Ejercicio 1

Ejercicio: Predecir la Producción (Ton) con Regresión Lineal

Contexto

Una empresa agrícola registra mensualmente su superficie cultivada, el tipo de cultivo, la región, y variables económicas básicas. Buscamos estimar la producción en toneladas para un nuevo periodo en función de estas variables.

Datos

Archivo: ejercicio_1_datos_regresion_lineal_agricola

Hoja Datos: registros 2023–2024 por mes, región y cultivo.

Hoja Diccionario: descripción de campos.

Variable objetivo (y): `Producción_Ton`

Variables predictoras (X):

- Numéricas: `Año`, `Mes`, `Superficie_Ha`, `Precio_Ton`, `Costos_Insumos`
- Categóricas: `Región`, `Cultivo`

Importante: `Ingresos` y `Utilidad` derivan de `Producción_Ton`. No usarlas como predictoras (evitar fuga de información).

1) Preparación y revisión inicial

1. Descarga y abre el archivo Excel. Lee la hoja Diccionario para entender cada columna.

2. Importa el archivo en tu Notebook (pandas) y muestra las primeras filas (`head()`).

3. Revisa tipos de dato, valores nulos y rangos (`info()`, `describe()`).

Ejercicio 1

Preguntas guía

- ¿Hay valores faltantes en variables clave?
- ¿Los rangos de `Superficie_Ha` y `Precio_Ton` son coherentes?

2) Análisis exploratorio breve (EDA)

1. Calcula estadísticas por `Cultivo` y `Región` (promedio y desviación de `Producción_Ton` y `Superficie_Ha`).

2. Grafica (opcional) la distribución de `Producción_Ton` y relación básica con `Superficie_Ha`.

Preguntas guía

¿Qué cultivos/regiones presentan mayor producción promedio?

¿Se observa relación positiva entre `Superficie_Ha` y `Producción_Ton`?

3) Definición de variables del modelo

1. Define `y = Producción_Ton`.

2. Define `X` con estas columnas: `Año`, `Mes`, `Región`, `Cultivo`, `Superficie_Ha`, `Precio_Ton`, `Costos_Insumos`.

3. No incluyas `Ingresos` ni `Utilidad` como predictores.

Preguntas guía

- ¿Por qué `Ingresos`/`Utilidad` generarían fuga de información?

Ejercicio 1

4) División en entrenamiento y prueba

1. Divide en `train` (80%) y `test` (20%) con `random_state` fijo.

2. Verifica el tamaño de ambos conjuntos.

Preguntas guía

- ¿Por qué separamos train/test?
- ¿Qué riesgo aparece si evaluamos en los mismos datos de entrenamiento?

5) Preprocesamiento

1. Crea un `ColumnTransformer` que:

- Para numéricas (`Año`, `Mes`, `Superficie_Ha`, `Precio_Ton`, `Costos_Insumos`): imputación por mediana y escalado estándar.
- Para categóricas (`Región`, `Cultivo`): imputación por moda y One-Hot Encoding.

2. Integra el preprocesamiento en un `Pipeline` junto con el modelo.

Preguntas guía

- ¿Por qué conviene imputar y codificar dentro del `Pipeline`?

Ejercicio 1

6) Entrenamiento del modelo (Regresión Lineal)

1. Agrega al `Pipeline` el estimador `LinearRegression`.

2. Ajusta el modelo con `fit(X_train, y_train)` y genera predicciones `y_pred` en `X_test`.

Preguntas guía

- ¿Qué aprende la regresión lineal (intuición de coeficientes)?

7) Evaluación del desempeño

1. Calcula MAE, RMSE (√MSE) y R^2 en el conjunto de prueba.

2. Interpreta las métricas en el contexto del negocio (toneladas).

Preguntas guía

- ¿El R^2 obtenido sugiere un modelo útil?
- ¿El RMSE es razonable frente al rango típico de producción?

8) Gráfico “Predicho vs Real”

1. Grafica `y_test` (eje X) vs `y_pred` (eje Y) y añade la línea diagonal `y=x`.

2. Describe visualmente si los puntos se acercan a la diagonal.

Ejercicio 1

Preguntas guía

- ¿El modelo sobrestima valores bajos y subestima valores altos?
- ¿Ves patrones que sugieran no linealidad?

9) Diagnóstico de residuos (opcional)

1. Calcula residuos: `res = y_test - y_pred`.

2. Grafica residuos vs predicción.

3. Comenta si hay heterocedasticidad (abanico) o patrones.

10) Interpretación de coeficientes (opcional)

1. Extrae nombres de features tras el one-hot.

2. Muestra coeficientes ordenados por magnitud (positivos y negativos).

3. Comenta qué variables parecen más influyentes y si el signo es coherente.

11) Entregables

Notebook con:

- Carga y EDA breve.
- Pipeline (preprocesamiento + modelo).
- Métricas y gráfico Predicho vs Real.
- (Opcional) residuos, coeficientes, y validación cruzada.
- Un breve informe (5–8 líneas) interpretando resultados y sugerencias de mejora.