Taller 1: Análisis Predictivo y Aprendizaje Estadístico

Caso de Estudio: Advertising Data

Ingeniería Industrial - Sistemas Avanzados de Producción

Febrero 2026

1. Contexto y Motivación

En el ámbito del análisis de negocios y la toma de decisiones basada en datos, surge una pregunta recurrente: *¿Cuál es el impacto real de la inversión publicitaria en las ventas?*. El conjunto de datos **Advertising** nos permite explorar esta relación a través de registros de ventas en 200 mercados diferentes, acompañados de los presupuestos destinados a tres canales de comunicación distintos.

Este taller tiene como propósito desarrollar habilidades para identificar si existe una asociación significativa entre la publicidad y el éxito comercial, cuantificar la magnitud de dicha relación y evaluar la capacidad predictiva de distintos modelos estadísticos.

2. Descripción de las Variables

Para este estudio, se han definido las siguientes variables, expresadas en unidades específicas para facilitar el modelamiento:

- **TV**: Inversión en publicidad por televisión (expresada en miles de dólares). Se considera el predictor principal en modelos simples.
- **Radio**: Inversión en publicidad por radio (expresada en miles de dólares).
- **Newspaper**: Inversión en publicidad por periódicos impresos (expresada en miles de dólares).
- **Sales (Ventas)**: Volumen de ventas del producto (expresado en miles de unidades). Esta actúa como nuestra variable respuesta o dependiente (Y).

3. Estructura de Evaluación

El taller se divide en tres fases fundamentales que cubren el ciclo de vida de un proyecto de analítica:

3.1. Fase 1: Estadística Descriptiva y Análisis Exploratorio

Esta fase constituye la .$^{\text{et}}$apa detectivesca"del análisis estadístico. El objetivo primordial es buscar indicios o "pistas"sobre el patrón de los datos antes de proceder a cualquier modelamiento formal. Para el desarrollo de esta sección, utilice como guía el notebook `Semana3.ipynb` disponible en el repositorio: GitHub: SAPP-IIND-2026.

Realice las siguientes tareas de manera secuencial:

1. **Carga y Preparación de Datos:** Cargue el archivo `Advertising.csv` directamente desde la URL oficial. Verifique la estructura de las variables y confirme que no existen datos faltantes que puedan sesgar el análisis.

2. **Caracterización Numérica:** Elabore una tabla que resuma las estadísticas descriptivas para las variables *TV*, *Radio*, *Newspaper* y *Sales*. Esta tabla debe reportar:

   - Media y Mediana (Medidas de tendencia central).
   - Desviación estándar, Mínimo y Máximo (Medidas de dispersión).
   - Sesgo (*Skewness*) y Curtosis (Medidas de forma).

3. **Análisis de Distribución y Atípicos:** Construya un **histograma** y un **diagrama de caja (Box Plot)** para cada variable.

   - Determine la simetría de las distribuciones.
   - Responda: ¿Se identifican datos atípicos o observaciones que se alejan significativamente de la masa de los datos? Explique cómo podrían influir estas observaciones en la trayectoria de una línea de regresión.

4. **Exploración de la Nube de Puntos:** Genere los dispersogramas (*Scatter Plots*) de la variable respuesta (*Sales*) frente a cada covariable. Identifique visualmente si la nube de puntos sugiere una relación lineal simple o si existen patrones más complejos (parábolas, segmentos o datos influyentes).

5. **Evaluación de Asociación:** Calcule y presente las matrices de correlación de **Pearson** y **Spearman**.

   - El coeficiente de Pearson evaluará la fuerza de la asociación lineal.
   - El coeficiente de Spearman permitirá identificar relaciones monótonas basadas en el rango de los datos.

6. **Interpretación de Resultados:** A partir de las métricas obtenidas en el punto anterior, responda:

   - ¿Qué significa un coeficiente cercano a 1 o -1 en el contexto de la inversión publicitaria y el retorno en ventas?
   - ¿Cómo se interpreta un valor de correlación cercano a 0?
   - Compare ambos coeficientes (Pearson y Spearman). ¿Sugieren estos resultados que las relaciones son estrictamente lineales? Justifique su respuesta basándose en la forma de la nube de puntos observada.

3.2. Fase 2: Regresión Lineal y Diagnóstico

En esta fase se evalúa la construcción y validación de modelos donde se admite que los factores que influyen en la variable respuesta se dividen en un grupo explicativo y un grupo de error o perturbación. El modelo estructural viene dado por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Utilizando el dataset `Advertising` y el entorno de Python, realice las siguientes tareas:

1. **Modelamiento Múltiple:** Estime un modelo de regresión lineal múltiple donde las ventas (*Sales*) dependan de la inversión en *TV*, *Radio* y *Newspaper*. Reporte los valores numéricos del intercepto ($\hat{\beta}_0$) y los coeficientes de pendiente ($\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$)).

2. **Interpretación de Parámetros:** Explique con sus propias palabras el significado de los resultados obtenidos:

   - Interprete el intercepto $\hat{\beta}_0$ indicando si tiene sentido conceptual que sea diferente de cero.
   - Interprete los coeficientes $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_3$ bajo el concepto de variación de la media de Y cuando hay un incremento unitario de la covariable, manteniendo la otra constante (*ceteris paribus*).

3. **Análisis de Bondad de Ajuste:** Calcule el coeficiente de determinación R^2. Indique qué proporción de la variabilidad total de las ventas es explicada por el modelo y analice si la inclusión de la variable *Radio* y *Newspaper* mejoró significativamente el ajuste respecto al modelo simple de la Fase 1.

4. **Estimación Matricial:** Considere el siguiente conjunto de 5 observaciones para una regresión lineal simple. Utilizando la formulación matricial $\hat{\beta} = (X'X)^{-1}X'Y$, encuentre los estimadores $\hat{\beta}_0$ (intercepto) y $\hat{\beta}_1$ (pendiente).

   Dadas las matrices de diseño X y el vector de respuesta Y:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}, \quad Y = \begin{pmatrix} 2 \\ 4 \\ 6 \\ 7 \\ 9 \end{pmatrix}$$

   Tarea:

   - Calcule la matriz transpuesta X'.
   - Calcule el producto $X'X$ y su respectiva matriz inversa $(X'X)^{-1}$.
   - Calcule el producto $X'Y$.
   - Obtenga el vector de parámetros $\hat{\beta}$ y escriba la ecuación de la recta estimada $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

3.3. Fase 3: Árboles de Decisión y Comparación de Modelos

En esta fase final, se explorarán métodos no paramétricos para capturar relaciones no lineales y posibles efectos de interacción (sinergia) entre los medios publicitarios. A diferencia de la regresión lineal, que asume una estructura funcional predefinida, los árboles de decisión segmentan el espacio de los predictores para identificar patrones complejos.

Para el desarrollo de esta sección, utilice el notebook `Semana4.ipynb` disponible en el repositorio oficial: GitHub: SAPP-IIND-2026.

Realice las siguientes actividades:

1. **Entrenamiento del Modelo:** Utilizando las variables *TV* y *Radio*, entrene un árbol de regresión para predecir las ventas (*Sales*). Grafique la estructura del árbol y explique bajo qué criterio se realizan las particiones en los nodos.

2. **Cálculo Manual de Incertidumbre (Clasificación):** Suponga una versión simplificada del problema donde las ventas se categorizan en "Altas" y "Bajas" Con los siguientes datos de 4 mercados, calcule la reducción de la incertidumbre:

Mercado	Inversión TV	Ventas (Clase)
1	Alta	Altas
2	Alta	Altas
3	Baja	Bajas
4	Baja	Altas

   Tareas:

   - Calcule la **Entropía del Nodo Padre**: $H(S) = -\sum p_i \log_2(p_i)$.
   - Calcule el **Índice de Gini** para el nodo padre: $Gini = 1 - \sum p_i^2$.
   - Determine la **Ganancia de Información** tras realizar una partición por la variable *Inversión TV*.

3. **Importancia de Predictores:** Identifique qué variable tiene mayor peso en el árbol de decisión y compárela con los coeficientes obtenidos en la regresión múltiple de la Fase 2.

4. **Diagnóstico Comparativo:** Discuta en qué escenarios sería preferible utilizar un modelo de regresión lineal frente a un árbol de decisión, considerando la interpretabilidad y la precisión del pronóstico.

4. Instrucciones de Entrega y Rúbrica

4.1. Especificaciones del Entregable

El trabajo debe ser entregado en un único archivo en formato **PDF**, el cual debe ser cargado en el repositorio de GitHub asignado a cada grupo. El documento debe seguir una estructura profesional que incluya:

- **Introducción:** Breve descripción del problema y los objetivos del análisis.
- **Desarrollo Técnico:** Evidencia de los cálculos, gráficas y resultados obtenidos en las Fases 1, 2 y 3.
- **Análisis Crítico:** El núcleo del informe debe ser la interpretación de cada resultado. No se aceptarán capturas de código sin una explicación detallada de lo que los números representan para el negocio.
- **Conclusiones:** Reflexión sobre qué modelo (Regresión vs. Árboles) es más robusto para este caso particular.

Criterio	Peso	Descripción
Correctitud Técnica	30 %	Ejecución correcta de los modelos en Python, cálculos manuales (escalares y matriciales) precisos y gráficas bien etiquetadas.
Interpretación Estadística	40 %	Explicación profunda de los coeficientes β_j, el significado del R^2, y el análisis de pureza (Gini/Entropía). Se evalúa el uso correcto de conceptos como Entropía y ganancia de información.
Análisis de Diagnóstico	20 %	Capacidad para detectar datos atípicos, evaluar la linealidad y discutir las limitaciones de cada modelo propuesto.
Calidad del Informe	10 %	Redacción clara, ortografía y organización del repositorio en GitHub.

4.2. Rúbrica de Evaluación

La calificación se basará en los siguientes criterios, con un énfasis especial en la capacidad de traducción de resultados estadísticos a lenguaje de toma de decisiones:

Nota Importante: Los resultados numéricos correctos sin una interpretación adecuada se penalizarán con el 50 % del valor del punto correspondiente.