Análisis Predictivo y Aprendizaje Estadístico
Caso de Estudio: Advertising Data

Presentado por

July Andrea Cabral rodríguez
Sebastián Alexander Borbón Hernández
Wilson Enrique Torres Burgos

Facultad de Ingeniería, Universidad ECCI

Sistemas Avanzados de Producción
Grupo 9AN

Docente
Fredy Alexander Orjuela López

Marzo 2026

INTRODUCCIÓN

En este trabajo se analiza el conjunto de datos *Advertising*, el cual contiene información de 200 mercados diferentes, incluyendo la inversión en TV, radio y newspaper, así como el volumen de ventas generado en cada uno.

El objetivo principal del análisis es identificar si existe una relación significativa entre la inversión publicitaria y las ventas, determinar cuál es el medio que tiene mayor impacto y evaluar qué modelo estadístico describe mejor este comportamiento. El trabajo se divide en tres fases: La Fase 1 contiene un análisis exploratorio inicial, en la Fase 2 se hizo un modelo de regresión lineal y en la Fase 3 se hizo un análisis a través de un árbol de decisión.

Los resultados muestran que la televisión es el medio que más se relaciona con el aumento de las ventas. La radio también tiene un efecto positivo, pero no tan fuerte como la TV, mientras que el periódico tiene una influencia menor. En general, la regresión lineal ayuda a entender cuánto aumentan las ventas cuando se invierte más en cada medio, y el árbol de decisión permite identificar qué niveles de inversión funcionan mejor en los distintos mercados.

Fase 1: Estadística Descriptiva y Análisis Exploratorio

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 Kurtosis (Medidas de forma).

Tabla 1. Resumen estadístico descriptivo para las variable TV, Radio, Newspaper y Sales

Variable	Media Mean	Mediana Median	Desv. Estándar	Min	Max	Asimetría Skewness	Curtosis Kurtosis
TV	147,04	149,75	85,85	0,70	296,40	-0,07	-1,23
Radio	23,26	22,90	14,85	0,00	49,60	0,09	-1,26
Newspaper	30,55	25,75	21,78	0,30	114,00	0,89	0,65
Sales	14,02	12,90	5,22	1,60	27,00	0,41	-0,41

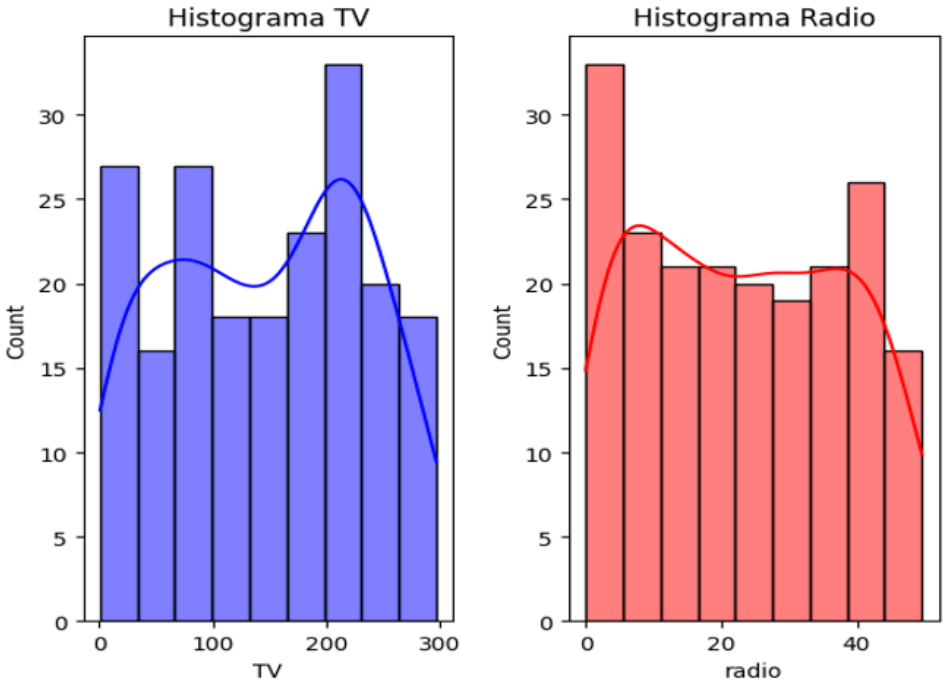
En la **Tabla 1** se muestra el resumen estadístico para las variables TV, Radio, Newspaper y Sales.
Las inversiones en TV y Radio muestran medias y medianas muy cercanas, esto quiere decir que las distribuciones son aproximadamente simétricas, además de asimetrías cercanas a cero.
En el caso de Newspaper, la media es mayor que la mediana y tiene una asimetría positiva (0,89), lo que sugiere una ligera concentración de valores altos en algunos mercados.
La variable Sales también tiene una asimetría positiva (0,41), indicando que existen algunos mercados con ventas superiores al promedio.
La inversión en TV muestra la mayor variabilidad (desviación estándar de 85,85), lo que evidencia diferencias importantes en los niveles de inversión entre los mercados.

3. Análisis de distribución y atípicos

Construya un histograma y un diagrama de caja (Box Plot) para cada variable.

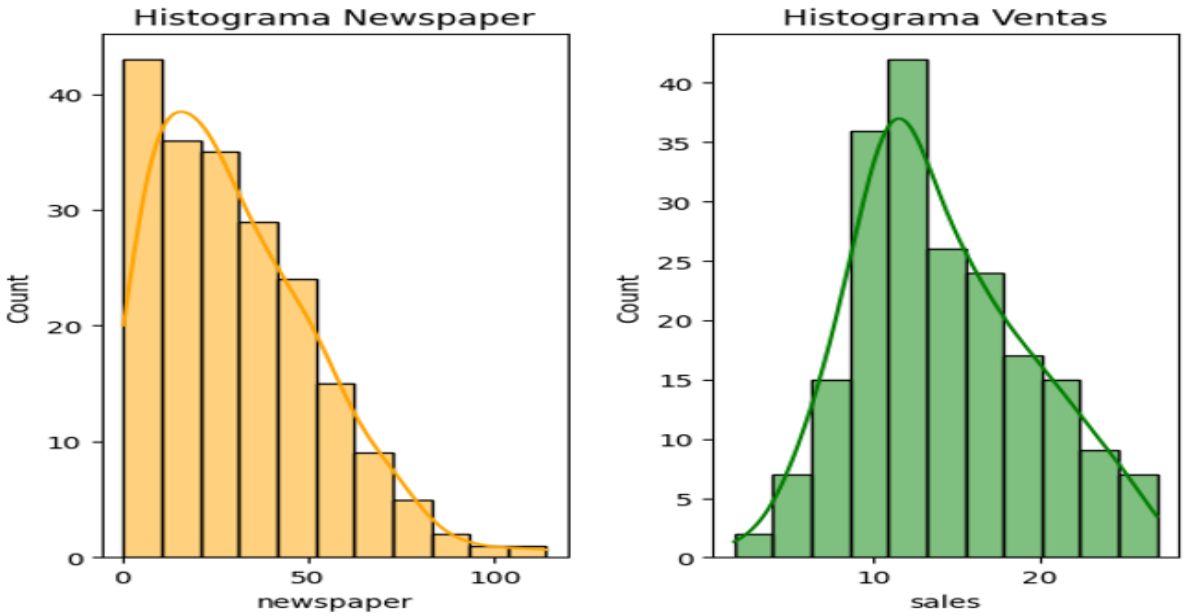
Histogramas y Diagramas de caja (Box Plot)

Figura 1. Histogramas TV y Radio



La **Figura 1** muestra los histogramas para TV y Radio y se observa que estas dos variables presentan distribuciones aproximadamente simétricas. Los datos están distribuidos de manera relativamente equilibrada alrededor de los valores centrales, y no se evidencia una cola pronunciada hacia ninguno de los extremos.

Figura 2. Histogramas Newspaper y Ventas



La **Figura 2** muestra los histogramas para Newspaper y Ventas. Se observa que la variable Newspaper tiene una asimetría positiva (sesgo hacia la derecha), la mayor concentración de datos se encuentra en valores bajos de inversión y existe una cola extendida hacia valores más altos
La variable Sales también muestra una ligera asimetría positiva, y una mayor concentración de los datos alrededor de valores medios y algunos valores más altos que se alejan del centro

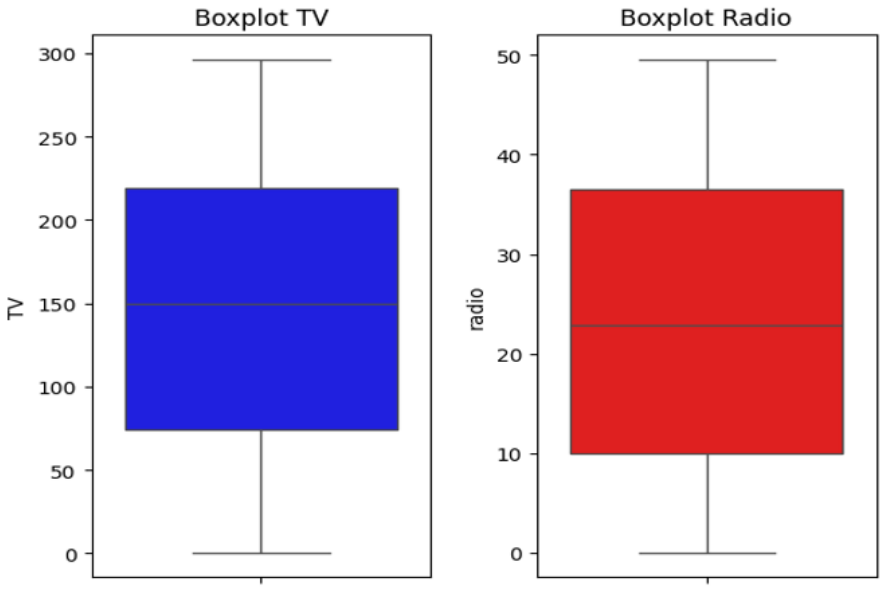
- **Determine la simetría de las distribuciones.**

Las distribuciones de TV y Radio se observan aproximadamente simétricas, ya que los datos se distribuyen alrededor del centro y no presentan una cola marcada hacia ninguno de los extremos.
Newspaper muestra una asimetría positiva, ya que la mayoría de los valores se concentran en niveles bajos de inversión y existe una cola hacia la derecha con algunos valores más altos.
Sales tiene una ligera asimetría positiva, con mayor concentración en valores medios y algunos datos que se extienden hacia niveles altos de ventas.

- **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.**

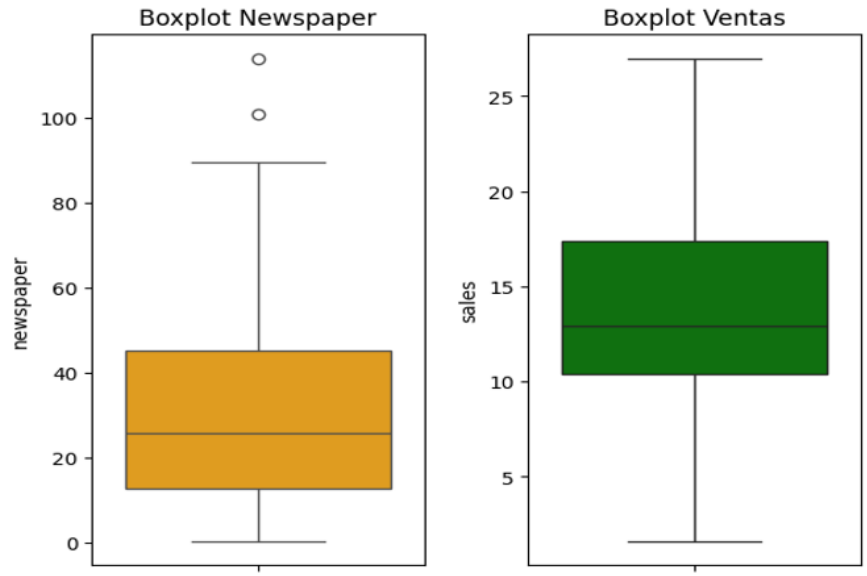
En los histogramas no se observan valores extremadamente aislados en las variables TV y Radio, sin embargo, en Newspaper sí aparecen algunos valores altos que se alejan del grupo principal de datos. Estos podrían considerarse atípicos. Este tipo de observaciones puede influir en la trayectoria de una línea de regresión, ya que al estar alejadas del resto tienden a jalar la recta hacia ellas, modificando la pendiente y afectando la estimación de los coeficientes.

Figura 3. Boxplot TV y Radio



La **Figura No 3** muestra los Boxplot de las variables TV y Radio, en donde se observa que la mediana se encuentra aproximadamente en el centro de la caja y no se observan puntos fuera de los bigotes, lo que indica distribuciones relativamente simétricas y sin valores atípicos.

Figura 4. Boxplot Newspaper y Ventas



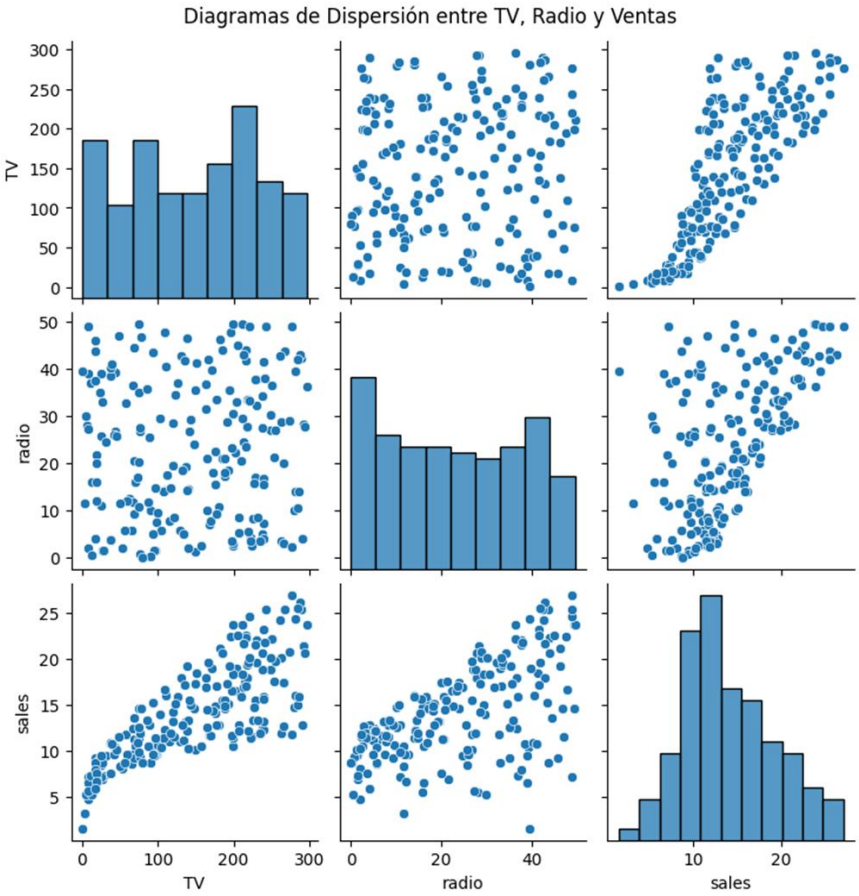
La **Figura No 4** muestra los Boxplot de las variables Newspaper y Ventas. En el Boxplot de Newspaper, se observa que la mayor parte de los datos se concentra en valores bajos y medios, aparecen algunos puntos por encima del límite superior, lo que quiere decir que hay valores atípicos altos.
En el Boxplot de Sales, se observa una distribución más equilibrada, con una mediana centrada y sin outliers marcados.

Los boxplots confirman lo que ya se había observado en los histogramas. En TV y Radio, ambos gráficos muestran distribuciones bastante simétricas y sin valores atípicos. En el caso de Newspaper, el histograma mostraba una asimetría positiva, y el boxplot lo confirma al presentar algunos puntos por encima del límite superior, lo que indica valores atípicos altos. Para Sales, tanto el histograma como el boxplot reflejan una distribución relativamente equilibrada.

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).

Figura 5. Diagrama de Dispersión de las variables TV, Radio y Ventas.



La **Figura 5** muestra los diagramas de dispersión de las variables TV, Radio y Ventas. Estos diagramas muestran la relación entre las variables de inversión y las ventas.

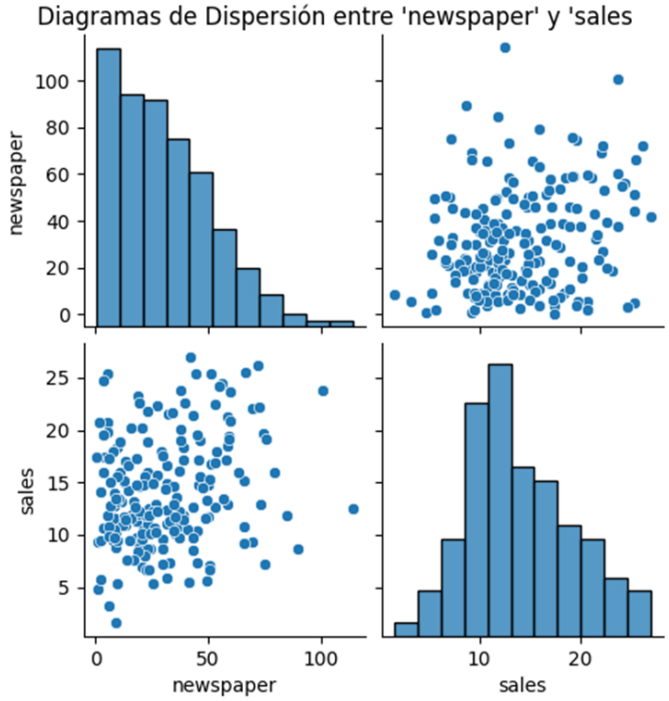
Visualmente, se observa una relación positiva entre TV y Sales, ya que a medida que aumenta la inversión en televisión, las ventas tienden a incrementarse de manera casi lineal, formando una nube de puntos con tendencia ascendente. Esto significa que cuando aumenta la inversión en televisión, las ventas también tienden a incrementarse.

En el caso de Radio y Sales, también se evidencia una relación positiva, aunque con mayor dispersión, lo que indica que la asociación es moderada pero menos fuerte que con TV. Esto significa que su efecto sobre las ventas es moderado pero no tan fuerte como el de la TV.

Entre TV y Radio no se observa un patrón lineal evidente, lo que sugiere que no existe una relación fuerte entre ambas inversiones. Esto significa que los mercados que invierten mucho en televisión no necesariamente invierten lo mismo en radio.

Las nubes de puntos no muestran patrones curvos pronunciados ni formas parabólicas, por lo que la relación parece principalmente lineal, aunque pueden existir algunos puntos más alejados que podrían influir ligeramente en el ajuste de una recta de regresión

Figura 6 Diagrama de dispersión de la variable Newspaper y Ventas

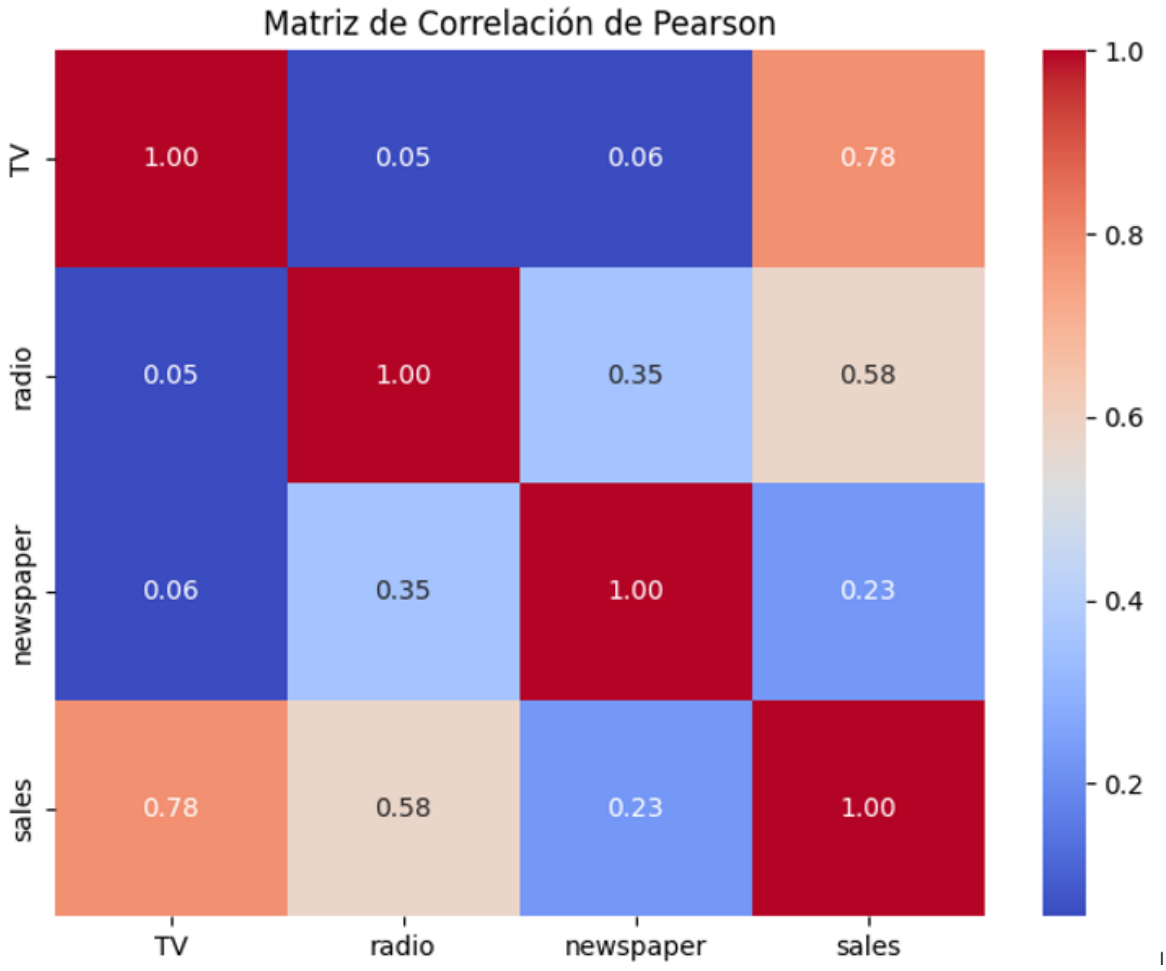


La **Figura 6** muestra los diagramas de dispersión de las variables Newspaper y Ventas. En este diagrama se observa una nube de puntos bastante dispersa, sin una tendencia ascendente o descendente. Existe una ligera relación positiva, lo que quiere decir, que a mayor inversión en periódico las ventas tienden a aumentar un poco, pero la asociación no es fuerte ni claramente lineal. Se observan algunos puntos más alejados, especialmente en niveles altos de inversión en periódico, que podrían actuar como datos influyentes. La relación entre Newspaper y Sales parece débil, lo que sugiere que la inversión en este medio no explica de manera clara el comportamiento de las ventas en comparación con TV o Radio.

5. Evaluación de Asociación

Calcule y presente las matrices de correlación de Pearson y Spearman

Figura 7 Matriz de correlación de Pearson para las variables TV, Radio, Newspaper, Sales



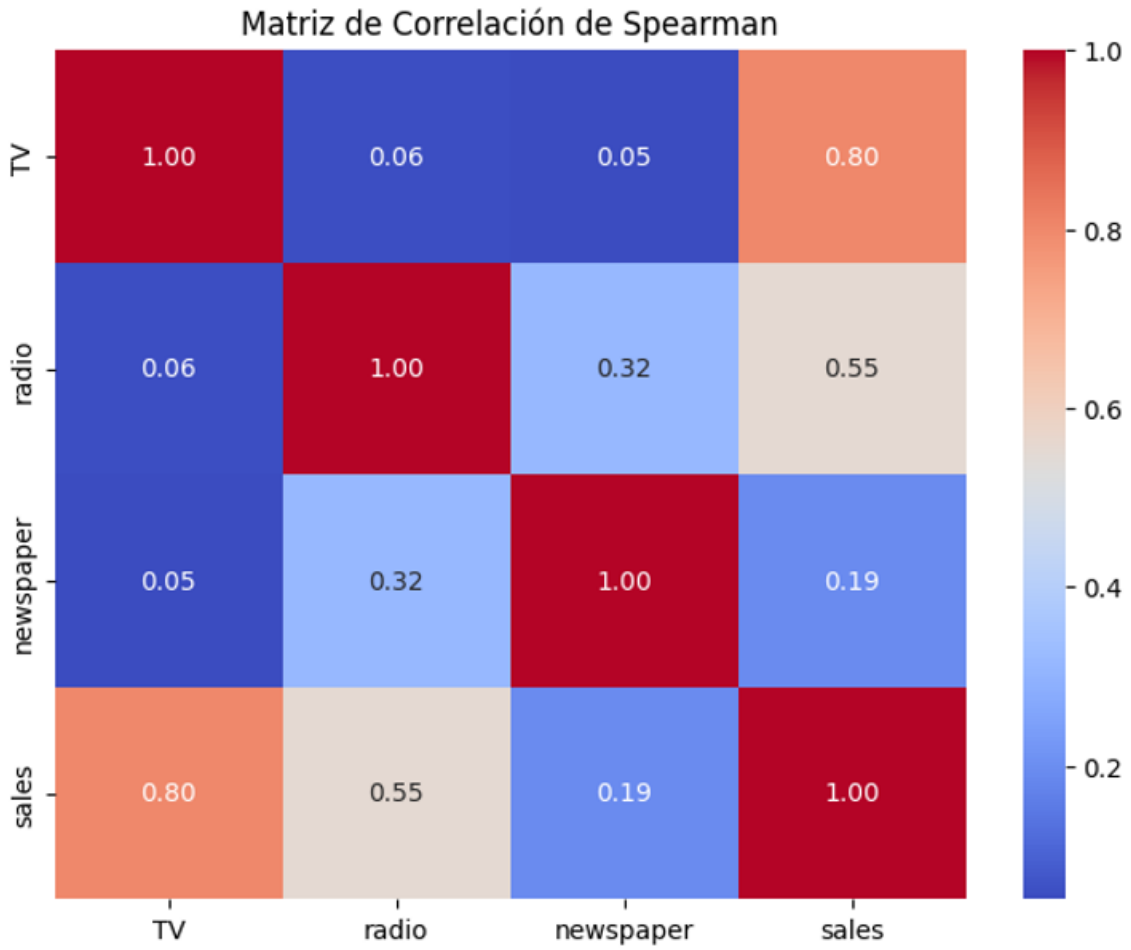
En la **Figura 7** se muestra la matriz de correlación de Pearson. Esta matriz muestra la fuerza y dirección de la relación lineal entre las variables.

Se observa que TV y Sales tienen una correlación alta y positiva (0.78), lo que indica una relación lineal fuerte: a mayor inversión en televisión, mayores ventas.

Radio y Sales presentan una correlación positiva moderada (0.58), lo que sugiere que la radio también influye en las ventas, aunque en menor medida que la TV.

Newspaper y Sales muestran una correlación baja (0.23), lo que indica una relación débil. Además, las correlaciones entre los medios publicitarios son bajas (TV–Radio = 0.05 y TV–Newspaper = 0.06) o moderadas (Radio–Newspaper = 0.35), lo que sugiere que las decisiones de inversión en cada medio no están fuertemente relacionadas entre sí.

Figura 8 Matriz de correlación de Spearman para las variables TV, Radio, Newspaper, Sales



En la **Figura 8** se muestra la matriz de correlación de Spearman. Esta matriz muestra cómo las variables cambian juntas en términos de orden o tendencia, sin asumir que la relación sea estrictamente lineal.

Se observa que TV y Sales tienen una correlación alta y positiva (0.80), lo que indica que cuando la inversión en televisión aumenta, las ventas tienden a aumentar de manera consistente

Radio y Sales presentan una correlación positiva moderada (0.55), lo que sugiere una relación favorable pero menos fuerte que la de TV.

Newspaper y Sales muestran una correlación baja (0.19), lo que indica una asociación débil.

Las correlaciones entre los medios publicitarios son bajas o moderadas, lo que sugiere que las decisiones de inversión en cada medio no siguen exactamente el mismo patrón.

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?**

Un coeficiente de correlación cercano a 1 significa que existe una relación positiva entre la inversión publicitaria y las ventas. Eso quiere decir que cuando la inversión en un medio (por ejemplo, TV) aumenta, las ventas también aumentan casi de manera proporcional. Es decir, los mercados que más invierten tienden a registrar mayores niveles de ventas.

Un coeficiente cercano a −1 indicaría una relación negativa, lo que significa que a mayor inversión, menores ventas. Podría interpretarse como que la inversión en ese medio no está siendo efectiva o que existe algún efecto contrario al esperado.

En ambos casos, un valor cercano a 1 o −1 indica una asociación muy fuerte, que puede ser positiva o negativa, mientras que valores cercanos a 0 indican que no existe una relación clara entre la inversión y el retorno en ventas.

- **¿Cómo se interpreta un valor de correlación cercano a 0?**

Un valor de correlación cercano a 0 se interpreta como la ausencia de relación lineal entre dos variables.
Por ejemplo, si la correlación entre las variables Newspaper y Sales es cercana a 0, esto significa que los mercados que invierten más en periódico no necesariamente venden más, ni los que invierten menos venden menos.
Pero es importante aclarar que una correlación cercana a 0 no siempre significa que no exista ningún tipo de relación. Lo que indica es que no hay una relación lineal clara. Podría existir una relación más compleja o no lineal que la correlación de Pearson no detecta.

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.**

Al comparar los coeficientes de Pearson y Spearman, se observa que los valores son muy similares en todas las variables, especialmente en la relación entre TV y Sales, que es alta y positiva en ambos casos, y entre Radio y Sales que es moderada y positiva.
Esto indica que no solo existe una relación lineal fuerte, que es lo que muestra la matriz de Pearson, sino que también hay una relación entre las variables en la que tienden a moverse en una sola dirección respecto a otra, que es lo que muestra la matriz de Spearman.
Como los coeficientes de Pearson y Spearman son parecidos, se puede concluir que las relaciones observadas son principalmente lineales.

En relación con las nubes de puntos en los diagramas de dispersión, se puede evidenciar una tendencia ascendente, especialmente en TV y Sales, y bastante alineada, lo que confirma que hay una relación lineal fuerte.
Para Radio y Sales, aunque la dispersión es mayor, también se observa una tendencia positiva relativamente recta, lo que sugiere una relación lineal moderada.
En el caso de las variables Newspaper y Sales, la nube de puntos es más dispersa y no presenta una forma definida, lo que indica una relación débil y poco lineal.
Entre TV y Radio no se observa un patrón lineal, lo que muestra que hay una baja correlación entre ambas inversiones.

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:

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

1. Modelamiento Múltiple

Estime un modelo de regresión lineal múltiple donde las ventas (Sales) dependan de la inversión en TV y Radio.
Reporte los valores numéricos del intercepto (βˆ 0) y los coeficientes de pendiente (βˆ 1, βˆ 2)).

Intercepto (beta_0): 2.9211

Pendiente (beta_1): 0.0458

Pendiente (beta_2): 0.1880

2. Interpretación de parámetros

Explique con sus propias palabras el significado de los resultados obtenidos:

a. Intercepto (β_0 = 2.9211**)**

Aquí el valor del intercepto beta sub cero nos indica que si no se decide hacer inversión alguna en televisión (TV) ni en radio, es decir invertir cero dólares en ambas variables, las ventas base en promedio serían 2,92 miles de unidades. Es decir, sin tener nada de inversión, se empieza con unas ventas de casi tres mil unidades.

b. Pendiente (β1 = 0.0458) - Variable 1 (TV)

La interpretación de esta variable beta sub 1 nos quiere indicar que por cada 1000 mil dólares adicionales invertidos en publicidad de televisión (TV), se estima que el volumen en ventas se incrementa en 45,8 unidades, pero teniendo constante la variable radio (ceteris paribus).

c. Pendiente (β2 = 0.1880) - Variable 2 (radio)

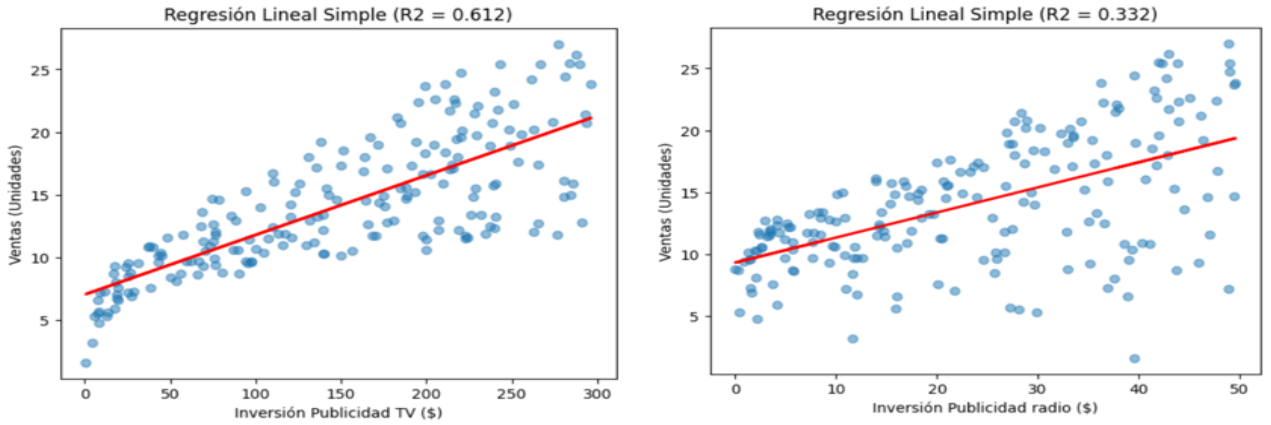
Asimismo, para la segunda variable beta sub 2, nos indica que por cada incremento adicional de mil dólares de publicidad en radio, el volumen en las ventas crecerían 188 unidades de producto, manteniendo constante la variable TV (ceteris paribus).

Como podemos notar, es mucho más rentable y eficiente invertir en publicidad para radio, ya que esto puede generar alrededor de cuatro veces más volumen en ventas si lo comparamos con la inversión en TV. Esto se debe gracias a la versatilidad de obtener la publicidad por radio, en comparación con la televisión.

3. Análisis de Bondad de Ajuste:

Calcule el coeficiente de determinación R2 . 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

Figura 9. Regresiones lineales calculadas de TV y radio versus Ventas



-- Bondad de Ajuste ---

R-cuadrado (R2): 0.8972

Si bien en la figura 9, se muestran por separado cada R2, cuando se calcula por medio de la plataforma el resultado de las dos al tiempo, el R2 nos da como indicador un 0,8972. Así las cosas y para nuestro ejercicio, el 89.72% de la variación en las ventas se puede explicar gracias a las inversiones de la publicidad en televisión y radio, y quiere decir que tiene un ajuste sólido, ya que sólo el 10,28% de las variaciones en las ventas se deben a otros factores que no están aquí contemplados tales como la competencia, los precios, la demanda y etc. Así las cosas, se puede mencionar que el modelo es

confiable con las dos variables interpretadas y que sugieren que tienen un impacto alto en el volumen de las ventas del producto.

Ahora bien, si resultó que la inclusión de la variable radio mejoró significativamente el ajuste con respecto al modelo simple, dado que la variación en las ventas se reduce un 28.5 % con relación a tener las dos variables al mismo tiempo, y ya que el modelo determina que la variable radio es mucho mejor que la variable TV, sustenta lo anteriormente dicho.

4. **Estimación Matricial:**

Considere el siguiente conjunto de 5 observaciones para una regresión lineal simple. Utilizando la formulación matricial βˆ = (X'X) −1X'Y, encuentre los estimadores βˆ 0 (intercepto) y βˆ 1 (pendiente).

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

a. (

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

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

b. Cálculo del producto X'X y su respectiva matriz inversa (X'X)−**1**

Producto X'X

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix} = \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix}$$

Se halla el determinante **|x'x| = (5 x 55) - (15 x 15) = 275 - 225 = 50**

Luego la inversa

$$(X'X)^{-1} = \frac{1}{50}\begin{pmatrix} 55 & -15 \\ -15 & 5 \end{pmatrix} = \begin{pmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{pmatrix}$$

c. Cálculo del producto X'Y

$$X'Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \\ 6 \\ 7 \\ 9 \end{pmatrix} = \begin{pmatrix} 2+4+6+7+9 \\ 2+8+18+28+45 \end{pmatrix} = \begin{pmatrix} 28 \\ 101 \end{pmatrix}$$

d. Obtenga el vector de parámetros $\hat{\beta}$ y escriba la ecuación de la recta estimada $\hat{y} = \hat{\beta}0 + \hat{\beta}1x$.

Primero se multiplica la inversa hallada (X'X)-1 por X'Y

$$\hat{\beta} = \begin{pmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{pmatrix} \begin{pmatrix} 28 \\ 101 \end{pmatrix}$$

El resultado es:

- $\hat{\beta}_0 = (1.1 \times 28) + (-0.3 \times 101) = 30.8 - 30.3 = \mathbf{0.5}$
- $\hat{\beta}_1 = (-0.3 \times 28) + (0.1 \times 101) = -8.4 + 10.1 = \mathbf{1.7}$

Vector de parámetros:

$$\hat{\beta} = \begin{pmatrix} 0.5 \\ 1.7 \end{pmatrix}$$

Ecuación de la recta estimada:

$$\hat{y} = 0.5 + 1.7x$$

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.

1. **Entrenamiento del Modelo:**

Utilizando las variables TV y Radio, entrene un árbol de regresión para predecir las ventas (Sales).

El árbol de regresión realiza las particiones utilizando el criterio de minimización del error cuadrático medio (MSE). Esto significa que en cada nodo el algoritmo busca el punto de corte que reduzca al máximo la variabilidad de las ventas dentro de cada grupo. En otras palabras, intenta que los datos dentro de cada segmento sean lo más homogéneos posible.

En el árbol obtenido, la primera partición se realiza en:
TV ≤ 122.05

Esto indica que la variable TV es la que más reduce el error en el primer nivel, por lo tanto, es la variable con mayor capacidad inicial de segmentación.

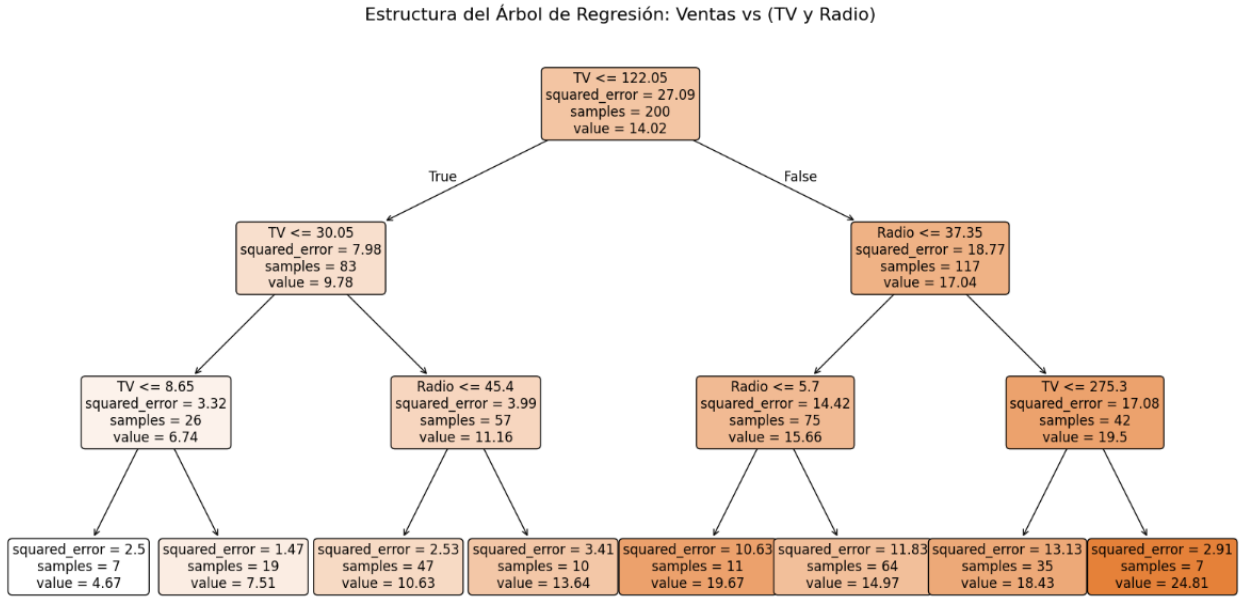
Cada nodo terminal muestra:

- El número de observaciones
- El error cuadrático
- El valor promedio de ventas en ese segmento

Por ejemplo, cuando TV es alto y Radio también es alto, las ventas alcanzan valores cercanos a 21.95, que corresponden a los niveles más altos del modelo

Grafique la estructura del árbol y explique bajo qué criterio se realizan las particiones en los nodos.

Figura. 11. Estructura de árbol de regresión: Ventas VS Tv y Radio




```
--- RESULTADOS COMPARATIVOS ---
Importancia en el Árbol (TV):        0.8241
Importancia en el Árbol (Newspaper): 0.1759
-----------------------------------
Coeficiente Regresión (TV):         0.0469
Coeficiente Regresión (Newspaper):  0.0442
```

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

Calcule la Entropía del Nodo Padre:

Proporciones:

- Altas = 3/4 = 0.75
- Bajas = 1/4 = 0.25

Fórmula:

$$H(S) = -\sum p_i \log_2(p_i)$$

$$H(S) = -(0.75 \log_2 0.75 + 0.25 \log_2 0.25)$$

$$H(S) - (0.75(-0.415) + 0.25(-2))$$

$$H(S)0.811$$

La entropía inicial es 0.811, lo que indica que existe incertidumbre moderada.

Calcule el Índice de Gini para el nodo padre:

$$Gini = 1 - \sum p_i^2$$

$$Gini = 1 - (0.75^2 + 0.25^2)$$

$$Gini = 1 - (0.5625 + 0.0625)$$

$$Gini = 0.375$$

El nodo no es puro, ya que el Gini es distinto de 0.

Determine la Ganancia de Información tras realizar una partición por la variable Inversión TV

TV Alta → 2 Altas

TV Baja → 1 Alta, 1 Baja

Entropía TV Alta:

Es nodo puro → H = 0

Entropía TV Baja:

- Altas = 1/2
- Bajas = 1/2

H = - (0.5\log2 0.5 + 0.5\log2 0.5)

H = 1

Entropía ponderada:

H{nueva} = (2/4)(0) + (2/4)(1) = 0.5

Ganancia de Información:

IG = 0.811 - 0.5 = 0.311

La partición por TV reduce la incertidumbre en 0.311, por lo tanto sí aporta información relevante.

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.

Resultados del árbol:

- Importancia TV: 0.6425
- Importancia Radio: 0.3575

Esto indica que TV explica aproximadamente el 64% de la reducción total del error, siendo la variable más influyente en la estructura del árbol.

Comparación con regresión:

- Coeficiente TV = 0.0458
- Coeficiente Radio = 0.1880

En la regresión, Radio tiene mayor efecto marginal por unidad invertida.

Esto no contradice el árbol, porque:

- La regresión mide el efecto promedio lineal.
- El árbol mide capacidad de segmentación y reducción de error.

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.

La regresión lineal es preferible cuando se busca medir de manera clara y cuantitativa cuánto influyen las variables en las ventas. Este modelo es más fácil de interpretar, ya que permite expresar resultados concretos como “por cada mil dólares adicionales invertidos, las ventas aumentan en cierta cantidad”. Por esta razón, es útil cuando el objetivo principal es entender el impacto promedio de cada medio publicitario y tomar decisiones basadas en ese efecto estimado.

El árbol de decisión es más adecuado cuando se sospecha que la relación entre las variables no es completamente lineal o cuando se desea identificar niveles específicos de inversión que generen mejores resultados. Este modelo permite detectar combinaciones entre variables y establecer reglas claras, como umbrales de inversión en TV o Radio que aumenten significativamente las ventas.

En este caso particular, como las relaciones observadas son principalmente lineales y no se identifican patrones curvos fuertes en las nubes de puntos, la regresión lineal resulta una herramienta adecuada y suficiente. Sin embargo, el árbol también complementa el análisis al permitir visualizar cómo se segmentan los mercados según sus niveles de inversión.

CONCLUSIONES

Con base en los resultados obtenidos en el análisis, la regresión lineal y el árbol de decisión, podemos concluir que la inversión en televisión (TV) es el principal factor asociado al incremento en las ventas. Esto se evidencia en su alta correlación con Sales (0.78 en Pearson y 0.80 en Spearman), así como en su mayor importancia dentro del árbol de decisión (64% de reducción del error).

Aunque en la regresión el coeficiente de Radio es mayor por cada unidad invertida, esto no significa que sea más importante en general. La televisión explica una mayor parte del comportamiento de las ventas y tiene mayor influencia en el modelo.

La variable Newspaper muestra una relación débil con las ventas, lo que indica que su impacto es menor en comparación con TV y Radio.

En cuanto a los modelos utilizados, la regresión lineal permite medir cuánto aumentan las ventas cuando se incrementa la inversión en cada medio, mientras que el árbol de decisión ayuda a identificar niveles de inversión que generan mejores resultados. En este caso, ambos modelos coinciden en que la televisión es el medio más influyente.

Aunque el modelo explica una gran parte de la variación en las ventas, todavía existen otros factores externos que pueden influir, como la competencia o los precios, por lo que los resultados deben interpretarse como una guía para la toma de decisiones y no como una predicción exacta.