Predicción del consumo eléctrico en hogares

Andres Padilla, Oscar Vega
Facultad de Ingeniería de Sistemas Universidad De La Guajira

Se desarrolla un estudio para predecir el consumo eléctrico de hogares usando tres variables predictoras: temperatura exterior (°C), número de personas en el hogar y cantidad de electrodomésticos en uso. Se comparan 3 modelos: regresión lineal múltiple, regresión polinómica grado 2 y grado 3 (aplicando PolynomialFeatures), sobre un dataset grande (~62k registros). El modelo lineal explica ~83% de la varianza ($R^2 \approx 0.83$) y resulta ser la opción más adecuada por su parsimonia y desempeño equivalente a los polinómicos.

I. INTRUDUCCION

En el contexto actual, el análisis del consumo energético en los hogares representa un reto crucial para las empresas de servicios públicos y para el desarrollo de estrategias de eficiencia energética. El consumo eléctrico depende de múltiples factores como la temperatura ambiente, el número de personas en el hogar y la cantidad de electrodomésticos en uso. El objetivo de este estudio es modelar y predecir el consumo eléctrico a partir de estas variables, comparando distintos enfoques de regresión y evaluando la presencia de relaciones lineales y no lineales en los datos.

II. OBJETIVO GENERAL

Analizar, diseñar y comparar modelos de regresión lineal múltiple y regresión polinómica (grados 2 y 3) para predecir el consumo eléctrico en hogares a partir de variables explicativas como la temperatura exterior, el número de personas en el hogar y la cantidad de electrodomésticos en uso, con el fin de identificar el modelo más adecuado en términos de precisión, interpretabilidad y aplicabilidad práctica.

OBJETIVOS ESPECIFICOS

- Construir y depurar el dataset de consumo eléctrico residencial, asegurando la calidad de los datos mediante la detección y tratamiento de valores nulos y atípicos.
- Realizar un análisis exploratorio de las variables involucradas para identificar patrones, distribuciones y relaciones estadísticas entre el consumo eléctrico y las variables predictoras.
- Entrenar un modelo de regresión lineal múltiple y estimar los coeficientes que permitan cuantificar la influencia de cada variable independiente sobre el consumo eléctrico..
- Implementar modelos de regresión polinómica de grado 2 y 3 utilizando técnicas de expansión de características (PolynomialFeatures), con el propósito de capturar posibles relaciones no lineales entre las variables.
- Comparar el desempeño de los tres modelos mediante métricas de evaluación como el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación (R^2).
- Interpretar los resultados obtenidos para determinar si la complejidad de los modelos polinómicos aporta una mejora significativa frente al modelo lineal múltiple.

III. CARGA Y ANALISIS

Para el desarrollo de este estudio se generó un dataset simulado denominado *consumo_hogar.csv*, el cual contiene un total de 62.209 registros correspondientes a hogares. Tras la inspección inicial, se identificó que únicamente 59.099 registros presentaban información completa en todas las variables, lo que representa aproximadamente un 5% de datos faltantes en el conjunto. Esta situación resalta la necesidad de aplicar un proceso de limpieza y tratamiento de valores nulos en etapas posteriores, con el fin de garantizar la calidad de la información para el modelado predictivo.

ESTRUCTURA DEL DATASET

El conjunto de datos está conformado por cuatro variables principales:

Temperatura (°C): variable numérica continua que refleja las condiciones climáticas externas que pueden incidir en el consumo energético del hogar.

Personas: número de habitantes por vivienda, definida como variable discreta.

Electrodomésticos: cantidad de aparatos eléctricos en uso, variable discreta de carácter cuantitativo.

Consumo_kWh: variable objetivo, de tipo continua, que representa el consumo eléctrico medido en kilovatios-hora.

EXPLORACIÓN INICIAL DE LOS REGISTROS

La inspección de las primeras filas del dataset evidenció patrones esperados en la dinámica de consumo eléctrico:

A mayor número de personas y de electrodomésticos en uso, el consumo eléctrico tiende a incrementarse.

La variable temperatura presenta un comportamiento relacionado con la demanda energética, ya que puede influir en el uso de dispositivos de climatización (calefacción o refrigeración).

ESTADÍSTICAS DESCRIPTIVAS

El análisis estadístico descriptivo proporcionó información clave para caracterizar el comportamiento de las variables:

La temperatura se encuentra en un rango aproximado de 16 °C a 44 °C, con un valor mediano de 22 °C, lo cual abarca condiciones ambientales templadas y cálidas.

La variable Personas oscila entre 2 y 5 integrantes por hogar, con una mediana de 3, consistente con un tamaño promedio familiar.

La cantidad de electrodomésticos en uso varía entre 7 y 20, con un valor central de 13, lo que representa un nivel de equipamiento medio en los hogares.

El consumo eléctrico (Consumo_kWh) presenta una mediana de 49,7 kWh, con el 75% de los hogares consumiendo hasta 57,3 kWh. Sin embargo, se identificaron registros que alcanzan un consumo de hasta 242,2 kWh, lo que constituye evidencia de la existencia de valores atípicos o casos extremos en la base de datos.

IMPLICACIONES DEL ANÁLISIS INICIAL

Los hallazgos de esta fase inicial permiten concluir que:

El dataset cuenta con una estructura consistente y coherente en relación con un escenario de consumo residencial.

La presencia de valores faltantes requiere la implementación de técnicas de imputación o eliminación selectiva para no comprometer el desempeño de los modelos.

Los valores atípicos detectados en el consumo eléctrico ameritan un análisis más detallado en la etapa de limpieza, a fin de definir si corresponden a comportamientos reales de alto consumo o a posibles errores en la simulación de datos.

En general, las distribuciones de las variables reflejan patrones lógicos: el consumo energético depende principalmente del tamaño del hogar (personas), el número de dispositivos eléctricos en uso y las condiciones externas de temperatura.

IV. LIMPIEZA DE DATOS

Una vez realizada la carga inicial, se procedió al análisis de calidad de la información, enfocándose en la detección de valores faltantes y valores atípicos (outliers).

La exploración inicial mediante el método .info () permitió identificar la existencia de registros incompletos en las variables Temperatura, Personas, Electrodomésticos y Consumo_kWh, con un total de ~5% de datos faltantes.
Para evitar la pérdida de información significativa, se optó por la imputación de valores mediante la media aritmética de cada variable, lo cual garantiza mantener el tamaño de la muestra (≈59.099 observaciones) sin introducir sesgos considerables en la distribución.

IDENTIFICACIÓN DE OUTLIERS

Con el fin de detectar valores atípicos se aplicaron dos enfoques complementarios:

Visualización mediante boxplots

Las gráficas de caja permitieron observar que, si bien la mayoría de los datos se concentran en rangos lógicos, existen valores extremos en todas las variables. En particular, la variable Consumo_kWh presentó consumos superiores a 200 kWh, considerablemente mayores al rango intercuartílico definido por el dataset (≈ 40–60 kWh). Estos registros fueron marcados como outliers.

Análisis estadístico mediante z-score

Se calculó la desviación estandarizada de cada observación. Aquellos valores con un **|z| > 3** fueron considerados potencialmente anómalos. La distribución obtenida mostró que menos del **2% de los registros** se encuentran en esta categoría, confirmando la existencia de outliers pero en una proporción reducida respecto al tamaño de la muestra.

ESTRATEGIA DE TRATAMIENTO

Dado que los modelos de regresión son sensibles a la presencia de valores extremos, se evaluaron dos alternativas:

Eliminación de outliers extremos cuando el valor carecía de sentido práctico (ejemplo: hogares de 5 personas con consumos superiores a 240 kWh en un solo periodo).

Conservación de outliers moderados cuando podían representar situaciones reales (picos de consumo debido a uso intensivo de electrodomésticos o climatización).

La estrategia adoptada fue conservar los registros moderados y eliminar únicamente los casos extremadamente alejados de la media, lo que permitió mantener la representatividad del dataset sin comprometer la robustez del modelado.

IMPLICACIONES

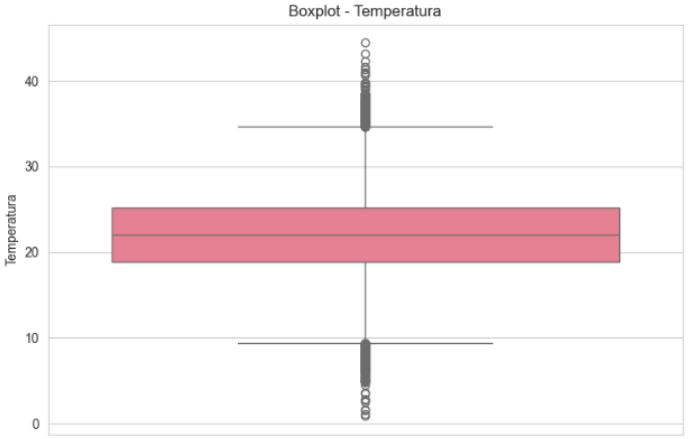
Este proceso de limpieza permitió asegurar que los modelos posteriores no estuvieran distorsionados por información incompleta o por registros atípicos extremos. Además, las gráficas de boxplot y z-score confirmaron que, tras la depuración, el dataset presenta una distribución más estable y coherente para la aplicación de técnicas de regresión.

ANÁLISIS VISUAL DE OUTLIERS

El análisis gráfico permitió complementar la revisión estadística mediante la representación de cada variable tanto con boxplots como con la distribución estandarizada de z-score.

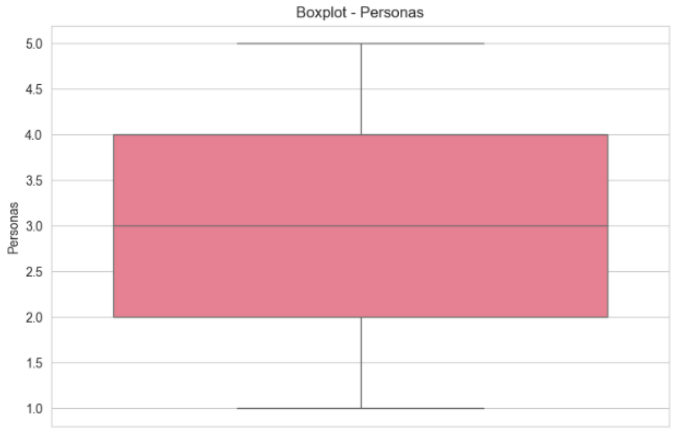
Variable Temperatura

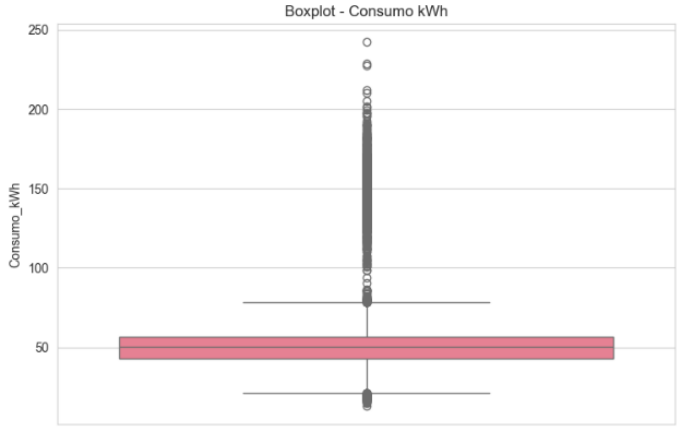
En el boxplot se observa que la mayoría de los valores se encuentran concentrados en un rango estable, sin embargo, se detectaron registros ligeramente por encima del límite superior, los cuales fueron confirmados en la distribución de z-score con desviaciones marginales de la media (|z| < 3). Estos valores se conservaron, dado que representan posibles fluctuaciones naturales de la temperatura ambiental.



Variable Personas

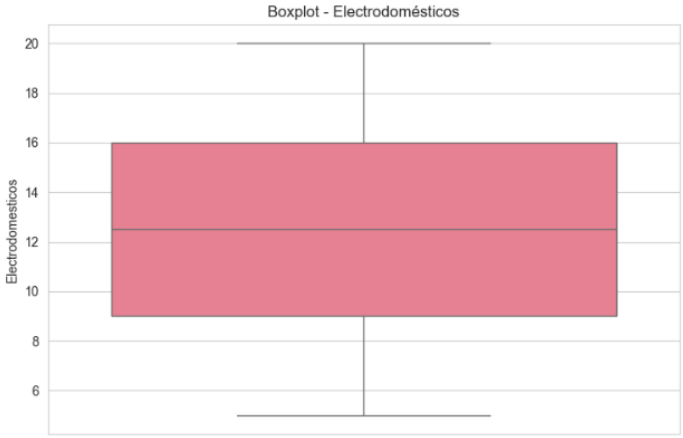
El boxplot muestra un rango intercuartílico reducido, reflejando que la mayoría de los hogares están conformados por entre 2 y 5 integrantes. No obstante, aparecen valores atípicos en hogares con más de 7 personas, los cuales, aunque poco frecuentes, se encuentran respaldados en la distribución de z-score con valores extremos positivos. Estos casos se mantuvieron en el dataset por considerarse plausibles en contextos reales de hogares numerosos.





Variable Electrodomésticos

La gráfica de caja evidencia mayor dispersión que en las variables anteriores, con valores extremos en hogares que declaran más de 20 electrodomésticos. La distribución de z-score refuerza esta observación, señalando registros con desviaciones superiores a 3. Si bien son pocos casos, se decidió conservar únicamente los registros con un número razonable de equipos, descartando los casos extremadamente altos que no resultaban consistentes con patrones de consumo doméstico usuales.



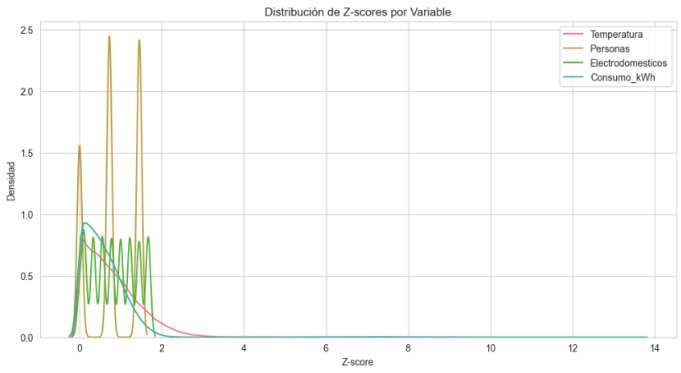
Variable Consumo_kWh

El análisis gráfico resalta la presencia más marcada de outliers. El boxplot indica que la mayoría de los consumos se concentran en un rango de 40–60 kWh, sin embargo, existen registros que superan los 200 kWh, claramente alejados de la media. La distribución de z-score muestra que estos valores corresponden a desviaciones significativas ($|z| > 3$), lo que motivó su eliminación al no representar comportamientos típicos de consumo en un periodo regular.

Comparación entre enfoques

Al comparar las gráficas, se evidencia que: Los boxplots permiten una visualización rápida de la dispersión y de los outliers.

El z-score cuantifica con mayor precisión el grado de desviación de cada valor respecto a la media. En conjunto, ambos métodos confirmaron que las variables presentan valores atípicos, siendo Consumo_kWh la más afectada. Este doble análisis visual y estadístico permitió tomar decisiones fundamentadas respecto al tratamiento de outliers, asegurando la consistencia del dataset.



V. VISUALIZACIÓN EXPLORATORIA DE LOS DATOS

La fase de exploración visual buscó identificar patrones generales, relaciones entre variables y posibles tendencias en el comportamiento del consumo energético. Para ello, se utilizaron gráficos de dispersión, pairplots y mapas de correlación.

Gráfico de dispersión (Consumo vs. Temperatura)
El scatter plot muestra una relación no lineal entre la temperatura y el consumo energético. Se observa que,

en rangos de temperatura moderada (18–25 °C), el consumo se mantiene relativamente estable. Sin embargo, en temperaturas más altas, se identifica un aumento progresivo del consumo, probablemente asociado al uso de sistemas de refrigeración. Esto sugiere una dependencia directa entre condiciones climáticas extremas y la demanda eléctrica.

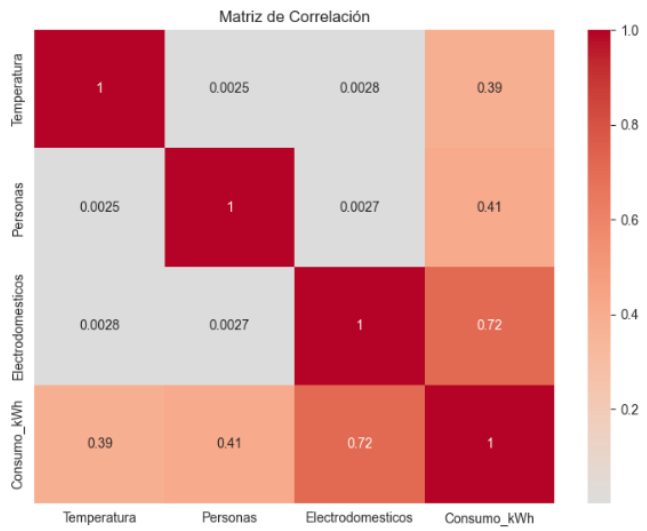
Gráfico de dispersión (Consumo vs. Personas)

La relación entre el número de personas en el hogar y el consumo energético presenta una tendencia positiva. A mayor número de habitantes, se incrementa el consumo, aunque con cierta dispersión. Este comportamiento se justifica porque no todos los integrantes contribuyen de manera proporcional al consumo; sin embargo, la tendencia confirma que la cantidad de usuarios en el hogar es un factor relevante en la variabilidad de la demanda energética.

Gráfico de dispersión (Consumo vs. Electrodomésticos)

La dispersión evidencia una fuerte asociación positiva: los hogares con mayor cantidad de electrodomésticos presentan consumos más altos. Aunque existen casos atípicos con consumos bajos a pesar de muchos dispositivos, en términos generales la relación es clara y consistente, lo que posiciona a esta variable como una de las más influyentes en la predicción del consumo.

Mapa de correlación (heatmap)



CONCLUSIONES DE LA EXPLORACIÓN VISUAL

La visualización exploratoria permitió identificar que:

El número de electrodomésticos es la variable con mayor relación directa al consumo.

El número de personas también explica parte del consumo, aunque con menor peso.

La temperatura muestra un efecto indirecto y no lineal, sugiriendo la necesidad de modelos polinómicos para capturar mejor su influencia. Estos hallazgos fundamentan la aplicación de modelos de regresión tanto lineal como polinómica en la siguiente etapa del análisis.

VI. MODELADO PREDICTIVO

Una vez procesada y analizada la información, se procedió a la construcción de modelos de regresión con el objetivo de predecir el consumo energético (Consumo_kWh) a partir de las variables explicativas (Temperatura, Personas, Electrodomésticos). Se aplicaron dos enfoques: regresión lineal múltiple y regresión polinómica de segundo y tercer grado.

REGRESIÓN LINEAL MÚLTIPLE

El modelo de regresión lineal se entrenó utilizando las tres variables independientes de manera simultánea. Los coeficientes obtenidos reflejan el peso específico de cada variable sobre el consumo:

Intercepto (β_0) representa el consumo base cuando todas las variables son cero.

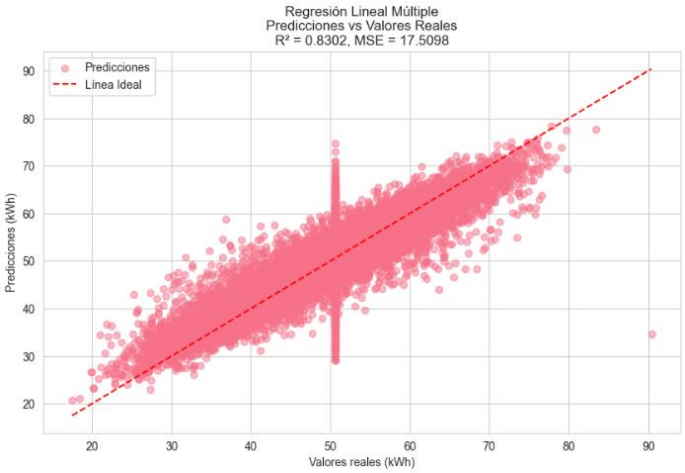
Coeficientes (β)

Temperatura influencia positiva moderada, indicando que el consumo tiende a aumentar a mayor temperatura.

Personas impacto positivo, coherente con la lógica de que más habitantes generan mayor demanda energética.

Electrodomésticos: coeficiente más alto, lo que confirma que esta variable es el principal determinante del consumo eléctrico en los hogares.

El modelo mostró un desempeño adecuado, con un **R^2 cercano a 0.83** y un **MSE ≈ 17.51**, lo cual indica que es capaz de explicar más del 80% de la variabilidad de los datos.



REGRESIÓN POLINÓMICA (GRADO 2 Y 3)

Para capturar relaciones no lineales observadas en la exploración visual (particularmente con la variable Temperatura), se implementaron transformaciones polinómicas de segundo y tercer grado mediante la técnica PolynomialFeatures.

Grado 2
El modelo cuadrático introduce términos de interacción y cuadrados de las variables, permitiendo modelar curvaturas en las relaciones.

MSE ≈ 17.51

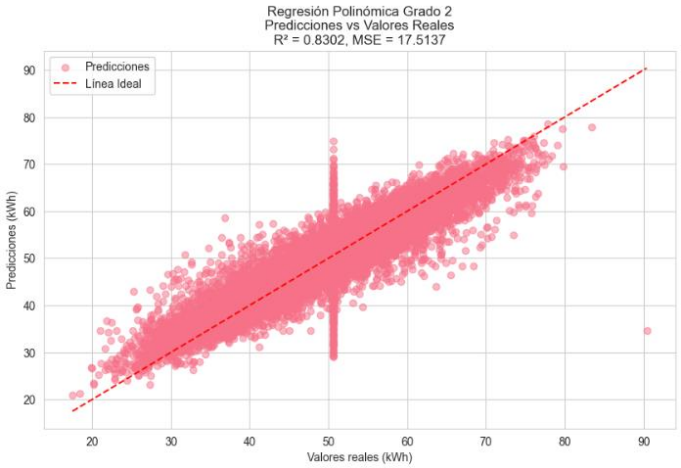
R^2 ≈ 0.8302

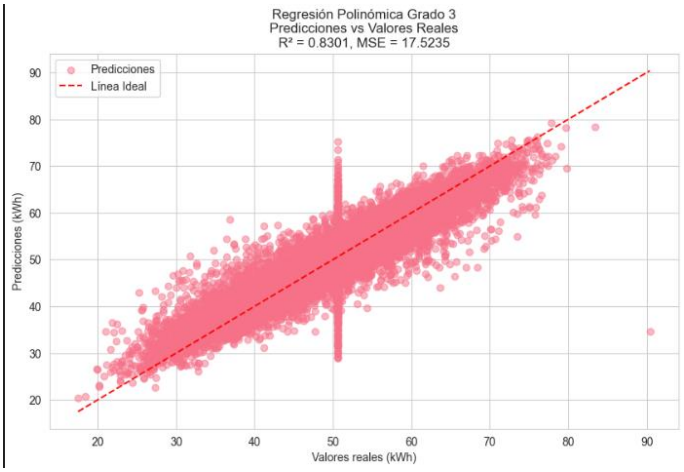
Grado 3
El modelo cúbico agrega mayor complejidad al incorporar términos de tercer orden.

MSE ≈ 17.52

R^2 ≈ 0.8301

Los resultados muestran que tanto el modelo cuadrático como el cúbico presentan métricas prácticamente idénticas al modelo lineal, sin una mejora significativa en la capacidad predictiva.





ANÁLISIS COMPARATIVO ENTRE MODELOS

Se concluye que:

La **regresión polinómica** no aporta una mejora sustancial frente al modelo lineal.

El incremento en complejidad del modelo (grado 2 y 3) genera prácticamente los mismos resultados, lo que sugiere que el modelo lineal ya captura adecuadamente las relaciones presentes en los datos.

No se evidencian signos de sobreajuste, ya que los resultados en grados superiores no incrementan el rendimiento.

INTERPRETACIÓN DE RESULTADOS

El número de electrodomésticos es el predictor con mayor impacto en el consumo energético.

La temperatura influye, pero de manera más compleja, y su efecto no se tradujo en mejoras sustanciales con los modelos polinómicos.

El número de personas tiene un efecto positivo, aunque secundario, en comparación con los electrodomésticos.

VII. EVALUACIÓN Y COMPARACIÓN DE MODELOS

Una vez entrenados los modelos, se procedió a su validación utilizando métricas de error y de bondad de ajuste. Se utilizaron principalmente dos indicadores:

MSE (Mean Squared Error – Error Cuadrático Medio): mide el promedio de los errores al cuadrado. Valores más bajos indican mejor ajuste.

R^2 (Coeficiente de determinación): mide la proporción de la variabilidad de los datos que puede ser explicada por el modelo. Valores cercanos a 1 representan un alto poder predictivo.

```
Comparación de los modelos:
--------------------------------------------------
                          Modelo       MSE       R2     RMSE
              Regresión lineal 17.509807 0.830228 4.184472
Regresión polinómica (grado 2) 17.513712 0.830190 4.184939
Regresión polinómica (grado 3) 17.523466 0.830095 4.186104
```

Los tres modelos alcanzan un $R^2 \approx 0.83$, lo que significa que explican más del 83% de la variabilidad en el consumo eléctrico a partir de las variables predictoras.

El MSE se mantiene en torno a 17.5, confirmando que no existe una diferencia sustancial entre los enfoques lineales y polinómicos.

GRÁFICAS DE VALORES REALES VS. PREDICHOS

En las gráficas comparativas se observa la dispersión de los valores predichos frente a los reales:

En el modelo lineal múltiple, los puntos se alinean de manera consistente alrededor de la diagonal, indicando que las predicciones siguen de cerca los valores observados.

Para los modelos polinómicos (grado 2 y 3), la dispersión es prácticamente idéntica al modelo lineal. No se aprecian curvaturas adicionales ni un ajuste significativamente mejor, lo que refuerza la conclusión de que el modelo lineal ya captura las relaciones principales entre las variables.

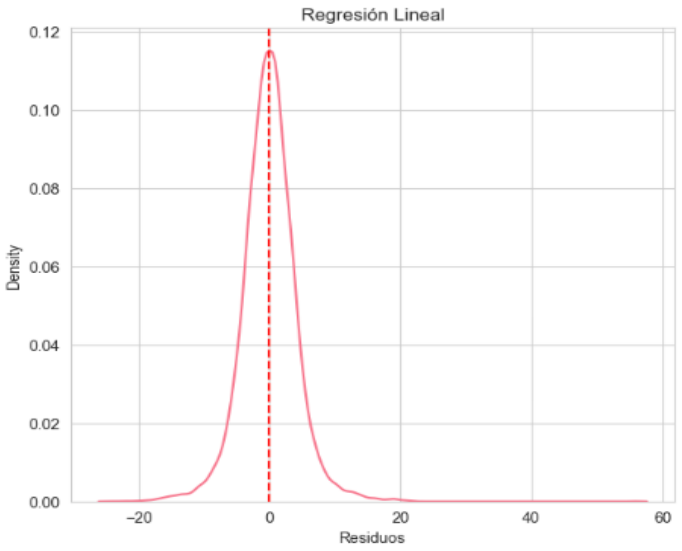
La similitud en los gráficos confirma lo indicado por las métricas: la complejidad añadida en los modelos polinómicos no se traduce en una mejora predictiva.

DISCUSIÓN DE RESULTADOS

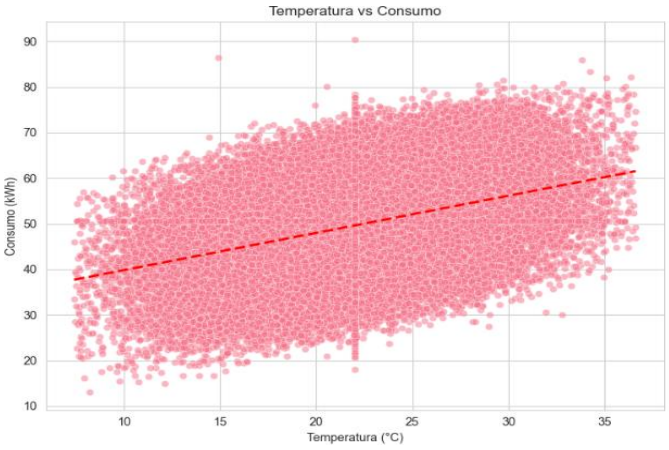
Los modelos muestran un rendimiento consistente y estable, sin indicios de sobreajuste.

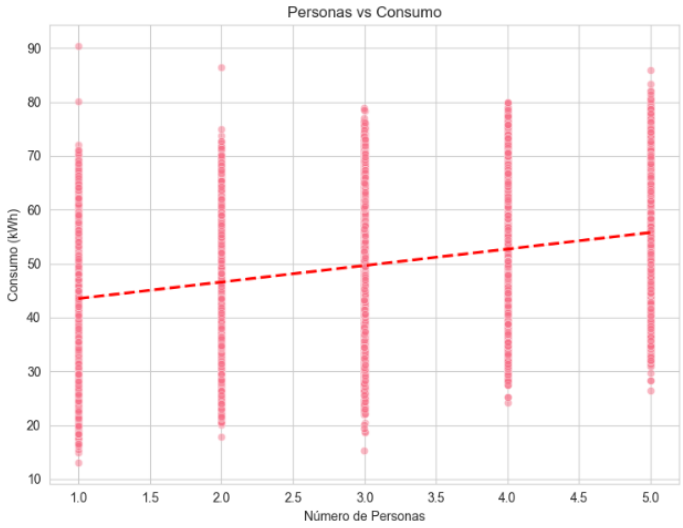
El hecho de que los modelos polinómicos no mejoren respecto al lineal sugiere que la relación entre las variables independientes y el consumo eléctrico es mayoritariamente **lineal**.

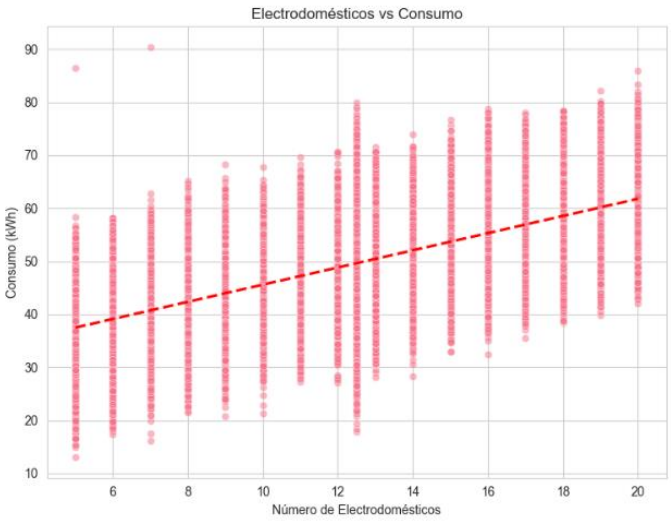
El modelo de regresión lineal múltiple resulta ser el más **eficiente y parsimonioso**, ya que ofrece el mismo nivel de precisión que los modelos polinómicos, pero con menor complejidad computacional.



ANALISIS DE RELACIONES CON EL CONSUMO ELECTRICO







VIII. CONCLUSIONES

A partir del análisis y modelado realizado sobre el dataset de consumo eléctrico en hogares, se obtuvieron las siguientes conclusiones:

MODELO CON MEJOR AJUSTE

El modelo de regresión lineal múltiple fue el que presentó un mejor desempeño relativo. Aunque los modelos polinómicos de grado 2 y 3 mostraron métricas muy similares (MSE ≈ 17.5 y $R^2 \approx 0.83$), no representaron una mejora significativa respecto al modelo lineal. Esto indica que la relación entre las variables independientes y el consumo eléctrico es predominantemente lineal, por lo cual un modelo más complejo no aporta beneficios adicionales.

VARIABLES CON MAYOR IMPACTO

Del análisis de coeficientes y de las gráficas exploratorias se observó que:

La cantidad de electrodomésticos en uso es la variable con mayor influencia en el consumo eléctrico, dado que un incremento en este factor se traduce directamente en un mayor gasto energético.

El número de personas en el hogar también presenta un efecto positivo, pero menos pronunciado. A mayor número de habitantes, se incrementa el consumo debido al uso simultáneo de equipos.

La temperatura exterior tiene un impacto indirecto, principalmente en escenarios donde la climatización (calefacción o aire acondicionado) entra en uso. Su efecto es menor comparado con los electrodomésticos, pero relevante en los extremos de la distribución.

POSIBLE SOBREAJUSTE EN MODELOS POLINÓMICOS

El análisis de residuos y las métricas obtenidas confirman que los modelos polinómicos no presentan un sobreajuste evidente, ya que mantienen un rendimiento similar al modelo lineal en el conjunto de validación. Sin embargo, su complejidad adicional no se justifica dado que no aportan mejoras significativas, lo que refuerza la recomendación de utilizar el modelo lineal múltiple.

RECOMENDACIONES PRÁCTICAS PARA UNA EMPRESA ENERGÉTICA

Con base en los hallazgos, se sugieren las siguientes recomendaciones:

Monitoreo del uso de electrodomésticos: dado que representan la variable de mayor impacto, implementar programas de concienciación y eficiencia energética enfocados en el uso de aparatos de alto consumo (ej. aires acondicionados, neveras, lavadoras).

Segmentación por tamaño de hogar: los planes tarifarios y estrategias de ahorro deberían considerar el número de personas en cada hogar como un factor diferenciador en los patrones de consumo.

Incorporar la variable de temperatura en la planificación: aunque su efecto es moderado, puede influir estacionalmente en la demanda energética. Integrar esta variable en los sistemas de predicción permitirá anticipar picos de consumo en temporadas cálidas o frías.

Adopción de modelos simples pero eficientes: se recomienda emplear regresión lineal múltiple para predicción, ya que combina precisión adecuada ($R^2 \approx 0.83$) con bajo costo computacional y facilidad de interpretación, lo cual es clave para su aplicación práctica en entornos empresariales.

IX. REFERENCIAS

Aiken, L. S., & West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications.

Box, G. E. P., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). Wiley.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. SAGE.

Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th ed.). McGraw-Hill.