Universidad Nacional de Ingeniería
Facultad de Ciencias
Escuela Profesional de Ciencia de la Computación

Examen Parcial

Parte Práctica

Minería de Datos

CC442 **06/02/2026** **Ciclo: 2025-III**

Puntaje: **12 puntos** . **Duración**. 100 minutos

PARTE I (2 puntos)

Dado el siguiente conjunto de datos para decidir si comprar un producto:

i	Precio	Calidad	Comprar
1	Alto	Alta	Sí
2	Alto	Baja	No
3	Medio	Alta	Sí
4	Bajo	Alta	Sí
5	Bajo	Baja	No
6	Alto	Alta	No
7	Medio	Baja	No
8	Bajo	Baja	No

Se pide:

a) Calcula la entropía inicial del conjunto completo $H(Y)$. **(0.3 puntos)**

b) Calcula la ganancia de información (Information Gain) para ambos atributos: Precio y Calidad. **(0.9 puntos)**

c) ¿Qué atributo elegirías como raíz del árbol y por qué? **(0.2 puntos)**

d) Construye el árbol de decisión completo usando el algoritmo ID3. Muestra todos los nodos internos y hojas. **(0.6 puntos)**

PARTE II (5 puntos)

CASO: Clasificación de Propietarios de Cortacéspedes

Dataset: ***RidingMowers.csv***

Contexto

Una empresa que fabrica cortacéspedes autopropulsados (Riding Mowers) desea identificar a los mejores prospectos de ventas para una campaña de marketing intensiva. En particular, el fabricante está interesado en clasificar los hogares como **propietarios (Owners)** o **no propietarios (Nonowners)** basándose en dos variables predictoras:

1. **Ingresos (Income):** Expresado en miles de dólares ($1000s).
2. **Tamaño del lote (Lot Size):** Expresado en miles de pies cuadrados (1000 ft^2).

El experto en marketing seleccionó una muestra aleatoria de 24 hogares, cuyos datos se encuentran en el archivo *RidingMowers.csv*.

Tareas a realizar

Utilice todos los datos para ajustar un modelo de **regresión logística** donde la variable dependiente es la propiedad (Ownership) y los predictores son el Ingreso y el Tamaño del Lote. Responda lo siguiente:

a. ¿Qué porcentaje de los hogares del estudio eran propietarios de un cortacésped?

b. Cree un gráfico de dispersión de *Ingresos* vs. *Tamaño del Lote* utilizando color o símbolos para distinguir a los propietarios de los no propietarios. A partir del gráfico, ¿qué clase parece tener un ingreso promedio más alto?

c. Entre los que realmente son "no propietarios", ¿cuál es el porcentaje de hogares clasificados correctamente por el modelo (usando un punto de corte de 0.5)?

d. Si el objetivo es aumentar el porcentaje de "no propietarios" clasificados correctamente, ¿el umbral de probabilidad (cutoff) debería aumentarse o disminuirse?

e. Calcule los *odds* (posibilidades) de que un hogar con un ingreso de $60,000 y un tamaño de lote de 20,000 pies cuadrados sea propietario.

f. ¿Cuál es la clasificación (propietario o no propietario) de un hogar con un ingreso de $60,000 y un tamaño de lote de 20,000 pies cuadrados, utilizando un punto de corte (cutoff) de 0.5?

g. ¿Cuál es el ingreso mínimo que debería tener un hogar con un tamaño de lote de 16,000 pies cuadrados para ser clasificado como propietario?

PARTE III (**5 puntos**)

CASO: Predicción de vuelos retrasados (Boosting)

Dataset: *FlightDelays.csv*

Contexto

El archivo *FlightDelays.csv* contiene información sobre todos los vuelos comerciales que partieron del área de Washington, D.C. (aeropuertos IAD, DCA, BWI) con destino a Nueva York (LGA, JFK, EWR) durante el mes de enero de 2004. Para cada vuelo, se dispone de información sobre los

aeropuertos de origen y destino, la distancia de la ruta, el día de la semana, la aerolínea y la hora programada de salida.

La variable objetivo a predecir es **Flight Status** (Estado del vuelo), la cual indica si un vuelo se retrasó o no. Para este estudio, un retraso se define como una llegada que ocurre al menos 15 minutos tarde respecto a lo programado.

Tareas de Preprocesamiento de Datos

Antes de entrenar el modelo, realice las siguientes transformaciones:

1. **Categorización:** Transforme la variable que contiene el día de la semana (day_week) en una variable de tipo categórico.
2. **Agrupación (Binning):** Divida la hora de salida programada (scheduled departure time) en ocho intervalos o "bins" (en Python, utilice la función pd.cut() del paquete pandas).
3. **Partición de Datos:** Divida el conjunto de datos en un set de **entrenamiento (60%)** y un set de **validación (40%)**.

Modelado

Ajuste un modelo de **Árbol de Clasificación con Boosting** (AdaBoost) para predecir los retrasos.

Utilice n_estimators = 500.

Establezca random_state = 1.

Mantenga la configuración por defecto para el clasificador base (DecisionTreeClassifier) y para el AdaBoostClassifier.

Preguntas a resolver

a. En comparación con un árbol de decisión único, ¿cómo se comporta el árbol con Boosting en términos de la **precisión (accuracy) global**?

b. En comparación con un árbol de decisión único, ¿cómo se comporta el árbol con Boosting en términos de su capacidad para **identificar específicamente los vuelos retrasados**?

c. Explique detalladamente por qué este modelo (Boosting) podría presentar un mejor rendimiento en comparación con otros modelos de clasificación que haya ajustado previamente para este problema.