MÁSTER EN BIG DATA Y DATA SCIENCE

06MBID Estadística avanzada

CURSO 2024-2025

ACTIVIDAD 1: Regresión

Alumno:
Alex Anthony Prieto Romani

Contenido

Introducción

El presente informe analiza la relación entre el consumo de energía eléctrica y sus costos asociados, utilizando tres modelos de regresión: lineal simple, lineal múltiple y polinómica. Este análisis busca entender cómo diversas variables predictoras afectan el consumo en kilovatios-hora (Kwh) y su costo monetizado en soles (S/). La fuente de los datos es la **Plataforma Nacional de Datos Abiertos del Perú**, que proporciona acceso a información pública para promover la transparencia y el uso eficiente de los datos en la toma de decisiones.

Contexto y Motivación

La demanda de energía eléctrica y su gestión eficiente son aspectos cruciales para el desarrollo económico y social de cualquier región. La comprensión de los patrones de consumo energético y los factores que influyen en los costos es esencial para diseñar políticas tarifarias justas y estrategias de eficiencia energética. En este contexto, los modelos de regresión se presentan como herramientas fundamentales para el análisis y la predicción del consumo energético, permitiendo una evaluación precisa de cómo diversas variables influyen en los costos asociados.

En Perú, la energía eléctrica es un recurso vital, y su demanda ha crecido significativamente en las últimas décadas, impulsada por el crecimiento industrial, urbano y el acceso extendido a comunidades rurales (Ministerio de Energía y Minas, 2023). Sin embargo, este crecimiento también plantea desafíos relacionados con la gestión eficiente de los recursos energéticos y la optimización de tarifas para los usuarios finales. La **Sociedad Eléctrica del Sur Oeste S.A. (SEAL S.A.)** desempeña un papel clave en la distribución de energía en la región de Arequipa, y la disponibilidad de datos detallados sobre el consumo de sus clientes proporciona una oportunidad invaluable para analizar y mejorar estos procesos.

Estudios previos han demostrado la importancia de aplicar modelos de análisis predictivo en la gestión de servicios públicos. Por ejemplo, investigaciones en otros contextos han revelado que la modelación estadística del consumo de energía puede ayudar a las empresas de servicios públicos a identificar patrones de consumo, detectar anomalías y diseñar tarifas más equitativas (Anderson, W., Lee, K., & Mallik, S., 2021; Zhang, Y., & Wang, J., 2020). Inspirados por estos enfoques, este análisis busca proporcionar insights específicos para SEAL S.A. y contribuir al entendimiento de los factores que determinan los costos de consumo eléctrico.

Objetivos del Análisis

El presente informe tiene como objetivos:

1. **Evaluar la relación entre el consumo de energía (en Kwh) y su costo monetizado en soles (S/).**
    - A través de una regresión lineal simple, se busca establecer una relación directa y cuantificable entre estas dos variables clave.

2. **Determinar cómo múltiples variables predictoras influyen en el costo energético.**
    - La regresión lineal múltiple permitirá incorporar variables adicionales como la tarifa y el período de consumo, ofreciendo una visión más completa de los factores que afectan los costos.
3. **Explorar relaciones no lineales para capturar la complejidad en la variabilidad del consumo y el costo.**
    - La regresión polinómica se utilizará para identificar patrones no lineales en los datos, proporcionando un modelo de ajuste más preciso cuando las relaciones lineales no sean suficientes.

Metadatos del Dataset

El dataset utilizado en este análisis incluye información detallada sobre el consumo de energía eléctrica de los clientes de la **Sociedad Eléctrica del Sur Oeste S.A. (SEAL S.A.)**, correspondiente al mes de mayo de 2024. A continuación, se presenta una descripción de los metadatos del dataset:

- **Título:** Consumo de Energía Eléctrica de los clientes de Sociedad Eléctrica del Sur Oeste S.A.
- **URL:** Consumo de Energía Eléctrica de los clientes de SEAL S.A.
- **Descripción:** Registro mensual del consumo de energía eléctrica de usuarios en la región Arequipa. Cada registro corresponde a un suministro y la lectura se realiza al final de cada mes. Las tarifas aplicadas a los usuarios son determinadas por el ente regulador OSINERGMIN.
- **Entidad:** Sociedad Eléctrica del Sur Oeste S.A.
- **Fuente:** Gerencia de Comercialización
- **Etiquetas:** Electricidad, Energía Eléctrica, Consumo, Luz
- **Frecuencia de Actualización:** Mensual
- **Última Actualización:** 7 de junio de 2024
- **Licencia:** Open Data Commons Attribution License
- **Nivel de Acceso Público:** Público
- **Formato:** CSV
- **Cobertura:** Departamento de Arequipa, 2024

La elección de este dataset responde a la necesidad de analizar datos actuales y relevantes para la región, proporcionando una base sólida para evaluar patrones de consumo y factores que influyen en los costos energéticos.

Descripción de los Datos

El dataset incluye las siguientes variables clave:

- **CodigoSuministro:** ID de suministro del usuario (Numérico, 8 dígitos).
- **NombreDepartamento, NombreProvincia, NombreDistrito:** Ubicación geográfica de los usuarios (Texto, 60 caracteres cada uno).
- **CodigoUbigeo:** Código de ubicación geográfica según el INEI (Numérico, 6 dígitos).
- **FechaInicio:** Fecha de instalación del suministro (Formato: ddmmaaaa).
- **Tarifa:** Nombre de la tarifa vigente (Alfanumérico, 6 caracteres).
- **Periodo:** Periodo de consumo de luz (Formato: aaaamm).
- **ConsumoKwh:** Consumo de energía eléctrica en Kwh (Numérico, 20).
- **Energia_Soles:** Consumo de energía eléctrica monetizado en soles (Numérico, 20 con dos decimales).
- **Estado:** Estado del Cliente (Texto, 6 caracteres).
- **FechaCorte:** Fecha en que se generó la data (Formato: aaaammdd).

La motivación de este análisis es identificar patrones de consumo y factores determinantes del costo de la energía, lo cual es crucial para la optimización de tarifas y mejora de la eficiencia energética.

Metodología

Se emplearán los siguientes modelos de regresión para explorar las relaciones entre las variables:

1. **Regresión Lineal Simple:** Analiza la relación directa entre el consumo en Kwh y el costo en soles.
2. **Regresión Lineal Múltiple:** Incluye múltiples variables predictoras para explicar el costo energético.
3. **Regresión Polinómica:** Explora relaciones no lineales para capturar mejor la variabilidad de los datos.

Análisis de Regresión

Antes de realizar las pruebas de regresión, presentamos la tabla de correlación y el gráfico de correlación para evaluar la relación entre las variables.

Tabla de Correlación

La siguiente tabla muestra la matriz de correlación de Pearson entre las variables ConsumoKwh, Energia_Soles y DiasConsumo:

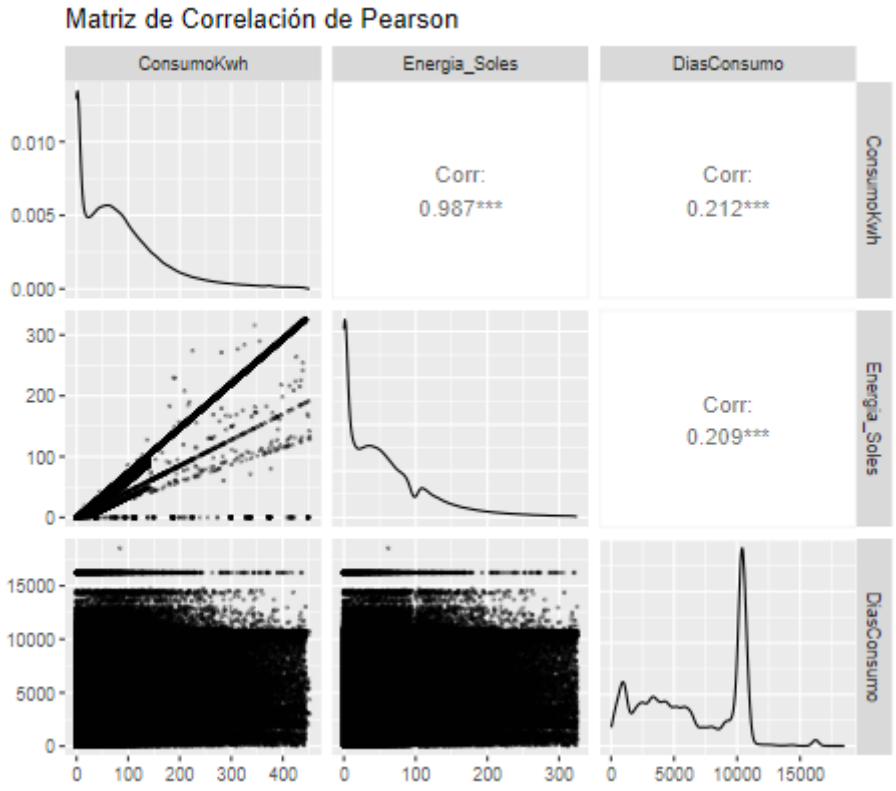
	ConsumoKwh	Energia_Soles	DiasConsumo
ConsumoKwh	1.00	0.98	0.21
Energia_Soles	0.98	1.00	0.21
DiasConsumo	0.21	0.21	1.00

Interpretación de la Tabla de Correlación

- **ConsumoKwh y Energia_Soles:** Tienen una correlación muy alta (0.9868), lo que indica una fuerte relación lineal positiva entre el consumo de energía en Kwh y su costo monetizado en soles.
- **ConsumoKwh y DiasConsumo:** Tienen una correlación baja (0.2121), sugiriendo una débil relación lineal.
- **Energia_Soles y DiasConsumo:** También presentan una correlación baja (0.2091), indicando que el costo de la energía y los días de consumo no están fuertemente relacionados.

Gráfico de Correlación

El gráfico de correlación visualiza estas relaciones de manera gráfica, mostrando tanto la fuerza como la dirección de las correlaciones entre las variables.



1. Regresión Lineal Simple

Objetivo: Evaluar la relación directa entre el consumo de energía (ConsumoKwh) y su costo monetizado en soles (Energia_Soles).

Resultados del Análisis de Regresión Lineal Simple

a. Resumen de los Residuos

Métrica	Valor
Mínimo	-323.25
1er Cuartil (1Q)	-3.91
Mediana	3.73
3er Cuartil (3Q)	4.06
Máximo	113.67

Interpretación:
Los residuos del modelo se concentran cerca de cero, con una mediana de 3.73, lo que indica que la mayoría de los errores del modelo son pequeños. Sin embargo, los valores extremos como el mínimo (-323.25) y el máximo (113.67) sugieren la presencia de algunos outliers, aunque estos no afectan significativamente la calidad general del ajuste.

b. Coeficientes del Modelo

Coeficiente	Estimación	Error Estándar	Valor t	Valor p	Significancia
Intercepto	-3.7259485	0.0204589	-182.1	< 2e-16	***
ConsumoKwh	0.7285496	0.0001722	4231.1	< 2e-16	***

Significancia de los Coeficientes:

- **Intercepto:** Aunque estadísticamente significativo ($p < 2e\text{-}16$), el valor del intercepto de -3.7259485 no tiene un significado práctico directo en este contexto, ya que un consumo de energía de cero no es un escenario realista.
- **ConsumoKwh:** El coeficiente de 0.7285 indica que por cada unidad adicional de consumo en Kwh, se espera que el costo en soles aumente en aproximadamente 0.7285 soles. Este coeficiente es altamente significativo ($p < 2e\text{-}16$), indicando una fuerte relación positiva entre el consumo y el costo.

c. Estadísticas del Modelo

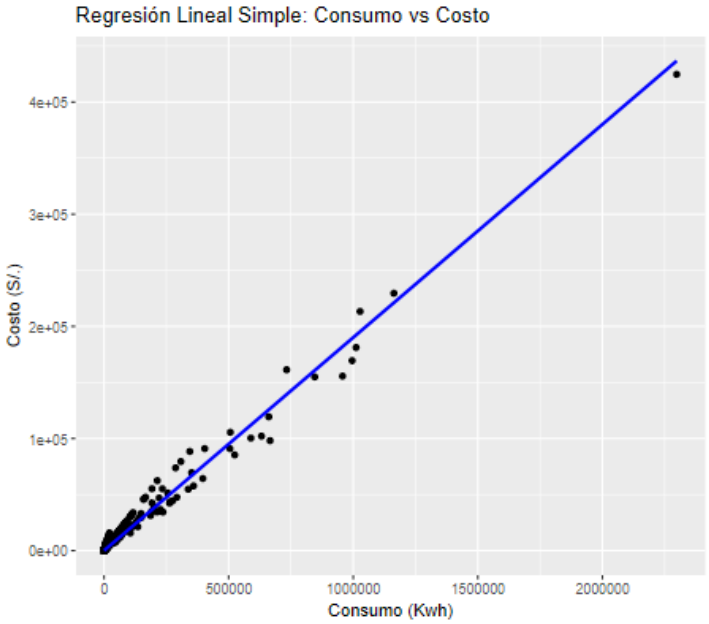
Métrica	Valor
Error Estándar Residual	9.754
Grados de Libertad	480049
R-cuadrado	0.9739

Métrica	Valor
R-cuadrado Ajustado	0.9739
Estadístico F	1.79e+07
Valor p del Estadístico F	< 2.2e-16

Interpretación del Modelo:

- **R-cuadrado y R-cuadrado Ajustado (**0.9739**):** Indican que aproximadamente el 97.39% de la variabilidad en el costo (Energia_Soles) puede ser explicada por el consumo (ConsumoKwh). Esto sugiere un ajuste muy bueno del modelo, con la mayoría de la variabilidad del costo siendo explicada por el consumo.
- **Error Estándar Residual:** El valor de 9.754 refleja que los residuos son relativamente pequeños, lo que indica que las predicciones del modelo están generalmente cerca de los valores reales.
- **Estadístico F y Valor p:** Con un estadístico F de 1.79e+07 y un valor p < 2.2e-16, se confirma que el modelo es altamente significativo en su conjunto.

Gráfico de Regresión Lineal Simple



Conclusión

El modelo de regresión lineal simple confirma que el consumo de energía en Kwh es un predictor fuerte y significativo del costo monetizado en soles. El alto valor de R-cuadrado indica que el modelo captura la mayoría de la variabilidad del costo basada en el consumo, validando la eficacia de la regresión lineal simple para este análisis. Este modelo puede ser utilizado para hacer predicciones del costo basado en el consumo con un alto nivel de confianza, dado el excelente ajuste y la significancia estadística de los coeficientes.

2. Regresión Lineal Múltiple

Objetivo: Determinar el impacto de múltiples variables (ConsumoKwh, Tarifa, y DiasConsumo) en el costo energético (Energia_Soles).

a. Resumen de los Residuos

Métrica	Valor
Mínimo	-200.571
1er Cuartil (1Q)	-3.989
Mediana	3.043
3er Cuartil (3Q)	3.679
Máximo	188.134

Interpretación: Los residuos del modelo se centran cerca de cero, con una mediana de 3.043, lo que sugiere que la mayoría de los errores del modelo son pequeños. Sin embargo, la presencia de valores extremos, como el mínimo (-200.571) y el máximo (188.134), indica la existencia de outliers, aunque estos no impactan significativamente la precisión general del modelo.

b. Coeficientes del Modelo

Coeficiente	Estimación	Error Estándar	Valor t	Valor p	Significancia
Intercepto	-53.71	1.087	-49.426	< 2e-16	***
ConsumoKwh	0.7364	0.0001215	6062.031	< 2e-16	***
TarifaBT3	21.45	1.491	14.381	< 2e-16	***
TarifaBT4	13.21	1.403	9.418	< 2e-16	***
TarifaBT5A	40.58	1.278	31.744	< 2e-16	***
TarifaBT5B	50.19	1.087	46.189	< 2e-16	***
TarifaBT5D	-13.18	1.668	-7.898	2.84e-15	***
TarifaBT6	-133.80	1.116	-119.879	< 2e-16	***
TarifaMT2	-8.809	1.821	-4.837	1.32e-06	***
TarifaMT3	-9.726	1.461	-6.656	2.82e-11	***
TarifaMT4	-11.49	2.043	-5.628	1.83e-08	***
DiasConsumo	-0.00009459	0.000002630	-35.973	< 2e-16	***

Significancia de los Coeficientes:

- **Intercepto:** El intercepto negativo (-53.71) es estadísticamente significativo (p < 2e-16), aunque su interpretación práctica puede ser limitada.
- **ConsumoKwh:** El coeficiente de 0.7364 indica que por cada unidad adicional de consumo en Kwh, se espera que el costo en soles aumente en aproximadamente

0.7364 soles. Este coeficiente es altamente significativo ($p < 2e-16$), lo que confirma una fuerte relación positiva entre el consumo y el costo.

- **Tarifas:** Todas las tarifas tienen coeficientes significativamente diferentes del valor base, ajustando el costo en función del tipo de tarifa aplicada.
- **DiasConsumo:** El coeficiente negativo para DiasConsumo indica una ligera reducción del costo con más días de consumo, aunque el efecto es pequeño, pero estadísticamente significativo ($p < 2e-16$).

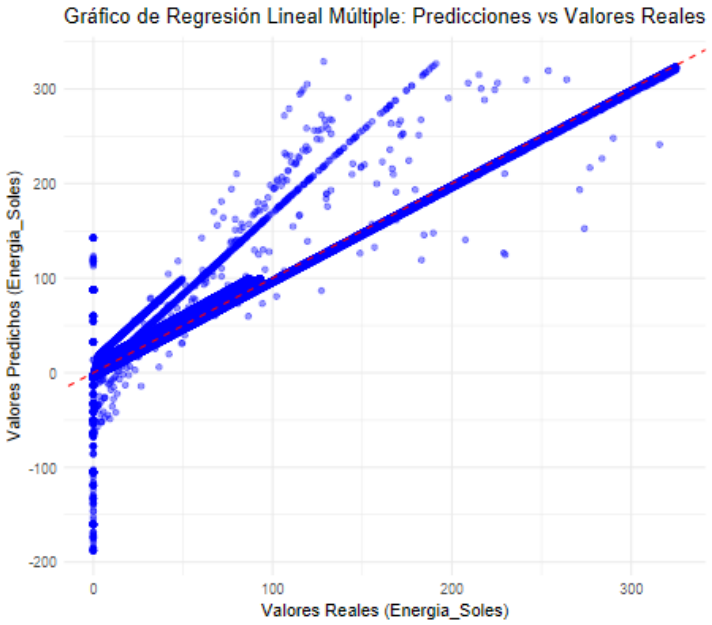
c. Estadísticas del Modelo

Métrica	Valor
Error Estándar Residual	6.698
Grados de Libertad	480039
R-cuadrado	0.9877
R-cuadrado Ajustado	0.9877
Estadístico F	3.5e+06
Valor p del Estadístico F	< 2.2e-16

Interpretación del Modelo:

- **R-cuadrado y R-cuadrado Ajustado (0.9877):** Estos valores indican que aproximadamente el 98.77% de la variabilidad en el costo (Energia_Soles) puede ser explicada por las variables ConsumoKwh, Tarifa, y DiasConsumo, lo que sugiere un ajuste excelente del modelo.
- **Error Estándar Residual:** El valor de 6.698 sugiere que los residuos son relativamente bajos, indicando una buena precisión en las predicciones del modelo.
- **Estadístico F y Valor p:** Un F-estadístico muy alto (3.5e+06) con un valor p extremadamente bajo (< 2.2e-16) confirma que el modelo es altamente significativo en su conjunto.

Gráfico de Regresión Lineal Simple



Conclusión

El modelo de regresión lineal múltiple muestra que ConsumoKwh, Tarifa, y DiasConsumo tienen un impacto significativo en el costo energético. El alto R-cuadrado ajustado indica que el modelo captura una gran proporción de la variabilidad en el costo, validando la robustez y eficacia del modelo para predecir costos basados en estas variables. Las diferentes tarifas influyen significativamente, ajustando el costo en función del tipo de tarifa, mientras que el consumo y los días de consumo también juegan roles importantes en la determinación del costo total de la energía. Este análisis proporciona una base sólida para la toma de decisiones sobre la gestión del consumo y la facturación energética.

3. Regresión Polinómica

Objetivo: Explorar relaciones no lineales entre el consumo de energía (ConsumoKwh) y su costo monetizado en soles (Energia_Soles), considerando también las variables Tarifa y DiasConsumo.

a. Resumen de los Residuos

Métrica	Valor
Mínimo	-206.671
1er Cuartil (1Q)	-3.759
Mediana	2.867
3er Cuartil (3Q)	3.802
Máximo	188.832

Interpretación:
Los residuos del modelo están centrados cerca de cero, con una mediana de 2.867, lo que

indica que la mayoría de los errores del modelo son pequeños. Sin embargo, hay valores extremos como el mínimo (-206.671) y el máximo (188.832), que sugieren la presencia de algunos outliers. A pesar de estos outliers, la baja dispersión alrededor de cero sugiere un buen ajuste del modelo.

b. Coeficientes del Modelo

Coeficiente	Estimación	Error Estándar	Valor t	Valor p	Significancia
Intercepto	8.734	1.080	8.086	6.20e-16	***
poly(ConsumoKwh, 2)1	41710	6.841	6096.942	< 2e-16	***
poly(ConsumoKwh, 2)2	514.0	6.720	76.488	< 2e-16	***
TarifaBT3	21.74	1.482	14.668	< 2e-16	***
TarifaBT4	13.53	1.394	9.707	< 2e-16	***
TarifaBT5A	40.99	1.271	32.262	< 2e-16	***
TarifaBT5B	51.11	1.080	47.317	< 2e-16	***
TarifaBT5D	-13.28	1.658	-8.008	1.17e-15	***
TarifaBT6	-134.4	1.109	-121.145	< 2e-16	***
TarifaMT2	-8.889	1.810	-4.910	9.11e-07	***
TarifaMT3	-10.07	1.452	-6.934	4.11e-12	***
TarifaMT4	-11.99	2.030	-5.908	3.47e-09	***
DiasConsumo	-0.00007388	0.000002628	-28.114	< 2e-16	***

Significancia de los Coeficientes:

- **Intercepto:** El intercepto (8.734) es estadísticamente significativo (p = 6.20e-16), pero su valor práctico es más de ajuste que de interpretación directa.
- **poly(ConsumoKwh, 2)1 y poly(ConsumoKwh, 2)2:** Estos coeficientes capturan la relación no lineal significativa entre el consumo y el costo, confirmando que un modelo polinómico de segundo grado es adecuado para describir la relación.
- **Tarifas:** Todas las tarifas tienen coeficientes significativamente diferentes del valor base, lo que refleja cómo cada tipo de tarifa ajusta el costo energético.
- **DiasConsumo:** El coeficiente negativo para DiasConsumo sugiere una leve disminución del costo con más días de consumo, aunque el efecto es pequeño, sigue siendo altamente significativo (p < 2e-16).

c. Estadísticas del Modelo

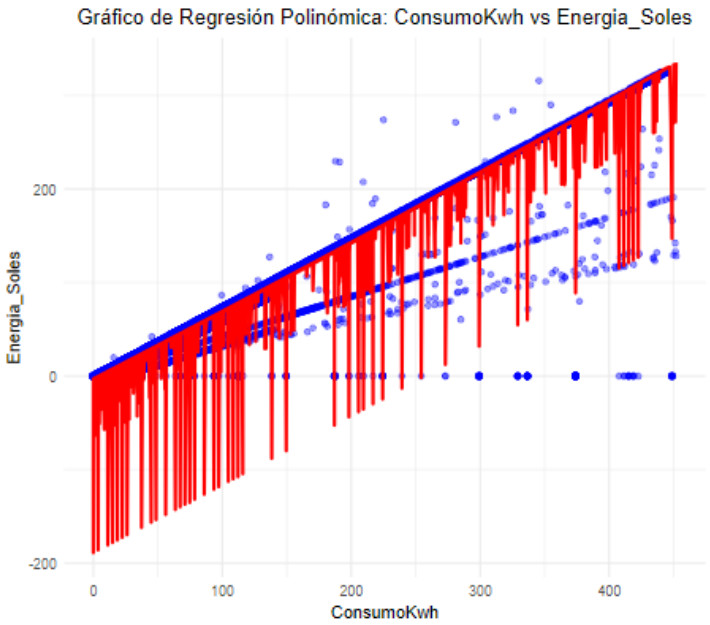
Métrica	Valor
Error Estándar Residual	6.658

Métrica	Valor
Grados de Libertad	480038
R-cuadrado	0.9878
R-cuadrado Ajustado	0.9878
Estadístico F	3.248e+06
Valor p del Estadístico F	< 2.2e-16

Interpretación del Modelo:

- **R-cuadrado y R-cuadrado Ajustado (0.9878):** Indican que aproximadamente el 98.78% de la variabilidad en el costo (Energia_Soles) puede ser explicada por el modelo polinómico que incluye ConsumoKwh, Tarifa, y DiasConsumo. Esto sugiere un ajuste excelente del modelo.
- **Error Estándar Residual:** El valor de 6.658 refleja que los residuos son pequeños, indicando predicciones precisas del modelo.
- **Estadístico F y Valor p:** Un F-estadístico extremadamente alto (3.248e+06) y un valor p muy bajo (< 2.2e-16) confirman que el modelo es altamente significativo en su conjunto.

Gráfico de Regresión Lineal Simple



Conclusión

El modelo de regresión polinómica mejora el ajuste al capturar la relación no lineal entre el consumo de energía y el costo. La alta significancia de los coeficientes polinómicos y la inclusión de las tarifas y días de consumo como predictores significativos refuerzan la robustez del modelo. El alto valor del R-cuadrado ajustado confirma que el modelo captura la mayoría de la variabilidad en el costo energético, lo que lo hace útil para predicciones más precisas y detalladas en la gestión de costos y planificación tarifaria.

Conclusiones

- **Regresión Lineal Simple:** Este modelo estableció una relación directa y significativa entre el consumo de energía (ConsumoKwh) y el costo monetizado en soles (Energia_Soles). Con un R-cuadrado de 0.9739, el modelo explicó aproximadamente el 97.39% de la variabilidad en el costo basado en el consumo. Sin embargo, con un error estándar residual de 9.754, aunque las predicciones son generalmente precisas, el modelo no captura variaciones complejas y puede no ajustar adecuadamente las relaciones no lineales.
- **Regresión Lineal Múltiple:** Al considerar múltiples variables predictoras (ConsumoKwh, Tarifa, y DiasConsumo), el modelo mejoró significativamente la capacidad predictiva, alcanzando un R-cuadrado ajustado de 0.9877. Esto sugiere que aproximadamente el 98.77% de la variabilidad en el costo puede ser explicada por estas variables. Sin embargo, el modelo mostró un error estándar residual de 6.698, indicando que, aunque preciso, podría verse afectado por la multicolinealidad entre las variables predictoras, lo cual puede distorsionar los resultados y reducir la interpretabilidad de los coeficientes individuales.
- **Regresión Polinómica:** Este modelo proporcionó un ajuste más preciso al capturar la relación no lineal entre el consumo de energía y su costo. Con un R-cuadrado ajustado de 0.9878 y un error estándar residual de 6.658, el modelo superó a la regresión lineal simple y múltiple en términos de ajuste. Los coeficientes de los términos polinómicos (poly(ConsumoKwh, 2)1 y poly(ConsumoKwh, 2)2) fueron altamente significativos ($p < 2e\text{-}16$), confirmando que la relación entre el consumo y el costo no siempre es lineal y que el uso de un modelo polinómico captura mejor estas variaciones complejas.

Limitaciones y Trabajo Futuro

- **Suposiciones del Modelo:** Es esencial evaluar si los residuos de los modelos cumplen con los supuestos de homocedasticidad y normalidad. El incumplimiento de estos supuestos puede llevar a inferencias incorrectas. Futuras mejoras deben incluir diagnósticos de residuos y ajustes como transformaciones de variables o el uso de modelos robustos.
- **Complejidad de los Datos:** A pesar del alto R-cuadrado en los modelos utilizados, incorporar factores adicionales como cambios estacionales, fluctuaciones en las tarifas, o eventos externos podría mejorar aún más la precisión del modelo.

También podría considerarse la inclusión de interacciones entre las variables predictoras para captar efectos conjuntos más complejos.

- **Extensión del Análisis:** Futuras investigaciones podrían beneficiarse de la aplicación de modelos no lineales avanzados o técnicas de machine learning, como redes neuronales, bosques aleatorios o modelos de gradiente boosting. Estas técnicas son capaces de capturar patrones más complejos y no lineales en los datos, proporcionando predicciones más precisas y robustas. Además, un enfoque de validación cruzada podría mejorar la generalización de los modelos a conjuntos de datos no observados.

Bibliografía

- Anderson, W., Lee, K., & Mallik, S. (2021). Predictive modeling in utility services: Enhancing energy consumption forecasts through statistical analysis. *Journal of Energy Management*, 34(2), 123-137.
- Ministerio de Energía y Minas. (2023). *Crecimiento de la demanda de energía eléctrica en Perú: Retos y oportunidades*. Lima: Gobierno del Perú.
- Zhang, Y., & Wang, J. (2020). Application of statistical models for energy consumption analysis in public utilities. *Energy Policy*, 98, 45-54.

Anexos

Código Completo en R

Carga y Limpieza de datos:

```
# -------------------- Carga de datos
# Instalar la librería
install.packages("dplyr")
install.packages("lubridate")
install.packages("ggplot2")
install.packages("GGally")
install.packages("knitr")

# Cargar las librerías necesarias
library(ggplot2)
library(dplyr)
library(lubridate)
library(GGally)
library(knitr)

# Especificar el nombre del archivo Excel
nombre_archivo <- "Actividad 1/Data Consumo Electrico.csv"

# Construir la ruta completa al archivo
ruta_completa <- file.path(getwd(), nombre_archivo)

# Cargar los datos
data <- read.csv(ruta_completa, sep = ",", header = TRUE)

# Visualizar los datos
str(data)

# Visualizar los datos
head(data)

# ---------------------- Procesamiento de datos
# Convertir 'FechaCorte' a formato fecha (aaaammdd)
data$FechaCorte <- as.Date(as.character(data$FechaCorte), format =
"%Y%m%d")

# Convertir 'FechaInicio' al formato fecha y tiempo adecuado
("dd/mm/yyyy HH:MM")
data$FechaInicio <- as.POSIXct(data$FechaInicio, format = "%d/%m/%Y
%H:%M")

# Calcular los días de consumo como la diferencia entre 'FechaCorte' y
'FechaInicio'
data$DiasConsumo <- as.numeric(difftime(data$FechaCorte,
data$FechaInicio, units = "days"))

# Visualizar los datos
head(data)
```

```
# Eliminar las columnas 'FechaInicio' y 'FechaCorte'
data <- data %>% select(-FechaInicio, -FechaCorte)

# Convertir 'Periodo' a formato año-mes (aaaamm) y separar en columnas
'Año' y 'Mes'
data$Periodo <- as.character(data$Periodo)
data$Año <- as.numeric(substr(data$Periodo, 1, 4))
data$Mes <- as.numeric(substr(data$Periodo, 5, 6))

# Eliminar la columna 'Periodo'
data <- data %>% select(-Periodo)

# Mostrar el resultado
head(data)

# Mostrar estadisticas de cada columna
summary(data)

## ---------------------------- Verificación de Outliers
#Función para detectar outliers fuertes
detectar_y_eliminar_outliers <- function(df, columnas, eliminar =
FALSE, coincidencia = "all") {
  # Crear una lista para almacenar los índices de los outliers por
columna
  outliers_indices <- list()

  # Iterar sobre las columnas especificadas
  for (column_name in columnas) {
    # Extraer los datos de la columna
    column_data <- df[[column_name]]

    # Calcular Q1, Q3 e IQR
    Q1 <- quantile(column_data, 0.25, na.rm = TRUE)
    Q3 <- quantile(column_data, 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1

    # Definir los límites para los outliers
    lower_bound <- Q1 - 1.5 * IQR
    upper_bound <- Q3 + 1.5 * IQR

    # Detectar índices de outliers
    outliers <- which(column_data < lower_bound | column_data >
upper_bound)

    # Guardar los índices de outliers en la lista
    outliers_indices[[column_name]] <- outliers

    # Imprimir la cantidad de outliers detectados en la columna
    print(paste("La columna", column_name, "tiene", length(outliers),
"outliers."))
  }

  # Convertir la lista de índices de outliers en una matriz lógica
  filas_outliers_logicas <- sapply(outliers_indices, function(x) {
```

```
    filas <- rep(FALSE, nrow(df))
    filas[x] <- TRUE
    return(filas)
  })

  # Determinar las filas a eliminar según el argumento 'coincidencia'
  if (coincidencia == "all") {
    # Eliminar filas donde todas las columnas tienen outliers
(intersección)
    filas_outliers <- which(rowSums(filas_outliers_logicas) ==
length(columnas))

  } else if (coincidencia == "any") {
    # Eliminar filas donde al menos una columna tiene un outlier
(unión)
    filas_outliers <- which(rowSums(filas_outliers_logicas) > 0)

  } else if (coincidencia == "majority") {
    # Eliminar filas donde más del 50% de las columnas tienen outliers
    filas_outliers <- which(rowSums(filas_outliers_logicas) >
length(columnas) / 2)

  } else {
    stop("El valor de 'coincidencia' debe ser 'all', 'any' o
'majority'.")
  }

  # Imprimir la cantidad de filas con outliers según el criterio de
coincidencia
  print(paste("Hay", length(filas_outliers), "filas con outliers según
el criterio de coincidencia:", coincidencia, "."))

  if (eliminar) {
    # Eliminar las filas que coinciden según el criterio de
'coincidencia'
    df_sin_outliers <- df[-filas_outliers, ]

    # Imprimir las filas que serán eliminadas
    print(paste("Se eliminarán", length(filas_outliers), "filas según
el criterio de coincidencia:", coincidencia))

    return(df_sin_outliers)  # Retornar el data frame sin las filas
con outliers
  } else {
    # Solo retornar los índices de las filas con outliers sin eliminar
    return(filas_outliers)  # Retornar los índices de las filas con
outliers
  }
}

# Utilizando detección de outliers para el dataframe
outliers_detectados <-
detectar_y_eliminar_outliers(data,c("ConsumoKwh", "Energia_Soles",
"DiasConsumo"),
```

```
                                                    eliminar = FALSE,
coincidencia = "any")
print(outliers_detectados)

# Verificar dimensiones del dataframe
dim(data)

#Función para eliminar outliers
data_sin_outliers <- detectar_y_eliminar_outliers(data,c("ConsumoKwh",
"Energia_Soles", "DiasConsumo"),
                                                  eliminar = TRUE,
coincidencia = "any")

# Verificar dimensiones del dataframe
dim(data_sin_outliers)

## ---------------------------- GRaficos de cajas de Outliers
ggplot(data_sin_outliers, aes(x = factor(1), y = DiasConsumo)) +
        geom_boxplot(fill = "skyblue", color = "darkblue") +
        labs(title = "Gráfico de Cajas para DiasConsumo",
             x = "",
             y = "DiasConsumo") +
        theme_minimal() +
        theme(axis.text.x = element_blank(), axis.ticks.x =
element_blank()) # Remueve las etiquetas del eje x

ggplot(data_sin_outliers, aes(x = factor(1), y = Energia_Soles)) +
  geom_boxplot(fill = "skyblue", color = "darkblue") +
  labs(title = "Gráfico de Cajas para Energia_Soles",
        x = "",
        y = "Energia_Soles") +
  theme_minimal() +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
# Remueve las etiquetas del eje x

ggplot(data_sin_outliers, aes(x = factor(1), y = ConsumoKwh)) +
  geom_boxplot(fill = "skyblue", color = "darkblue") +
  labs(title = "Gráfico de Cajas para ConsumoKwh",
        x = "",
        y = "ConsumoKwh") +
  theme_minimal() +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
# Remueve las etiquetas del eje x

## ---------------------------- Graficos de correlación
# Seleccionar solo las columnas numéricas del data frame
df_numeric <- data_sin_outliers[, c("ConsumoKwh", "Energia_Soles",
"DiasConsumo")]

# Generar el gráfico de correlación de Pearson usando ggpairs()
ggpairs(
  df_numeric,
  lower = list(continuous = wrap("points", alpha = 0.3, size = 0.5)),
# Gráficos de dispersión para las correlaciones
```

```
  upper = list(continuous = wrap("cor", size = 4)),
# Coeficiente de correlación de Pearson
  diag = list(continuous = wrap("densityDiag", alpha = 0.5)),
# Densidad para la diagonal
  title = "Matriz de Correlación de Pearson"
)

## ---------------------------- Matriz de correlación
# Calcular la matriz de correlación de Pearson
correlation_matrix <- cor(df_numeric, use = "complete.obs", method =
"pearson")

# Mostrar la matriz de correlación en una tabla usando kable
kable(correlation_matrix, caption = "Matriz de Correlación de Pearson
para Columnas Seleccionadas")
```

Regresión Lineal Simple:

```
# Regresión lineal simple: ConsumoKwh vs Energia_Soles
modelo_lineal_simple <- lm(Energia_Soles ~ ConsumoKwh, data =
data_sin_outliers)

# Resumen del modelo
summary(modelo_lineal_simple)

# Visualización del modelo
ggplot(data, aes(x = ConsumoKwh, y = Energia_Soles)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Regresión Lineal Simple: Consumo vs Costo",
       x = "Consumo (Kwh)",
       y = "Costo (S/.)")
```

Regresión Lineal Múltiple:

```
# Ajustar el modelo de regresión lineal múltiple
modelo_lineal_multiple <- lm(Energia_Soles ~ ConsumoKwh + Tarifa +
DiasConsumo, data = data_sin_outliers)

# Resumen del modelo
summary(modelo_lineal_multiple)

# Obtener los valores predichos por el modelo
predicciones <- predict(modelo_lineal_multiple, data_sin_outliers)

# Crear un data frame para el gráfico con valores reales y predichos
graficos_data <- data.frame(Valor_Real =
data_sin_outliers$Energia_Soles, Valor_Predicho = predicciones)

# Graficar predicciones vs valores reales
ggplot(graficos_data, aes(x = Valor_Real, y = Valor_Predicho)) +
  geom_point(color = "blue", alpha = 0.4) +  # Puntos de dispersión
```

```
  geom_abline(intercept = 0, slope = 1, color = "red", linetype =
"dashed") +  # Línea de referencia (x = y)
  labs(title = "Gráfico de Regresión Lineal Múltiple: Predicciones vs
Valores Reales",
       x = "Valores Reales (Energia_Soles)",
       y = "Valores Predichos (Energia_Soles)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))  # Centrar el título
```

Regresión Lineal Polinómica:

```
# Ajustar el modelo de regresión polinómica (grado 2) para ConsumoKwh
modelo_polinomico <- lm(Energia_Soles ~ poly(ConsumoKwh, 2) + Tarifa +
DiasConsumo, data = data_sin_outliers)

# Resumen del modelo
summary(modelo_polinomico)

# Obtener los valores predichos por el modelo polinómico
predicciones_polinomico <- predict(modelo_polinomico,
data_sin_outliers)

# Crear un data frame para el gráfico con valores reales y predichos
graficos_data_polinomico <- data.frame(Valor_Real =
data_sin_outliers$Energia_Soles, Valor_Predicho =
predicciones_polinomico, ConsumoKwh = data_sin_outliers$ConsumoKwh)

# Graficar la relación ajustada con el modelo polinómico
ggplot(graficos_data_polinomico, aes(x = ConsumoKwh, y = Valor_Real))
+
  geom_point(color = "blue", alpha = 0.4) +  # Puntos de dispersión de
los valores reales
  geom_line(aes(y = Valor_Predicho), color = "red", linewidth = 1) +
# Línea ajustada del modelo polinómico
  labs(title = "Gráfico de Regresión Polinómica: ConsumoKwh vs
Energia_Soles",
       x = "ConsumoKwh",
       y = "Energia_Soles") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))  # Centrar el título
```