

TALLER 1 ANÁLISIS PREDICTIVO Y APRENDIZAJE ESTADÍSTICO

SARA CANRO CAMPOS

SARA SOFIA VILLANUEVA GUARIN

KEVIN ALEJANDRO PEREZ ÑUSTES

SISTEMAS AVANZADOS DE PRODUCCIÓN

UNIVERSIDAD ECCI

FREDY ALEXANDER ORJUELA LOPEZ

FEBRERO 2026

INTRODUCCIÓN

El presente taller aborda el análisis predictivo aplicado al conjunto de datos Advertising, con el propósito de evaluar el impacto de la inversión publicitaria en distintos medios sobre el volumen de ventas. A través de un enfoque estructurado en tres fases estadística descriptiva, regresión lineal y árboles de decisión se desarrollan competencias relacionadas con la exploración, modelamiento y comparación de técnicas de aprendizaje estadístico. En la primera fase se realiza un análisis exploratorio que incluye medidas de tendencia central, dispersión, forma y asociación, con el fin de identificar patrones y posibles relaciones lineales entre las variables de inversión en televisión, radio y newspaper frente a las ventas. Posteriormente, se construye un modelo de regresión lineal múltiple para estimar la magnitud del efecto de cada medio publicitario, interpretando los coeficientes, el intercepto y el coeficiente de determinación como indicadores de ajuste y capacidad explicativa. Finalmente, se implementa un árbol de decisión para capturar posibles relaciones no lineales y comparar su desempeño.

Palabras clave: análisis predictivo; regresión lineal; árboles de decisión; aprendizaje estadístico; inversión publicitaria; análisis exploratorio de datos

DESARROLLO TÉCNICO

Fase 1: Estadística Descriptiva y Análisis Exploratorio

Variable	M	Mdn	DE	Mín	Máx	Asimetría	Curtosis
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

Tabla [1]. La tabla 1 muestra las medidas de la tendencia central y dispersión para las variables independientes y la variable Sales dependiente.

1.1 Análisis de Distribución y Atípicos

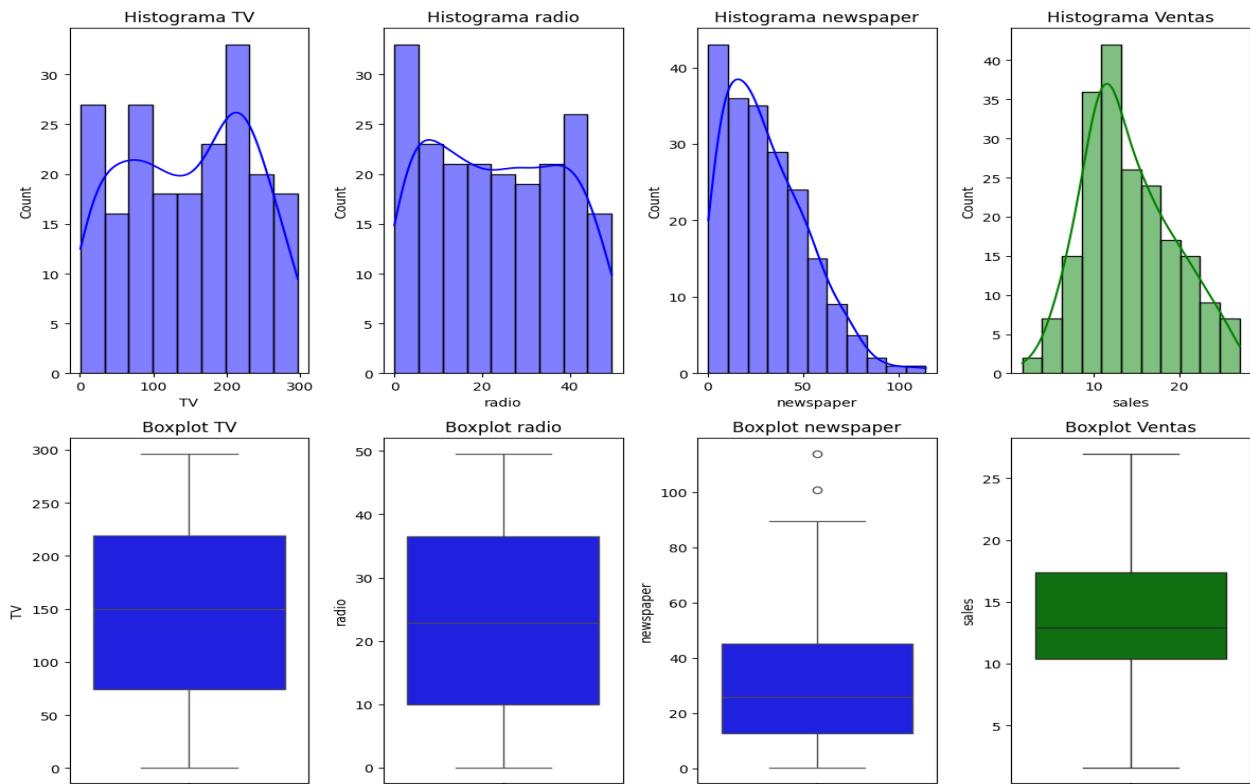


Figura [1]. Histogramas y diagramas de caja (Box Plot) de las variables TV, radio, newspaper y Sales.

Análisis de la simetría de las distribuciones

A partir de los histogramas presentados en la Figura 1, se observa que las variables TV y radio presentan distribuciones relativamente simétricas, aunque con ligera concentración de valores en ciertos rangos intermedios. No se evidencia una asimetría marcada, lo que sugiere una distribución cercana a la normalidad en ambos casos.

En contraste, la variable newspaper muestra una distribución claramente asimétrica positiva (sesgo hacia la derecha), caracterizada por una mayor concentración de observaciones en valores bajos de inversión y una cola extendida hacia valores altos. Esto indica que en la mayoría de los mercados la inversión en prensa es baja, pero existen algunos casos donde el gasto es considerablemente mayor.

Por su parte, la variable Sales presenta una distribución aproximadamente simétrica, con ligera asimetría positiva. La mayoría de las ventas se concentran en valores intermedios, lo cual es consistente con un comportamiento relativamente estable del mercado.

¿Se identifican datos atípicos u observaciones que se alejan significativamente de la masa de los datos?

- Identificación de datos atípicos**

El análisis mediante los diagramas de caja permite identificar posibles valores atípicos (outliers).

- TV y Radio no presentan valores atípicos evidentes fuera de los límites intercuartílicos.
- Newspaper sí muestra observaciones atípicas superiores, reflejando mercados con inversiones significativamente mayores al promedio.
- Sales no evidencia valores extremos pronunciados.

Los valores atípicos en la variable Newspaper podrían influir en el ajuste de una línea de regresión, especialmente si están asociados a niveles de ventas inusualmente altos o bajos. En modelos de regresión lineal, estas observaciones pueden:

- Aumentar la pendiente de la recta si se alinean con la tendencia.
- Distorsionar el intercepto.
- Incrementar la varianza residual.
- Reducir la estabilidad del modelo si son puntos influyentes (high leverage points).

- **Impacto en la trayectoria de una línea de regresión**

Los datos atípicos pueden afectar significativamente la trayectoria de una línea de regresión debido a que el método de mínimos cuadrados minimiza los errores al cuadrado. Esto implica que observaciones alejadas de la masa principal de datos tienen mayor peso en el cálculo de los coeficientes. En este caso, si las observaciones atípicas de Newspaper no están asociadas a un aumento proporcional en las ventas, podrían generar:

- Una sobreestimación o subestimación del efecto real de la publicidad en prensa
- Un aumento artificial del error estándar del coeficiente
- Una disminución del poder explicativo del modelo.

Por lo tanto, es recomendable evaluar la influencia de estas observaciones mediante análisis adicionales como medidas de leverage o distancia de Cook en fases posteriores del modelamiento.

1.2 Exploración de la Nube de Puntos.

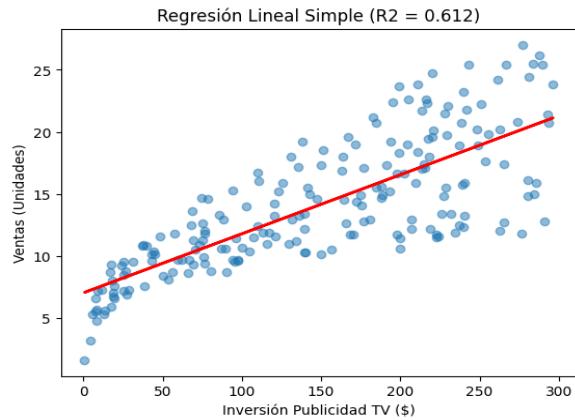


Figura [2]. Regresión lineal simple entre inversión en TV y ventas ($R^2 = 0.612$).

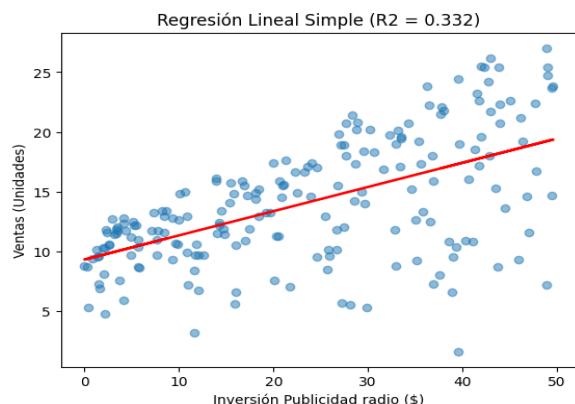


Figura [3]. Regresión lineal simple entre inversión en Radio y ventas ($R^2 = 0.332$).

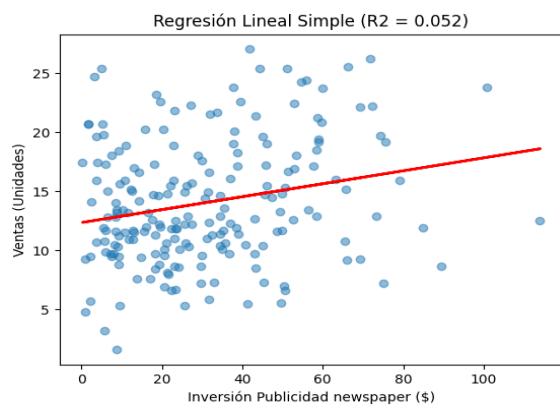


Figura [4]. Regresión lineal simple entre inversión en Newspaper y ventas ($R^2 = 0.052$).

- **Identificación visual de la relación entre variables**

A partir de los dispersogramas presentados en la Figura 2, Figura 3, Figura 4 se analiza visualmente la relación entre la variable respuesta Sales y cada una de las covariables (TV, Radio y Newspaper).

1. Sales vs. TV ($R^2 = 0.612$)

La nube de puntos sugiere una relación lineal positiva clara y consistente. A medida que aumenta la inversión en televisión, se observa un incremento en el volumen de ventas.

- **El coeficiente de determinación $R^2=0.612$**

Indica que aproximadamente el 61.2 % de la variabilidad en las ventas es explicada únicamente por la inversión en TV. Esto sugiere que la publicidad en televisión es un predictor fuerte dentro del modelo simple.

Visualmente:

- La tendencia ascendente es evidente.
- No se observan patrones curvilíneos.
- La dispersión aumenta ligeramente en valores altos de inversión, lo que podría sugerir leve heterocedasticidad.
- No se identifican puntos extremadamente influyentes que distorsionen la pendiente.

Conclusión: La relación parece predominantemente lineal.

2. Sales vs. Radio ($R^2 = 0.332$)

En este caso se observa una relación lineal positiva moderada, aunque con mayor dispersión alrededor de la recta estimada. El $R^2=0.332$ indica que aproximadamente el 33.2 % de la variabilidad en las ventas es explicada por la inversión en radio.

Visualmente:

- Existe una tendencia creciente.
- La nube de puntos es más dispersa que en TV.
- Se identifican algunos puntos alejados de la línea, lo que puede reducir la precisión del modelo.
- No se evidencian patrones no lineales marcados.

Conclusión: La relación es lineal, pero más débil que la observada en TV.

3. Sales vs. Newspaper ($R^2 = 0.052$)

En este dispersograma no se observa una relación lineal clara entre la inversión en prensa y las ventas. El $R^2 = 0.052$ indica que únicamente el 5.2 % de la variabilidad en las ventas es explicada por la inversión en periódicos, lo cual evidencia una capacidad predictiva muy baja.

Visualmente:

- La nube de puntos es altamente dispersa.
- No se aprecia una tendencia definida.
- La recta de regresión tiene pendiente muy leve.
- Existen observaciones que podrían considerarse influyentes en valores altos de inversión.

Conclusión: No se evidencia una relación lineal significativa.

- **Respuesta al punto solicitado**

De acuerdo con la exploración visual:

- TV sugiere una relación lineal simple fuerte.
- Radio sugiere una relación lineal moderada.
- Newspaper no presenta una relación lineal significativa.

No se identifican patrones parabólicos ni segmentaciones evidentes en los dispersogramas. Sin embargo, en el caso de Newspaper podrían existir observaciones influyentes que afectan la estimación del modelo simple.

En términos analíticos, estos resultados justifican avanzar hacia un modelo de regresión múltiple, donde se evalúe el efecto conjunto de los medios, ya que individualmente no todos presentan igual poder explicativo.

1.3 Evaluación de Asociación: Pearson y Spearman.

Se calcularon las matrices de correlación de Pearson y Spearman con el fin de evaluar la intensidad y naturaleza de la asociación entre la inversión en los distintos medios publicitarios y las ventas.

- **Correlación de Pearson (Asociación lineal)**

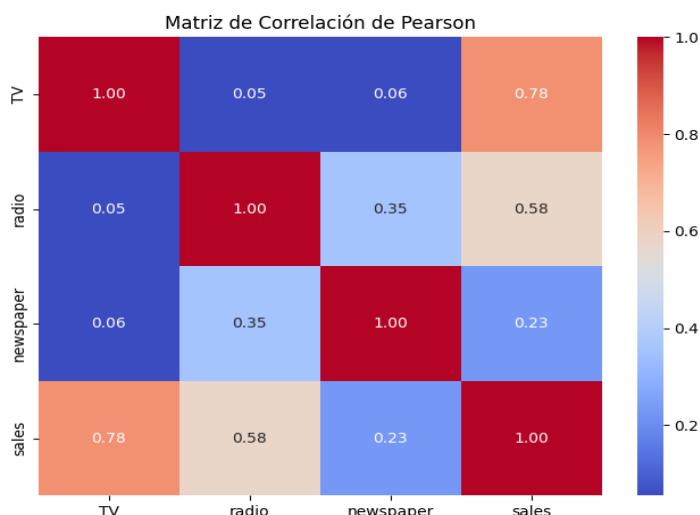


Figura [5]. Matriz de correlación de Pearson entre inversión publicitaria y ventas.

Los principales resultados son:

- TV – Sales: 0.78
- Radio – Sales: 0.58
- Newspaper – Sales: 0.23

El coeficiente de Pearson mide la fuerza de la relación lineal entre dos variables cuantitativas.

Interpretación:

- La inversión en TV presenta una correlación positiva fuerte con las ventas.
- La inversión en Radio muestra una correlación positiva moderada
- La inversión en Newspaper evidencia una correlación débil.

Esto confirma lo observado previamente en los dispersogramas: TV es el medio con mayor capacidad explicativa individual.

- **Correlación de Spearman (Asociación monótona)**

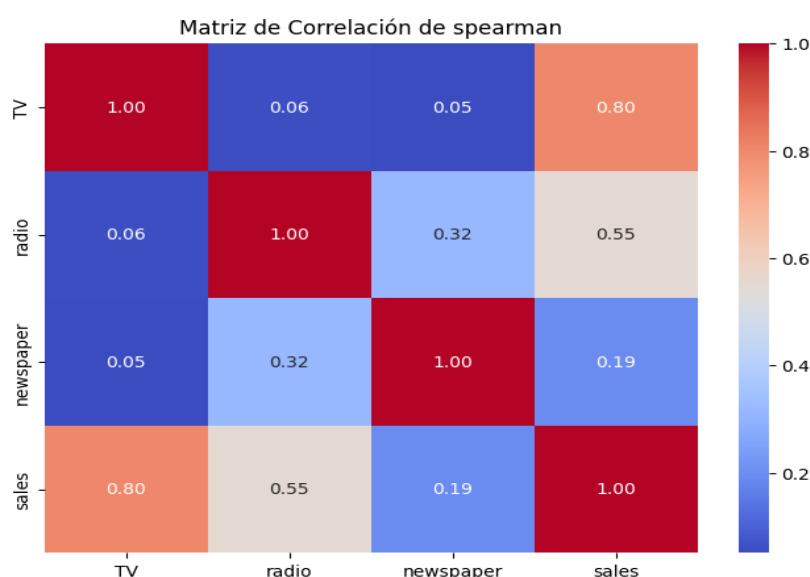


Figura 6. Matriz de correlación de Spearman entre inversión publicitaria y ventas.

Los resultados obtenidos fueron:

- TV – Sales: 0.80
- Radio – Sales: 0.55
- Newspaper – Sales: 0.19

El coeficiente de Spearman evalúa si existe una relación monótona, es decir, si al aumentar una variable la otra tiende a aumentar (o disminuir), sin requerir estrictamente linealidad.

Se observa que:

- TV mantiene una asociación fuerte.
- Radio conserva una relación moderada.
- Newspaper continúa mostrando una asociación débil.

La similitud entre los coeficientes de Pearson y Spearman sugiere que las relaciones observadas son predominantemente lineales y no presentan patrones no lineales significativos.

1.4 Interpretación de Resultados

¿Qué significa un coeficiente cercano a 1 o -1?

- Un coeficiente cercano a 1 indica una relación positiva fuerte: a mayor inversión publicitaria, mayores ventas.
- Un coeficiente cercano a -1 indicaría una relación negativa fuerte: a mayor inversión, menores ventas (lo cual no ocurre en este caso).
- En este estudio, el valor cercano a 0.80 entre TV y Sales indica que la inversión en televisión tiene un impacto directo y significativo en el aumento de ventas.

¿Cómo se interpreta un valor cercano a 0?

Un coeficiente cercano a 0 indica ausencia de relación lineal o monótona relevante.

El valor de 0.23 (Pearson) y 0.19 (Spearman) entre Newspaper y Sales indica que la inversión en prensa tiene una influencia muy limitada sobre las ventas, al menos cuando se analiza de manera individual.

¿Sugieren los resultados que las relaciones son estrictamente lineales?

Sí, en gran medida. La cercanía entre los coeficientes de Pearson y Spearman sugiere que las relaciones son mayormente lineales.

- No se identifican comportamientos curvilíneos importantes.
- La nube de puntos no evidencia patrones parabólicos ni segmentaciones claras.

Esto respalda la pertinencia del uso de modelos de regresión lineal en las fases posteriores del análisis.

Fase 2: Regresión Lineal y Diagnóstico

2.1. Modelamiento Múltiple

Se estimó un modelo de regresión lineal múltiple donde la variable dependiente Sales se explicó a partir de las inversiones en TV, Radio y Newspaper.

- Intercepto (β_0): 2.9389
- Coeficiente TV (β_1): 0.0458
- Coeficiente Radio (β_2): 0.1885
- Coeficiente newspaper (β_3): -0.0010
- R-cuadrado (R^2): 0.8972

2.2 Interpretación de Parámetros

Interpretación del intercepto ($\beta^0=2.9389$)

El intercepto representa el valor esperado de las ventas cuando la inversión en TV, Radio y Newspaper es igual a cero. En términos prácticos, esto significa que, aun sin inversión publicitaria, se estiman aproximadamente 2.94 mil unidades vendidas.

Conceptualmente, este valor puede interpretarse como el nivel base de ventas atribuible a factores externos no incluidos en el modelo, como reconocimiento de marca, demanda natural del mercado o distribución.

Interpretación de los coeficientes (ceteris paribus)

- TV ($\beta^1=0.0458$): Manteniendo constantes Radio y Newspaper:

Por cada incremento de 1 mil dólares en publicidad en TV, las ventas aumentan en promedio 0.0458 mil unidades (es decir, aproximadamente 45.8 unidades).

Esto confirma que TV tiene un efecto positivo y significativo sobre las ventas.

- Radio ($\beta^2=0.1885$): Manteniendo constantes TV y Newspaper:

Por cada incremento de 1 mil dólares en publicidad en Radio, las ventas aumentan en promedio 0.1885 mil unidades (aproximadamente 188.5 unidades).

Este coeficiente es mayor que el de TV en magnitud marginal, lo que sugiere que, por unidad adicional invertida, Radio tiene un impacto más fuerte sobre las ventas cuando se controla por los demás medios.

- Newspaper ($\beta^3=-0.0010$): Manteniendo constantes TV y Radio:

Por cada incremento de 1 mil dólares en publicidad en Newspaper, las ventas disminuyen en promedio 0.0010 mil unidades. Este coeficiente es muy cercano a cero y negativo, lo que sugiere que la inversión en prensa no tiene un impacto significativo sobre las ventas dentro del modelo múltiple.

Esto es coherente con lo observado en la matriz de correlación y en la regresión simple, donde Newspaper mostró baja capacidad explicativa.

2.3 Análisis de Bondad de Ajuste

El coeficiente de determinación obtenido para el modelo de regresión lineal múltiple fue:

- R-cuadrado (R^2): 0.8972

Interpretación del $R^2=0.8972$ indica que el 89.72 % de la variabilidad total de las ventas (Sales) es explicada por el modelo que incluye las variables TV, Radio y Newspaper.

En términos prácticos, esto significa que casi el 90 % del comportamiento de las ventas puede ser atribuido a la inversión publicitaria en estos tres medios, mientras que el 10.28 % restante corresponde a factores no observados o al error aleatorio del modelo.

Un R^2 cercano a 1 sugiere un alto poder explicativo y un buen ajuste del modelo a los datos observados.

- **Comparación con el modelo simple (Fase 1)**

En la fase anterior se evaluaron modelos de regresión simple:

TV sola: $R^2=0.612$

Radio sola: $R^2=0.332$

Newspaper sola: $R^2=0.052$

Al comparar estos resultados con el modelo múltiple:

$0.8972 > 0.612$

Se observa una mejora sustancial en la capacidad explicativa, pasando de explicar el 61.2 % de la variabilidad (con TV sola) a explicar el 89.72 % al incluir Radio y Newspaper.

- **¿Mejoró significativamente el ajuste?**

Sí, la inclusión de Radio mejora notablemente el modelo, ya que aporta información adicional relevante que incrementa el poder predictivo.

En contraste, la variable Newspaper, aunque se incluye en el modelo, presenta un coeficiente cercano a cero y negativo, lo que sugiere que su contribución individual es marginal cuando se controla por TV y Radio.

Sin embargo, su presencia no deteriora el ajuste global, pero tampoco aporta una mejora sustancial.

- **Conclusión del análisis**

El modelo múltiple presenta un ajuste considerablemente superior al modelo simple.

TV y Radio son los principales determinantes de las ventas.

Newspaper tiene baja relevancia estadística.

El modelo explica casi el 90 % de la variabilidad, lo que indica alta capacidad predictiva.

2.4. Estimación Matricial

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

Figura [7]. Matriz a resolver.

2.4 ESTIMACIÓN MATRICIAL

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

Inversa $(X^T X)^{-1}$

$$|X^T X| = (5)(55) - (15)(15) = 275 - 225 = 50$$

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

Producto $X^T y$

$$X^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \\ 6 \\ 7 \\ 9 \end{pmatrix} = \begin{pmatrix} 28 \\ 101 \end{pmatrix}$$

Calculo de $\hat{\beta}$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \frac{1}{50} \begin{pmatrix} 55 & -15 \\ -15 & 5 \end{pmatrix} \begin{pmatrix} 28 \\ 101 \end{pmatrix}$$

$$\hat{\beta}_0 = \frac{1}{50} (55 \cdot 28 - 15 \cdot 101) = \frac{1}{50} (1540 - 1515) = \frac{25}{50} = 0,5$$

$$\hat{\beta}_1 = \frac{1}{50} (-15 \cdot 28 + 5 \cdot 101) = \frac{1}{50} (-420 + 505) = \frac{85}{50} = 1,7$$

$$\hat{\beta} = \begin{pmatrix} 0,5 \\ 1,7 \end{pmatrix}$$

Ecación de la recta estimada

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \rightarrow \hat{y} = 0,5 + 1,7x$$

Figura [8]. Cálculos de la solución de la matriz.

La estimación matricial permitió obtener los parámetros del modelo de regresión lineal simple mediante la fórmula $\beta^{\wedge} = (X'X)^{-1}X'Y$, garantizando la solución óptima bajo el criterio de mínimos cuadrados. El resultado fue la ecuación $y^{\wedge} = 0.5 + 1.7x$, lo que indica que cuando $x=0$, el valor esperado de y es 0.5, y que por cada incremento de una unidad en x , la variable respuesta aumenta en promedio 1.7 unidades. Esto confirma la existencia de una relación lineal positiva entre las variables y demuestra que el procedimiento matricial proporciona una solución única y consistente para el modelo.

Fase 3: Arboles de Decisión y Comparación de Modelos

3.1 Entrenamiento del Modelo

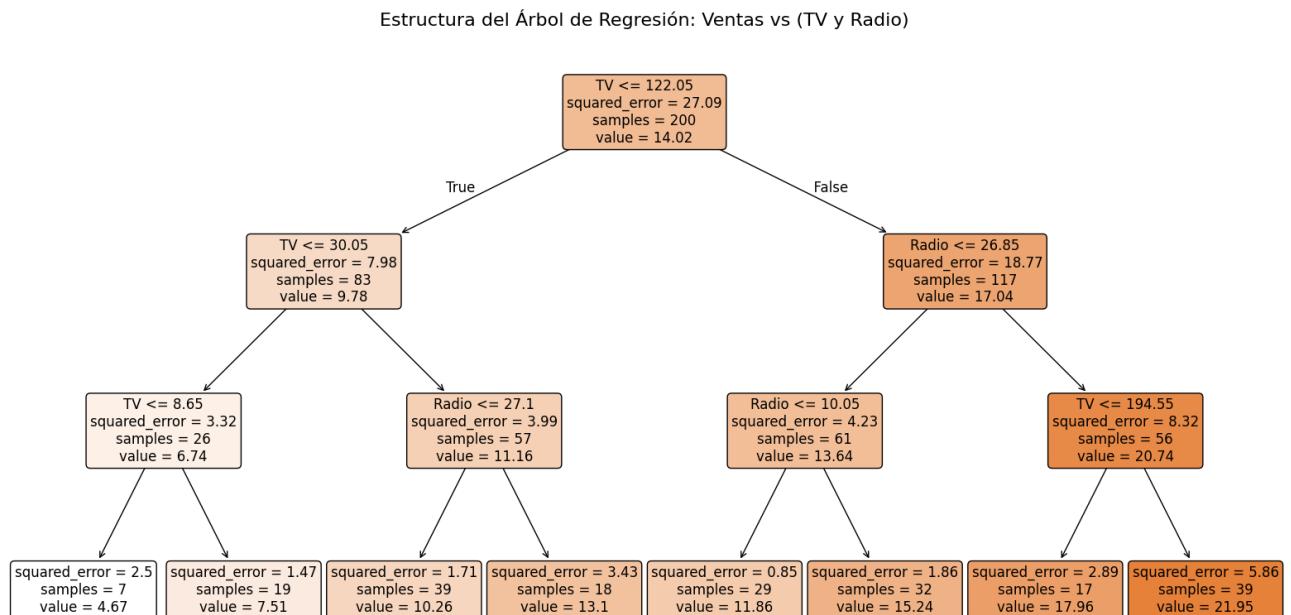


Figura [9]. Árbol de regresión de Tv y Radio frente a las ventas

El árbol de regresión muestra cómo las ventas se segmentan según la inversión en TV y Radio, iniciando por TV, lo que indica que es la variable más influyente. La primera división en $TV \leq 122.05$ separa los mercados en niveles de ventas más bajas y más altas; cuando la inversión en TV es mayor, las ventas tienden a aumentar. En niveles bajos o intermedios de TV, la variable Radio ayuda a ajustar la predicción, evidenciando que las mayores ventas se logran cuando ambas inversiones son altas. En cada nodo, samples indica la cantidad de datos en ese grupo y value es el promedio de ventas, que corresponde a la predicción. Las particiones se realizan minimizando el error cuadrático (MSE), es decir, el modelo divide los datos buscando que las ventas dentro de cada grupo sean lo más similares posible y así reducir el error de predicción.

3.2 Cálculo Manual de Incertidumbre

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

Entropía del Nodo padre

$$p(Altas) = \frac{3}{4} = 0.75$$

$$p(Bajas) = \frac{1}{4} = 0.25$$

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

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

$$H(S) = 0.811$$

Índice de Gini

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

$$Gini = 0.375$$

Ganancia de Información

$$IG = H(S) - \sum \frac{|Sv|}{|S|} H(Sv)$$

Cada grupo tiene dos datos

Grupo 1: Mercado 1 y 2

$$H(S_{Altas}) = 0$$

Nodo puro

Grupo 2: Mercado 3 y 4

$$p = 0.5$$

$$H(S_{Bajas}) = -(0.5 \log_2 0.5 + 0.5 \log_2 0.5)$$

$$H(S_{Bajas}) = 1$$

$$H_N = \frac{2}{4}(0) + \frac{2}{4}(1)$$

$$H_N = 0.5$$

$$IG = 0.811 - 0.5 = 0.311$$

3.3 Importancia de Predictores

En el árbol de decisión, la variable más importante es TV, porque aparece en la primera división del modelo. Eso significa que es la que más ayuda a reducir el error al inicio y la que más influye en cómo se segmentan las ventas. Radio aparece después, ayudando a ajustar la predicción en ciertos niveles de inversión, mientras que Newspaper no tiene un papel relevante.

En la regresión múltiple, el coeficiente de Radio (0.1885) es mayor que el de TV (0.0458), lo que indica que, por cada unidad invertida, Radio tiene mayor efecto marginal. Sin embargo, TV mostró mayor correlación y mayor poder explicativo individual, por eso en el árbol resulta ser la variable principal. En resumen, TV domina la estructura del modelo y Radio complementa el impacto.

3.4 Diagnóstico Comparativo

La regresión lineal es más conveniente cuando la relación entre variables es lineal y se busca mayor precisión en el pronóstico. En este caso, el modelo tuvo un R^2 de 0.8972, lo que muestra un ajuste bastante alto y buena capacidad predictiva. El árbol de decisión es útil cuando se quieren identificar niveles específicos de inversión o reglas claras para tomar decisiones, ya que no depende de una relación estrictamente lineal. Aunque puede ser un poco menos preciso, es más fácil de interpretar. En este análisis, la regresión ofrece mejor precisión general, mientras que el árbol ayuda a entender mejor cómo se comportan las ventas según distintos niveles de inversión.

CONCLUSIONES

El análisis comparativo entre la regresión lineal múltiple y el árbol de decisión permitió evaluar cuál modelo resulta más adecuado para explicar el comportamiento de las ventas en función de la inversión publicitaria.

La regresión lineal múltiple presentó un ajuste superior, con un $R^2=0.8972$, lo que indica que el 89.72 % de la variabilidad de las ventas es explicada por las variables TV, Radio y Newspaper. Además, las fases exploratorias mostraron que las relaciones entre las variables son predominantemente lineales, lo que respalda el uso de un modelo paramétrico. La regresión permite interpretar claramente el efecto marginal de cada medio bajo el principio *ceteris paribus*, facilitando la toma de decisiones cuantitativas.

Por su parte, el árbol de decisión aporta valor en términos de segmentación y visualización de reglas, identificando a TV como la variable más influyente. Sin embargo, dado que no se evidencian patrones no lineales complejos, su ventaja predictiva no supera a la regresión en este caso.

En conclusión, para este conjunto de datos, la regresión lineal múltiple se considera el modelo más robusto, debido a su mayor precisión y capacidad explicativa, mientras que el árbol de decisión cumple un papel complementario de interpretación estratégica.