Objetivo principal del curso:

Aprender y aplicar modelos de **aprendizaje automático**.
Dentro de **aprendizaje automático** hay dos grandes paradigmas de aprendizaje:

- **Aprendizaje supervisado:** Son los problemas en donde queremos aprender una variable objetivo que nosotros conocemos y se la otorgamos a nuestro modelo para que junto con un par de variables predictoras la pueda aprender a predecir.
    - Se suele hacer la distinción entre:
        - **Regresión:** La variable objetivo es un número real.
        - **Clasificación**: Se busca predecir una clase. Queremos aprender a definir a qué clase pertenecen las distintas instancias del problema. *Por ejemplo:* Otorgar un crédito, es una clasificación binaria de sí o no.
        - La diferencia entre estos es la naturaleza de la variable objetivo.
- **Aprendizaje no supervisado:** Es cuando a priori no sabemos, o no tenemos definido, el concepto al cual queremos llegar. Lo interesante de estos modelos es que uno no sabe cuántos grupos hay ni qué es lo que hace que un grupo sea uniforme y diferente al resto. Por ejemplo, el clustering.

Regresión lineal:

➔ Tenemos una variable objetivo, que por lo general la llamaremos **Y**.
➔ Una serie de variables predictoras que van a estar dentro del eje **X**.

Machete:

- La **población** es inaccesible o muy costosa de acceder.
- Por lo general, si bien queremos acceder a una **población**, no accedemos a ella pero si a una **muestra** que intenta replicar a esa **población**, queremos que la **muestra** sea representativa.
    - La estadística nos dice qué tan confiable es esa **muestra** respecto a la **población** en función de ciertas medidas.
    - Se trata de arribar a una **población** a través de una **muestra** por que es más fácil de manejar (es más accesible).
- μ es la **media** de la variable que nos interesa medir a nivel poblacional.
    - $\overline{X}$ es la **media muestral**. Es el promedio que calculamos con la **muestra**.
- La **desviación estándar**, nos dice qué tan representativa es la **media**, en qué medida las observaciones se acercan o no a ese promedio muestral. Es el desvío respecto a la tendencia central, promedio.
    - Si todas las observaciones son parecidas al promedio, entonces el desvío es bajo.
    - Si todas las observaciones se encuentran dispersas entonces el desvío es alto.

Librerías:

- **Pandas** se usa para levantar conjuntos de datos estructurados.
- **Numpy** es una librería de cálculo numérico.
- **Seaborn** y **Matplotlib** son librerías para visualización.
- **Sklearn** es una librería se usa para aprendizaje automático, invoca modelos junto con todo su ecosistema. Además que también podemos invocar las métricas.

Covarianza y Correlación:

- Son medidas de asociación entre variables.
- Miden el nivel de asociación lineal entre variables.
- **Covarianza**: Es una medida que es medida entre dos variables
    - Mide qué tan asociadas están estas dos variables.
    - No nos dice nada respecto de qué tan fuerte es la asociación entre las dos variables.
        - A la hora de medir la fuerza de la relación, el valor numérico de la **covarianza** no nos va a interesar tanto porque depende de la escala y las magnitudes en las que se mueven las variables de medición.
    - Es un paso previo para llegar a la **correlación**.

Signo de la Covarianza:

- Será positivo cuando ambas variables tengan un mayor valor que la media respectiva o que ambas tengan un valor menor a la media respectiva.
- Será negativa cuando pasa lo contrario y los puntos tienden a concentrarse en los cuadrantes opuestos (una variable mayor que su media y la otra menor que su media respectivamente).
- La **covarianza** será aproximadamente 0 cuando los puntos tienden a concentrarse de forma casi igualitaria en todos los cuadrantes.

La **correlación** se utiliza para estudiar la fuerza de la asociación lineal entre dos variables.

- La **correlación** es la covarianza normalizada por los desvíos estándares.
- El **coeficiente de correlación** varía entre 1 y -1, cualquiera sea X o Y sin importar sus unidades.
- Nos da un resultado estandarizado, independientemente del dominio de las variables.
- La correlación es más interpretable que la covarianza.
    - Siendo -1 correlación lineal perfecta negativa.
    - Siendo 1 correlación lineal perfecta positiva.

Coeficiente de Correlación de Pearson:

- Es un estimador muestral.
- Sirve para medir la relación lineal entre dos variables.
- Es una versión estandarizada de la covarianza, lo que significa que siempre queda entre -1 y 1 independientemente de las unidades de medida.
- Valores entre -1 y 1 indican distintos grados de asociación, más cerca de +- ± 1, más fuerte.

Transparencia con Matplot:

- Poner un buen nivel de **transparencia** de los puntos nos permite detectar dónde tenemos mayor acumulación de los mismos.
- Está bueno fijar la aleatoriedad para que los resultados sean reproducibles y que no cambien por cada vez que corremos el notebook porque es poco práctico.

Cálculo de la matriz covarianza:

- Hacemos uso de *numeric_only* = *True* para que solo tenga en cuenta a los datos numéricos.
- Nos devuelve una **matriz de covarianza simétrica**, donde en la diagonal (\) tenemos la varianza.
- Recordemos que la **covarianza** nos dice el sentido de esa asociación, pero no nos dice qué tan fuerte es.

Cálculo de la matriz de correlación

- Nos da un resumen de **correlación** entre las variables que estamos comparando.
- La **correlación** de una variable consigo mismo siempre va a ser 1.
- Al ser una gran cantidad de resultados que se muestran en pantalla, se complica la extracción de información/conocimiento, por esta razón, es que a estas matrices las solemos poner en un mapa de calor (*heatmap*).

La correlación no es causalidad.

Que haya dos variables con el **coeficiente de correlación de Pearson** alto no quiere decir que una esté causando la otra.

Modelo Lineal Simple:

- ➢ Proponemos una recta.
- ➢ Es un modelo de **Regresión Lineal**.
- ➢ Va a ser una sola variable a la que vamos a estar tomando como input (variable predictora).
- ➢ Si tuviéramos más de una variable, pasaremos al **Modelo Lineal Múltiple**.

- ➢ **b_0** es la ordenada al origen.
    - A veces es interpretable pero no siempre es útil.
- ➢ **b_1** es el coeficiente que acompaña a la variable X (es la pendiente de la recta).
- ➢ Este modelo nos va a estar explicando la sensibilidad de Y respecto de X. Es decir, cómo es que afecta la variable X a la variable Y.
- ➢ El proceso de aprendizaje es que a partir de los puntos ver el **b_0** y **b_1** que podemos obtener.
- ➢ El mejor **b_0** y **b_1** son aquellos que minimizan el error cuadrático medio.
- ➢ Cuando uno encuentra una tendencia lineal en los datos es conveniente usar regresión lineal, aunque la regresión lineal uno la puede usar con o sin tendencia lineal.

- Modelos de **caja negra**, son aquellos que uno no sabe realmente lo que está pasando o cómo es que el modelo logra predecir.
- Los modelos de **caja blanca**, son aquellos donde uno sabe cómo está implementado.

¿Cómo se interpreta este modelo?

- Tenemos un **B_0** poblacional el cual no accedemos, pero los estimamos a partir del $\overline{B}_{0.}$

Ventas por radio

- Como tienen una baja correlación entre diario y ventas, en consecuencia, tendremos una peor predicción.

Métricas:

R^2 es una métrica que nos dice qué también una recta se ajusta a los datos de entrenamiento de ajuste.

- Varía entre 0 y 1. Vamos a desear que se acerque más a 1, porque eso implica un ajuste lineal perfecto.
- Nos explica la variabilidad de los datos
- Se suele usar con la data de entrenamiento.

Para tests se usan más **MSE** y **RMSE**.

- **MSE** no nos suele interesar tanto porque es poco interpretable con lo que buscamos predecir (más que nada en magnitudes grandes).

En **aprendizaje automático** lo que nos interesa es el poder de generalización de estos modelos para ver qué tan bien predicen fuera de la muestra.

- Cuando tengamos modelos más complejos debemos tener cuidado de cuánta flexibilidad les damos, ya que podemos caer en un caso de overfitting y al

momento de probarlos con los test data obtendremos resultados con muy mala performance.