



Proyecto Deep Learning, Informe Final

Santiago Rivera Montoya
Santiago Alexander Losada
Emanuel López Higueta

1. Introducción

Este informe ejecutivo resume el desarrollo del proyecto de predicción del consumo energético mediante redes neuronales LSTM, centrado en países que formaron parte de la Gran Colombia: Colombia, Venezuela, Ecuador y Panamá.

El objetivo principal es prever el consumo energético a partir de datos históricos, apoyando la toma de decisiones en planificación energética, políticas públicas y sostenibilidad. La solución propuesta emplea técnicas de Deep Learning para modelar series de tiempo multivariadas, anticipando patrones complejos de comportamiento energético.

2. Exploración de datos

A través del dataset “World Energy Consumption” disponible en Kaggle, se ha puesto a disposición un conjunto de datos globales sobre el consumo de energía. Este dataset incluye registros históricos de consumo energético, producción, importaciones y exportaciones de energía para varios países desde 1900 hasta 2022.

En total, cuenta con 22,012 filas y 129 columnas. Contiene datos de consumo de energía en múltiples regiones, permitiendo la creación de modelos predictivos que consideren las variaciones en las necesidades energéticas de cada país. Algunas de las variables claves del conjunto de datos incluyen:

- Consumo total de energía
- Producción de energía
- Importación y exportación de energía
- Población
- Año
- País y región

Se filtraron únicamente los datos correspondientes a Colombia, Venezuela, Ecuador y Panamá. Tras la limpieza inicial, se seleccionaron columnas relevantes y se reorganizaron las variables por país y año.

3. Estructura de los Notebooks

En el primer notebook, titulado “01 - EDA”, se lleva a cabo un Análisis Exploratorio de Datos, cuyo objetivo es comprender la estructura y características del conjunto de datos. En esta etapa se visualiza el dataset, se identifican y examinan los valores nulos, se seleccionan los países con los que se trabajará posteriormente, y se generan diversas gráficas para analizar el comportamiento temporal de las variables más relevantes.

En el segundo notebook, titulado “02 - Modelos LSTM”, se lleva a cabo la limpieza del conjunto de datos, reemplazando los valores faltantes con ceros. Esta decisión se basa en que los datos provienen de fuentes históricas únicas para cada país, por lo que los valores ausentes no pueden considerarse comparables ni imputarse mediante promedios globales. Además, el dataset abarca información de 303 países, lo cual refuerza la necesidad de un tratamiento neutral que evite introducir sesgos.

Una vez finalizada la limpieza, se seleccionan los países de interés y se construye un modelo LSTM básico, inicialmente aplicado a un solo país, con el fin de predecir la energía total generada a partir de secuencias históricas de cinco años. Posteriormente, se escalan los datos para optimizar el rendimiento y se generaliza el modelo a todos los países seleccionados. Se generan visualizaciones comparativas para analizar los resultados tanto del modelo individual como del global, y se realizan predicciones futuras que permiten identificar tendencias en la generación de energía en la región de la Gran Colombia.

La solución se basa en una arquitectura de red neuronal recurrente LSTM bidireccional, especialmente diseñada para modelar secuencias temporales multivariadas como las del consumo y generación energética. El modelo recibe como entrada ventanas de 5 años consecutivos que incluyen 21 variables relevantes, como demanda eléctrica, producción de combustibles fósiles, participación de energías renovables, población y PIB, y predice la electricidad generada al año siguiente.

4. Descripción de la Solución

La solución implementada se basa en una arquitectura de red neuronal recurrente tipo LSTM bidireccional, adecuada para modelar secuencias temporales multivariadas como las del consumo y generación energética a lo largo de los años.

El modelo recibe como entrada ventanas de 5 años consecutivos de datos históricos energéticos, incluyendo 21 variables como demanda eléctrica, producción de combustibles fósiles, participación de energías renovables, población y PIB. A partir de estas secuencias, el modelo predice la electricidad generada al año siguiente.

Detalles técnicos de la arquitectura:

- **Entrada:** Secuencias temporales de 5 años \times 21 variables.
- **Capas:**
 - Bidirectional LSTM (128 unidades) con `return_sequences=True`.
 - Dropout (30%) para regularización.
 - Bidirectional LSTM (64 unidades) sin retorno de secuencias.
 - Otro Dropout (30%).
 - Dense (64 unidades) con activación ReLU.
 - Capa de salida Dense(1) para regresión univariada.
- **Optimización:** Adam con tasa de aprendizaje de 0.001.
- **Pérdida:** Huber Loss, adecuada para suavizar el impacto de valores atípicos.
- **Entrenamiento:**
 - EarlyStopping con paciencia de 10 épocas para evitar sobreajuste.
 - División de datos 80/20 de forma cronológica (sin aleatorización) para respetar la secuencia temporal.

Cada país se modeló **por separado**, utilizando sus propios datos históricos y escaladores personalizados, con el fin de capturar patrones únicos y evitar que las diferencias estructurales entre países afectaran el desempeño del modelo.

5. Iteraciones del Modelo

Durante el desarrollo del sistema, se realizaron diversas iteraciones y pruebas para mejorar la capacidad predictiva y adaptabilidad del modelo:

- **Ventana de tiempo fija (5 años):** Se optó por un tamaño de secuencia de 5 años como entrada. Esto permitió capturar suficiente contexto sin introducir demasiada variabilidad, funcionando como un buen balance entre historia y estabilidad.
- **Ajuste de arquitectura:** Se probaron distintas combinaciones de capas LSTM (simples, bidireccionales, con diferentes unidades). La arquitectura final, con dos capas bidireccionales y dropout, ofreció un buen compromiso entre capacidad de representación y generalización.
- **Entrenamiento por país:** En lugar de un modelo general para todos los países, se entrenaron modelos específicos por país (Colombia, Ecuador, Panamá y Venezuela). Esto permitió capturar patrones individuales y evitó que diferencias en escala o comportamiento entre países afectaran negativamente los resultados.
- **Normalización MinMax por país:** Dado que los datos energéticos varían considerablemente entre países, se aplicó MinMaxScaler de forma individual a cada país, garantizando que los valores estuvieran en una escala compatible con redes neuronales.
- **Validación temporal realista:** La separación entre entrenamiento y prueba se realizó manteniendo la secuencia cronológica (sin shuffle), simulando un escenario de predicción futura real.

Estas iteraciones no solo contribuyeron a una mejor comprensión del comportamiento energético en cada país, sino que también permitieron refinar los hiperparámetros y las decisiones de diseño para lograr un mejor ajuste del modelo a cada contexto nacional.

6. Métricas

Para valorar la efectividad del modelo de predicción de generación energética desarrollado con redes LSTM, se emplearon métricas tanto desde una perspectiva técnica como práctica. Estas métricas permiten no solo cuantificar el rendimiento del modelo, sino también entender su utilidad en el contexto de análisis energético por país.

Desde el punto de vista técnico, se utilizó el **Error Cuadrático Medio (MSE)**, que calcula el promedio de las diferencias elevadas al cuadrado entre los valores reales y las predicciones. Esta métrica es especialmente útil porque penaliza más fuertemente los errores grandes, lo cual ayuda a identificar modelos que generan predicciones inestables o con desviaciones significativas. Aunque el MSE no se encuentra en las mismas unidades que la variable objetivo (está expresado en unidades al cuadrado), resulta útil para comparar el comportamiento relativo de diferentes configuraciones del modelo.

Por otro lado, el **Error Absoluto Medio (MAE)** proporcionó una estimación directa de la magnitud del error promedio por observación. Esta métrica se expresa en las mismas unidades que la variable objetivo —en este caso, teravatios hora (TWh)—, lo que permite una interpretación sencilla: indica cuánta generación de electricidad se predice de forma incorrecta, en promedio, por año y por país.

Finalmente, se utilizó el **coeficiente de determinación R^2** , que mide la proporción de la variabilidad en la generación eléctrica que puede ser explicada por el modelo. Un valor de R^2 cercano a 1 indica que el modelo logra capturar correctamente la tendencia general de los datos, mientras que valores cercanos a cero o negativos evidencian una baja capacidad explicativa. Esta métrica es particularmente valiosa para comprender en qué medida el modelo aprende patrones temporales relevantes en cada país, más allá de acertar valores puntuales.

En conjunto, estas métricas permitieron no solo identificar el modelo con mejor desempeño, sino también entender sus limitaciones y fortalezas para cada país analizado.

7. Resultados

País	MSE	MAE	R^2	Análisis
Colombia	31.63	4.62	0.8342	Buen ajuste. El modelo capta correctamente la tendencia general. Picos altos suavizados, pero la predicción sigue el crecimiento de la serie.

Ecuador	29.32	4.46	0.5090	Ajuste moderado. El modelo sigue la tendencia pero no logra capturar completamente los picos crecientes.
Panamá	0.18	0.39	0.7762	Muy buen ajuste, aunque el set de prueba es pequeño (solo 7 muestras). El modelo capta correctamente la pendiente. Interpretar con cautela por poca data.
Venezuela	70.00	7.35	0.5783	Ajuste más difícil. La serie es muy volátil, con subidas y bajadas abruptas. El modelo tiende a suavizar esos picos, penalizando el R^2 .

Predicción de Electricidad Generada (2023-2027) utilizando los modelos previamente entrenados:

País	Años de Predicción	Valores Predichos (TWh)	Tendencia General
Colombia	2023 - 2027	73.49, 72.02, 72.10, 71.44, 71.91	Ligeramente decreciente
Ecuador	2023 - 2027	20.34, 20.07, 19.98, 19.59, 19.91	Decreciente con leve repunte
Panamá	2022 - 2026	10.72, 10.60, 10.49, 10.55, 10.57	Estable con pequeñas oscilaciones

Venezuela	2023 - 2027	78.83, 75.99, 76.87, 76.32, 77.46	Irregular pero con repunte final
------------------	-------------	--------------------------------------	-------------------------------------

En este proyecto se entrenó un modelo LSTM individual para cada país de la región Gran Colombia (Colombia, Ecuador, Panamá y Venezuela), lo cual permitió adaptar el modelo a los patrones energéticos propios de cada nación. Esta estrategia fue fundamental para obtener mejores resultados y evitar interferencias entre series temporales heterogéneas.

Colombia obtuvo el mejor desempeño, alcanzando un coeficiente de determinación (R^2) de 0.83 y un MAE bajo. Las predicciones futuras (2023–2027) muestran una tendencia leve hacia la estabilización, con valores cercanos a los 72 TWh. Panamá también presentó resultados destacados, con un R^2 de 0.77 y predicciones muy estables entre 10.5 y 10.7 TWh, lo que evidencia una serie energética predecible y controlada.

En Ecuador, el modelo logró un R^2 de 0.50. Aunque fue capaz de capturar la tendencia decreciente general, se observaron mayores errores al enfrentar cambios abruptos en la serie. Por último, Venezuela presentó la mayor dificultad debido a su alta volatilidad. Con un R^2 de 0.57, las predicciones reflejan oscilaciones y picos que hacen más complejo anticipar con precisión su comportamiento energético futuro.

8. Conclusiones

El enfoque de entrenar modelos LSTM de manera independiente por país resultó adecuado para capturar las dinámicas específicas de cada sistema energético. Esta decisión mejoró la capacidad predictiva del modelo y permitió reducir los riesgos de sobreajuste entre países con comportamientos muy distintos.

Los resultados muestran que en contextos con datos históricos estables, como en Colombia y Panamá, el modelo logra un alto rendimiento y ofrece proyecciones consistentes. En contraste, en países como Ecuador y especialmente Venezuela, donde hay mayor variabilidad o inestabilidad energética, la precisión del modelo se ve limitada, aunque aún permite identificar la dirección general del comportamiento energético.

Además, el uso de ventanas temporales de cinco años, normalización individual por país y validación temporal sin aleatorización fueron elementos clave para asegurar un entrenamiento robusto. La arquitectura bidireccional permitió capturar dependencias pasadas y futuras dentro de cada secuencia. En conjunto, estos modelos ofrecen una

herramienta valiosa para apoyar la planificación energética en la región, con base en proyecciones confiables y ajustadas a la realidad de cada país.

9. Referencias

- Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice.
- Chakraborty, P., & Sahoo, A. (2018). Energy consumption forecasting using machine learning algorithms: A review. *Energy Reports*, 4, 52–64.
- Zhao, Z., & Zhang, L. (2019). Deep learning for energy consumption prediction and optimization. *Energy and Buildings*, 199, 47–58.
- Li, G., & Li, X. (2020). Prediction of electricity consumption based on LSTM neural network. *Journal of Electrical Engineering & Technology*, 15(2), 765–772.
- Zhang, X., & Li, Y. (2021). Short-term energy consumption forecasting using machine learning models: A comparative study. *Energy*, 214, 118674.