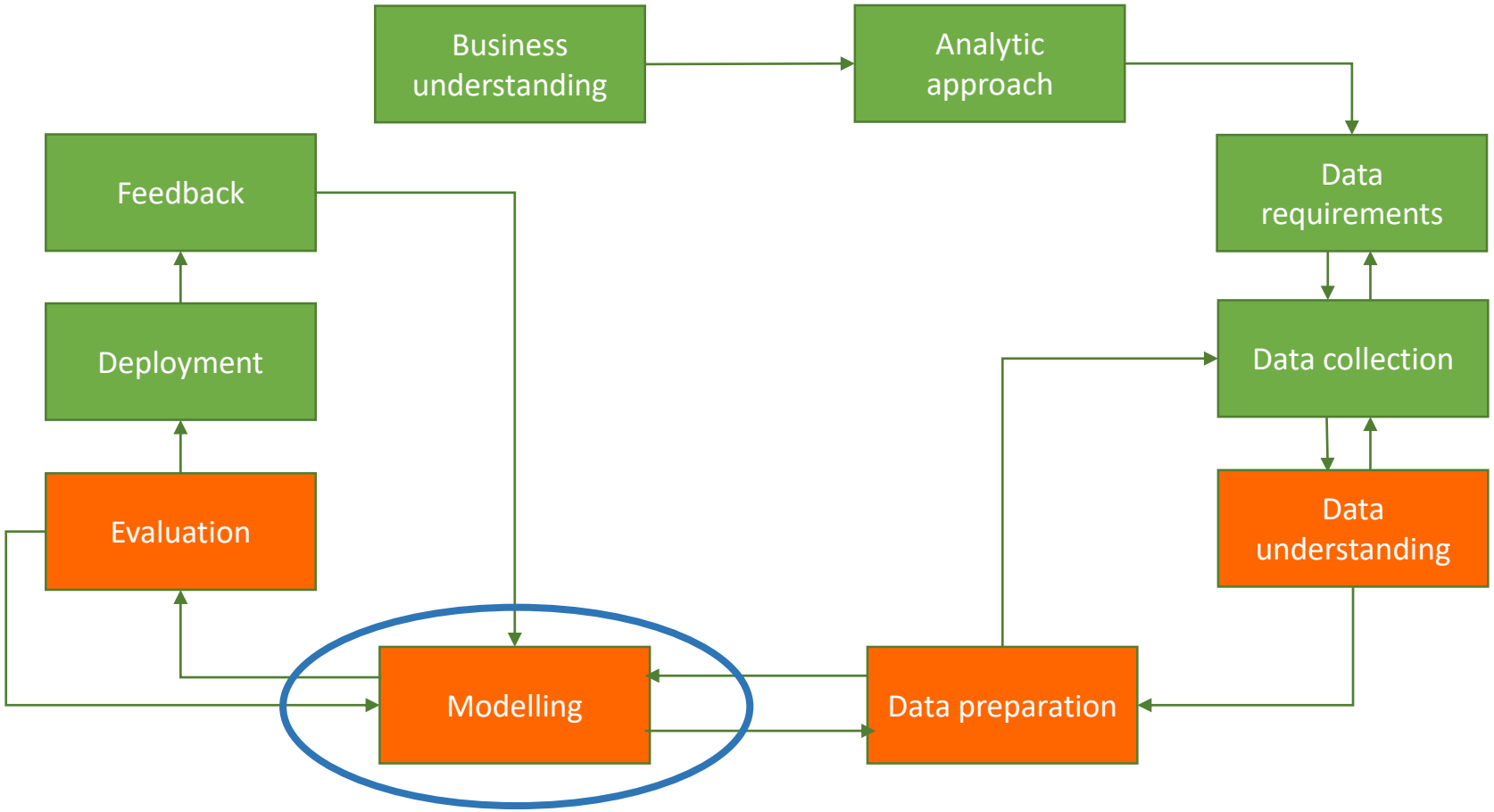
Modelado de datos

Análisis de Datos - Paola A. Sánchez-Sánchez

Metodología para el análisis de datos



Agenda

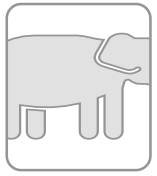
- Aprendizaje supervisado y no supervisado
- Predicción: Modelos de regresión
- Modelos de regression en Python

Tipos de Aprendizaje

Aprendizaje supervisado

Aprendizaje no supervisado

Aprendizaje semi-supervisado



Aprendizaje reforzado

Supervised Learning:

Predicting values. **Known** targets.
User inputs correct answers to learn from. Machine uses the information to guess new answers.

REGRESSION:
Estimate continuous values
(Real-valued output)

CLASSIFICATION:
Identify a unique class
(Discrete values, Boolean, Categories)

Unsupervised Learning:

Search for structure in data. **Unknown** targets.
User inputs data with undefined answers. Machine finds useful information hidden in data.

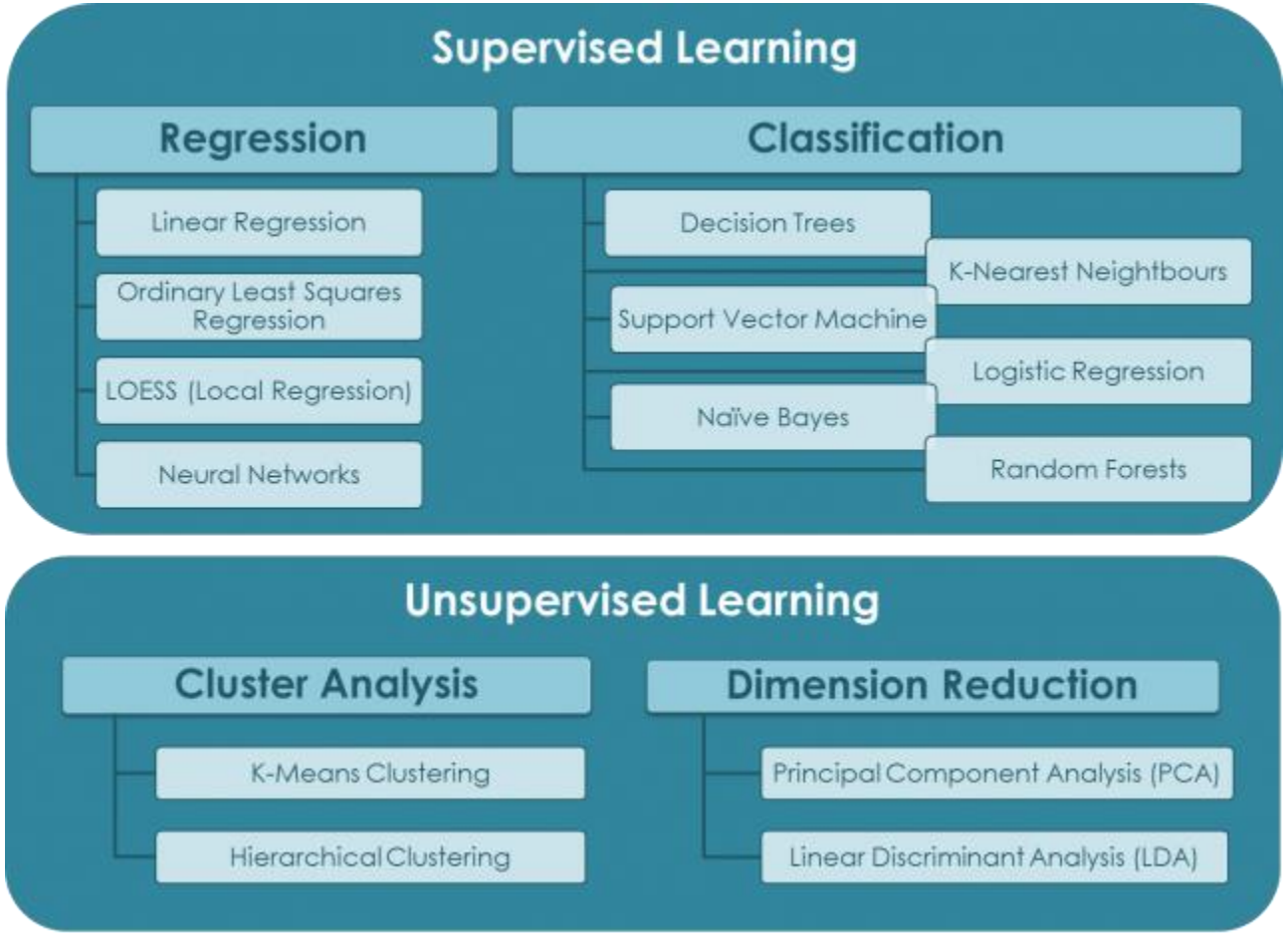
Cluster Analysis
Group into sets

Density Estimation
Approximate distributions

Dimension Reduction
Select relevant variables

https://quantdare.com/machine-learning-a-brief-breakdown/

Modelos AD



https://quantdare.com/machine-learning-a-brief-breakdown/

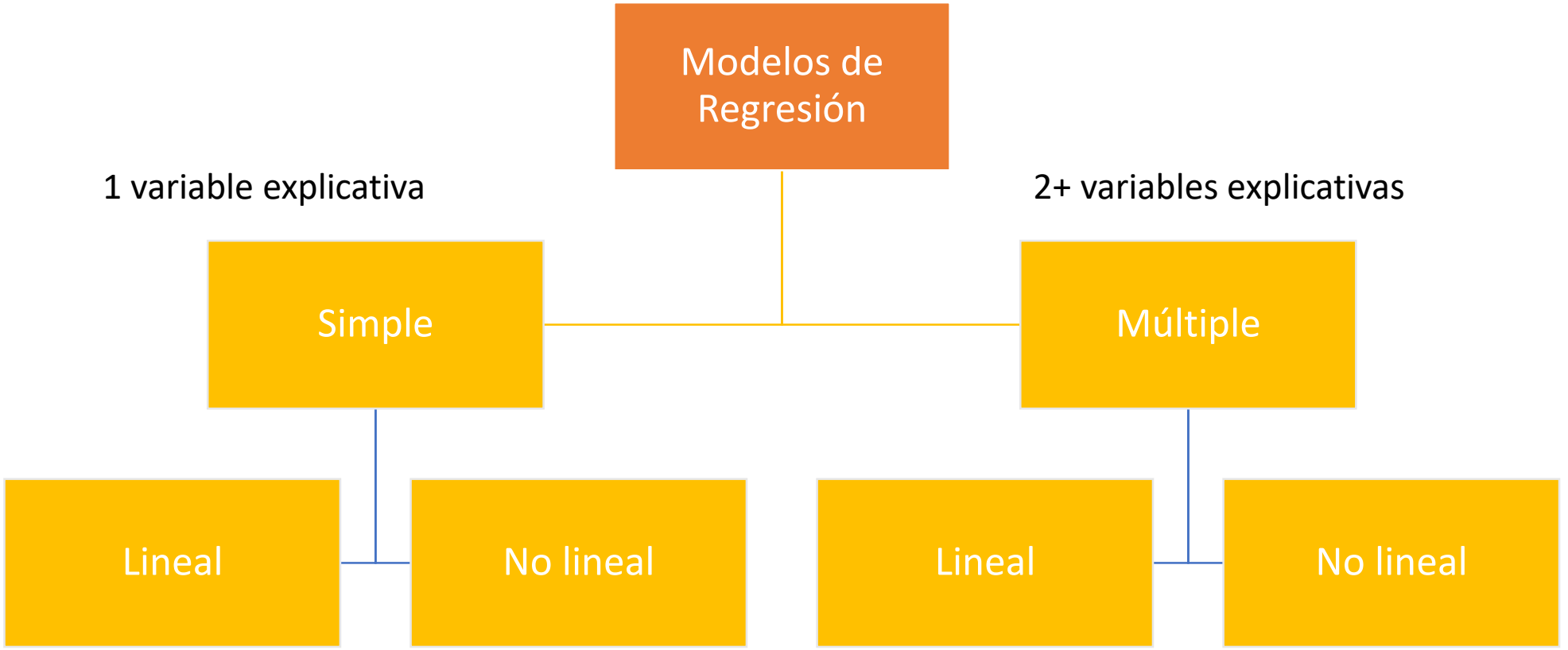
Modelos de Regresión

Técnica paramétrica para evaluar el nivel de asociación entre dos o más variables.

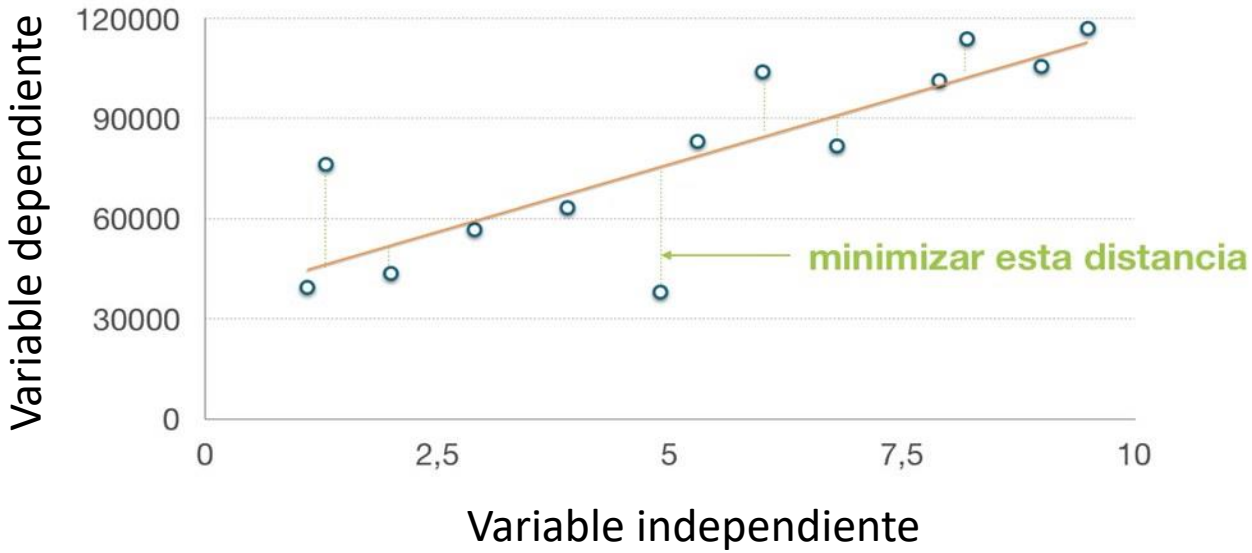
Objetivos:

- Estimar la relación entre una variable dependiente (respuesta) y una o más variables independientes (explicatorias).
- Determinar el efecto de cada variable explicatoria sobre la variable dependiente
- Predecir el comportamiento (valor) de una variable dependiente basado en valores pasados de variables explicatorias

Tipo de Modelos de Regresión



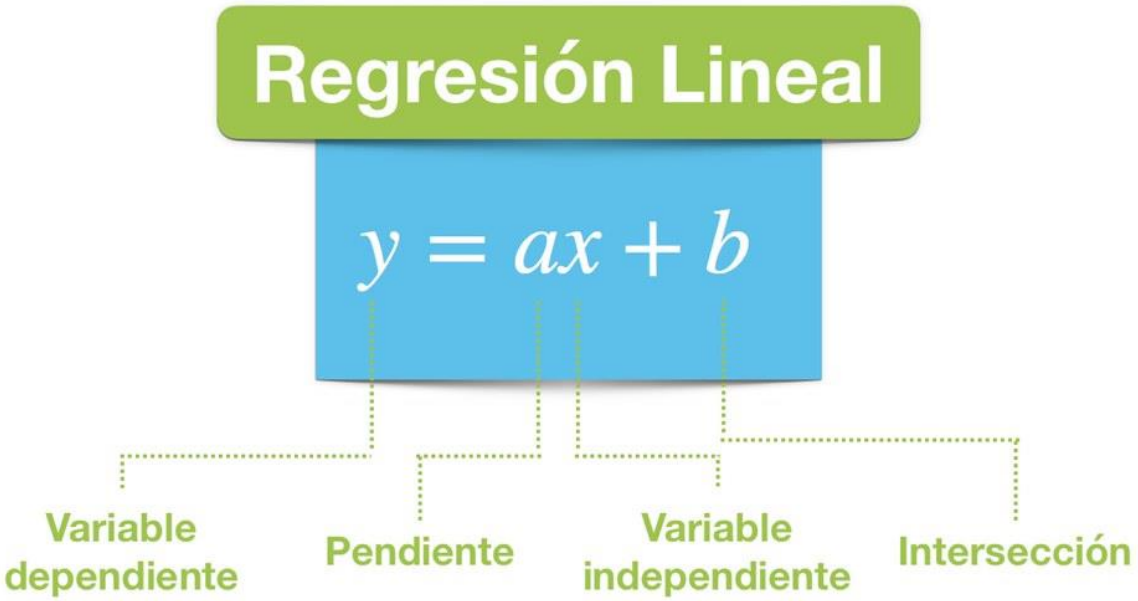
Regresión lineal simple



El objetivo de la regresión lineal simple es minimizar la distancia vertical entre todos los datos y la línea, por lo tanto, para determinar la mejor línea, debemos minimizar la distancia entre todos los puntos y la línea.

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



$$a = \frac{\sum(xi - xmedia)(yi - ymedia)}{\sum(xi - xmedia)^2}$$

$$b = ymedia - a(xmedia)$$

Supuestos sobre el modelo

Existe una relación lineal y aditiva, entre las variables dependientes e independientes.

No debe haber correlación entre las variables independientes.

Los términos de error deben poseer varianza constante, ni deben correlacionarse.

La variable dependiente y los términos de error deben tener una distribución normal

Implementación de modelo en Python

1. Importar librerías
2. Cargar Datos/selección/renombrar
3. Limpieza de Datos
4. Partición de conjunto de datos
5. Ajuste del modelo y predicción
6. Cálculo de parámetros, precisión
7. Gráfico de modelo

```
#Separo los datos de "train" en entrenamiento y prueba para probar los algoritmos
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
#Defino el algoritmo a utilizar
lr = linear_model.LinearRegression()

#Entreno el modelo
lr.fit(X_train, y_train)

#Realizo una predicción
Y_pred = lr.predict(X_test)
```

```
#Calculo de coeficiente
a = lr.coef_

#Calculo de intercepto
b = lr.intercept_

#Calculo de precisión R2
R2 = lr.score(X_test, y_test)
```

Resumen

Definir algoritmo linear_model.LinearRegression()	Conocer pendiente (a) coef_
Entrenar modelo fit(x, y)	Conocer intersección (b) intercept_
Predecir modelo predict(x)	Precisión modelo score(x)