Informe: Trabajo Árbol de Decisión - Grupal

1. Introducción.

El presente proyecto tiene como objetivo *predecir el posible abandono o continuidad académica* de los estudiantes en una carrera, empleando técnicas de minería de datos e inteligencia artificial.

Para ello, se desarrolló un modelo basado en *Árboles de Decisión*, utilizando métricas de *entropía y ganancia de información* para determinar los atributos más influyentes en el rendimiento estudiantil.

2. Descripción del Proceso.

El proceso se compone de seis etapas principales:

2.1 Carga y preparación de datos

Se importó la hoja de cálculo "*TablaPrediccionAbandono-DatosFinal.xlsx*" mediante la librería pandas la cual contenía los datos para entrenar el modelo. Se estandarizó los nombres de las columnas y se añadió una nueva columna denominada **"EstadoFinal"**, encargada de almacenar la predicción final del modelo (*Continua o Abandona*).

2.2 Criterio inicial y cálculo de entropía

Para generar etiquetas de ejemplo, se estableció un criterio simple:

Si el promedio del primer cuatrimestre es mayor o igual a 7 → *Continua*.
En caso contrario → *Abandona*.

Con estas etiquetas, se calcularon:

- El número total de positivos (Continua)
- El número total de negativos (Abandona)
- La **entropía total** del conjunto de datos mediante la función externa "*funciones_entropia"*.

Este valor indica el ***grado de desorden*** o ***incertidumbre*** presente antes de aplicar cualquier división.

2.3 Evaluación de atributos y ganancia de información

Cada columna fue evaluada como posible "nodo raíz" del árbol.
Para ello, el código realiza un bucle que:

1. Detecta si el atributo es numérico o categórico.
2. Si es numérico, lo divide en tres rangos (cuartiles) mediante ***pd.qcut***.
3. Si es categórico, usa directamente sus categorías.
4. Para cada grupo, obtiene la cantidad de estudiantes que **continúan** y los que **abandonan**.
5. Con estos valores, calcula la **ganancia de información** utilizando la función externa *funciones_ganancia*.

La ganancia mide cuánto *reduce la entropía* ese atributo al ser usado para dividir el conjunto. Mientras que el atributo con mayor ganancia fue identificado como el más relevante para comenzar el árbol.

2.4 Entrenamiento del modelo

Se empleó el algoritmo **DecisionTreeClassifier** de scikit-learn con el criterio "entropy". Las variables categóricas fueron codificadas con LabelEncoder para convertirlas en valores numéricos.
Posteriormente, el modelo fue entrenado utilizando todos los atributos disponibles.

2.5 Predicción y exportación de resultados

Una vez entrenado, el modelo predijo el valor de "*EstadoFinal*" para cada registro del conjunto de datos. Las predicciones fueron revertidas a sus etiquetas originales (Continua / Abandona) y exportadas a un nuevo archivo: ***"TablaPrediccionAbandono-Final.xlsx"***

3. Resultados Obtenidos:

- Se imprimió en consola la **entropía inicial** del conjunto.
- Se listaron las **ganancias de información** por atributo, mostrando para cada uno la distribución de estudiantes según su estado final.

- Se identificó el **atributo con mayor capacidad predictiva**, base para construir el árbol.
- El modelo generó correctamente la columna "EstadoFinal" con los valores estimados para cada estudiante.

4. Conclusión:

El desarrollo de este modelo de predicción permitió comprender de manera estructurada, los factores que influyen en la permanencia o abandono de los estudiantes dentro de una carrera. A través del uso de *entropía* y *ganancia de información* se logró identificar cuáles atributos aportan mayor relevancia en la reducción de la incertidumbre del sistema, es decir, cuáles variables son más determinantes al momento de predecir el desempeño académico. Este análisis no solo brinda un enfoque estadístico, sino también interpretativo, ya que permite visualizar el comportamiento de los estudiantes en función de sus características individuales.

La implementación del árbol de decisión, respaldada por la librería en Python *scikit-learn*, facilitó la creación de un modelo capaz de generalizar patrones y emitir predicciones concretas sobre el estado final de cada estudiante. Al mismo tiempo, el uso de funciones externas para el cálculo de la *entropía* y la *ganancia de información* aportó claridad conceptual y una estructura eficiente al código, facilitando su entendimiento.

El resultado final —la generación automática de un archivo con la predicción del estado académico— representa una herramienta práctica que puede emplearse tanto para la toma de decisiones institucionales como para la detección temprana de casos de riesgo. En suma, el trabajo demuestra cómo la combinación entre métodos estadísticos, programación y aprendizaje automático puede transformarse en una herramienta para anticipar comportamientos y optimizar estrategias.

Trabajo hecho por: Gil Lascano Lorenzo, Bayaslian Santiago, Núñez Mauro, Buchholz Ariel.