Predicción del Déficit, Demanda y Disponibilidad de Datos Eléctricos en Cuba Utilizando Técnicas de Machine Learning

Amalia González Ortega
Universidad de la habana
MATCOM

8 de julio de 2025

Índice general

Capítulo 1

Introducción

La situación del sistema eléctrico de Cuba es "graveçon "largas horas de apagón"debido al mal estado de sus termoeléctricas, la falta de financiamiento para su reparación y la escasez de combustible. Durante las últimas semanas los cubanos han enfrentado cortes de energía de hasta 21 horas al día en algunas provincias de la isla, cuyo sistema eléctrico solo está produciendo 45 gigawatts de los 63 que consume en cada jornada, según el director de la Unión Eléctrica de Cuba (UNE), Alfredo López. Cuba, en una profunda crisis económica desde hace más de cuatro años, cuenta con un ruinoso sistema eléctrico, que colapsó en marzo por cuarta vez en menos de seis meses. Esta red está integrada por ocho desgastadas termoeléctricas, algunas centrales eléctricas flotantes alquiladas a una empresa turca y generadores, que funcionan fundamentalmente con diésel que Cuba tiene grandes dificultades para importar.

Esta situación pone de manifiesto la necesidad urgente de herramientas que permitan prever con mayor precisión los comportamientos del sistema eléctrico nacional, particularmente en lo referente a la demanda, el déficit y la disponibilidad energética.

En este contexto, el uso de técnicas de *Machine Learning* (ML) se presenta como una alternativa para modelar dinámicamente el comportamiento energético del país, a partir del análisis de datos de consumo eléctrico.La situacion que se describe abre paso al deseo y la necesidad de hacer una predicción precisa de la demanda energética y los posibles déficits del sistema eléctrico

1.1. Objetivos

Desarrollar y comparar varios modelos de *Machine Learning* para predecir la demanda energética, el déficit y la disponibilidad en Cuba, utilizando los datos de consumo eléctrico, con el fin de proporcionar herramientas de apoyo para la planificación energética y optimización del sistema eléctrico nacional.

- Analizar patrones históricos de consumo energético en Cuba.
- Identificar variables que influyen en la demanda energética.
- Implementar y entrenar algoritmos de *Machine Learning*, incluyendo SVR, *Random Forest*, *Gradient Boosting*, así como modelos basados en series temporales.
- Evaluar y comparar el rendimiento de los modelos utilizando métricas apropiadas.
- Generar predicciones a corto y mediano plazo de la demanda energética nacional.

Capítulo 2

Marco Teórico

2.1. Sistemas eléctricos en Cuba

El sistema eléctrico cubano depende principalmente de combustibles fósiles importados y enfrenta desafíos constantes. La infraestructura, construida hace décadas, genera frecuentes apagones y déficits energéticos que afectan a toda la población. La demanda varía mucho según la estación del año, especialmente en verano cuando aumenta el uso de aires acondicionados.

2.2. Aprendizaje automático en energía

El machine learning ha cambiado la forma de abordar los problemas energéticos. Los algoritmos de ML pueden encontrar patrones complejos en los datos de consumo y generación eléctrica que los métodos tradicionales no logran captar. Estas técnicas se usan para predecir demanda eléctrica, pronosticar generación renovable, detectar anomalías en el consumo y optimizar la distribución. Los resultados son mucho más precisos que los métodos estadísticos convencionales.

Aunque la investigación actual muestra que no hay un algoritmo perfecto para todo. Tiwari et al. (2021) demostraron que los modelos deben adaptarse a cada situación específica. Su trabajo con redes inteligentes reveló la importancia de combinar múltiples fuentes de datos. Los estudios comparativos confirman que la efectividad depende del contexto, la calidad de los datos y qué tan lejos quieres predecir. La tendencia ahora es hacia modelos híbridos que combinan diferentes enfoques y técnicas de deep learning para patrones más complejos. Para Cuba, estas tecnologías representan una oportunidad real de mejorar la eficiencia del sistema eléctrico y reducir los déficits mediante predicciones más precisas. Atendiendo a esto, en este trabajo estaremos usando los siguientes modelos:

- XGBoost y Gradient Boosting XGBoost se ha convertido en el estándar para predicción energética. Nabatchian (2024) demostró cómo este algoritmo puede capturar efectivamente los patrones estacionales en el consumo de energía. Funciona construyendo muchos árboles de decisión simples donde cada uno corrige los errores del anterior. Esto permite modelar relaciones complejas entre variables como temperatura, día de la semana y consumo histórico.
- Support Vector Regression (SVR) SVR es útil para predicciones a corto plazo donde necesitas mucha precisión. Maneja bien datos complicados y es resistente a valores extremos que podrían confundir otros algoritmos.

- Random Forest Este método combina múltiples árboles de decisión para crear un modelo robusto. Es especialmente bueno cuando trabajas con diferentes tipos de datos mezclados, como información meteorológica y patrones de consumo.

Capítulo 3

Metodología

3.1. Obtención de datos

Los datos se extrajeron de Cubadebate scrapeando los reportes diarios del SEN desde septiembre de 2022 hasta mayo de 2025. Aunque el objetivo era obtener todos los días de este período, el proceso de scraping logró extraer información de 697 días entre estos 4 años. Cada artículo del SEN fue procesado usando un LLM a través de la API de https://firework.ai. Le dijimos al modelo que extraiga automáticamente toda la información eléctrica de cada reporte y la estructure en JSON. Esto nos ahorró buena parte de trabajo manual. El LLM identificó variables importantes como fecha, año, mes, dia, dia de la semana, si es fin semana, demanda maxima, disponibilidad total, afectacion predicha, deficit predicho, respaldo, disponibilidad 07am, demanda 07am, plantas averiadas, plantas mantenimiento, mw de limitacion termica, mw de motores problemas, horas de afectacion, max afectacion en mw y otros datos clave del sistema eléctrico.

3.2. Preprocesamiento

Los datos JSON se procesaron para crear un CSV limpio. Eliminamos duplicados, corregimos formatos de fecha, tratamos valores faltantes y nos aseguramos de que todo estuviera bien estructurado. El CSV final tiene todas las variables del JSON que necesitamos para predecir la situación energética de Cuba.

3.3. Análisis Exploratorio de los datos

El dataset final comprende un período de análisis desde diciembre de 2022 hasta mayo de 2025, con un total de 696 observaciones distribuidas en 689 días únicos. Esta información proporciona una base sólida para el análisis del comportamiento del sistema eléctrico cubano durante más de dos años.

Cuadro 3.1: Estadísticas descriptivas de variables energéticas

Estadística	Demanda	Disponibilidad	Déficit	Afectación
Count	692.00	693.00	641.00	675.00
Mean	2986.58	2292.08	744.15	791.23
Std	229.26	357.80	496.30	507.34
Min	2450.00	1460.00	2.00	14.00
25 %	2800.00	1980.00	305.00	337.50
50 % (Mediana)	3000.00	2326.00	661.00	710.00
75 %	3162.50	2583.00	1208.00	1264.50
Max	3500.00	3109.00	1800.00	1870.00

3.3.1. Análisis de Correlación

Con el objetivo de seleccionar las variables más relevantes para la predicción de la **demanda máxima**, el **déficit** y la **disponibilidad energética**, se calculó la correlación de Pearson entre estas variables objetivo y el resto de variables disponibles en el conjunto de datos.

Los coeficientes de correlación permiten identificar relaciones lineales entre variables. Aquellas con mayor valor absoluto indican una asociación más fuerte (ya sea positiva o negativa), y por tanto, son candidatas a ser utilizadas como *features* en los modelos.

A continuación, se presentan las correlaciones más significativas para cada variable objetivo. Se destacan en **negrita** aquellas correlaciones positivas, ya que pueden indicar relaciones directas útiles en modelos de regresión.

Cuadro 3.2: Correlaciones de variables con la Demanda Máxima, Déficit y Disponibilidad

Variable	Demanda Máxima	Déficit	Disponibilidad
deficit_real	**0.828**	—	-0.934
demanda_07am	**0.755**	**0.466**	-0.201
año	**0.655**	**0.761**	-0.696
disponibilidad_total	-0.573	-0.934	—
disponibilidad_07am	-0.518	-0.861	**0.929**
mw_motores_problemas	-0.410	-0.432	**0.370**
mw_limitacion_termica	**0.192**	-0.001	**0.124**
plantas_averiadas	-0.108	-0.054	**0.010**
mes	**0.073**	-0.006	**0.055**
plantas_mantenimiento	-0.066	**0.243**	-0.397
dia	-0.050	-0.001	-0.030
dia_semana	-0.032	-0.008	-0.008
demanda_maxima	—	**0.828**	-0.573

Capítulo 4

Diseño Experimental y Selección de Modelos

4.1. Cantidad de Datos y División del Conjunto

El conjunto de datos disponible abarca un total de 696 observaciones diarias correspondientes a un período de aproximadamente 4 años. Dado el tamaño relativamente reducido del conjunto, se optó por estrategias cuidadosas de división de los datos para garantizar una evaluación robusta de los modelos.

El conjunto se dividió en tres partes:

- **Entrenamiento (70 %)**: utilizado para ajustar los modelos.
- **Validación (15 %)**: para la selección de hiperparámetros.
- **Prueba (15 %)**: para evaluar el rendimiento final del modelo.

4.2. Selección de Variables Predictoras

Los *features* utilizados en los modelos fueron seleccionados en base al análisis de correlación de Pearson entre las variables predictoras disponibles y las tres variables objetivo: **demanda máxima**, **déficit energético** y **disponibilidad total**. Se eligieron aquellas variables con correlaciones absolutas altas, al considerar tanto relaciones positivas como negativas, con énfasis en aquellas relaciones positivas que pudieran mejorar el ajuste en modelos lineales o no lineales.

4.3. Modelos de Machine Learning Utilizados

Dado que el problema de predicción se basa en series de tiempo con una cantidad limitada de datos históricos, se eligieron modelos supervisados de bajo a mediano riesgo de sobreajuste, pero con buena capacidad de aprendizaje no lineal. Se usaron los mismos modelos y configuración para las tres tareas de predicción (demanda, déficit y disponibilidad), con el objetivo de mantener coherencia y facilitar la comparación de resultados. Los modelos utilizados fueron:

- **Support Vector Regression (SVR)**: eficiente en espacios de alta dimensión y robusto frente a outliers. Utiliza un *kernel RBF* para capturar relaciones no lineales. Es apropiado en contextos con pocos datos y buena regularización.
- **Random Forest Regressor**: modelo de ensamble basado en árboles de decisión, útil para capturar relaciones no lineales complejas y manejar interacciones entre variables sin necesidad de normalización. También ayuda a reducir el riesgo de sobreajuste mediante el promedio de múltiples árboles.
- **Gradient Boosting Regressor (GBR)**: modelo de ensamble que construye árboles secuencialmente, minimizando el error residuo en cada etapa. Tiene mejor capacidad de predicción que Random Forest en muchos casos, aunque es más sensible al sobreajuste si no se controla adecuadamente.

4.4. Justificación de Hiperparámetros Utilizados

Los hiperparámetros fueron seleccionados manualmente mediante búsqueda guiada y validación cruzada debido a la limitación de datos. Los mismos se aplicaron en las tres tareas de predicción.

- **SVR**:
  - `kernel='rbf'`: permite capturar relaciones no lineales.
  - `C=1000`: controla el margen de tolerancia del error; valores altos penalizan errores con más severidad.
  - `epsilon=0.1`: define la zona de tolerancia alrededor de la predicción sin penalización.
- **Random Forest**:
  - `n_estimators=300`: número de árboles en el bosque; un valor suficiente para estabilizar la predicción.
  - `max_depth=10`: limita la profundidad de cada árbol para evitar sobreajuste en un conjunto pequeño.
- **Gradient Boosting**:
  - `n_estimators=300`: número de etapas de boosting.
  - `learning_rate=0.1`: controla la contribución de cada árbol; un valor moderado para evitar oscilaciones.
  - `max_depth=5`: menor profundidad que RF para mejorar la generalización.

4.5. Consideraciones Finales

El uso de múltiples modelos permite evaluar distintos enfoques de aprendizaje y comparar su rendimiento en un entorno de datos reales, escasos y altamente variables. Esta comparación es clave para identificar la mejor estrategia para predecir con precisión eventos críticos del sistema eléctrico nacional, como la ocurrencia de déficits o caídas en la disponibilidad.

Capítulo 5

Resultados y Evaluación de Modelo

5.1. Métricas de Evaluación

Para evaluar el rendimiento de los modelos en cada tarea de predicción se utilizaron las siguientes métricas:

- **MSE (Error Cuadrático Medio)**: penaliza errores grandes, útil para detectar grandes desviaciones.
- **RMSE (Raíz del Error Cuadrático Medio)**: interpretable en las mismas unidades que la variable objetivo.
- **MAE (Error Absoluto Medio)**: representa el promedio del error absoluto.
- **MedAE (Error Absoluto Mediano)**: menos sensible a outliers que el MAE.
- **R^2 (Coeficiente de Determinación)**: indica la proporción de la varianza explicada por el modelo.
- **MAPE (Error Porcentual Absoluto Medio)**: expresa el error relativo en porcentaje.

A continuación se presentan los resultados obtenidos para cada variable objetivo.

5.2. Predicción de la Demanda Máxima

Cuadro 5.1: Rendimiento de los modelos para la predicción de la demanda máxima (MW)

Modelo	MSE	R^2	MAE	RMSE	MAPE	MedAE
Random Forest	4138.62	0.928	46.54	64.33	1.57 %	32.58
SVR	4734.53	0.917	48.84	68.81	1.65 %	33.00
Gradient Boosting	4734.53	0.917	48.84	68.81	1.65 %	33.00

Análisis: El modelo Random Forest obtuvo los mejores resultados en todas las métricas, destacándose especialmente en MAE, RMSE y R^2. SVR y Gradient Boosting mostraron rendimientos muy similares entre sí, ligeramente por debajo de Random Forest.

5.3. Predicción del Déficit Energético

Cuadro 5.2: Rendimiento de los modelos para la predicción del déficit energético (MW)

Modelo	MSE	R^2	MAE	RMSE	MAPE	MedAE
Random Forest	34499.84	0.880	147.77	185.74	–	123.43
SVR	36606.48	0.873	156.58	191.33	–	146.90
Gradient Boosting	36606.48	0.873	156.58	191.33	–	146.90

Nota: Los valores de MAPE son extremadamente altos y poco interpretables debido a posibles divisiones por cero o valores muy pequeños en el denominador, por lo que no se muestran.

Análisis: De nuevo, Random Forest queda como el mejor modelo, con menor error absoluto y mayor capacidad explicativa. SVR y Gradient Boosting presentaron errores ligeramente superiores.

5.4. Predicción de la Disponibilidad Total

Cuadro 5.3: Rendimiento de los modelos para la predicción de la disponibilidad total (MW)

Modelo	MSE	R^2	MAE	RMSE	MAPE	MedAE
Random Forest	12914.41	0.898	86.51	113.64	3.83 %	71.58
SVR	15959.94	0.874	94.27	126.33	4.24 %	66.34
Gradient Boosting	15959.94	0.874	94.27	126.33	4.24 %	66.34

Análisis: En la tarea de predicción de la disponibilidad, Random Forest nuevamente demuestra un mejor rendimiento global. SVR y Gradient Boosting comparten exactamente los mismos resultados, siendo ligeramente menos precisos.

5.5. Resumen Comparativo

En las tres tareas de predicción, **Random Forest** se posiciona como el modelo más robusto y confiable, obteniendo los mejores resultados en términos de error absoluto, precisión y capacidad explicativa. Aunque SVR y Gradient Boosting muestran rendimientos aceptables, no superan al enfoque basado en árboles aleatorios.

Estos resultados refuerzan la elección de modelos de ensamble como una buena alternativa para contextos con datos limitados y relaciones no lineales como los del sistema eléctrico cubano.

Capítulo 6

Predicciones Basadas en Series Temporales

Además del enfoque tradicional basado en modelos de aprendizaje supervisado utilizando variables predictoras (features), se ha incorporado una estrategia complementaria que trata los datos como una **serie temporal**. Este enfoque reconoce la naturaleza secuencial del consumo eléctrico y la evolución del sistema energético cubano en el tiempo.

6.1. Motivación

Dado que los datos disponibles están organizados cronológicamente a lo largo de cuatro años, con observaciones diarias, resulta razonable suponer que existen dependencias temporales en variables como la demanda energética, el déficit y la disponibilidad.

El enfoque de series temporales busca aprovechar estas dependencias internas para mejorar la capacidad predictiva, especialmente en el corto plazo, donde las condiciones recientes influyen fuertemente en el comportamiento futuro.

6.1.1. Ventajas del Enfoque Temporal

- Permite capturar patrones estacionales, tendencias y ciclos diarios, semanales o mensuales.
- No requiere necesariamente de muchas variables externas si el comportamiento pasado contiene suficiente información.
- Se adapta mejor a la predicción secuencial o continua (rolling forecasts).

Capítulo 7

Conclusiones y Trabajo Futuro

La crisis energética que enfrenta Cuba requiere herramientas analíticas que permitan una planificación más precisa y eficiente del sistema eléctrico nacional. En este trabajo, se abordó el problema mediante el desarrollo de modelos de predicción de tres variables clave: **demanda energética máxima**, **déficit energético** y **disponibilidad total del sistema**.

- Se utilizaron datos históricos energéticos recolectados durante un período de cuatro años, analizando las correlaciones entre variables para seleccionar características relevantes para la predicción.
- Se entrenaron y compararon tres modelos de aprendizaje automático: **Random Forest**, **SVR** y **Gradient Boosting**, utilizando las mismas características y procedimientos de validación.
- En todas las tareas de predicción, **Random Forest** demostró ser el modelo más preciso y robusto, superando sistemáticamente a los demás en métricas como MAE, RMSE y R^2.
- Se identificaron limitaciones en el uso de MAPE para el déficit energético, debido a la presencia de valores extremadamente bajos que distorsionan esta métrica porcentual.
- Se incorporó un enfoque adicional basado en **series temporales**, reconociendo la naturaleza secuencial de los datos. Este enfoque permitirá realizar predicciones automáticas y adaptativas a futuro.

7.0.1. Líneas Futuras

Este trabajo puede ampliarse en las siguientes direcciones:

- Incorporar variables adicionales como temperatura, precipitaciones, eventos políticos o económicos.
- Ampliar la base de datos teniendo en cuenta que no falten en todo lo posible dias de por medio

Bibliografía

[1] Ehsan Nabatchian. *Exploración de la predicción de series temporales del consumo de energía mediante XGBoost y validación cruzada*. Publicado el 10 de enero de 2024.

[2] George Kamtziridis. *Time Series Forecasting with XGBoost and LightGBM: Predicting Energy Consumption*. Publicado el 27 de febrero de 2023.

[3] Shamik Tiwari, Anurag Jain, Kusum Yadav, Rabie Ramadan. *Machine Learning-Based Model for Prediction of Power Consumption in Smart Grid*. Recibido el 18 de septiembre de 2020; aceptado el 31 de agosto de 2021.

[4] Autores no especificados. *Evaluating Machine Learning Algorithms for Energy Consumption Prediction in Electric Vehicles: A Comparative Study*.