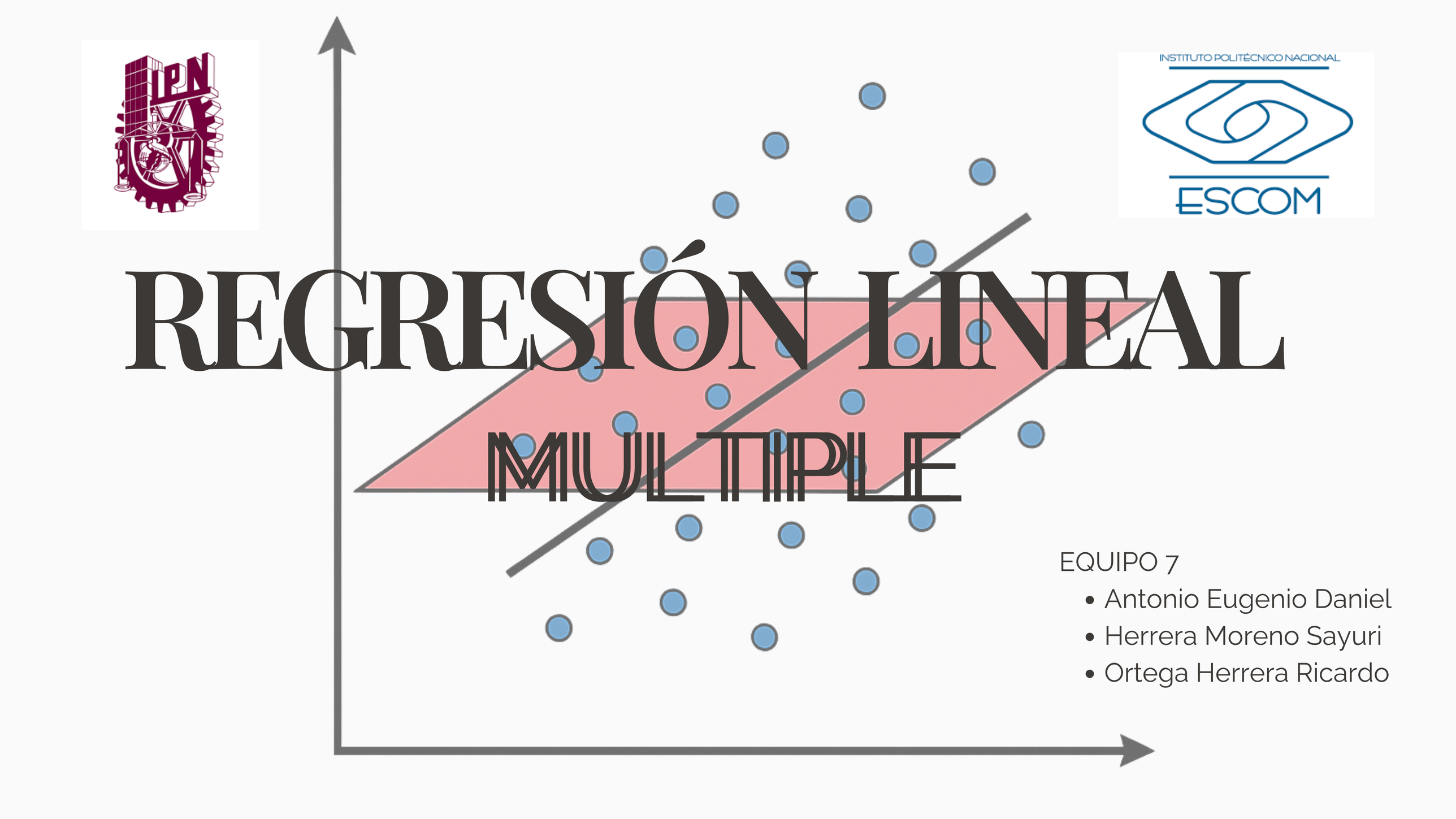
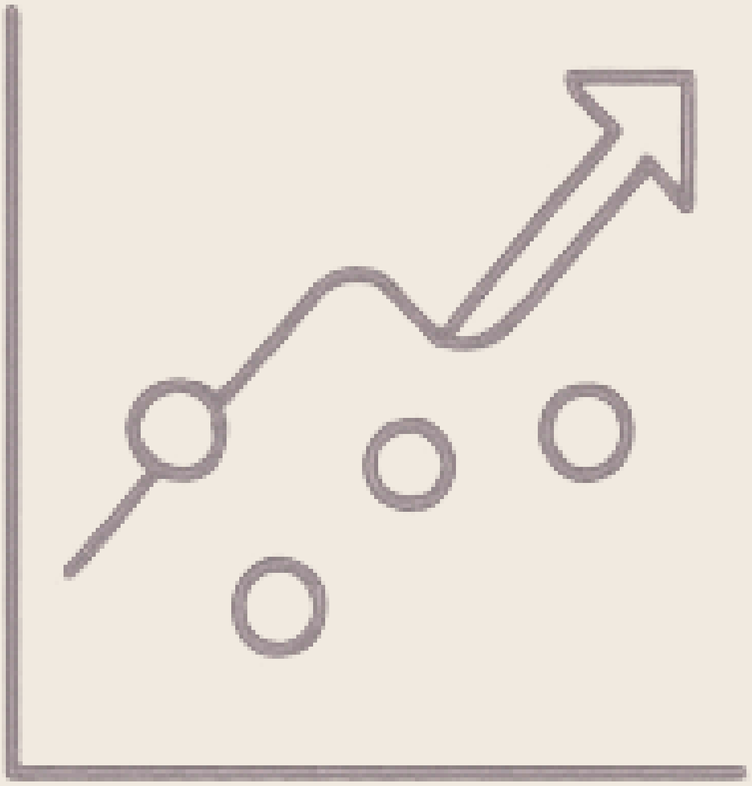


REGRESIÓN LINEAL MULTIPLE

EQUIPO 7

- Antonio Eugenio Daniel
- Herrera Moreno Sayuri
- Ortega Herrera Ricardo

La regresión lineal múltiple trata de ajustar modelos lineales o linealizables entre una variable dependiente y más de una variable independiente (Montero 2016).



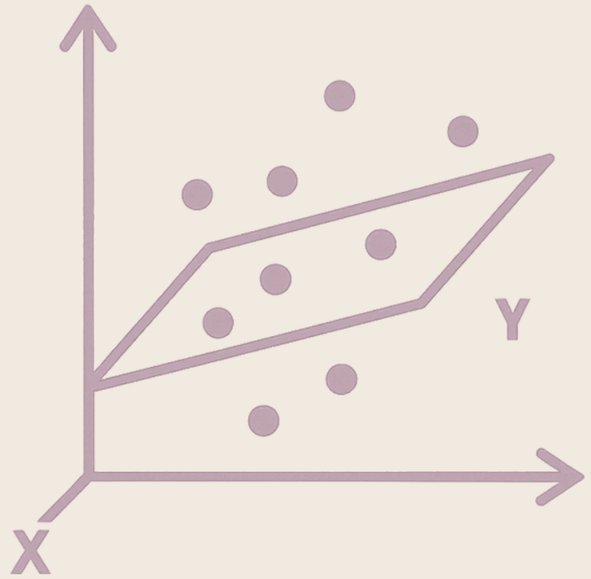
REGRESIÓN LINEAL MULTIPLE

La regresión lineal múltiple (RLM) es una técnica de aprendizaje supervizado fundamental en el análisis predictivo. Permite predecir una variable dependiente (Y) a partir de dos o más varibales independientes.
Su objetivo es modelar la relación lineal entre multiples factores y el resultado de interés.

Ecuación general:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- Y: variable dependiente
- Xi: variables independientes
- βi: coeficientes de regresión
- ε: error o residuo



IMPORTANCIA EN MINERA DE DATOS

La RLM es un pilar en el ámbito de la Minería de Datos y el Machine Learning:

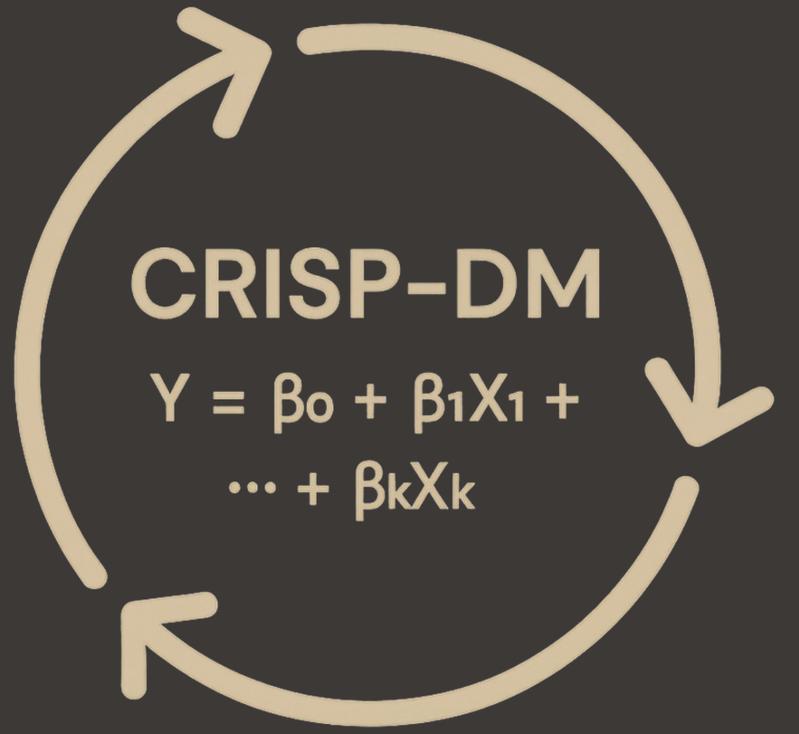
- **Modelo de Referencia (Baseline):** Es la línea base simple contra la que se comparan modelos más complejos como árboles de decisión, Random Forest, XGBoost y redes neuronales.
- **Predictor Fundamental:** Sirve como modelo supervisado básico para la predicción de valores continuos.
- **Interpretabilidad:** Una de sus ventajas es que facilita la interpretación de las relaciones causales o asociativas entre las variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_y + e$$

PIPELINE CRISP-DM Y LA REGRESIÓN MÚLTIPLE

El modelo se integra en varias fases de la metodología de Minería de Datos CRISP-DM (Cross-Industry Standard Process for Data Mining):

- **Preparación de Datos:** Se abordan problemas como la detección de multicolinealidad (evaluada con VIF), el manejo de outliers y el escalamiento (normalización/estandarización) de las variables.
- **Modelado:** Implica el ajuste de la regresión (Mínimos Cuadrados Ordinarios) y la selección adecuada de variables.
- **Evaluación:** Se utilizan métricas específicas para valorar la capacidad predictiva y la bondad del ajuste del modelo (R^2, MAE, MSE, RMSE, F-test).



VALIDACIÓN DE LA LRM

Para que los resultados de la RLM sean válidos y no sesgados, se deben verificar ciertos supuestos estadísticos:

LINEALIDAD

La relación entre las Xi y la Y debe ser lineal. Si no lo es, se pueden aplicar transformaciones de variables (logarítmicas, polinomiales).

NORMALIDAD DE RESIDUOS

Los errores ε deben seguir una distribución norma.

HOMOCEDASTICIDAD

La varianza de los residuos debe ser constante para todos los niveles de las variables predictoras (dispersión uniforme). La heterocedasticidad es un problema común.

VALIDACIÓN DE LA LRM

Para que los resultados de la RLM sean válidos y no sesgados, se deben verificar ciertos supuestos estadísticos:

INDEPENDENCIA DE ERRORES

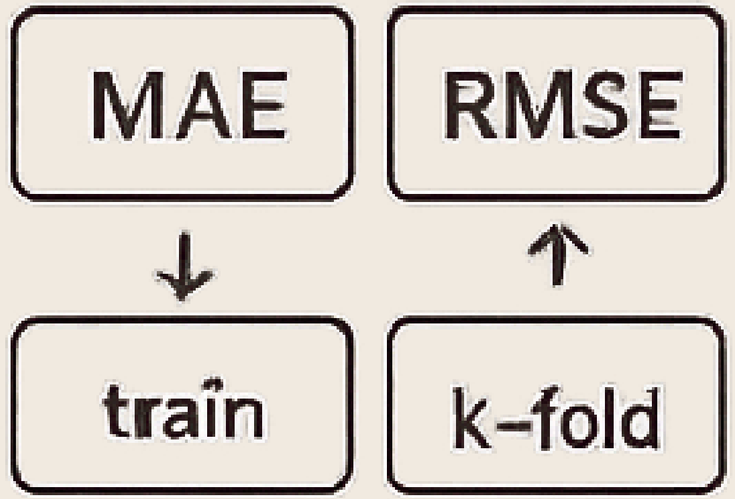
Los errores no deben estar correlacionados entre sí (importante en series de tiempo).

AUSENCIA DE MULTICOLINEALIDAD

Las variables independientes no deben estar fuertemente correlacionadas entre sí. Esto distorsiona la interpretación de los coeficientes.

EVALUACIÓN Y MÉTRICAS CLAVE

- **R^2:** Proporción de la varianza en Y que es explicada por el modelo. Un valor cercano a 1 indica un buen ajuste.
- **R^2 ajustado:** Similar a R^2, pero penaliza la adición de variables irrelevantes al modelo. Se prefiere al seleccionar modelos.
- **MAE (Error Medio Absoluto):** Mide la magnitud promedio de los errores en unidades de Y. Es robusto a outliers.
- **MSE / RMSE (Error Cuadrático Medio):** Penaliza los errores grandes elevándolos al cuadrado. El RMSE está en las mismas unidades que Y y es más interpretable.
- **F-test y p-values:** El F-test evalúa la calidad del modelo general (si es mejor que un modelo sin predictores). Los p-values de los coeficientes βi indican la relevancia estadística de cada predictor individual.



PROBLEMAS COMUNES Y SOLUCIONES

Multicolinealidad: Ocurre cuando dos o más Xi están altamente correlacionadas.

- Detección: Se utiliza el Factor de Inflación de Varianza (VIF). Un VIF > 5 o > 10 es preocupante.
- Solución: Eliminar la variable con alto VIF o utilizar técnicas de reducción de dimensionalidad como PCA (Análisis de Componentes Principales) o regularización.

Sobreajuste (Overfitting): El modelo ajusta muy bien los datos de entrenamiento pero falla con datos nuevos.

- Solución: Validación Cruzada (Cross-Validation) para estimar el error de generalización y reducir la cantidad de variables.

Outliers: Valores atípicos que pueden afectar significativamente el plano de regresión.

- Solución: Detección y tratamiento.

APLICACIONES

- Predicción de ventas por región.
- Estimar precios de casas
- Predicción de demanda de productos.
- Cálculo de riesgo crediticio.
- Estimar rendimiento escolar.
- Predicción de costos operativos.

EJERCICIO

Vamos a usar Regresión Lineal Múltiple para predecir el Precio (MDP) usando varias variables al mismo tiempo.

Objetivo: Predecir el precio de un terreno en millones de pesos usando:

- Terreno (m²)
- Frente (m)
- Profundidad (m)
- Índice_Desarrollo

Sistema de ecuaciones

$$Ecuación1: \sum Y = n(b_0) + b_1(\sum X_1) + b_2(\sum X_2)$$

$$Ecuación2: \sum X_1Y = b_0(\sum X_1) + b_1(\sum X_1^2) + b_2(\sum X_1X_2)$$

$$Ecuación3: \sum X_2Y = b_0(\sum X_2) + b_1(\sum X_1X_2) + b_2(\sum X_2^2)$$

EJERCICIO

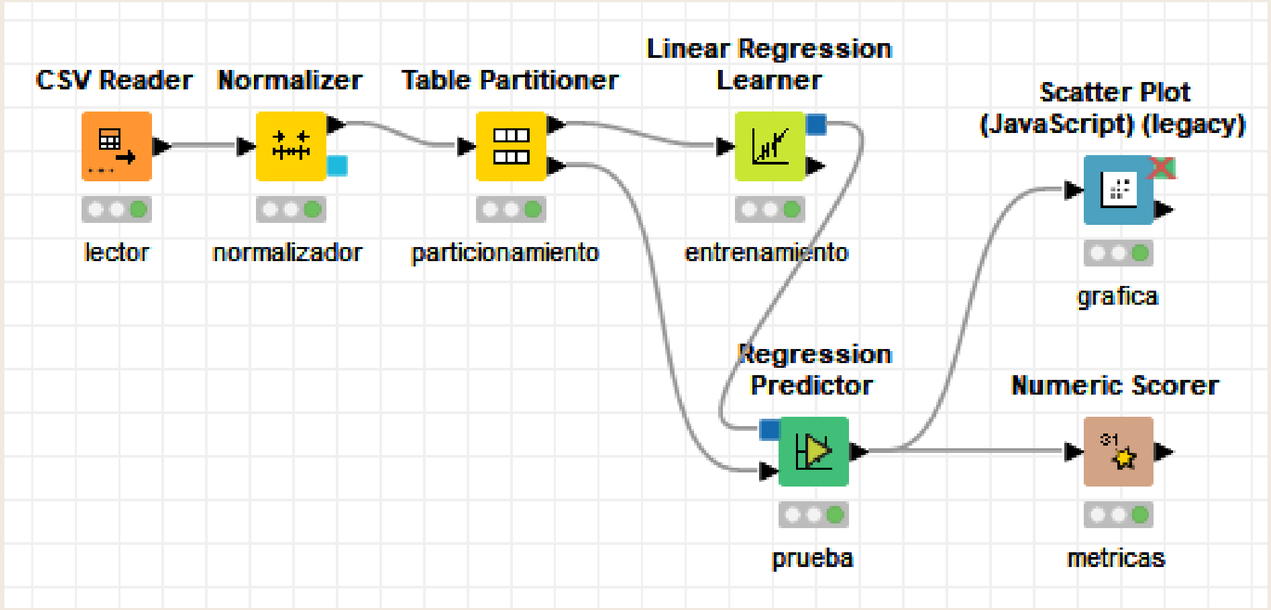
Terreno (m²)	Precio (MDP)	Frente (m)	Profundidad (m)	Índice_Desarrollo
440	1.01	12	36.67	2
616	1.42	14	44	3
381	0.88	11	34.64	1
963	2.21	18	53.5	3
431	0.99	12	35.92	2
255	0.59	10	25.5	1
594	1.37	14	42.43	2
625	1.44	15	41.67	3
708	1.63	16	44.25	3
468	1.08	12	39	2

CONCLUSIÓN

La regresión lineal multiple es una variante de la regresión lineal clasica con una sola variable pero mucho mas poderosa que predice con mayor precisión a diferencia de regresión lineal con una variable, y que dara las bases mas adelante para utilizar otro tipo de regresión como regresión logistica.

TAREA

Usar el siguiente diagrama de KNIME para hacer la regresión lineal multiple del dataset proporcionado a la variable “Profit”, poner metricas y describir la grafica final: https://raw.githubusercontent.com/krishnaik06/Multiple-Linear-Regression/master/50_Startups.csv



TAREA

Debera incluir:

- Portada con integrantes y número de equipo.
- Diccionario de las variable
- Diagrama del KNIME (normalizar con parametro Z, usar 70-30 para entrenamiento y prueba).
- Resultados de metricas obtenidas.
- Poner la grafica generada.
- Explicar brevemente las metricas y la grafica generada.

REFERENCIAS

- Montero Granados, R. (s.f.). Modelos de regresión lineal múltiple. Universidad de Granada. Recuperado el 25 de noviembre de 2025, de Autor desconocido. (s.f.). Regresión lineal. Universidad de Granada. Recuperado el 25 de noviembre de 2025 de Autor desconocido. (s.f.). Regresión lineal. Universidad de Granada. Recuperado el 20 de noviembre de 2025 de https://www.ugr.es/~montero/matematicas/regresion_lineal.pdf

GRACIAS POR SU ATENCIÓN