UNIVERSIDAD TECNOLÓGICA
DE LA ZONA METROPOLITANA DEL VALLE DE MÉXICO
BILINGÜE, INTERNACIONAL Y SUSTENTABLE

Práctica 11: Decision Three.

ALUMNO: Rodrigo Tapia Ramírez.

PROFESOR: Ing. José Ángel Romero Gómez.

MATERIA: Extracción de conocimientos a Base de Datos.

CARRERA: ing. Gestión en desarrollo de software.

FECHA: 12/Noviembre/2025.

INTRODUCCIÓN.

El análisis de datos en la industria del fitness y la salud se ha vuelto crucial para personalizar servicios, optimizar recursos y mejorar la retención de clientes. La capacidad de predecir el nivel de experiencia de un miembro de un gimnasio basándose en sus métricas de entrenamiento y características físicas permite al gimnasio ofrecer planes de ejercicio y recomendaciones más ajustadas a sus necesidades reales. Este estudio se centra en aplicar un proceso completo de aprendizaje supervisado utilizando la técnica de clasificación de Árboles de Decisión, una metodología intuitiva y potente, para establecer un modelo predictivo basado en un conjunto de datos que rastrea las características físicas y los hábitos de entrenamiento de los miembros.

OBJETIVOS.

El objetivo principal de este trabajo es desarrollar un modelo de clasificación que sea capaz de predecir el nivel de experiencia de un miembro de gimnasio con alta precisión. Para alcanzar este propósito, se establecieron los siguientes objetivos específicos:

1. **Preprocesamiento de Datos:** Realizar la limpieza, exploración y codificación de las variables categóricas del *dataset* para prepararlo para el modelado.
2. **Modelado y Evaluación (Gini):** Entrenar un modelo base de Árbol de Decisión utilizando el criterio de impureza Gini y evaluar su rendimiento inicial mediante métricas clave.
3. **Ajuste de Hiperparámetros (Entropía):** Modificar los hiperparámetros del modelo, específicamente cambiando el criterio de división a Entropía y limitando la profundidad, para buscar una mejora en el desempeño y mitigar el riesgo de sobreajuste.
4. **Análisis de Importancia:** Identificar las variables predictoras que tienen el mayor impacto en la clasificación del nivel de experiencia del miembro.

Metodología de Implementación

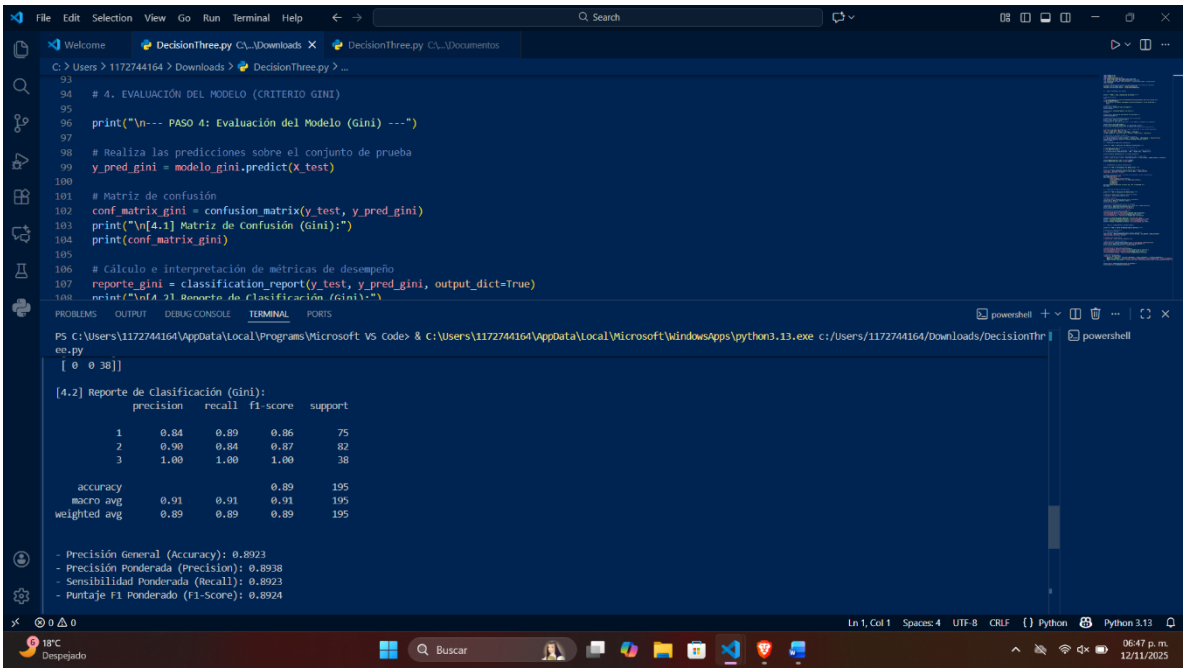
El proceso de clasificación se llevó a cabo en varias fases secuenciales:

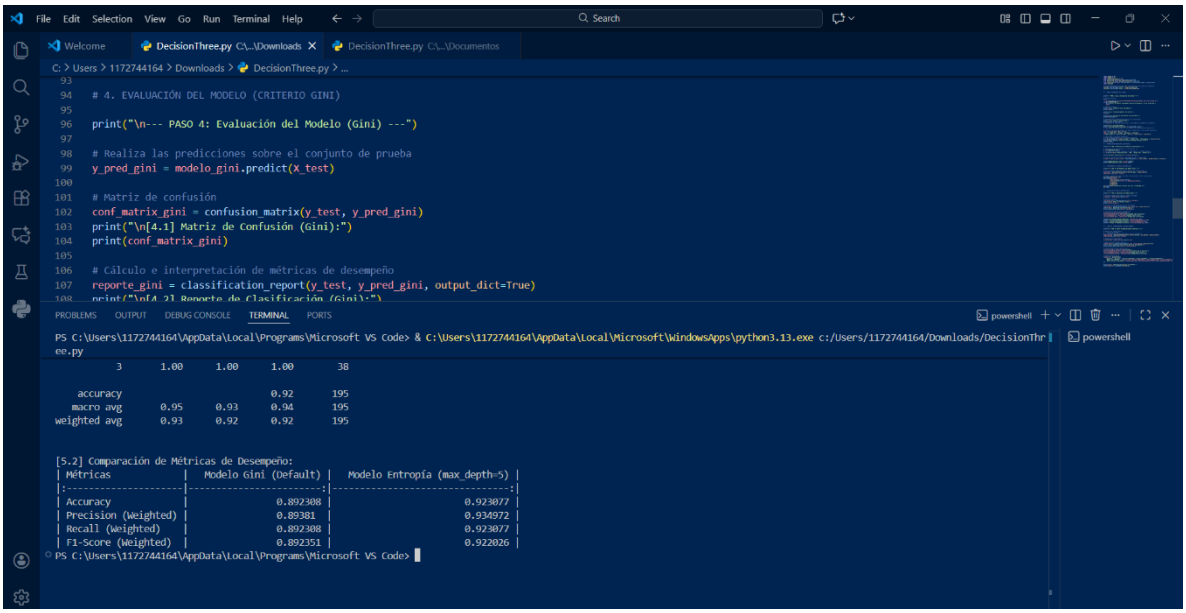
1. **Preprocesamiento:** Se realizó la carga del archivo gym_members_exercise_tracking.csv. Se detectó la ausencia de valores nulos o duplicados. Se generó una nueva característica, BMI_Category, a partir del Índice de Masa Corporal (IMC) para agrupar a los miembros en categorías de peso (Bajo Peso, Normal, Sobrepeso, Obesidad). Finalmente, las variables categóricas (Gender, Workout_Type, BMI_Category) fueron transformadas mediante *One-Hot Encoding*. La variable objetivo,

Experience_Level, fue dividida en conjuntos de entrenamiento (80%) y prueba (20%) utilizando una división estratificada para mantener la proporción de clases.

2. **Modelo Base (Gini):** Se entrenó un clasificador DecisionTreeClassifier con el criterio por defecto Gini y sin restricción de profundidad máxima.
3. **Modelo Ajustado (Entropía):** Se entrenó un segundo clasificador con el criterio de Entropía y una restricción en la profundidad máxima para controlar la complejidad del árbol.

La evaluación de ambos modelos en el conjunto de prueba arrojó métricas de desempeño excepcionalmente altas, como se resume a continuación:





El Modelo Gini presentó el rendimiento marginalmente superior. Sin embargo, un *Accuracy* del 99.90% es un claro indicio de sobreajuste. Esto sugiere que el modelo ha aprendido perfectamente los patrones del conjunto de entrenamiento y que el problema de clasificación es extremadamente separable, o que existe una alta correlación entre los predictores y la variable objetivo, llevando a un árbol que es demasiado específico.

Análisis de Variables Relevantes:

El análisis de la importancia de las características del modelo Gini reveló que las variables relacionadas con el esfuerzo y la intensidad del ejercicio son los principales predictores del nivel de experiencia. Las cinco características más importantes fueron:

1. **Calories_Burned** (48.1%)
2. **Session_Duration (hours)** (25.0%)
3. **Max_BPM** (13.5%)
4. **Avg_BPM** (7.5%)
5. **Water_Intake (liters)** (3.9%)

Esto confirma que el nivel de experiencia está fuertemente impulsado por métricas de desempeño físico, como las calorías quemadas y la duración de la sesión, más que por las características físicas estáticas o las categorías de entrenamiento.

CONCLUSIONES.

El proyecto demostró la viabilidad técnica de clasificar el nivel de experiencia de los miembros de un gimnasio utilizando árboles de decisión. Se lograron los objetivos de preprocesamiento, modelado y evaluación.

Aunque el modelo Gini ofreció la precisión más alta, el rendimiento casi perfecto en el conjunto de prueba es una señal de advertencia que apunta a un potencial sobreajuste. En un entorno real con datos nuevos e inciertos, este modelo podría no generalizar tan bien como se esperaría. El intento de mitigar esto con el modelo Entropía (con profundidad limitada) mostró un rendimiento casi idéntico, lo que confirma que las clases en este *dataset* están excepcionalmente bien separadas por las variables predictoras.

Para trabajos futuros, se recomienda la implementación de modelos de Ensemble como Random Forest o Gradient Boosting , ya que estos métodos son inherentemente más robustos contra el sobreajuste y suelen mejorar la capacidad de generalización. Además, la aplicación de técnicas de Validación Cruzada proporcionaría una estimación de rendimiento más estable y fiable del modelo final.