

INFORME FINAL DE AVANCE PROYECTO APLICADO II

MAESTRÍA EN CIENCIA DE DATOS

FECHA: 08 de 12 de 2025

1. Datos generales del Proyecto Aplicado

| |
|--|
| Título del Proyecto: Desarrollo de un Modelo de Análisis de Carga no Intrusiva para Eficiencia Energética Basado en Técnicas de Machine Learning |
| Nombre del/los Estudiante(s): Meidy Lizeth Moreno, Jose Luis Munévar Diaz, Julian Javier Jaimes Otero |
| Director del Proyecto Aplicado: David Arango Londoño |
| Codirector/ Asesores (si aplica): |
| Grupo de Investigación o Empresa que apoya (si aplica): |
| Fecha de Inicio del proyecto: junio 1 de 2025 |

2. Resumen del Trabajo de Grado

Objetivo General:

Desarrollar un modelo de análisis de carga no intrusivo basado en técnicas de machine learning que permita desagregar el consumo energético de los equipos.

Objetivos Específicos:

1. Procesar los datos de carga energética y consumo eléctrico a través del análisis exploratorio y correlaciones.
2. Evaluar diferentes modelos de machine learning para la desagregación e implementación de un modelo capaz de identificar y categorizar las cargas de los equipos eléctricos.

3. Desarrollar un mecanismo de reconciliación para alinear las predicciones de consumo generadas por distintos modelos según las categorías de carga, para coherencia y precisión en la desagregación.
4. Evaluar la precisión y eficiencia del modelo considerando diferentes tipos de cargas eléctricas.

Cronograma de Actividades propuesto:

| <u>Actividad</u> | <u>Octubre</u> | <u>Noviembre</u> | <u>Diciembre</u> | <u>Enero</u> | <u>Febrero</u> |
|---|----------------|------------------|------------------|--------------|----------------|
| <u>Objetivo 1: Procesar los datos</u> | | | | | |
| <u>Recolección de datos</u> | 13 | | | | |
| <u>Limpieza de datos</u> | 13 | | | | |
| <u>Análisis exploratorio (EDA)</u> | 26 | | | | |
| <u>Cálculo de correlaciones</u> | | 3 | | | |
| <u>Objetivo 2: Evaluar diferentes modelos de machine learning</u> | | | | | |
| <u>Selección de técnicas de machine learning</u> | | 10 | | | |
| <u>Implementación inicial de modelos base</u> | | 17 | | | |
| <u>Comparación de modelos</u> | | 24 | | | |
| <u>Ajuste del modelos seleccionados modificando los parámetros de entrada</u> | - | | 3 | | |
| <u>Entrenamiento de modelos, implementación de técnicas de regularización</u> | | | | 30 | |
| <u>Validación del modelo</u> | | | | | 15 |

Frecuencia de Reuniones con el director: Quincenal

3. Resultados, avances y estado del Proyecto Aplicado

3.1. Procesamiento y recolección de datos

3.1.1. Recolección de datos

El conjunto de datos contiene 43825 registros y presenta mediciones horarias del consumo energético desagregado por tipo de carga eléctrica. La variable Time indica la fecha y hora de cada registro. Las variables Ventilation, Sockets plug, Lighting, Other electricity, Cooling y Heating representan el consumo de energía kWh correspondiente a los sistemas de ventilación, los equipos conectados a los enchufes, la iluminación, otros consumos eléctricos no clasificados, la refrigeración y la calefacción, respectivamente. En conjunto, estas variables permiten analizar el comportamiento y la distribución del uso de energía a lo largo del tiempo, identificando cuáles sistemas son los principales responsables del consumo en cada hora.

3.1.2. Limpieza de datos

Se subdivide la variable Time, creando variables adicionales con la información de la hora, el número del día del mes, el día de la semana, mes y año, en la Imagen 1. Resumen set de datos Se puede observar un resumen del set de datos.

Durante la exploración no se observan datos faltantes, pero sí se encuentra que el 35% de los datos del set contiene valores atípicos distribuidos en las variables, como se ve en la Imagen 2. Además, se encontraron registros duplicados o desfase de 2 horas en la variable Time; se aplica el método de interpolación a los valores atípicos del set de datos, se borran registros duplicados y se ajustan ambigüedades en la variable Time con el fin de mantener la coherencia temporal y evitar distorsiones en el análisis o el modelado; esta transformación se observa en la Imagen 3. Set de datos sin valores atípicos.

3.1.3. Análisis exploratorio (EDA)

Se realiza un análisis exploratorio de datos sobre las variables Ventilation, Sokets plug, Lighting, Other electricity, Cooling y Heating, considerando además la variable Time, segmentada por día-hora y año-mes, como se observa en la Imagen 4. Consumos por día-hora Y la Imagen 5. Consumos por año-mes. respectivamente.











Data summary

| | |
|------------------------|-------|
| Name | df |
| Number of rows | 43824 |
| Number of columns | 12 |
| Column type frequency: | |
| factor | 1 |
| numeric | 10 |
| POSIXct | 1 |
| Group variables | |
| | None |

Variable type: factor

| skim_variable | n_missing | complete_rate | ordered | n_unique | top_counts |
|---------------|-----------|---------------|---------|----------|--|
| dia | 0 | 1 | FALSE | 7 | mié: 6264, jue: 6264, vie: 6264, sáb: 6264 |

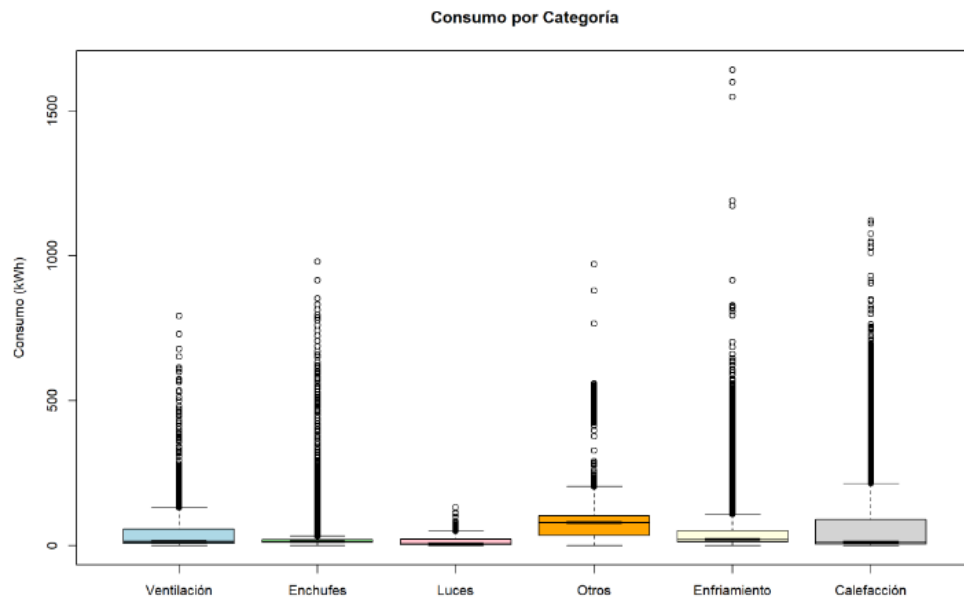
Variable type: numeric

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---------------|-----------|---------------|---------|--------|------|---------|--------|---------|------|---|
| Ventilación | 0 | 1 | 32.43 | 36.21 | 0 | 7.00 | 14.1 | 56.80 | 793 |  |
| Enchufes | 0 | 1 | 18.41 | 27.67 | 0 | 11.40 | 13.6 | 20.10 | 980 |  |
| Luces | 0 | 1 | 14.32 | 15.06 | 0 | 3.69 | 6.5 | 22.50 | 132 |  |
| Otros | 0 | 1 | 87.85 | 94.95 | 0 | 35.80 | 79.6 | 103.00 | 971 |  |
| Enfriamiento | 0 | 1 | 47.08 | 71.42 | 0 | 13.60 | 21.5 | 51.60 | 1641 |  |
| Calefacción | 0 | 1 | 68.99 | 113.33 | 0 | 5.50 | 12.3 | 89.00 | 1120 |  |
| hora | 0 | 1 | 12.50 | 6.92 | 1 | 6.75 | 12.5 | 18.25 | 24 |  |
| mes | 0 | 1 | 6.52 | 3.45 | 1 | 4.00 | 7.0 | 10.00 | 12 |  |
| anno | 0 | 1 | 2019.01 | 1.42 | 2017 | 2018.00 | 2019.0 | 2020.00 | 2022 |  |
| dia_mes_1 | 0 | 1 | 15.73 | 8.80 | 1 | 8.00 | 16.0 | 23.00 | 31 |  |

Variable type: POSIXct

| skim_variable | n_missing | complete_rate | min | max | median | n_unique |
|---------------|-----------|---------------|---------------------|---------------------|---------------------|----------|
| Time | 0 | 1 | 2017-01-03 13:00:00 | 2022-01-03 12:00:00 | 2019-07-05 13:30:00 | 43819 |

Imagen 1. Resumen set de datos



Cálculo de valores atípicos de cada variable.

[Show](#)

| | | | | | | |
|-------------|-------------|----------|-------|-------|--------------|--|
| Time | Ventilación | Enchufes | Luces | Otros | Enfriamiento | |
| NA | 448 | 3537 | 1334 | 2154 | 4556 | |
| Calefacción | día | hora | mes | anno | día_mes_1 | |
| 4929 | NA | 0 | 0 | 0 | 0 | |

Imagen 2. Distribución de valores atípicos.

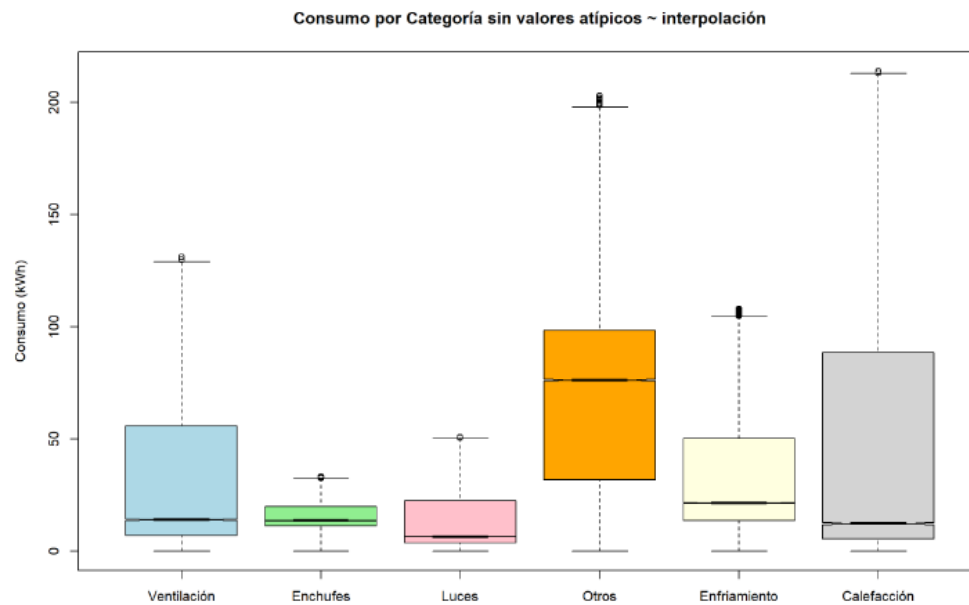


Imagen 3. Set de datos sin valores atípicos.

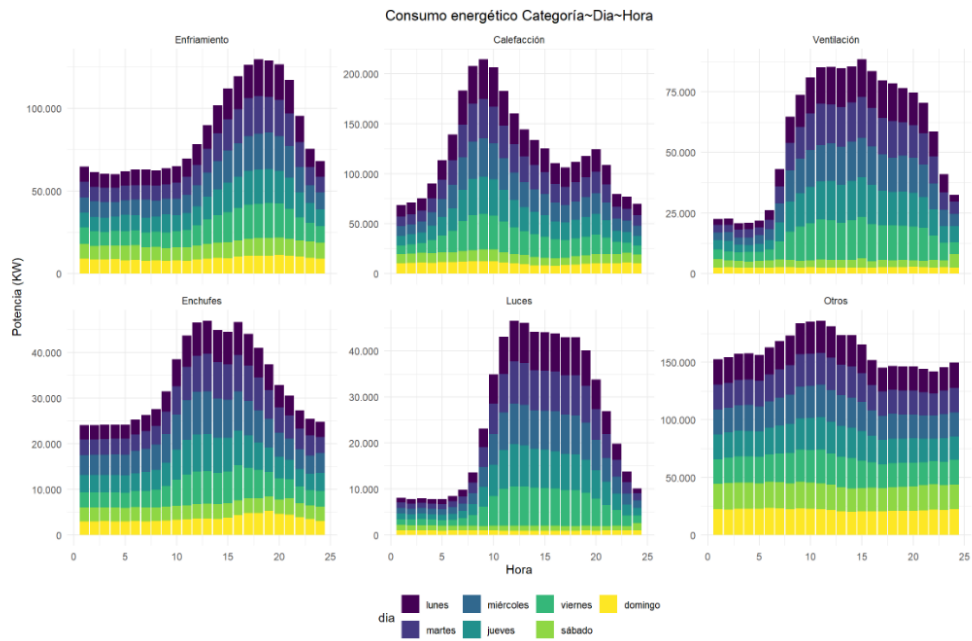


Imagen 4. Consumos por día-hora

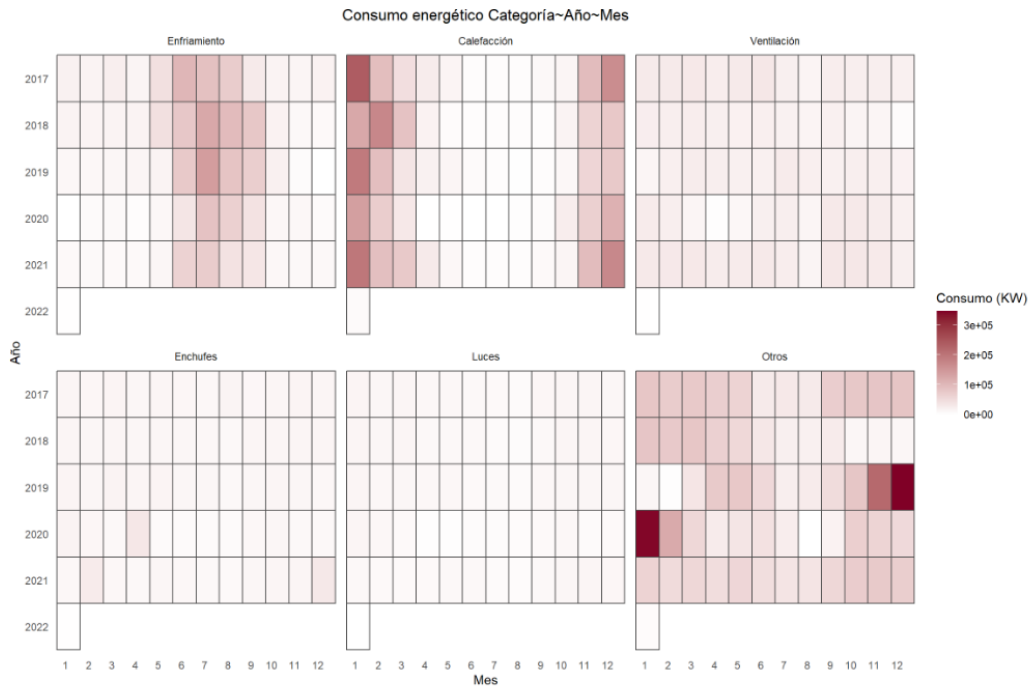


Imagen 5. Consumos por año-mes.

3.1.4. Conclusiones

- Se observa que el consumo de cada categoría contiene una tendencia marcada por la hora y el día de la semana; las categorías calefacción y enfriamiento muestran los picos más altos, seguidos de la categoría otros. Los consumos más bajos ocurren los fines de semana.
- La categoría “enfriamiento” intensifica los consumos en los meses comprendidos entre junio y septiembre, mientras que la categoría “calefacción” muestra los mayores consumos en los meses de enero, febrero, octubre y diciembre; esta tendencia podría estar asociada a cambios climáticos o a hábitos comerciales. La categoría “otros” muestra fluctuaciones sin una tendencia aparente.
- Las categorías de ventilación, iluminación y enchufes no presentan una tendencia temporal claramente definida, lo que sugiere que su comportamiento a lo largo del tiempo es más bien estable o está influenciado por variaciones aleatorias en lugar de patrones sistemáticos.

4. Evaluar diferentes modelos de machine learning para desagregar el consumo energético.

4.1. Implementación inicial de modelos

4.1.1. Regresión Lineal

Se construye un modelo de regresión lineal y se evalúan los resultados utilizando con dos conjuntos de datos. En el set de datos 1 se eliminaron los registros duplicados y se ajustaron aquellos en los que la variable Time presentaba una diferencia distinta de una hora. En el set de datos 2 se trataron los registros atípicos mediante el método de interpolación.

Los resultados de la Tabla 1. muestran el desempeño del modelo de regresión lineal para los dos sets de datos, se observan ligeras mejoras cuando se utilizan los datos del Set de datos 2, donde se trataron los valores atípicos mediante interpolación. En general, los valores de R^2 son bajos en la mayoría de las variables, lo que indica que el modelo explica solo una pequeña parte de la variabilidad del consumo energético, especialmente en variables como Sockets plug, Cooling y Other electricity. Sin embargo, Lighting y Heating presentan los mejores niveles de ajuste dentro del conjunto, aunque siguen siendo moderados. La reducción del RMSE y MAE en casi todas las variables en el Set de datos 2 sugiere que la corrección de atípicos contribuyó a mejorar la precisión de las predicciones, aunque el modelo sigue mostrando limitaciones importantes para capturar adecuadamente la dinámica del consumo eléctrico.

Tabla 1. Resultados modelo regresión

| Variable | Set de datos 1 | | | Set de datos 2 | | |
|-------------------|----------------|-------|-------|----------------|-------|-------|
| | R2 | RMSE | MAE | R2 | RMSE | MAE |
| Ventilation | 0,31 | 29,16 | 19,65 | 0,36 | 23,82 | 18,41 |
| Sokets plug | 0,07 | 28,53 | 7,31 | 0,34 | 5,64 | 4,30 |
| Lighting | 0,45 | 11,26 | 8,35 | 0,45 | 10,61 | 8,01 |
| Other electricity | 0,36 | 75,53 | 50,13 | 0,06 | 43,33 | 35,93 |
| Cooling | 0,06 | 69,11 | 42,53 | 0,09 | 26,48 | 21,06 |
| Heating | 0,60 | 72,54 | 50,89 | 0,47 | 48,06 | 36,04 |

4.1.2. Máquinas de Soporte Vectorial (SVM)

La implementación del sistema de desagregación mediante Máquinas de Soporte Vectorial (SVR) se configuró utilizando un kernel de Base Radial (RBF), el cual es idóneo para capturar relaciones no lineales en los datos. La configuración de los hiperparámetros se estandarizó para todas las cargas con un parámetro de regularización $C=100$, un coeficiente de kernel $\gamma=0.1$ y un margen de error $\epsilon=0.1$. A diferencia del enfoque con redes neuronales, esta metodología incluyó una etapa previa rigurosa de limpieza de datos, donde se eliminaron valores atípicos (outliers) mediante el rango intercuartílico (IQR) y se aplicó interpolación temporal, además de un escalado estandarizado (StandardScaler) tanto para las variables predictoras (tiempo y consumo total) como para las variables objetivo.

Los resultados obtenidos evidencian una capacidad predictiva notablemente superior y más estable en comparación con modelos previos. La carga de Calefacción (Heating) presentó el mejor desempeño del conjunto, con un coeficiente de determinación $R^2=0.891$ y un Error Cuadrático Medio (RMSE) de 21.689 W. Este valor indica que el modelo logra explicar el 89% de la variabilidad del consumo de calefacción, lo cual sugiere que, tras la limpieza de datos, el SVR logró capturar eficazmente la estacionalidad y la dependencia horaria de esta carga. De manera similar, la Iluminación (Lighting) mostró una robustez alta con un $R^2=0.872$ y un error reducido de 5.123 W, confirmando que los patrones de luz son altamente predecibles mediante variables temporales.

En un nivel de rendimiento intermedio-alto se encuentran las cargas de Refrigeración (Cooling) y Enchufes (Sokets plug). Para el caso de la refrigeración, se obtuvo un $R^2=0.764$ y un RMSE de 13.557 W, mientras que los enchufes alcanzaron un $R^2=0.757$ con un error de 3.427. Es particularmente destacable el rendimiento en la categoría de enchufes, la cual suele ser estocástica; el éxito aquí puede atribuirse a que la eliminación de picos atípicos suavizó la señal, permitiendo al SVR modelar la tendencia base del comportamiento del usuario.

Finalmente, las cargas con mayor dificultad de predicción fueron Otros aparatos (Other electricity) y Ventilación, con coeficientes R^2 de 0.611 y 0.558 respectivamente.

Aunque estos valores son los más bajos del experimento actual, representan una mejora sustancial frente a modelos lineales simples o redes sin pre-procesamiento avanzado, indicando que el modelo explica más de la mitad de la variabilidad de estos consumos erráticos. En conclusión, el modelo SVR, potenciado por una limpieza de datos efectiva, demuestra ser una herramienta altamente competente para este problema, aunque se reitera la recomendación de integrar variables meteorológicas exógenas (temperatura exterior)

para refinar aún más la precisión en las cargas de climatización (Heating y Cooling), las cuales mostraron una alta sensibilidad a las condiciones ambientales no incluidas explícitamente en el dataset.

El modelo SVR explica casi el 76% de la variabilidad en el consumo de refrigeración. Esto indica que el SVR puede hacer predicciones confiables para esta categoría.

```
##### Máquinas de Soporte Vectorial (SVR) Cooling/Enfriamiento #####

# Variables predictoras y objetivo
X = df[["Dia", "Mes", "Anno", "DiaSemanalNum", "Total", "hora"]]
y = df["Cooling"]

# Escalado de las variables (muy importante para SVM)
scaler_X = StandardScaler()
scaler_y = StandardScaler()

X_scaled = scaler_X.fit_transform(X)
y_scaled = scaler_y.fit_transform(y.values.reshape(-1, 1)).ravel()

# Dividir datos en entrenamiento y prueba (80/20)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_scaled, test_size=0.2, random_state=42)

# Crear modelo SVM con kernel RBF (común en regresión)
model = SVR(kernel='rbf', C=100, gamma=0.1, epsilon=0.1)

# Entrenar el modelo
model.fit(X_train, y_train)

# Predecir en el conjunto de prueba
y_pred_scaled = model.predict(X_test)

# Invertir la escala de la predicción para comparar con valores reales
y_pred = scaler_y.inverse_transform(y_pred_scaled.reshape(-1, 1))
y_test_real = scaler_y.inverse_transform(y_test.reshape(-1, 1))

# Evaluar el modelo
rmse = np.sqrt(mean_squared_error(y_test_real, y_pred))
r2 = r2_score(y_test_real, y_pred)

print(f"RMSE: {rmse:.3f}")
print(f"R²: {r2:.3f}")

... RMSE: 13.557
    R²: 0.764
```

Imagen 6. Extracto código Máquinas de soporte vectorial variable Cooling

Tabla 2 Resultados modelo Bosques Máquinas de Soporte Vectorial

| Variable | R2 | RMSE | MAE |
|-------------------|-------|--------|--------|
| Heating | 0,89 | 16 | 8,333 |
| Lighting | 0,87 | 4 | 2,762 |
| Cooling | 0,76 | 9 | 5,596 |
| Sockets plug | 0,75 | 3 | 1,872 |
| Other electricity | 0,61 | 21,839 | 13,779 |
| Ventilation | 0,558 | 21,952 | 10,36 |

4.1.3. Bosques Aleatorios (Random Forest)

Para la predicción del consumo energético en seis categorías: Ventilation, Sockets plug, Lighting, Other electricity, Cooling y Heating, se implementó un modelo Random Forest Regressor multisalida. Este algoritmo fue seleccionado por su capacidad para modelar relaciones no lineales entre las variables, su robustez ante valores atípicos y su buen desempeño en contextos donde las variables predictoras presentan diferentes escalas y distribuciones.

El coeficiente de determinación promedio ($R^2 = 0.83$) evidencia que el modelo logra explicar más del 80% de la variabilidad total de los consumos energéticos. Este desempeño puede considerarse altamente satisfactorio para un modelo de aprendizaje automático de tipo base, sin optimización de hiperparámetros. Las categorías Heating, Lighting y Cooling obtuvieron los mayores niveles de ajuste, lo que indica que el modelo fue capaz de capturar patrones consistentes en su comportamiento. En contraste, las variables Ventilation y Other electricity presentaron valores de R^2 más bajos, lo que sugiere una mayor complejidad o la influencia de factores externos no representados en las variables independientes.

En términos de error, los valores de MAE y RMSE muestran una magnitud de desviación moderada entre los valores predichos y los observados, siendo menores en las categorías con mayor regularidad temporal. Estos resultados confirman la capacidad del Random Forest para generalizar adecuadamente sobre datos energéticos heterogéneos, especialmente cuando los consumos presentan comportamientos no lineales y fluctuaciones abruptas.

El uso del Random Forest resulta adecuado como modelo base, ya que permite establecer una línea de referencia sólida frente a técnicas tradicionales como la regresión lineal múltiple. Su desempeño confirma la viabilidad del uso de modelos de Machine Learning en procesos de desagregación y predicción de consumo energético, aportando una primera aproximación robusta y explicativa al fenómeno.

No obstante, el modelo presenta limitaciones estructurales derivadas de su naturaleza no secuencial:

No incorpora la dependencia temporal inherente a los datos horarios de consumo.

No capta patrones estacionales o recurrentes más allá de las variables explícitamente incluidas (hora, día, mes, etc.).

Su capacidad predictiva podría mejorarse ajustando los hiperparámetros o empleando modelos que integren la memoria temporal, como las redes neuronales recurrentes (RNN, LSTM, GRU) o algoritmos de boosting como XGBoost o LightGBM, los cuales suelen ofrecer un mejor equilibrio entre sesgo y varianza en contextos con alta variabilidad temporal.

Por tanto, aunque el Random Forest proporciona resultados satisfactorios y demuestra la pertinencia del enfoque, se recomienda explorar modelos secuenciales y de optimización de hiperparámetros como parte de una segunda etapa de experimentación, con el fin de aumentar la precisión y mejorar la capacidad del sistema para anticipar fluctuaciones energéticas más complejas.

```
# Entrenamiento del modelo Random Forest
# =====
from sklearn.ensemble import RandomForestRegressor
from sklearn.multioutput import MultiOutputRegressor

# Crear el modelo base de Random Forest
rf = MultiOutputRegressor(
    RandomForestRegressor(
        n_estimators=200, # número de árboles
        max_depth=10,    # profundidad máxima
        random_state=42,
        n_jobs=-1        # usa todos los núcleos disponibles
    )
)

# Entrenar el modelo
rf.fit(X_train, y_train)

# Realizar predicciones
y_pred = rf.predict(X_test)

print("✅ Entrenamiento completado y predicciones generadas.")
print("Dimensión de y_pred:", y_pred.shape)
```

```
✅ Entrenamiento completado y predicciones generadas.
Dimensión de y_pred: (8765, 6)
```

Imagen 7. Extracto código modelo Bosques Aleatorios.

Tabla 3. Resultados modelo Bosques Aleatorios.

| Variable | R2 | RMSE | MAE |
|-------------------|--------|---------|---------|
| Ventilation | 0,6127 | 21,9515 | 10,3599 |
| Sockets plug | 0,8377 | 2,7989 | 1,8717 |
| Lighting | 0,908 | 4,3469 | 2,7616 |
| Other electricity | 0,7642 | 21,8394 | 13,7783 |
| Cooling | 0,8892 | 9,2904 | 5,5959 |
| Heating | 0,9437 | 15,5918 | 8,333 |

4.1.4. XGBoost (Extreme Gradient Boosting)

Como punto de partida, se desarrolló un modelo de regresión utilizando XGBoost (Extreme Gradient Boosting) con el fin de establecer una línea base que permitiera evaluar el desempeño inicial del sistema de desagregación de consumo energético. XGBoost es un algoritmo basado en ensambles de árboles de decisión, donde cada árbol corrige los errores cometidos por los árboles entrenados previamente. Esta estructura secuencial le permite capturar relaciones no lineales, interacciones entre variables y patrones

complejos que modelos tradicionales no logran identificar.

XGBoost es un modelo de árboles potenciado que aprende patrones complejos y no lineales. En la línea base se utilizó XGBoost sin memoria temporal, lo que permitía un desempeño aceptable, pero incompleto.

La división de los datos se realizó respetando el orden temporal (80% entrenamiento, 20% prueba), manteniendo la coherencia cronológica de las series. Los modelos baselines mostraron resultados aceptables en varias categorías, pero con limitaciones evidentes en términos de capacidad predictiva, especialmente en variables con fuerte comportamiento cíclico como Ventilation o Cooling.

El modelo XGBoost mostró un comportamiento consistente y adecuado para la desagregación del consumo eléctrico por categorías. Los mejores desempeños se observan en Heating, Lighting y Cooling, con valores de R^2 de 0.94, 0.83 y 0.80, respectivamente, indicando una excelente capacidad del modelo para capturar las variaciones temporales y de magnitud en estas variables.

En contraste, las categorías Ventilation y Other electricity presentaron desempeños moderados, con R^2 entre 0.54 y 0.57, lo cual sugiere que estas categorías pueden depender de factores externos no presentes en los datos o tener un comportamiento más errático. No obstante, el modelo sigue logrando una aproximación funcional para propósitos operativos y de análisis energético. Estos resultados iniciales confirmaron la necesidad de incorporar información temporal más profunda en el modelo.

Con el fin de mejorar la capacidad del modelo para capturar la dinámica propia de las series temporales, se construyó una segunda versión del modelo incorporando características temporales derivadas, específicamente:

- Lags (desplazamientos temporales)
- Rolling mean de 24 horas (media móvil de un día)
- Rolling std de 24 horas (variabilidad móvil de un día)

Mejora real:

- RMSE disminuye 44%
- MAE disminuye 40%
- R^2 aumenta de 0.56 \rightarrow 0.86 (incremento del +52%)

Las características temporales (lags y medias móviles) son esenciales para explicar el consumo eléctrico altamente variable como Ventilation.

Comparación de resultados de XGBoost Base Line vs XGBoost con Lags de series de tiempo:

La inclusión de lags (lag1, lag24) y variables de suavizamiento temporal (rolling mean y rolling std) produjo una mejora significativa en todas las categorías de consumo eléctrico. Su impacto es muy importante:

- Reducción del RMSE entre 20% y 45%
- Reducción del MAE entre 25% y 40%
- Incremento del R^2 entre +0.10 y +0.30

Esto valida empíricamente que la dinámica temporal del consumo eléctrico tiene un papel fundamental para explicar y predecir cada componente.

XGBoost combinado con información temporal (lags y rolling features) es altamente efectivo para desagregar el consumo eléctrico por categorías. Su desempeño es significativamente superior al modelo base sin información de memoria temporal.

Tabla 4. Resultados modelo XGBoost.

| Variable | R ² Baseline | RMSE Baseline | MAE Baseline | R ² Lags | RMSE Lags | MAE Lags |
|-------------------|-------------------------|---------------|--------------|---------------------|-----------|----------|
| Ventilation | 0,567 | 25,687 | 13,155 | 0,862 | 14,519 | 7,873 |
| Sockets plug | 0,664 | 3,917 | 2,723 | 0,824 | 2,837 | 1,895 |
| Lighting | 0,827 | 5,215 | 3,694 | 0,948 | 2,820 | 1,775 |
| Other electricity | 0,547 | 24,576 | 15,092 | 0,720 | 19,330 | 11,941 |
| Cooling | 0,800 | 9,473 | 7,112 | 0,921 | 5,967 | 4,109 |
| Heating | 0,940 | 17,425 | 10,296 | 0,964 | 13,502 | 7,421 |

4.1.5. Redes Neuronales Recurrentes (RNN lstm)

El modelo implementado corresponde a una arquitectura RNN basada en LSTM de dos capas, diseñada para capturar dependencias temporales en series de tiempo. La primera capa LSTM cuenta con 128 unidades y devuelve toda la secuencia de salida, permitiendo que el modelo aprenda patrones complejos a lo largo de los 168 pasos temporales. Posteriormente se aplica una capa de dropout para reducir el sobreajuste. La segunda capa LSTM, con 64 unidades, resume la información temporal en un único vector representativo. A continuación, una capa densa de 32 neuronas refina esta representación mediante transformaciones no lineales y, finalmente, una capa de salida con una sola neurona genera el valor predicho.

La evaluación del modelo RNN-LSTM con el Set de datos 2 evidencia un desempeño diverso según la variable analizada. Las categorías Heating, Cooling, Lighting y Ventilation presentan los valores de R² más altos (0,77; 0,72; 0,62 y 0,57, respectivamente), lo que indica que el modelo logra capturar una parte significativa de la dinámica temporal de estas series. Esto sugiere que estos consumos poseen patrones temporales más estables o predecibles, facilitando el aprendizaje de la LSTM. Por el contrario, las variables Sockets plug (R² = 0,02) y especialmente Other electricity (R² = -0,87) muestran un desempeño deficiente, lo que implica que el modelo no logra representar adecuadamente su variabilidad; en el caso de Other electricity, el valor negativo de R² indica que la predicción es incluso peor que usar el promedio histórico como estimador.

| Layer (type) | Output Shape | Param # |
|-------------------|------------------|---------|
| lstm (LSTM) | (None, 168, 128) | 69120 |
| dropout (Dropout) | (None, 168, 128) | 0 |
| lstm_1 (LSTM) | (None, 64) | 49408 |
| dense (Dense) | (None, 32) | 2080 |
| dense_1 (Dense) | (None, 1) | 33 |

Imagen 8. Arquitectura RNN lstm

Tabla 5. Resultados modelo RNN lstm.

| Variable | R2 | RMSE | MAE |
|-------------------|-------|-------|-------|
| Ventilation | 0,57 | 25,54 | 14,54 |
| Sockets plug | 0,02 | 47,99 | 8,62 |
| Lighting | 0,62 | 8,05 | 5,26 |
| Other electricity | -0,87 | 50,11 | 39,12 |
| Cooling | 0,72 | 30 | 14,73 |
| Heating | 0,77 | 54,4 | 30,65 |

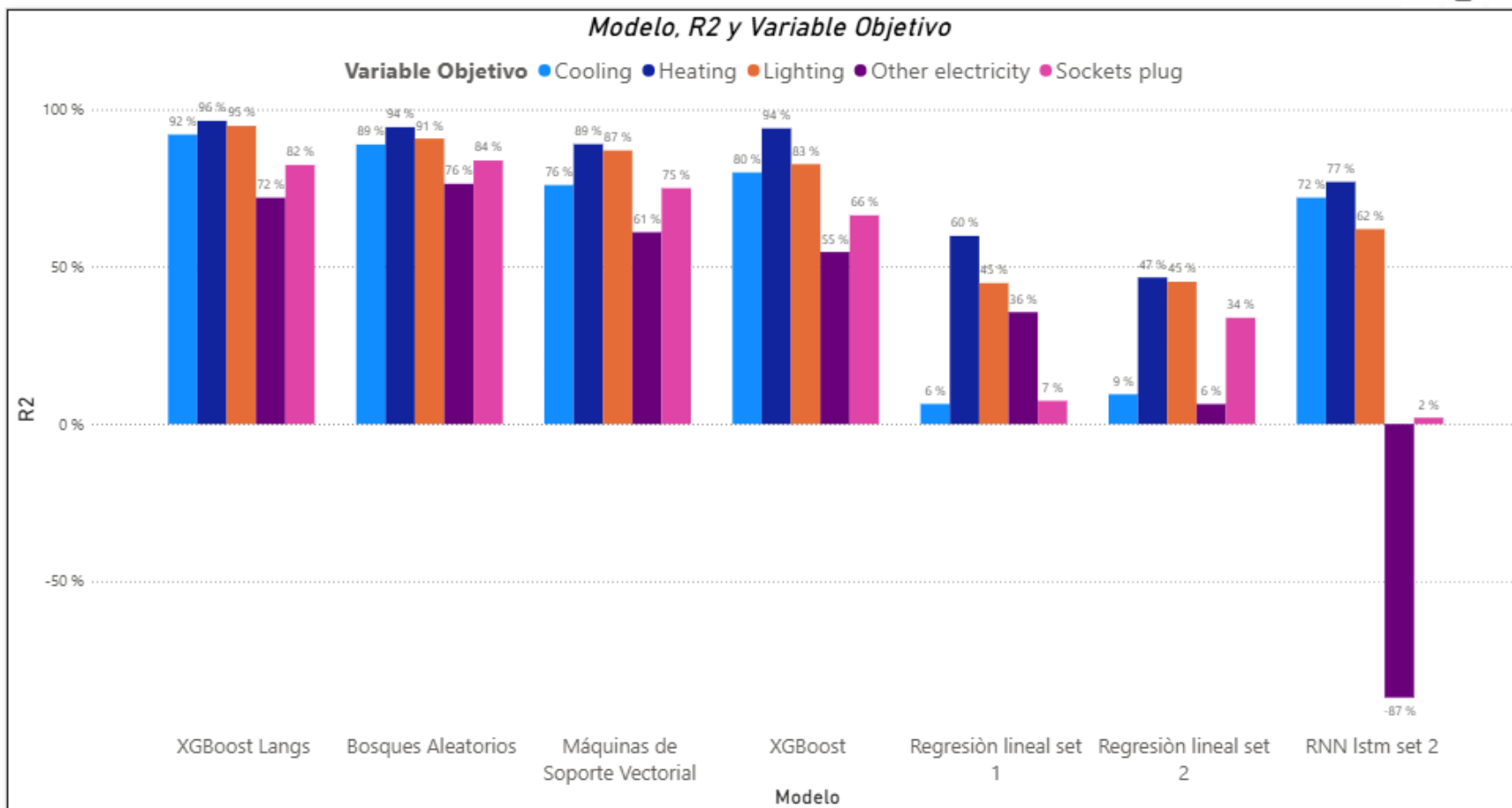


Imagen 9. comparación de resultados de los modelos implementados

5. Conclusiones

1. Los modelos basados en árboles de decisión, específicamente XGBoost Langs, Bosques Aleatorios (Random Forests), Máquinas de Soporte Vectorial (SVM) y XGBoost (segunda instancia), demostraron ser las soluciones más robustas y efectivas. Estos modelos alcanzan valores de R2 consistentemente superiores al 80% y a menudo por encima del 90% para la mayoría de las variables, indicando una excelente capacidad para capturar las relaciones no lineales y complejas de las dinámicas de consumo de energía, lo que los hace ideales para la predicción de Cooling y Heating.
2. Las variables de consumo energético asociadas al clima, Cooling (refrigeración) y Heating (calefacción), son las que presentan el mayor R2 en los modelos superiores (con picos de 96%), sugiriendo que la variabilidad de su consumo está fuerte y lógicamente correlacionada con las variables de entrada (ej. factores ambientales o temporales) utilizadas en los modelos. Por otro lado, la variable Lighting también muestra un alto R2 (hasta 94.8% con XGBoost Langs), lo que confirma su predictibilidad.
3. Los modelos más simples (Regresión Lineal) y el modelo de series temporales RNN Lstm muestran un rendimiento predictivo significativamente inferior. Los modelos de Regresión Lineal fallan drásticamente en predecir variables como Lighting y Sockets plug, con R2 en el rango de un solo dígito (6% - 9%), lo que confirma que la relación subyacente de estos consumos es predominantemente no lineal.
4. El modelo RNN Lstm presentó un rendimiento inaceptable para las variables Lighting y Sockets plug, con coeficientes de determinación 2 negativos (hasta -87). Este resultado implica que el modelo es peor para la predicción que simplemente tomar el promedio histórico del consumo, lo que sugiere un sobreajuste severo o una configuración inapropiada del modelo para este conjunto de datos específico de variables secundarias.

6. Porcentaje de avance

1. Procesar los datos de carga energética y consumo eléctrico a través del análisis exploratorio y correlaciones: Avance 100%
2. Evaluar diferentes modelos de machine learning para la desagregación e implementación de un modelo capaz de identificar y categorizar las cargas de los equipos eléctricos: Avance 85%
3. Desarrollar un mecanismo de reconciliación para alinear las predicciones de consumo generadas por distintos modelos según las categorías de carga, para coherencia y precisión en la desagregación: Avance 0%
4. Evaluar la precisión y eficiencia del modelo considerando diferentes tipos de cargas eléctricas: Avance 0%

Firmas del/los estudiante(s), director del Proyecto y codirector si aplica:

Atentamente,



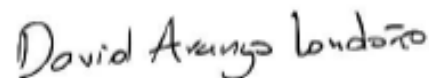
Meidy Lizeth Moreno Guzman
C.C. 1014223269 de Bogotá.



Jose Luis Munévar Diaz
C.C. 1036646650 de Medellín.



Julian Javier Jaimes Otero
C.C 91505906 de Bucaramanga.



David Arango Londoño
C.C. 1036646650 de Cali.