INFORME DEL MODELO DE PREDICCIÓN DE CONSUMO ELÉCTRICO EN HOGARES

Luis Guillermo Fernandez Suarez

MAESTRÍA CIENCIA DE DATOS - USFQ

Tabla de contenido

1. Título y Objetivo de Proyecto

1.1. Objetivo

Determinar un modelo de predicción de le demanda de energía eléctrica horaria a lo largo de un año para la estimación de los requerimientos de abastecimiento en nuevos proyectos inmobiliarios.

1.2. Visión del negocio

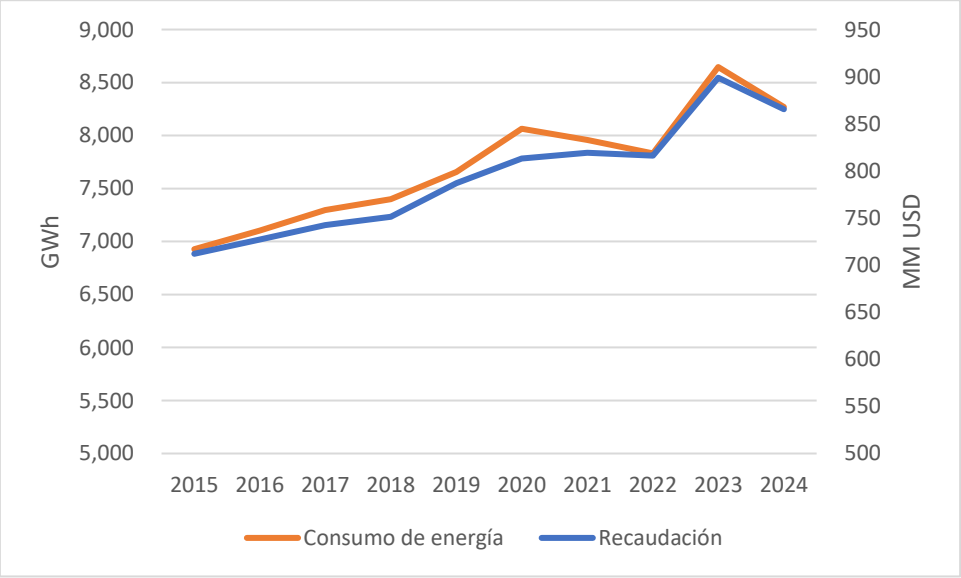
El modelo buscar optimizar el abastecimiento anual de la demanda eléctrica a través de la predicción horaria para evitar problemas de suministro de energía eléctrica como cortes en el servicio, flickers (parpadeos o fluctuaciones de la intensidad) u otros.

2. Contexto y Alcance

2.1. Antecedentes (Business Understanding – CRISP-DM):

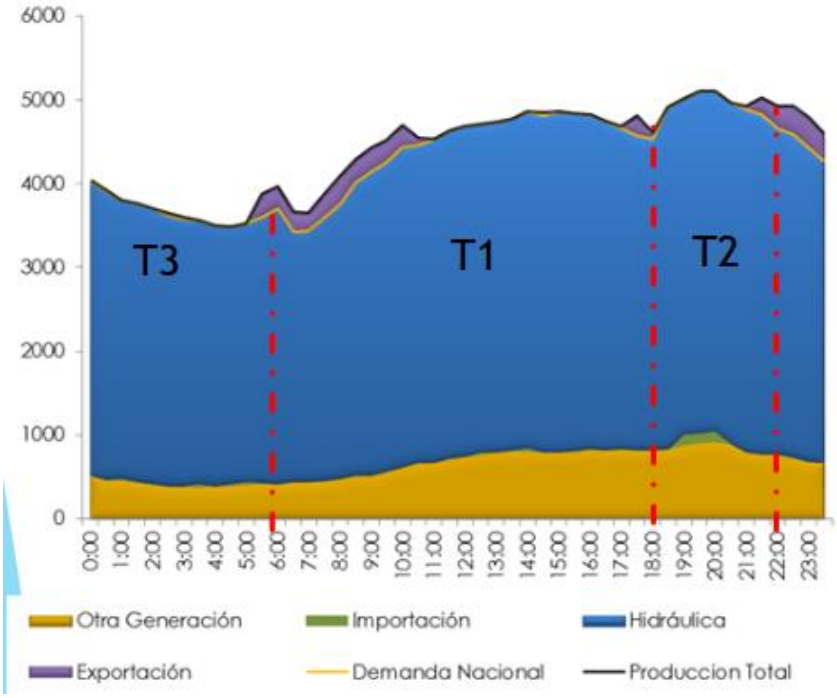
En Ecuador el consumo de energía en el sector residencial ha ido evolucionando en los últimos 10 años desde aproximadamente los 6900 GWh hasta los 8200 GWh registrándose un aumento del 22%. De la misma manera como se observa en la Figura 1, se ha incrementado el valor de la recaudación desde aproximadamente los 700 MM USD hasta los 850 MMUSD con un porcentaje de 19% de crecimiento. (ARCONEL, 2025)

Figura 1. Energía facturada y recaudación en el grupo de consumo residencial del Ecuador



En el Ecuador, a pesar de que el sistema tarifario en el grupo de consumo residencial tiene una sola tarifa en el pago final. Internamente la tarifa se encuentra cubierta de manera referencia como se muestra en la siguiente figura. (CENACE, 2025)

Figura 2. Demanda de energía a nivel Nacional



Dentro de las tarifas referenciales se puede mencionar qué, que la Tarifa 1 (T1) y la Tarifa 2 (T2), se encuentran en lo que se conoce horas valle y se pueden definir como horas en donde el sistema de generación eléctrica nacional, no se encuentra en un estado de pico de demanda eléctrica y su costo promedio de generación es 10 ctvs/kwh. Mientras que en la Tarifa 2(T2) de 18:00 a 22:00 al ser hora punta el costo de generación de la energía es de aproximadamente 17 ctv/kwh. Específicamente en los hogares se asume que el rango de ocupación, el uso de iluminación y otros electrodomésticos, genera estos picos de energía. Teniendo en cuenta el costo de generación en horas pico, los modelos de predicción de la demanda de energía eléctrica se vuelven indispensables para una planificación de abastecimiento de esta energía al menor costo posible para el país. Por tanto, se busca determinar un modelo de predicción de demanda en el sector residencial en el que se pueda obtener información de para estimar la demanda de nuevos proyectos inmobiliarios

2.2. Alcance

Elaboración de un modelo de predicción de demanda eléctrica basado en los features propuestos en el data set "Energy consumption prediction" disponible en https://www.kaggle.com/datasets/ajinilpatel/energy-consumption-prediction/data.

2.3. Limitaciones

El data set en el cual se basa el modelo de predicción (https://www.kaggle.com/datasets/ajinilpatel/energy-consumption-prediction/data) no especifica el número de hogares, el país y en los cuales se levantó la información. Por lo que se tiene se debería actualizar a la realidad ecuatoriana los datos.

3. Entendimiento de los Datos

Para este modelo se está utilizando el data set obtenido de el repositorio kaggle disponible en el siguiente enlace: https://www.kaggle.com/datasets/ajinilpatel/energy-consumption-prediction/data. Este data set mas provine de una recopilación de información en campo de 1000 incidencias, y adicionalmente se generaron 4000 incidencias adicionales con data sintética, además de esta información, se detalla en la descripción de la data set qué su tiempo de actualización es cada cuatrimestre.

La siguiente figura muestra la estructura del data set, en el cual se puede observar que el mismo consta con 5000 incidencias, 11 features de las cuales 4 son tipo “object”, 3 de tipo “int64” y los restantes del tipo “float64”.

Figura 3. Estructura de datos

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Month              5000 non-null   int64
 1   Hour               5000 non-null   int64
 2   DayOfWeek          5000 non-null   object
 3   Holiday            5000 non-null   object
 4   Temperature        5000 non-null   float64
 5   Humidity           5000 non-null   float64
 6   SquareFootage      5000 non-null   float64
 7   Occupancy          5000 non-null   int64
 8   HVACUsage          5000 non-null   object
 9   LightingUsage      5000 non-null   object
 10  RenewableEnergy    5000 non-null   float64
 11  EnergyConsumption  5000 non-null   float64
dtypes: float64(5), int64(3), object(4)
memory usage: 468.9+ KB
```

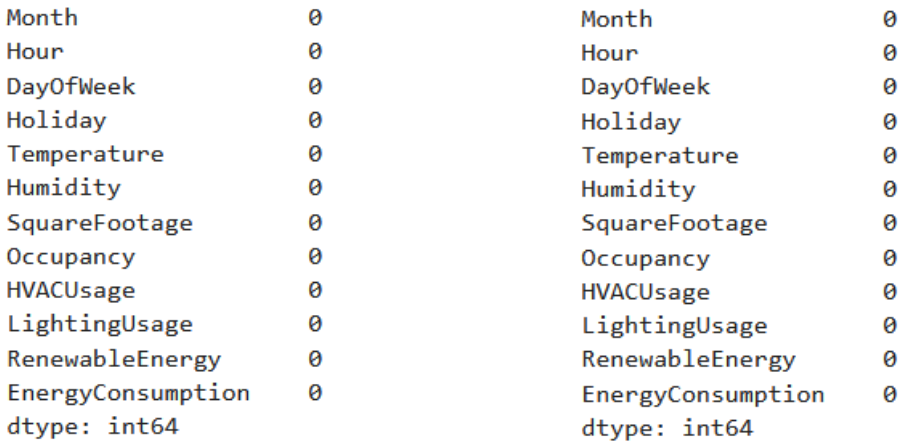
Además, en la siguiente tabla se puede mostrar el diccionario de variables de data set.

Tabla 1. Diccionario de variables del data set

Features	**Valores**
Month	1 – 12 Valores numéricos que representan los meses del año
Hour	1 – 24 Valores numéricos que representan las horas
DayOfWeek	Monday, Tuesday, Wednesday, Thrusday, Friday, Saturday, Sunday
Holiday	Yes No
Temperature	20.01-30.00
Humidity	30.02-59.97
SquareFootage	1000.51-1999.98
Occupacy	1-9
HVACUsage	On Off
LightiningUsage	On Off
RenewableEnergy	0.0066-53.36
EnergyConsuption	29.96-99.20

Se realizaron los análisis de valores nulos (isnull) y valores NaN (Non a number en ingles), dentro de los resultados, no se pudieron identificar valores que cumplan con estos requisitos. (figura 4).

Figura 4. Valores Is nulos (izquierda) y las valores no números (derecha)



Sobre los valores duplicados, se pueden mencionar los valores reportados en la figura 5.

Figura 5. Valores duplicados en el data set

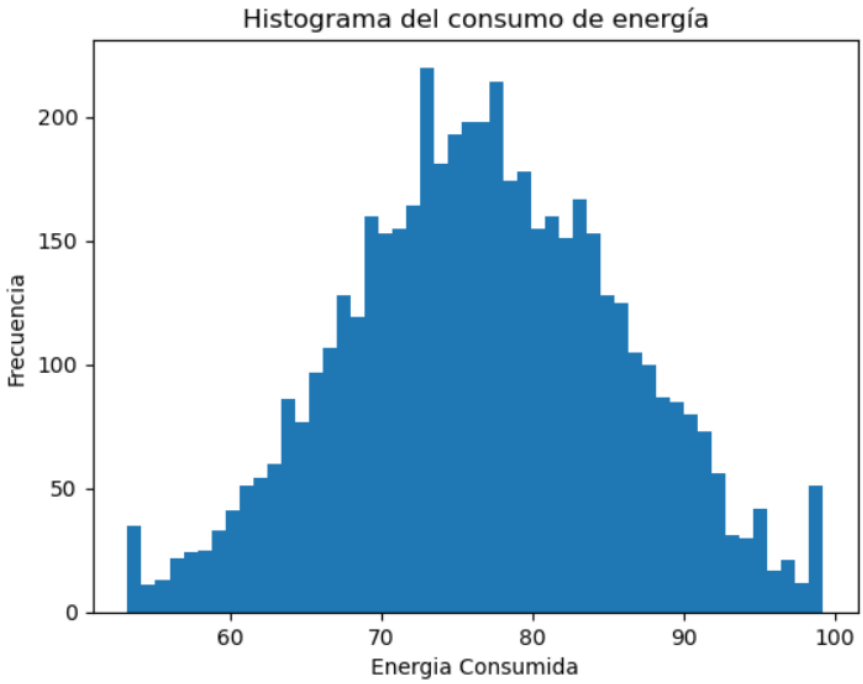
```
Month: 4988 duplicados
Hour: 4976 duplicados
DayOfWeek: 4993 duplicados
Holiday: 4998 duplicados
Temperature: 591 duplicados
Humidity: 511 duplicados
SquareFootage: 290 duplicados
Occupancy: 4990 duplicados
HVACUsage: 4998 duplicados
LightingUsage: 4998 duplicados
RenewableEnergy: 525 duplicados
EnergyConsumption: 63 duplicados
```

Con la información presentada en la figura 5, se puede mencionar lo siguiente sobre los valores duplicados:

- Month, Hour, DayofWeek, Holiday, HVACUsage, LightingUsage: esta información representa valiables categoricas, es por eso que se pueden identificar valores duplicados, pero la combinación independiente de cada una de ellas puede agrega valor a cada una de las incidencias.

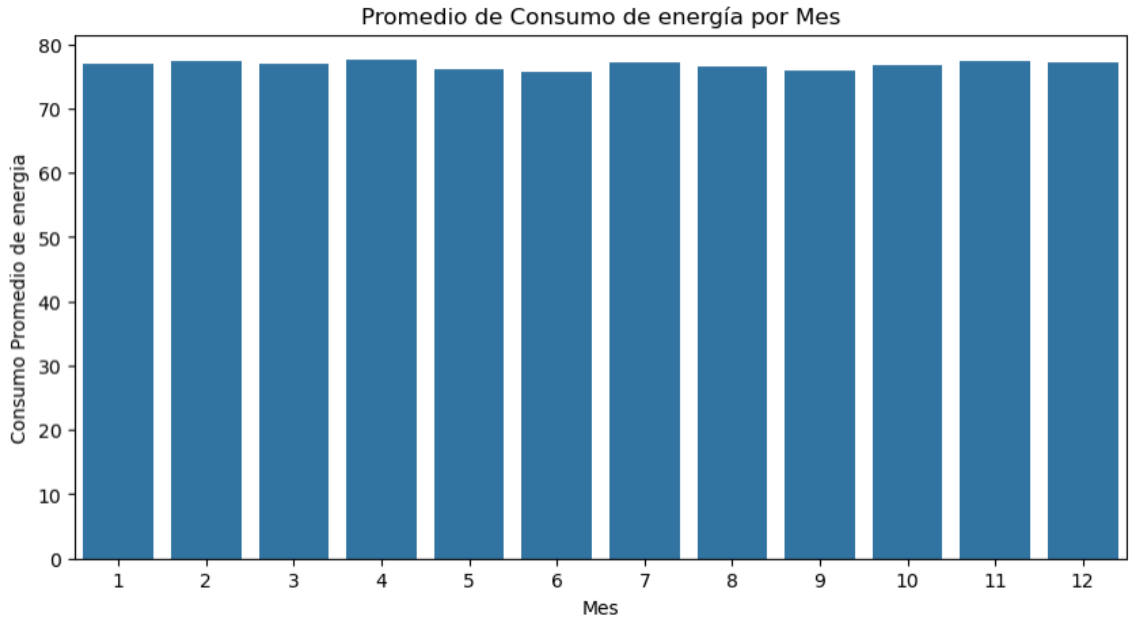
Una vez descritas las variables se procede con el análisis estadístico de los datos, la figura 6 muestra el histograma del consumo de energía eléctrica en todas las incidencias. Se puede observar que estas siguen una distribución normal entre los 29.96-99.20 Kwh

Figura 6. Histograma del consumo de energía



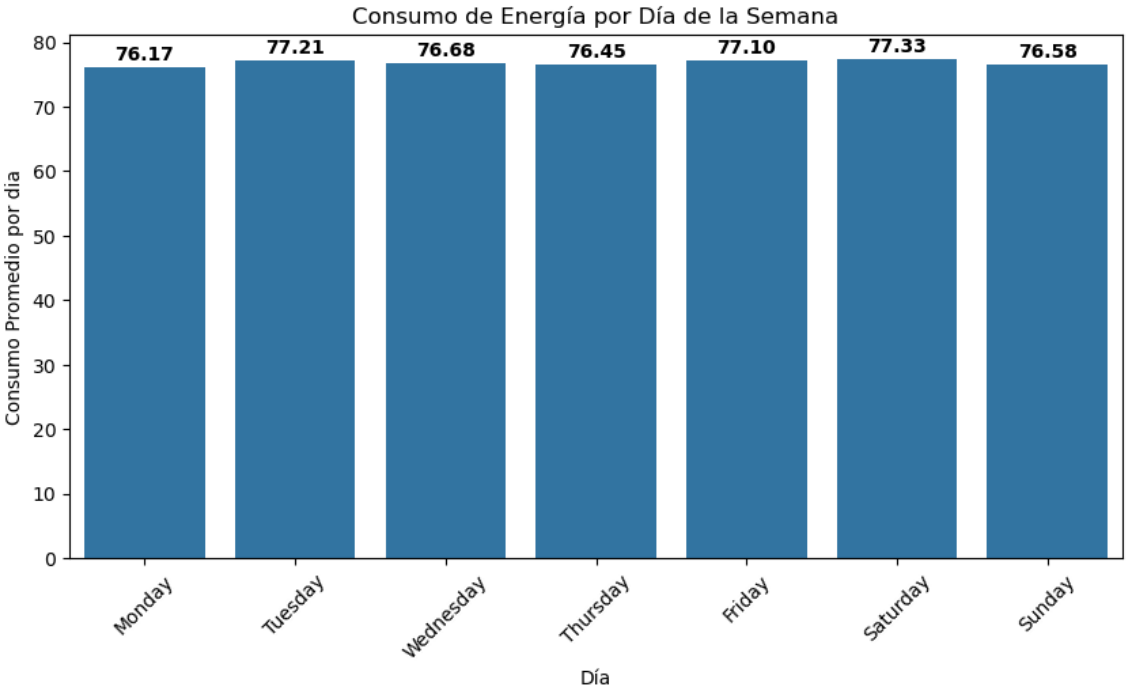
Filtrando el consumo promedio de energía a nivel mensual, se puede observar en la siguiente figura, que no existen variaciones significativas, ni se muestran signos de estacionalidad. Lo que plantea preguntas sobre en que tipo de hogares se realizó el levantamiento de información.

Figura 7.Promedio de consumo de energía en hogares por mes



Realizando un análisis más detallado se puede observar de igual manera que en la figura anterior, que, no existe una variación significativa de el consumo de energía a nivel de día, abriendo otro punto mas de inquietud sobre el origen del data set.

Figura 8. Consumo de energía diaria

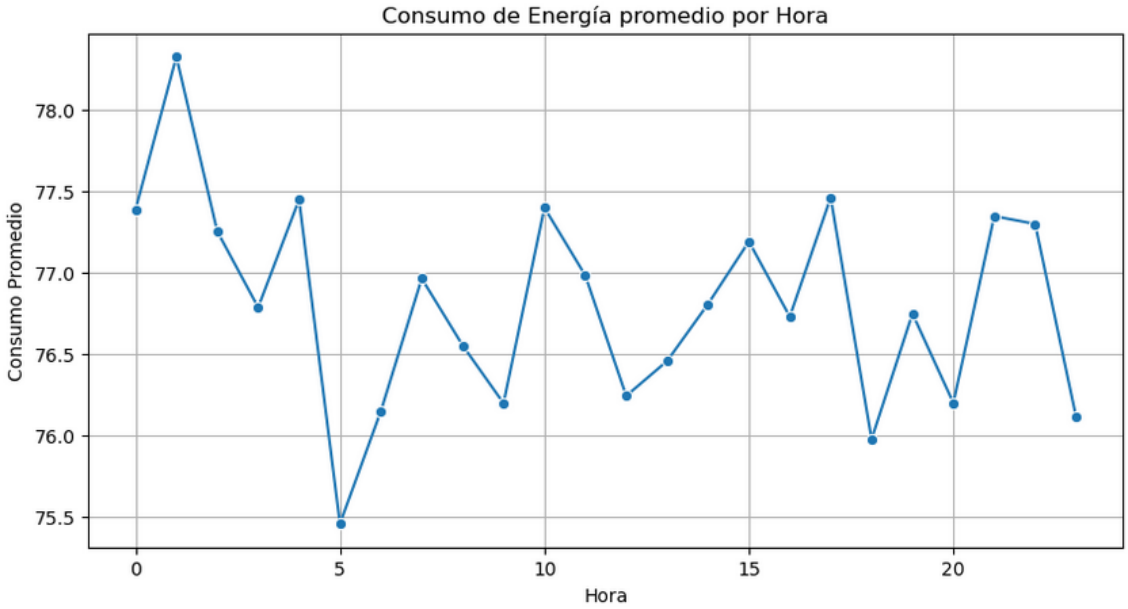


C

on el fin de obtener mas información desagregada del origen de estas variaciones, se puede observar que, a nivel horario se presentan variaciones para analizar. La más importante de mencionar es el consumo de energía en horario 1:00, que viene a ser la hora punta del gráfico. Se puede interpretar de manera contra intuitiva ya que a esta hora en general los hogares

duermes, si bien se debe analizar más a fondo el origen del data set, se puede plantear la hipótesis de que tiene un origen europeo y ciertos electrodomésticos pueden programarse para el consumo de energía en horas de la madrugada por su costo.

Figura 9. Consumo de energía diaria del data set



Con el fin de analizar el impacto del consumo de energía eléctrica en los días feriados, y el consumo de energía eléctrica en aires acondicionados se plantean las siguientes figuras, en las cuales se puede identificar que hay un ligero incremento (0.7%) en los días feriados y por uso de aire acondicionado.

Figura 10. Consumo de energía en feriados

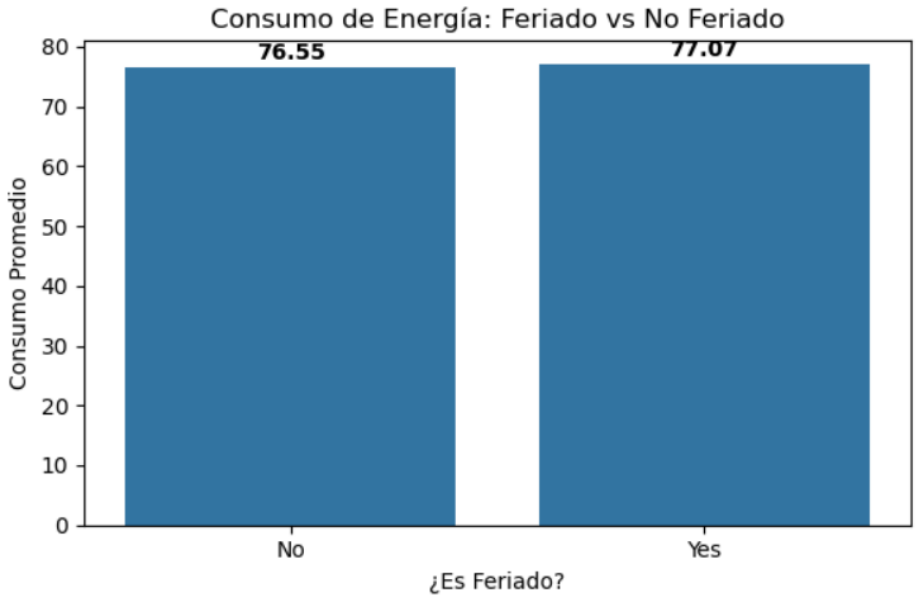
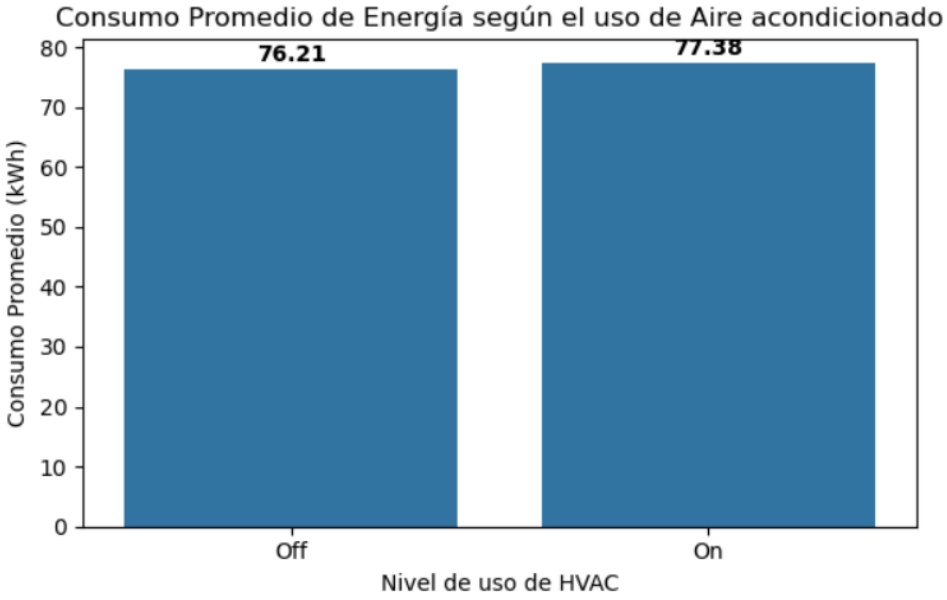
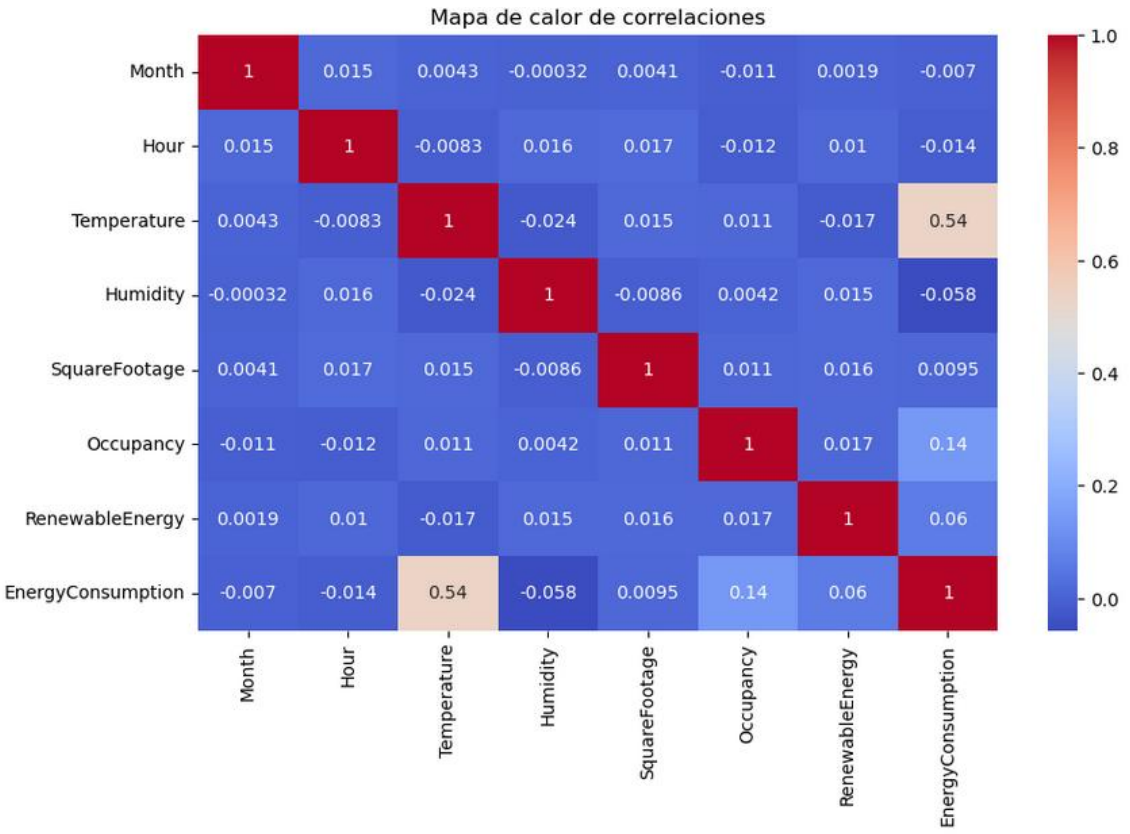


Figura 11. Consumo de energía por aire acondicionado



Con el conocimiento de que el consumo de energía (EnergyConsumtion) es el feature que se busca proyectar, a través del mapa de correlación podemos ver cual es la feature que más influye. La temperatura se presenta como la variable que más podría afectar al consumo de energía, seguido por la ocupación y el autoabastecimiento del sistema. Además, se puede observar que la humedad es una de las variables que menos afecta.

Figura 12. Mapa de correlación de variables



4. Preparación de Datos

Debido a que no se identificaron inconsistencias grandes dentro del Data ser se realizó el siguiente proceso:

1. Eliminación del feature “Humidity” ya que no tiene influencia significativa en el consumo de energía
2. Normalización de datos para optimizar el cálculo del modelo. Por ejemplo, la figura 12 muestra el diagrama de dispersión entre la temperatura y consumo de energía eléctrica donde la temperatura no se encuentra normalizada (rango de 19.99 a 30), mientras la figura 13 muestra la misma figura ya con la información normalizada (rango de -1.7 a 1.7)

Figura 13. Dispersión entre energía vs temperatura (sin normalizar)

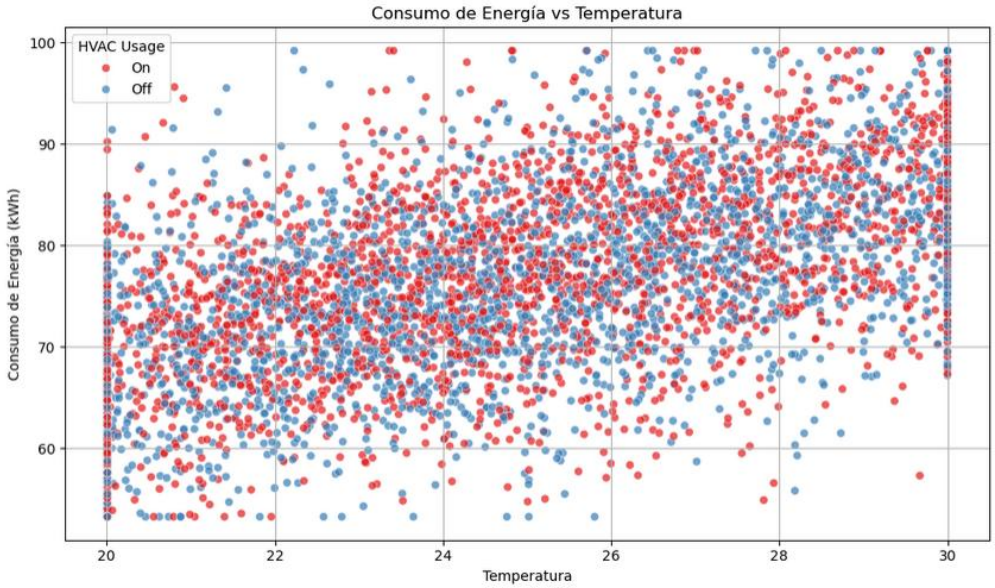
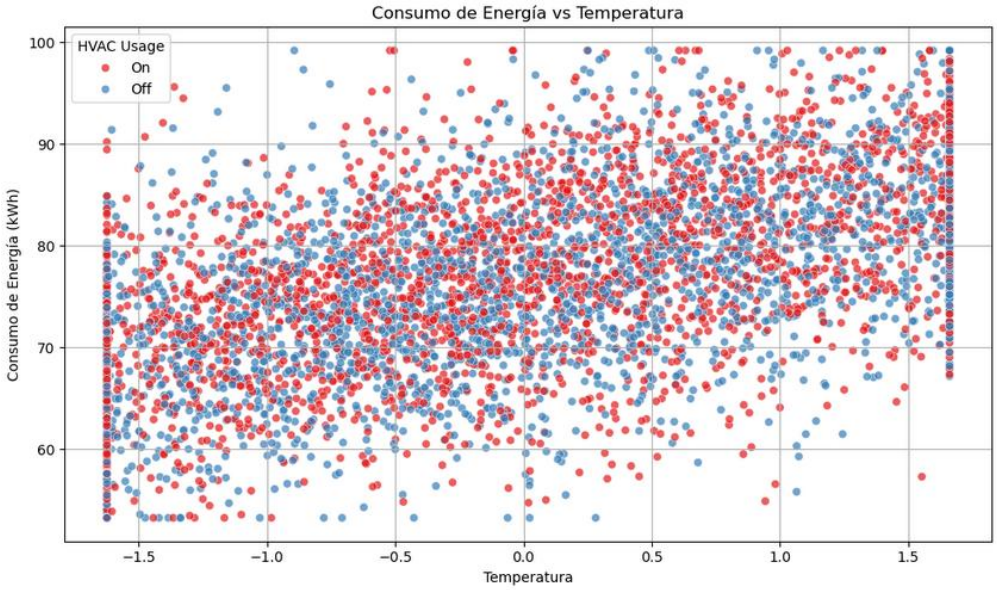


Figura 14. Dispersión entre energía vs temperatura (normalizado)



5. Modelado

Debido a que el consumo de energía se presenta como una variable de predicción se decide utilizar modelos de regresión para obtener la información requerida.

En ese sentido, se plantea utilizar 4 modelos de regresión lineal los cuales provienen de la biblioteca “sklearn” y los cuales son “Linear Regression”, “Random Forest” y “Bagging Regressor”. Para estos modelos se propone dividir el data set en 80% para entrenamiento y 20% para validación.

En la primera interacción con los modelos en estándar sin personalización se obtuvieron los siguientes resultados del modelo.

Tabla 2. Resultados de modelos de regresión Estándar

```
Tabla resumen de modelos de regresión:
                Model          RMSE   R2
0   Linear Regression  7.150484e-16  1.0
1       Random Forest  0.000000e+00  1.0
2   Gradient Boosting  1.859250e-05  1.0
3   Bagging Regressor  0.000000e+00  1.0
```

Debido a los resultados obtenidos específicamente en RMSE y R^2, se plantea una personalización de los hiperparámetros de los modelos, después de varias se utiliza los siguientes:

Tabla 3. Hiperparametros elegidos para los modelos de regresión

Modelo	**n_estimators**	**max_depth**	**learning_rate**	**max_features**	**max_samples**
Linear Regression					
Random Forest	20	5			
Gradient Boosting	30	3	0.1		
Bagging Regressor	10			0.5	0.5

En la tabla 4 se muestra los resultados de los modelos de regresión personalizados.

Tabla 4. Resultados de modelos de regresión personalizados

```
Tabla resumen de modelos de regresión:
                Model          RMSE        R2
0   Linear Regression  4.308466e-16  1.000000
1       Random Forest  0.000000e+00  1.000000
2   Gradient Boosting  2.967304e-02  0.998203
3   Bagging Regressor  4.324697e-01  0.618278
```

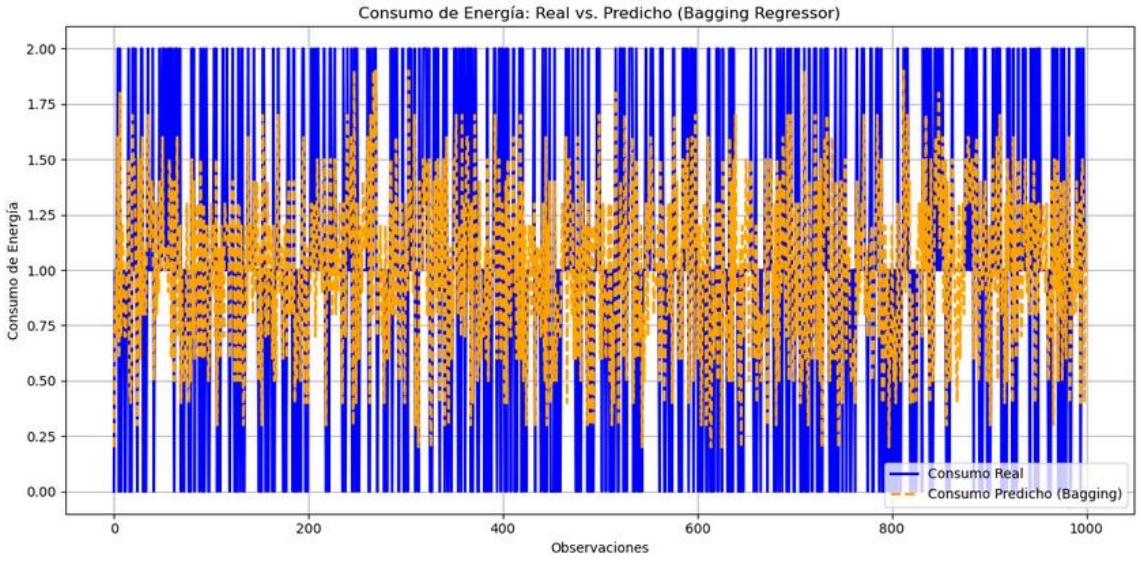
6. Evaluación e Interpretación de Resultados

En la Tabla 2 se puede observar que los coeficientes de R^2 se obtienen valores de 1, y en RMSE se obtinen valores muy cercanos a 0. Estos valores nos pueden dar signos de sobre entrenamiento del modelo (over fitting) lo que significa que nuestro modelo no se encuentra aprendiendo, sino ha memorizado los resultados.

Por otro lado se puede observar en la Tabla3 que a pesar de la personalización de los hiperparámetros, los modelos de Linear Regression, Random Forest y Gradient Boosting muestran signos de sobre entrenamiento con un R^2 mayores a 0.99. Por ultimo, el Bagging Regressor muestra un R2 con valores de 0.61 que indica que el modelo se encuentra aprendiendo. Es por eso que se decidió el análisis de resultados con este modelo.

En la figura 14 se puede observar de manera normalizado el consumo de energía eléctrica real reportado en el data set, así como el el consumo predicho por el modelo de Bagging Regressor. Debido a el contexto de esta predicción, donde los picos de consumo de energía eléctrica son los mas importantes, se puede observar que el modelo con los hiperparámetros que se están utilizando, esta logrando captar la información necesesaria.

Figura 15. Consumo de energía real vs Consumo de energía Predicho con Bagging Regressor normalizados



Por lo que, en un futuro se puede analizar el tener un data set mas grandes, procurar que la mayoría de las incidencias sea recopilada de manera real, ya que los datos sintéticos pueden estar sesgando la información. Además, es importante tener en cuenta el contexto sobre la recopilación de esta información para poder ubicar los picos de demanda diaria mostrados en las secciones anteriores.

7. Plan de Implementación

Debido a que actualmente el modelo no presenta los resultados esperados, se esperar poder realizar iteraciones y pruebas de otros modelos para regresión. Una vez terminado este proceso y se logre obtener valores en un umbral en R2 de 0.8. Se plantea el siguiente proceso de implementación.

Se propone una arquitectura en una maquina virtual dentro del servidor de una la empresa. Que debe seguir las siguientes consideraciones:

Requisitos: 4 cores, 8 GB memoria RAM

Versión Anaconda Navigator: 2.6.5

Además, se tiene pensado en la elaboración de un dashboard en Power BI con indicadores de energía consumida, numero de ocupación y uso de aire acondicionado.

Se plantea que la actualización del modelo se debe hacer cada 6 meses.

8. Conclusiones, Próximos Pasos y Recomendaciones

- Debido a que el contexto de este desarrollo y que los picos de uso de energía su predicción es fundamental para este modelo, se debe trabajar en mejorar el modelo de predicción como hiperparámetros o experimentación de otros modelos
- Teniendo en cuenta que en el mapa de correlación se observa que la temperatura es la feature que tienen más influencia el consumo de energía, se plantea que los próximos pasos deben trabajar en data sets de proyecciones climáticas (temperatura) y proyectos de viviendas para poder predecir la demanda de energía eléctrica.
- Debido a los hallazgos encontrados en la demanda diaria de energía eléctrica, se debe profundizar más en la recopilación de información para actualización del data set.

9. Apéndices

Bibliografía

ARCONEL. (2025). *www.controlelectrico.gob.ec*. Obtenido de https://controlelectrico.gob.ec/

CENACE. (2025). Obtenido de https://www.cenace.gob.ec/