Predicción del consumo energético en la zona de Gran Buenos Aires

Carrera: Ingeniería en Computación
Título de la Tesis: Magister en Inteligencia de Datos orientada a Big Data
Autor: Franco Ariel Demare
Director/Codirector: Prof. Dr. Aurelio F. Bariviera
e-mail de contacto del tesista: frandemare@gmail.com
github:
https://github.com/francodemare/Energy-consumption-forescanting-in-the-Greater-Buenos-Aires-area/tree/main

Biografía académica/profesional resumida

Inicié mis estudios en la carrera de Ingeniería en computación en el año 2011 en la Universidad Nacional de La Plata y me recibí en el año 2017. Desde que terminé la carrera, me encuentro trabajando en empresas de informática, puntualmente en automatización de procesos de negocio con diversos software de low code, python y las últimas tendencias de IA.

Palabras Claves: *Aprendizaje automático; Modelo estadístico; Consumo energético; Redes neuronales; Series Temporales; Predicción*

Motivación

La motivación principal de este trabajo radica en la creciente necesidad de predecir con precisión la demanda de energía eléctrica debido a la rápida expansión económica y el aumento del consumo energético, tanto industrial como doméstico. A medida que las sociedades y economías se desarrollan, la demanda energética crece exponencialmente, lo que plantea el reto de garantizar un suministro eficiente y sostenible de electricidad. La predicción precisa del consumo es fundamental para optimizar la gestión de recursos, reducir costos y minimizar el impacto ambiental, en línea con los Objetivos de Desarrollo Sostenible (ODS).

Este trabajo de tesis se centra en aplicar y comparar modelos predictivos, tanto estadísticos como de aprendizaje automático, con el objetivo de identificar el modelo que ofrezca la menor tasa de error en la predicción del consumo diario de electricidad en el Área Metropolitana de Buenos Aires.

Aportes de la tesis

El aporte central de la tesis reside en la comparación y desarrollo de modelos predictivos avanzados para la estimación del consumo de energía eléctrica en la región del Gran Buenos Aires (GBA), utilizando un enfoque innovador con modelos estadísticos y de redes neuronales. El objetivo principal del trabajo es seleccionar y proponer el modelo más adecuado, basándose en el menor error de predicción y su capacidad para ser escalado computacionalmente en tamaño de datos otras.

El trabajo introduce la implementación de modelos tradicionales como SARIMA, conocido por su uso en series temporales con estacionalidad, y modelos más avanzados como la regresión de vectores de soporte (SVM), con resultados competitivos en términos de precisión y tiempos de procesamiento. Posteriormente, se exploran modelos más complejos basados en redes neuronales como CNN y LSTM, que ofrecen mejoras notables en la predicción del consumo energético.

El valor más destacado del trabajo es la introducción de una metodología híbrida, combinando redes neuronales con técnicas de descomposición empírica de modos (EMD), lo que permite una mejora considerable en la precisión de la predicción. En particular, el modelo LSTM-EMD se consolida como el mejor, con un MAPE (**error porcentual absoluto medio**) del 4,1%, lo que demuestra su capacidad para predecir con alta precisión el consumo eléctrico.

En la tesis también se destaca el enfoque en la escalabilidad y el tiempo de procesamiento, variables cruciales para la implementación de estos modelos en entornos reales con grandes volúmenes de datos. La metodología de validación utilizada —una grilla de validación— aporta estabilidad a los resultados, permitiendo mayor seguridad en la evaluación del mejor modelo. Este enfoque, menos tradicional, evita la simple separación de datos entrenamiento y prueba, lo que asegura resultados más robustos.

En conclusión, este trabajo no solo avanza el campo de la predicción del consumo energético mediante técnicas de machine learning, sino que también propone un modelo escalable, eficiente en tiempo y costo computacional, adecuado para futuras aplicaciones en otros contextos geográficos.

Datos

En el sitio web de la Secretaría de Energía (http://datos.energia.gob.ar/dataset), se puede encontrar el conjunto de datos que se utilizó.

Dentro del sitio, se realizó un filtro por el campo “Demanda de Energía" y se descargó el dataset llamado “Base Demanda Diaria 2017 a 2023.xlsx"

El conjunto de datos básicamente se puede ver como diferentes series de tiempo. Estas series contienen, por día, el consumo energético expresado en MW (megavatio) de distintas regiones del país. En este trabajo, se utilizarán únicamente los valores de una serie temporal, la de consumo de la región del Gran Buenos Aires.

Estacionariedad

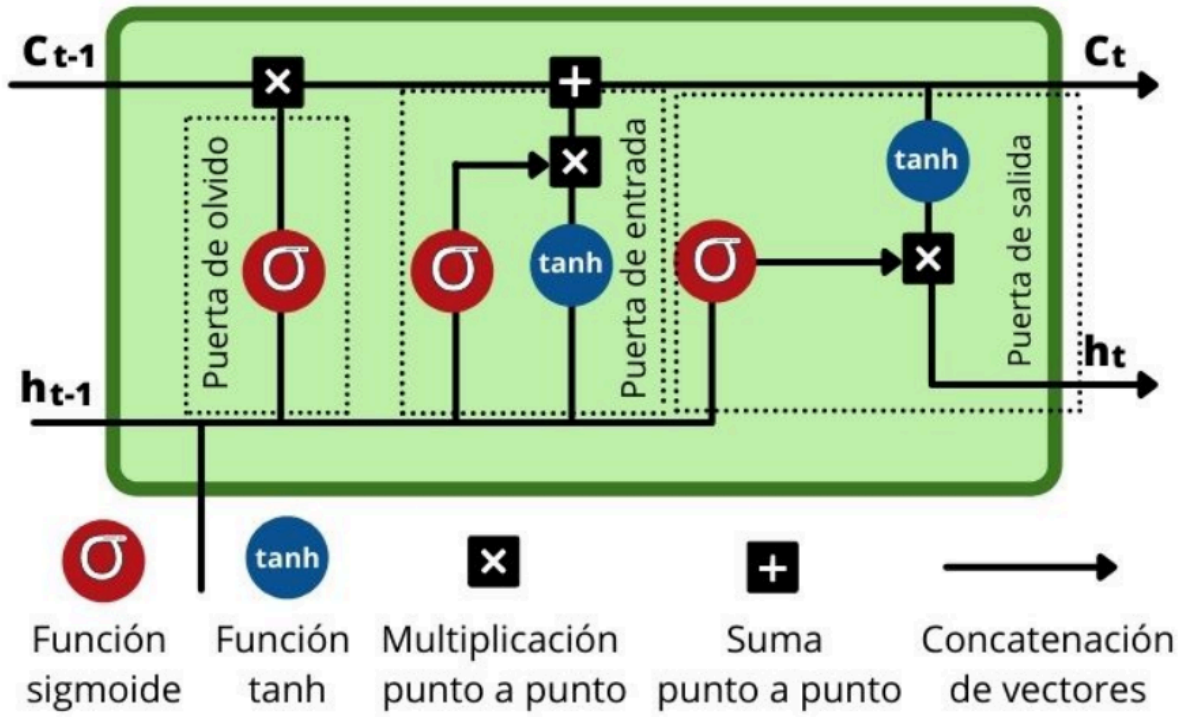
Para verificar la estacionariedad de los datos, se realizó un test de Dickey-Fuller Aumentado (ADF) en un conjunto de datos determinado. La prueba ADF se usa para detectar estadísticamente la presencia de conducta tendencial estocástica en las series temporales de las variables. En esta prueba, la hipótesis nula es que la serie temporal contiene una raíz unitaria y, por tanto, no es estacionaria. Por otro lado, la hipótesis alternativa es que no hay una raíz unitaria en la serie temporal. Por tanto, para poder concluir que una serie temporal es estacionaria se deberá rechazar la hipótesis nula.

Modelos de predicción

Modelo de Redes neuronales de memoria de corto-largo plazo (LSTM)

Las LSTM son un tipo de redes neuronales recurrentes donde cada célula de memoria o

memory cell tiene un grupo de operaciones muy específicas que permiten controlar el flujo de información. Estas operaciones, llamadas puertas permiten decidir si cierta informaci´on es recordada u olvidada. Dentro de la célula de memoria nueva información es añadida a la que proviene de secuencias anteriores, es decir, de pasos de tiempo previos. La información relevante nueva es
agregado al flujo gracias a una operación de adición.



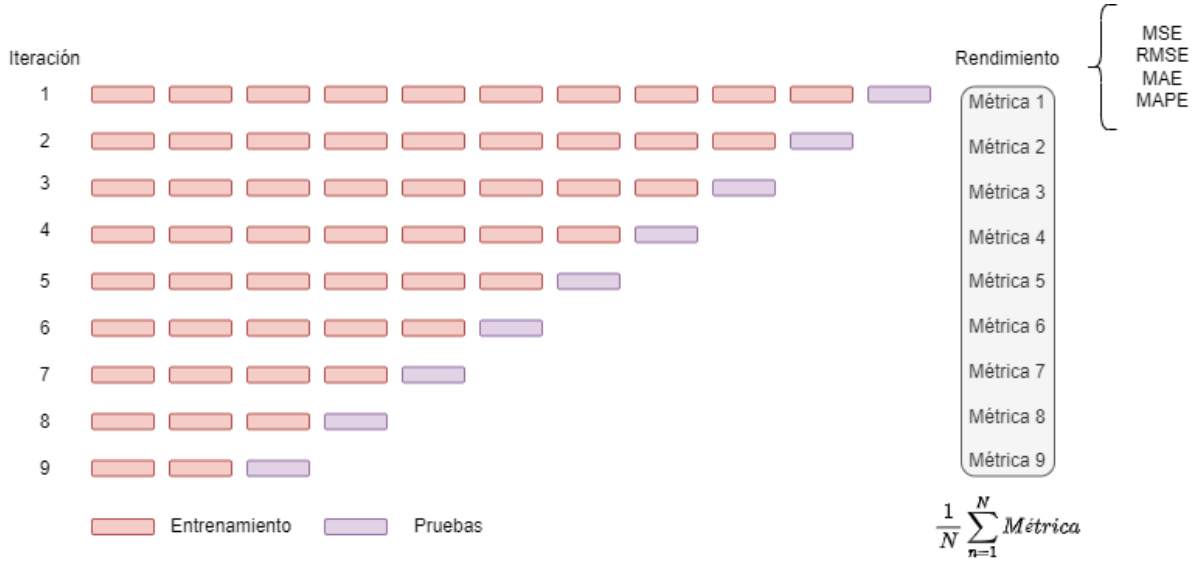
Descomposición de modos empíricos

La descomposición de modos empíricos (EMD) es una técnica de multi-resolución adaptativa de datos para descomponer una señal en componentes físicamente significativos. La particularidad de esta técnica es que no utiliza ninguna función prescrita, sino que se adapta automáticamente en función de la señal analizada y de ahí el adjetivo “empírica".

En este caso, se decidió aplicar dos variaciones de la descomposición de modos empíricos: el algoritmo original EMD y el CEEMDAN. Para ello, se utilizó la librería EMD.

Grilla de Validación

Se generó una validación de los modelos por lote.



Cada modelo se lo evalúo haciendo este mismo tipo de validación. El método consiste en tomar un lote inicial de datos, en este caso se inició por tomar el total de los datos. En esta primera etapa, el lote se separa en 80% datos de entrenamiento y 20% datos de pruebas. Se ejecuta el modelo y luego se obtienen las cuatro medidas de error detalladas anteriormente.

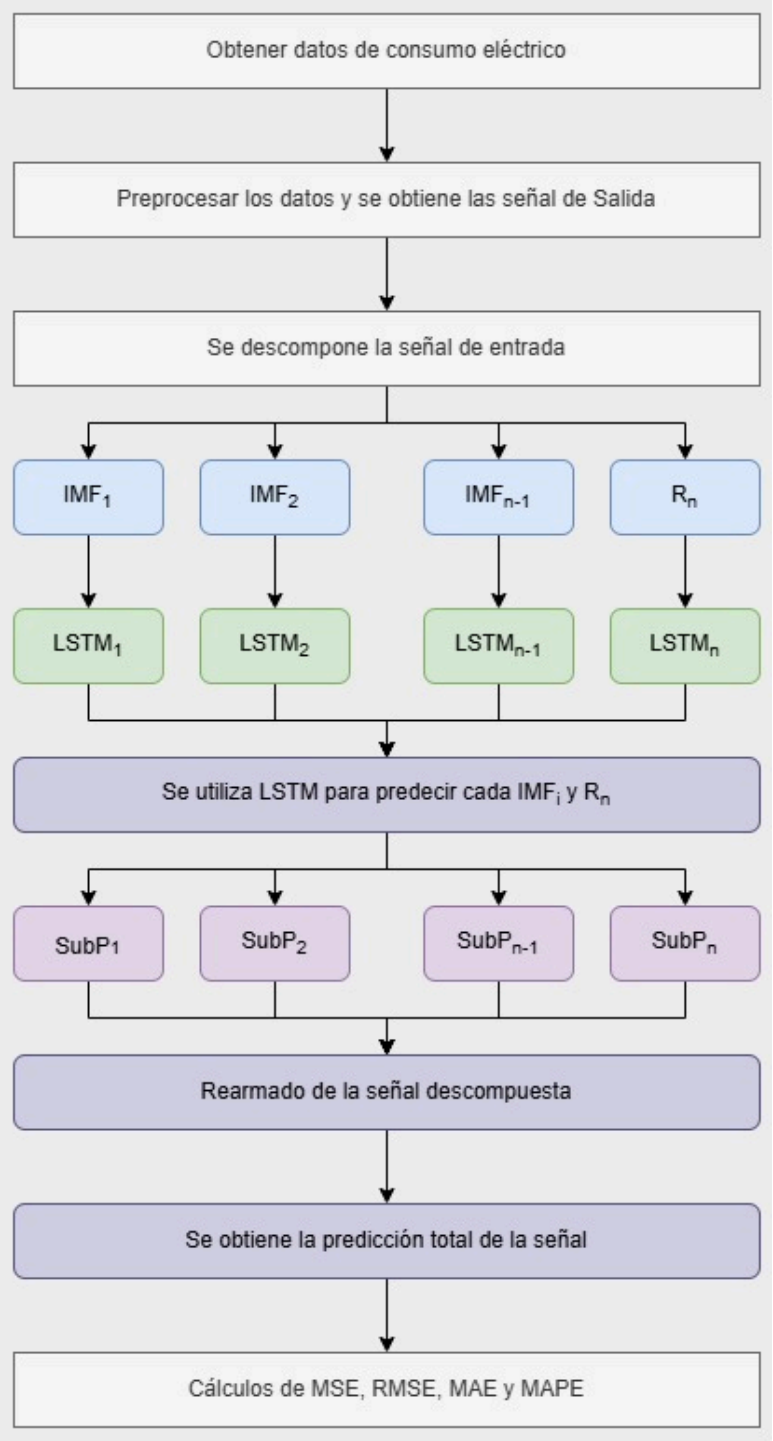
Luego en la siguiente iteración, se toma nuevamente otro lote pero ahora con un conjunto reducido de datos, para este caso en concreto, se realizó una reducción del 20% del lote inicial. Se realiza nuevamente la ejecución del modelo separando los datos remanentes en 80% para datos de entrenamiento y 20% para datos de prueba, y se vuelven a tomar las medidas de error. De esta manera, las medidas de error se van acumulando en variables y los lotes, por cada iteración, se van reduciendo.

Estos pasos se van a repetir hasta que el lote alcance un tamaño mínimo de datos. Para este desarrollo el tamaño mínimo es el 15% del set de datos.

Arquitectura final del modelo propuesto

Como primera instancia se obtienen los datos del consumo eléctrico a partir del set de datos descargado, luego de procesar esos datos, como se comentó en la sección.

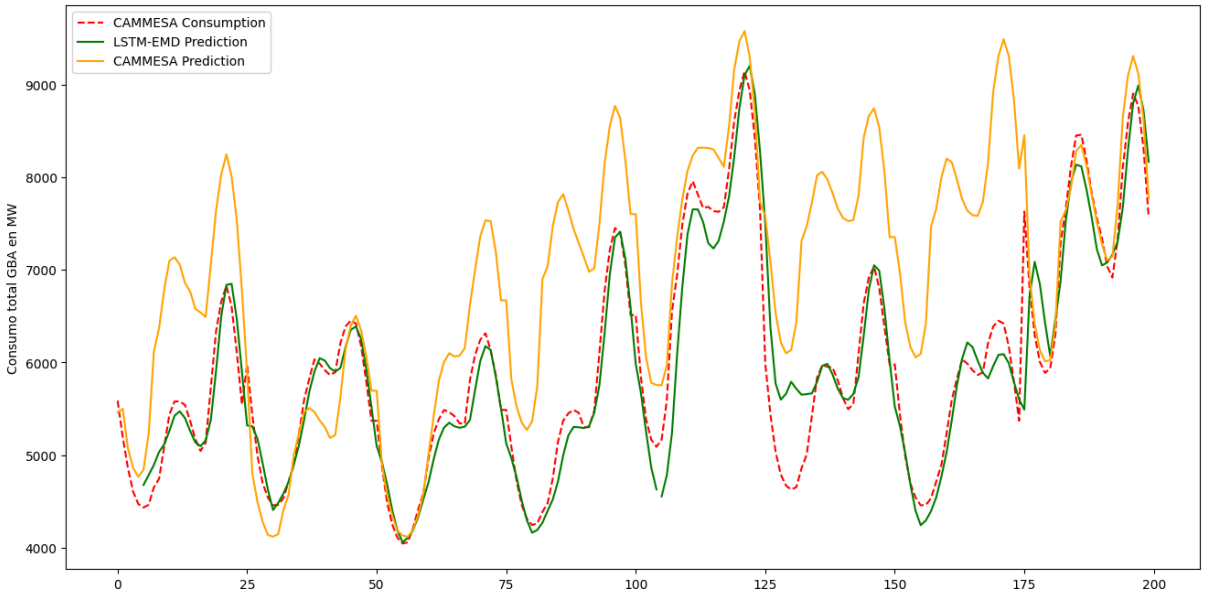
Luego se aplica la descomposición en modos empíricos, lo que nos da una serie de nuevas funciones generadas, para luego aplicarles el modelo LSTM a cada una de ellas. Finalmente, las predicciones de cada una de las funciones se rearma para obtener la función y predicción final. Como paso adicional, se realizan los cálculos de los errores definidos anteriormente, para comprobar el rendimiento del modelo.



Predicción del modelo

Se obtuvieron datos intradiarios oficiales de CAMMESA (ente regulador de energía eléctrica). Con estos datos intradiarios se realizó el cálculo de los errores entre el consumo real y la predicción realizada por CAMMESA: RMSE 1336,17; MAE 1044,9; MAPE 14%. Estos errores servirán de comparación contra el modelo implementado en esta tesis.

Se realizó la predicción del con el modelo LSTM-EMD, donde solamente se utilizaron 2 funciones IMF. Los resultados de los errores en nuestro experimento fueron: RMSE 501; MAE 182; MAPE 3.1%. El tiempo de procesamiento fue de 37 segundos.



En la figura anterior se puede ver la línea roja discontinua representando el consumo real por hora para cada una de las 201 observaciones horarias, en verde la predicción realizada por el modelo LSTM-EMD y finalmente en naranja el consumo predicho por CAMMESA. Este último, se puede ver que si bien acompaña la estacionalidad de la curva roja, tiende a dar resultados de consumo mucho más altos de los que resultan finalmente.

Comparando los errores reportados por CAMMESA y el de nuestro modelo, se puede observar que los errores de predicción de CAMMESA son entre 3 y 4 veces superiores al modelo LSTM-EMD. Si bien no se dispone de información acerca de cómo está implementado el modelo predictivo por CAMMESA, se puede intuir que realiza una predicción estadística de los datos. Una posible explicación al error de sobreestimación de consumo puede deberse a la propia función de la empresa: planificar las necesidades de capacidad de energía, coordinar las operaciones de despacho y regular las transacciones económicas del mercado eléctrico mayorista. Esto hace que, para garantizar una producción suficiente para cubrir el consumo, prefiera sobreestimar las predicciones. Como consecuencia, induce a una producción mayor para evitar una eventual falta de suministro.

Se justifica ejecutar el modelo simplemente con dos funciones IMF porque es el número mínimo de funciones generadas, con las que se obtienen buenos resultados. Se puede mejorar la predicción generando la cantidad de funciones IMF óptimas, pero el costo computacional no justifica el resultado de la predicción.