Ejercicio 2

Ejercicio: Clasificar Producción Alta vs Baja con Random Forest

Contexto

Una empresa agrícola registra mensualmente su superficie cultivada, el tipo de cultivo, la región y variables económicas básicas. Buscamos clasificar si la producción de un registro será Alta (1) o Baja (0).

Datos

Archivo: `1. datos_agricolas.xlsx` (hoja `Inicio`)

Variable objetivo (y): `Produccion_Alta` (binaria), definida como 1 si `Producción_Ton` > mediana del conjunto, 0 en caso contrario.

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.

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 2

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` (recuento por clase, medias y desvíos de variables numéricas).

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

Preguntas guía

- ¿Qué cultivos/regiones muestran mayores niveles de producción?
- ¿Se observa relación positiva entre `Superficie_Ha` y `Producción_Ton`?

3) Construcción de la variable objetivo binaria

1. Calcula la mediana de `Producción_Ton`.

2. Crea `Produccion_Alta = 1(Producción_Ton > mediana)`.

3. Revisa el balance de clases (`value_counts(normalize=True)`).

Preguntas guía

- ¿Las clases están balanceadas? Si no, ¿qué implicancias tiene para la evaluación?

Ejercicio 2

4) Definición de variables del modelo

1. Define `y = Produccion_Alta`.

2. Define `X` con: `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?

5) División en entrenamiento y prueba

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

2. Verifica tamaños y proporción de clases en ambos subconjuntos.

Preguntas guía

- ¿Por qué estratificar por la clase objetivo?
- ¿Qué riesgo aparece si evaluamos en los mismos datos de entrenamiento?

6) 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 (con `handle_unknown="ignore"`).

2. Integra el preprocesamiento en un `Pipeline`.

Ejercicio 2

Preguntas guía

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

7) Entrenamiento del modelo (Random Forest Classifier)

1. Agrega al `Pipeline` el estimador `RandomForestClassifier` (p. ej., `n_estimators=300`, `random_state=42`, `n_jobs=-1`).

2. Ajusta con `fit(X_train, y_train)` y obtén predicciones `y_pred` y probabilidades `y_proba`.

Preguntas guía

- ¿Qué ventajas ofrece Random Forest frente a un clasificador lineal en datos tabulares?

8) Evaluación del desempeño

1. Calcula accuracy y reporta precision, recall y F1-score (`classification_report`).

2. Construye y comenta la matriz de confusión.

Preguntas guía

- Si las clases están desbalanceadas, ¿es suficiente el accuracy?
- ¿Qué te dice la matriz de confusión sobre falsos positivos/negativos?

Ejercicio 2

9) Curva ROC y AUC (opcional)

1. Calcula ROC-AUC y grafica la curva ROC (con `RocCurveDisplay`).

2. Compara AUC con el valor de referencia 0.5.

Preguntas guía

- ¿La curva ROC y el AUC corroboran el desempeño observado en las métricas por clase?

10) Importancia de variables

1. Recupera los nombres de features finales tras el One-Hot.

2. Obtén `feature_importances_` del bosque y muestra el Top 10–15 en un gráfico de barras.

3. Interpreta qué variables aportan más a la predicción.

Preguntas guía

- ¿Qué variables parecen más influyentes? ¿Tiene sentido con el dominio?

11) Validación cruzada (opcional)

1. Aplica `StratifiedKFold(n_splits=5, shuffle=True, random_state=42)` y `cross_val_score` con `scoring="accuracy"` (o `f1`).

2. Reporta media ± desviación de la métrica.

Ejercicio 2

Preguntas guía

- ¿El desempeño es estable entre folds? ¿Qué variabilidad observas?