Práctica 4.4:

Reentrenamiento del Modelo Incorporando Preprocesamiento Avanzado

Contexto

En el Proyecto Final de Enero se desarrolló un sistema de predicción de fuga laboral utilizando el dataset **IBM HR Analytics Employee Attrition**, entrenando distintos modelos y evaluándolos en varios escenarios.

En ese momento aún no se había impartido el bloque de **Preprocesamiento de Datos**, por lo que muchas decisiones se tomaron con conocimiento parcial visto en algunas prácticas o por autoaprendizaje.

Ahora, tras haber estudiado el **Flujo Completo del Preprocesamiento (punto 4.4 del tema)** y las distintas técnicas disponibles en cada fase, se propone revisar el trabajo realizado y analizar cómo influye realmente el preprocesamiento en el rendimiento del modelo.

Objetivo

Evaluar experimentalmente cómo afecta el preprocesamiento al rendimiento del modelo en:

ESCENARIO 2: Control de costes del programa de retención

En este escenario:

- Se prioriza **PRECISION**
- Se busca reducir falsos positivos

Condiciones iniciales

- Se deberá fijar:
  - El **algoritmo seleccionado en el proyecto original para el Escenario 2**
  - Los **hiperparámetros óptimos ya obtenidos (en caso de haberlo hecho)**

En esta práctica **NO se modifican hiperparámetros**.
El foco está exclusivamente en el **impacto del preprocesamiento**.

Fases del experimento

Se deberán realizar tres entrenamientos comparativos:

1.Entrenamiento sin preprocesamiento avanzado

- Aplicar únicamente el preprocesamiento mínimo obligatorio y necesario para que el algoritmo funcione.
- Ejemplo: codificación imprescindible o escalado obligatorio para que funcione el modelo elegido.

No incluir:

- Ninguna Fase de Preprocesamiento.

Sólo registrar métricas, sin preprocesamiento extra.

2.Entrenamiento con el preprocesamiento original del proyecto

- Reproducir exactamente el pipeline usado en el proyecto inicial.
- No modificar decisiones.

Registrar métricas con el preprocesamiento original.

3.Nuevo diseño de preprocesamiento (basado en el punto 4.4)

Se deberá rediseñar el pipeline considerando **todas las fases del flujo de preprocesamiento visto en clase**.

Se deberá decidir:

- Qué fases incluir.
- Qué técnicas probar o descartar.
- Qué combinaciones comparar.

No se indica qué fases aplicar: deben considerarse todas las del punto 4.4 y decidir experimentalmente cuáles aportan valor.

Metodología obligatoria

Cada Fase del Preprocesamiento y con la técnica o estrategia que se incorpore deberá justificarse mediante pruebas comparativas:

- **CON técnica X**
- **SIN técnica X**
- o comparando **técnica X vs técnica Y vs Sin técnica**

Ejemplo correcto:

"Se prueba el modelo con y sin escalado. La precisión pasa de 0.78 a 0.83 con RobustScaler. Se mantiene el escalado."

Ejemplo incorrecto:

"Se incluye SMOTE porque mejora el modelo." (sin mostrar evidencia)

Métricas obligatorias

Dado que se trabaja en el Escenario 2, se deberá:

- **Priorizar:**
  - Precision
- **Mostrar también:**
  - Matriz de Confusión
  - Recall
  - F1-Score
  - Accuracy
  - AUC-ROC

La interpretación deberá centrarse en el impacto sobre la precisión y los falsos positivos.

Comparativa Final Obligatoria

Se deberá incluir una tabla comparativa con las 3 configuraciones y las métricas obtenidas

Análisis y Conclusiones

1. **Comparativa Final Obligatoria**

Se deberá incluir una tabla comparativa con las 3 configuraciones y las métricas obtenidas

2. **Preguntas a responder:**

Se deberá responder explícitamente a las siguientes cuestiones:

1. ¿Mejora realmente el modelo con un preprocesamiento más completo?
2. ¿El preprocesamiento original era adecuado?
3. ¿Qué técnicas nuevas aportaron mayor mejora?
4. ¿Hubo técnicas que empeoraron el rendimiento?
5. ¿Qué decisiones cambiarías respecto al proyecto original?

Las respuestas deberán estar basadas en resultados experimentales.

Entregables

- Notebook estructurado y organizado
- Implementación correcta mediante pipeline
- Tablas comparativas
- Conclusiones argumentadas