos que tienen miles de características.

La interpretación es mucho más fácil que otras técnicas de regresión.

# Regresión lineal simple

Comprendamos en detalle como usar estas formulas con un ejemplo:

Se nos da un conjunto de datos con 100 observaciones y 2 variables, altura y peso. Necesitamos predecir el peso dada la altura. La ecuación sería el de Regresión Lineal simple ya que solamente cuenta con una variable independiente y se puede escribir de la siguiente forma:

$$y = ax + b$$

Donde:

y – es el peso

x – es la altura

a, b son los coeficientes a ser calculados

# Regresión lineal simple

Al usar Python o cualquier lenguaje de programación no necesitas saber cómo se calculan estos coeficientes e inclusive el error, razón por la cual a la mayoría de las personas no les importa el cómo calcularla, pero es mi consejo que debes por lo menos tener un conocimiento al respecto para de esta forma te acerques a ser un maestro en estos temas.

La formula para calcular estos coeficientes es fácil inclusive si solamente tienes los datos y no tienes acceso a ninguna herramienta estadística para el cálculo podrás hacer la predicción.



# Regresión lineal simple

Comprendamos en detalle como usar estas formulas con un ejemplo:

Se nos da un conjunto de datos con 100 observaciones y 2 variables, altura y peso. Necesitamos predecir el peso dada la altura. La ecuación sería el de Regresión Lineal simple ya que solamente cuenta con una variable independiente y se puede escribir de la siguiente forma:

$$y = ax + b$$

Donde:

y – es el peso

x – es la altura

a, b son los coeficientes a ser calculados

# scikit-learn

¿Qué es Scikit-Learn?

Scikit-Learn es una de estas librerías gratuitas para Python. Cuenta con algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad. Además, presenta la compatibilidad con otras librerías de Python como NumPy, SciPy y matplotlib.

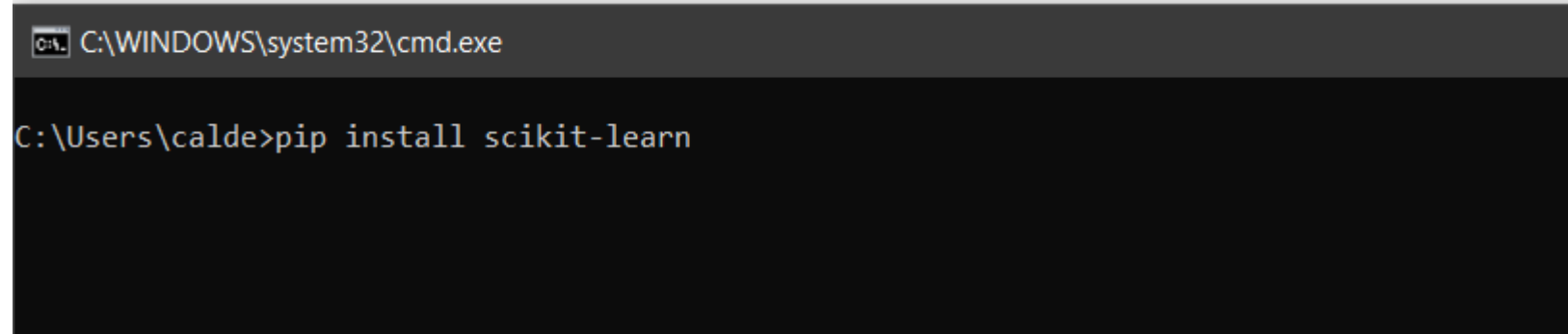
La gran variedad de algoritmos y utilidades de Scikit-learn la convierten en la herramienta básica para empezar a programar y estructurar los sistemas de análisis datos y modelado estadístico. Los algoritmos de Scikit-Learn se combinan y depuran con otras estructuras de datos y aplicaciones externas como Pandas o PyBrain.

# scikit-learn

## Instalación de Scikit-Learn

Se utiliza el comando de pip en la terminal:

**pip install -U scikit-learn**



```
C:\WINDOWS\system32\cmd.exe  
  
C:\Users\calde>pip install scikit-learn
```

# Scikit-learn

**Scikit-learn** es una biblioteca de análisis de datos de código abierto y el estándar de oro para el aprendizaje automático (ML) en el ecosistema de Python. Los conceptos y características clave incluyen

- Métodos algorítmicos de toma de decisiones, que incluyen:
  - 1. Clasificación:** identificación y categorización de datos basados en patrones.
  - 2. Regresión :** predecir o proyectar valores de datos basados en la media media de los datos existentes y planificados.
  - 3. Agrupación :** agrupación automática de datos similares en conjuntos de datos.
- Algoritmos que admiten análisis predictivos que van desde la regresión lineal simple hasta el reconocimiento de patrones de redes neuronales.
- Interoperabilidad con bibliotecas NumPy, pandas y matplotlib.

# Scikit-learn

## ¿Por qué utilizar Scikit-Learn para el aprendizaje automático?

Ya sea que solo esté buscando una introducción al aprendizaje automático, desee comenzar a trabajar rápidamente o esté buscando la última herramienta de investigación de aprendizaje automático, encontrará que scikit-learn está bien documentado y es fácil de aprender / usar. Como biblioteca de alto nivel, le permite definir un modelo de datos predictivo en solo unas pocas líneas de código y luego usar ese modelo para ajustar sus datos. Es versátil y se integra bien con otras bibliotecas de Python, como matplotlib para trazar , numpy para vectorización de matrices y pandas para marcos de datos .

# Sobre el Aprendizaje Automático

## Conceptos Básicos De Aprendizaje Automático

Para usar scikit-learn, primero debe estar familiarizado con parte de la terminología que se usa normalmente en los proyectos.

**Precisión** : la fracción de predicciones que acertó un modelo de clasificación. En la clasificación de clases múltiples, la precisión se define de la siguiente manera:  $\text{Exactitud} = \text{Predicciones correctas} / \text{Número total de ejemplos}$

En la clasificación binaria, la precisión tiene la siguiente definición:

$\text{Precisión}^* = (\text{Positivos verdaderos} + \text{Negativos verdaderos}) / \text{Número total de ejemplos}$

# Sobre el Aprendizaje Automático

**Datos de ejemplo** : instancia particular (característica) de datos, definida como  $x$ . Hay dos categorías de ejemplos de datos:

- Datos etiquetados : incluye tanto las características como la etiqueta, definida como  $\{\text{características}, \text{etiqueta}\}$ :  $(x, y)$
- Datos sin etiqueta : contiene características pero no la etiqueta, definida como:  $\{\text{características}, ?\}$ :  $(X, ?)$

**Característica**: una variable de entrada. Es una característica o propiedad medible de una cosa que se observa. Cada proyecto de AA tiene una o más funciones.

# Sobre el Aprendizaje Automático

**Agrupación:** una técnica que agrupa puntos de datos en función de sus similitudes. Cada grupo se llama Cluster.

**Regresión vs Clasificación:** ambos son modelos que le permiten hacer predicciones que responden preguntas, como qué equipo ganará un evento deportivo.

- Los modelos de **regresión** proporcionan un valor numérico o continuo.
- Los modelos de **clasificación** proporcionan un valor categórico o discreto.

**Modelo:** define la relación entre características y una etiqueta. Por ejemplo, un modelo de detección de rumores que asocia ciertas características asociadas con los rumores.



# Sobre el Aprendizaje Automático

**Aprendizaje supervisado:** el algoritmo utiliza un conjunto de datos etiquetado para "aprender" cómo reconocer las respuestas correctas, que luego puede aplicar a los datos de entrenamiento. Luego, se evalúa y refina la precisión del algoritmo. La mayoría de los proyectos de AA utilizan el aprendizaje supervisado.

**Aprendizaje no supervisado:** el algoritmo intenta dar sentido a los datos sin etiquetar "aprendiendo" características y patrones por sí solo.

# Sobre el Aprendizaje Automático

## Algoritmos ML

Para que las computadoras aprendan sin estar programadas explícitamente, se requieren algoritmos. Los algoritmos son simplemente conjuntos de reglas aplicadas a la computación.

## Conceptos básicos del algoritmo ML:

**Representación:** es una forma de configurar los datos de manera que se puedan evaluar. Los ejemplos incluyen árboles de decisión, conjuntos de reglas, instancias, modelos gráficos, redes neuronales, máquinas de vectores de soporte, conjuntos de modelos y otros.

# Sobre el Aprendizaje Automático

**Evaluación:** dada una hipótesis, la evaluación es una forma de evaluar su validez. Los ejemplos incluyen precisión, predicción y recuperación, error al cuadrado, verosimilitud, probabilidad posterior, costo, margen, entropía kL divergencia y otros.

**Optimización:** el proceso de ajuste de hiperparámetros para minimizar los errores del modelo mediante el uso de técnicas como optimización combinatoria, optimización convexa, optimización restringida, etc.

# Ejemplos en Python

# RESUMEN DE CLASE



SECRETARÍA DE  
INNOVACIÓN