**Fases del Proyecto**

El pipeline utilizado en este modelo sigue una secuencia de pasos que son comunes en el preprocesamiento y modelado de datos para clasificación. A continuación, se explica cada componente del pipeline y cómo se conecta con el flujo general de la solución:

**1. Preprocesamiento de los Datos:**

* StandardScaler: Esta etapa se encarga de normalizar las características numéricas para que tengan una media de 0 y una desviación estándar de 1. Esto es importante cuando se utilizan algoritmos sensibles a la escala de los datos, como los árboles de decisión o modelos lineales.
* OneHotEncoder: Codifica las características categóricas en formato binario (0 o 1), de manera que puedan ser utilizadas por el modelo. Este paso es esencial para convertir variables categóricas en un formato compatible con los algoritmos de aprendizaje automático.

**2. Modelo:**

* DecisionTreeClassifier: Este es el modelo principal utilizado para la clasificación. En este caso, se utilizó un árbol de decisión con un criterio de "entropía" y una profundidad máxima (max\_depth) de 5. Los árboles de decisión son intuitivos y ofrecen interpretabilidad, al dividir los datos en función de características relevantes hasta llegar a decisiones finales.

**3. Entrenamiento y Validación:**

* Durante el proceso de entrenamiento, el pipeline aplica las transformaciones (escalado y codificación) a los datos y luego entrena el modelo de clasificación. Este flujo integrado asegura que las mismas transformaciones se apliquen a los datos de prueba y a cualquier dato nuevo, manteniendo la coherencia del proceso.

**4. Evaluación del Modelo:**

* Matriz de confusión: Se observa que el modelo tiene un alto desempeño en la clase mayoritaria (clase 0, que representa "no deserción") con 49,451 predicciones correctas, pero es menos preciso al clasificar la clase minoritaria (clase 1, "deserción"), con 3,835 falsos negativos.
* Reporte de Clasificación: El modelo alcanza una precisión global del 91%, con un desempeño equilibrado en términos de recall y f1-score para la clase minoritaria.
* ROC AUC: El área bajo la curva ROC de 0.92 indica un buen rendimiento global del modelo en distinguir entre clases.

**5. Importancia de Características:**

* El modelo identificó como más importantes características como CICLO\_RELATIVO, EDAD, EDAD\_INICIO\_UNI, y las calificaciones en el CEPRE. Estas variables influyen fuertemente en la predicción de deserción, indicando que factores como el ciclo de estudios y la edad relativa son claves en este contexto.

**Comparación con otras metodologías:**

* Si bien se podría haber optado por modelos más complejos como Random Forests o modelos de ensamble, se eligió un árbol de decisión por su capacidad de interpretabilidad. Aunque estos modelos más avanzados podrían mejorar ligeramente el rendimiento, el árbol de decisión brinda una ventaja en la transparencia y facilidad para explicar las decisiones.

**Decisión Final:**

* La elección de este pipeline fue motivada por un equilibrio entre la precisión del modelo, su interpretabilidad y la simplicidad para implementar un flujo reproducible y escalable. Las métricas obtenidas validan que el modelo es efectivo en la tarea, aunque se pueden explorar ajustes adicionales para mejorar la detección de la clase minoritaria.

**Resumen Explicativo del Modelo Utilizado**

**1. Modelo Utilizado**

Se implementó un modelo de **árbol de decisión** con el criterio de entropía y una profundidad máxima de 5 niveles. Este enfoque fue seleccionado por su capacidad de interpretar fácilmente las reglas de clasificación, lo que permite visualizar cómo se toman decisiones en función de las características más importantes. Además, se aplicó un pipeline con técnicas de preprocesamiento como escalado estándar para variables numéricas y codificación one-hot para variables categóricas.

**2. Lógica de la Decisión**

El modelo prioriza la importancia de ciertas características para predecir la deserción estudiantil. Según los resultados, las variables más relevantes fueron:

* **CICLO\_RELATIVO** (38.1%)
* **EDAD** (16.1%)
* **EDAD\_INICIO\_UNI** (12.4%)
* **Calificaciones en CEPRE (mínima, máxima y promedio)** y **domicilio por macroregión**

Estas variables dominan el proceso de clasificación, sugiriendo que el momento en que un estudiante inicia o avanza en su carrera, su edad y su desempeño académico previo son determinantes para predecir si persistirá o desertará.

**3. Comparativa con Otras Metodologías**

El árbol de decisión fue comparado con otros modelos como regresión logística y SVM, pero se optó por el árbol debido a su balance entre interpretabilidad y precisión. Modelos más complejos como los bosques aleatorios podrían ofrecer mayor precisión, pero a costa de la interpretabilidad.

**4. Resultados y Métricas de Evaluación**

El desempeño del modelo fue evaluado con las siguientes métricas:

* **Matriz de Confusión:**
  + Verdaderos Negativos: 49,451
  + Falsos Positivos: 3,835
  + Falsos Negativos: 1,327
  + Verdaderos Positivos: 3,104
* **Reporte de Clasificación:**
  + Precisión (Clase 0): 97% (baja deserción)
  + Precisión (Clase 1): 45% (alta deserción)
  + Recall (Clase 1): 70% (capacidad de identificar desertores)
  + F1-Score (Clase 1): 55%
* **Curva ROC (Área bajo la curva):** 0.92

El modelo muestra un buen desempeño general (91% de precisión), pero su precisión para la clase 1 (desertores) es moderada. El recall elevado para la clase 1 sugiere que el modelo identifica a la mayoría de los desertores, aunque con algunos falsos positivos.

**5. Decisión Final e Interpretación**

Aunque la precisión en la clase 1 es moderada, el objetivo principal es minimizar la deserción, por lo que se prioriza un alto recall para detectar posibles desertores. El modelo logra este objetivo al identificar correctamente el 70% de los casos de deserción, lo que es crucial para diseñar intervenciones preventivas. La importancia de las variables sugiere que el análisis de los ciclos y la edad pueden ser los principales indicadores de riesgo de deserción, permitiendo focalizar recursos en los estudiantes más vulnerables.

En resumen, el modelo es útil y proporciona insights clave para la toma de decisiones estratégicas, con margen de mejora en la reducción de falsos positivos si se implementan modelos más avanzados o combinaciones de técnicas de ensamble.