Maestría en

INTELIGENCIA ARTIFICIAL APLICADA

Trabajo previo a la obtención de título de Magister en Inteligencia Artificial Aplicada

AUTOR/ES:

Roberto Santiago Toapanta Paredes

Ricardo Andrés Herrera Andrade

Robert Antonio Piza Reasco

TUTOR/ES:

Gabriela Estefania Chiliquinga Jiménez

Andrés Roberto Navas Castellanos

TEMA

Sistema de modelado basado en inteligencia artificial de los factores determinantes en la generación e intercambio energético del Sistema Nacional Interconectado del Ecuador.

Quito - Ecuador

2026

QUITO – ECUADOR | 2026

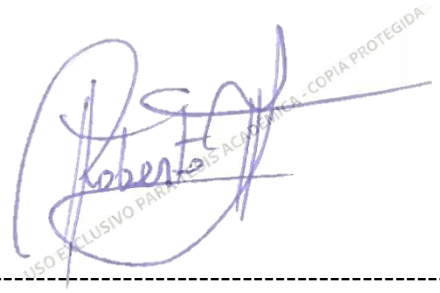
REINVENTEMOS EL FUTURO

Certificación de autoría

Nosotros, **Roberto Santiago Toapanta Paredes, Ricardo Andrés Herrera Andrade**

Robert Antonio Piza Reasco, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



Firma
Roberto Santiago Toapanta Paredes

Firma
Ricardo Andrés Herrera Andrade

Firma
Robert Antonio Piza Reasco

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Roberto Santiago Toapanta Paredes, Ricardo Andrés Herrera Andrade, Robert Antonio Piza Reasco,** en calidad de autores del trabajo de investigación titulado ***Sistema de modelado basado en inteligencia artificial de los factores determinantes en la generación e intercambio energético del Sistema Nacional Interconectado del Ecuador***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, agosto 2026

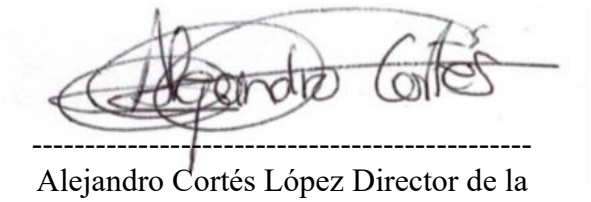
Firma
Roberto Santiago Toapanta Paredes

Firma
Ricardo Andrés Herrera Andrade

Firma
Robert Antonio Piza Reasco

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés López y Karla Estefanía Mora Cajas**, declaramos que: **Roberto Santiago Toapanta Paredes, Ricardo Andrés Herrera Andrade, Robert Antonio Piza Reasco** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López Director de la Maestría en Inteligencia Artificial Aplicada

Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Inteligencia Artificial Aplicada

DEDICATORIA

Este trabajo de investigación va dedicado a nuestros padres y seres queridos,

quienes nos apoyan y nos inspiran siempre para que logremos nuestras metas académicas.

AGRADECIMIENTOS

En primer lugar, agradecemos a nuestros seres queridos y a nuestras familias, quienes nos han apoyado sin condiciones en nuestro camino profesional y de manera particular a los maestros en general, al director del programa de maestría, Alejandro Cortés López.

Por último, queremos expresar nuestro agradecimiento a nuestros colegas de maestría, con los que en algún momento colaboramos en trabajos grupales y entregas. Gracias por su contribución y su respaldo en este camino de titulación.

RESUMEN

El Sistema Nacional Interconectado (SNI) de Ecuador enfrenta episodios recurrentes de crisis de seguridad energética, vulnerable ante fenómenos climáticos y cambios en los ciclos de precipitación. Ahora bien, se construyó un dataset con datos diarios para 2009-2025 incluyendo variables energéticas y climáticas provenientes de fuentes públicas.

Los modelos predictivos tradicionales no explican adecuadamente la interrelación entre la demanda, la disponibilidad hídrica y la capacidad térmica. Por ello el presente trabajo propone un enfoque analítico alternativo mediante modelado basado en aprendizaje automático con el fin de conocer la sensibilidad del SNI a los factores determinantes y convertir datos históricos en conocimiento estratégico. Así la investigación se enmarca en un enfoque cuantitativo, con un diseño no experimental de tipo ex post facto y un alcance exploratorio-explicativo, pues se trabaja sobre series históricas secundarias sin intervención directa sobre el sistema eléctrico.

La metodología descrita se basa en cuatro técnicas de inteligencia artificial con un módulo de recomendación: Primero se emplea un algoritmo de aprendizaje no supervisado llamado K-Means para identificar los regímenes operativos cuyo resultado es contrastado con los racionamientos oficiales para validar su coherencia. Segundo se emplea el clasificador Random Forest que a partir de variables físicas externas evalúa las condiciones hidroclimáticas y demográficas. Tercero XGBoost modela flujos de generación y emplea valores SHAP para explicar la contribución de cada variable. Por último, se emplea K-Nearest Neighbors (KNN) para análisis de analogías históricas en situaciones críticas. El trabajo se basa únicamente en el análisis posoperativo a nivel nacional sin utilizar la

optimización en tiempo real ni los pronósticos y crea la base para la evaluación de la resiliencia energética del país.

Los regímenes del sistema eléctrico encontrados por el algoritmo K-Means coincidieron hasta en un 73% con meses que fueron oficialmente decretados como racionamiento sin que ese dato formara parte del entrenamiento, el clasificador Random Forest logró un F1-macro de 0,396 y los cuatro modelos XGBoost mostraron patrones consistentes donde la generación térmica se explicó con $R^2 = 0,947$ y la exportación de energía eléctrica solo parcialmente con un $R^2 = 0,216$, lo que evidencia un componente contractual y comercial no capturado por las variables físicas disponibles.

El aumento más revelador se dio en el modelo de importación, donde el R^2 pasó de 0,315 a 0,649 al incorporar rezagos temporales, mostrando que las condiciones hidroclimáticas acumuladas explican la dependencia de importaciones con más solidez que las puntuales del día.

Palabras clave: Sistemas interconectados, Aprendizaje Automático, Algoritmos de Aglomeración, K-Means, Bosques Aleatorios, Potenciación del gradiente, Inteligencia Artificial Explicable, K-Nearest Neighbors, Intercambio Energético.

ABSTRACT

Ecuador's National Interconnected System (SNI) faces recurrent episodes of energy security crisis, being vulnerable to climate phenomena and shifts in precipitation cycles. A daily dataset for 2009–2025 was built, incorporating energy and climate variables from public sources. Traditional predictive models do not adequately explain the interrelationship between demand, water availability and thermal capacity. Therefore, this work proposes an alternative analytical approach using machine learning-based modeling to understand the sensitivity of the National Interconnected System (SNI) to determining factors and transform historical data into strategic knowledge. Thus, the research is framed within a quantitative approach, with a non-experimental, ex post facto design and an exploratory-explanatory scope, as it works with secondary historical series without direct intervention in the electrical system.

The described methodology is based on four artificial intelligence techniques with a recommendation module: First, an unsupervised learning algorithm called K-Means is used to identify operating regimes, the results of which are compared with official rationing schedules to validate their consistency. Second, the Random Forest classifier is used, which evaluates hydroclimatic and demographic conditions based on external physical variables. Third, XGBoost models generation flows and uses SHAP values to explain the contribution of each variable. Finally, K-Nearest Neighbors is used for the analysis of historical analogies in critical situations. The work is based solely on post-operational analysis at the national level without using real-time optimization or forecasting and creates the basis for the assessment of the country's energy resilience.

The regimes identified by K-Means matched up to 73% of the months officially declared as rationing without that information being part of the training; the Random Forest classifier achieved an F1-macro of 0,396; and the four XGBoost models showed consistent patterns where thermal generation was explained with R^2 = 0,947 and electricity export only partially with R^2 = 0,216, evidencing a contractual and commercial component not captured by the available physical variables.

The most revealing improvement occurred in the imports model, where R^2 rose from 0,315 to 0,649 once temporal lags of the predictor variables were included, showing that accumulated hydroclimatic conditions explain import dependence more solidly than same-day conditions.

Keywords: Interconnected System, Machine Learning, Clustering Algorithms, K-Means, Random Forests, Gradient Boosting, Explainable AI, Nearest Neighbor Search, Energy Exchange.

Índice

Índice de Tablas

Índice de Figuras

Capítulo 1

1. Introducción

1.1. Definición del proyecto

Ecuador cuenta con una matriz energética diversa que combina fuentes renovables y no renovables, la cual se divide en cuatro grupos de generación. La generación hidroeléctrica aprovecha el caudal de los ríos para mover turbinas; en el caso ecuatoriano, la mayoría de estas centrales son “de pasada”, es decir, carecen de embalses de gran capacidad para almacenar agua en tiempo de estiaje (Vela, 2024). La generación termoeléctrica, que funciona como respaldo durante períodos de sequía, produce electricidad principalmente a partir de fuel oil, diésel y gas natural. La generación no convencional agrupa fuentes como biomasa, eólica, fotovoltaica y biogás, siendo la biomasa la de mayor participación dentro de este grupo. Finalmente, las interconexiones internacionales permiten importar energía desde Colombia y Perú cuando las condiciones de precio o disponibilidad lo requieren. En 2024, estos cuatro grupos representaron, respectivamente, el 64,87%, 21,53%, 9,49% y 4,11% (importaciones netas) de la producción neta total del país (CENACE, 2025).

Según CENACE (2025) esta composición de la matriz energética no ha sido constante en el tiempo; durante el año 2024 se registró una disminución en la producción hidroeléctrica respecto a los últimos seis años, mientras que la generación no convencional y termoeléctrica incrementaron respecto al año 2023.

Además se menciona que, el Ecuador enfrenta una crisis de seguridad energética ocasionada de los efectos naturales que atraviesa la región. Estos ocasionaron estiajes prolongados que afectaron a las principales centrales de generación eléctrica cuya fuente de generación es hídrica, alimentada por los caudales principales del Ecuador. Aunque la mayor afectación existente producto de dicha problemática es la escasez energética, se ha visibilizado una falta de estructuración de los modelos actuales utilizados para explicar y determinar la disponibilidad hídrica, capacidad de respuesta térmica y demanda diaria.

Con lo antes ya mencionado, estos efectos naturales, provocados por el cambio climático, generan una frecuencia prolongada de sequias afectando drásticamente a los sistemas dependientes de hidroeléctricas (Naranjo et al., 2026). Este fenómeno climático produce estiajes cuya medida de afectación es proporcional a la variación de los caudales de las cuencas hídricas principales del Ecuador; según un reporte de, (CENACE, 2025) el año donde se presentó una mayor afectación es el 2024, donde el caudal promedio del embalse de la hidroeléctrica Mazar alcanzó únicamente el 52% de su capacidad total. A esta problemática es importante agregar que existe una mayor demanda eléctrica a nivel nacional cuyas proyecciones apuntan un crecimiento anual del 8.38% explícitamente para el sector industrial (Icaza et al., 2023).

Ante la disminución del porcentaje de participación hidroeléctrica, el sistema se ve forzado a incrementar la generación térmica basada en combustibles de alto costo como vapor-bunker y diésel o recurrir a un intercambio energético con países vecinos, lo que eleva la dependencia de importación y representa costos económicos significativamente

más altos. Las consecuencias de esta dinámica se hicieron evidente durante los racionamientos de energía en el período 2023 – 2024 que tuvieron una duración de 14 horas diarias, con un impacto económico sobre el aparato productivo nacional estimado en USD 2.000 millones, equivalente al 2% del PIB (Naranjo et al., 2026).

Dentro de la literatura revisada predominan los estudios enfocados en la predicción de variables del sistema eléctrico mediante técnicas de inteligencia artificial, como redes neuronales y modelos de aprendizaje automático (Sánchez y Coto, 2022). Esta misma orientación predictiva se observa en el contexto ecuatoriano. Según CENACE (2025), en su reporte anual tiene planteado dentro de sus proyectos internos, el desarrollo de modelos de predicción de demanda de corto plazo mediante redes LSTM. Sin embargo, estos enfoques comparten una limitación común: se basan en modelos de caja negra que, pese a su buen desempeño predictivo, carecen de explicabilidad sobre los factores que determinan sus resultados (Machlev et al., 2022). En consecuencia, no existe actualmente un sistema que permita analizar de forma integral y explicativa cómo interactúan las variables que determinan el comportamiento del SNI ecuatoriano, ni en qué condiciones específicas se producen fenómenos como el incremento de la generación térmica, el uso de importaciones o la disponibilidad de excedentes para exportación.

Frente a esta brecha, el proyecto propone desarrollar un sistema de modelado explicativo basado en técnicas de aprendizaje automático; el cual se centrará en analizar e identificar las condiciones críticas que fuerzan el aumento en generación térmica, optar por importación o la disponibilidad de excedentes para exportación. Para transformar los datos

históricos del periodo 2009-2025 en conocimiento accionable se propone una arquitectura basada en 5 componentes:

Identificación automática de regímenes operativos: Mediante el uso de aprendizaje no supervisado (K-Means), el sistema identificará automáticamente los conjuntos operativos del SNI.

Validación automática de regímenes mediante clasificación supervisada: Mediante un clasificador Random Forest entrenado únicamente con variables exógenas físicas (caudal, ONI, precipitación y población), se evalúa hasta qué punto los regímenes identificados pueden reproducirse desde las condiciones externas del sistema.

Modelado de Causalidad y Relevancia: Mediante el uso de un motor de regresión avanzada XGBoost y técnicas de explicabilidad SHAP, permitirá cuantificar exactamente qué factores determinan el comportamiento del SNI.

Análisis de analogía histórica: Identificación de precedentes históricos con condiciones similares a las actuales mediante K-Nearest Neighbors.

Análisis de escenarios basado en patrones históricos: A partir del régimen detectado y de los precedentes recuperados, se describe cómo se comportó el sistema en periodos con condiciones parecidas, sin pretender predecir valores futuros.

El sistema no busca predecir el comportamiento futuro u optimizar la operación en tiempo real, sino proporcionar un marco de comprensión robusto que muestre las

dependencias ocultas del SNI, abriendo la posibilidad que la planificación energética deje de ser reactiva y pase a ser informada por la evidencia de factores determinantes.

1.2. Justificación e importancia del trabajo de investigación

Este problema es digno de ser estudiado porque el sistema eléctrico es fundamental para el desarrollo económico, industrial y social del país, y porque permitiría tomar decisiones basadas en evidencia científica, optimizando el uso de recursos y reduciendo dependencias de combustibles fósiles, especialmente en el contexto de transición hacia las energías renovables.

Este proyecto beneficia directamente a organismos de planificación, al brindarles herramientas que identifiquen los factores determinantes para la toma de decisiones informadas; y al sector productivo y la ciudadanía, al sentar las bases para operar una red más resiliente que reduzca la probabilidad de apagones no programados.

El proyecto es altamente viable ya que utiliza exclusivamente datos públicos: registros operativos diarios del CENACE, caudales de ríos mediante GloFAS/Copernicus, índices climáticos de la NOAA, registros de precipitaciones de ERA5, y proyecciones demográficas del INEC. Esta integración de fuentes asegura una base sólida para el entrenamiento de modelos. Las herramientas computacionales necesarias para desarrollar la parte técnica de este proyecto son de código abierto y la principal es el lenguaje de programación Python.

Ecuador no puede seguir gestionando la crisis energética de forma reactiva ya que como es de conocimiento general se han presentado diferentes problemas al tratar de integrar generadores y barcazas al SNI. (Vela, 2024)

La idea es transformar datos históricos en conocimiento estratégico, permitiendo entender no solo cuando falla el sistema, sino que factores específicos estuvieron presentes, así permitiendo una planificación preventiva.

1.3. Alcance

El presente proyecto de titulación tiene como objetivo desarrollar un sistema de modelado explicativo para el sistema nacional interconectado (SNI) del Ecuador. Se analizará el comportamiento de la red ecuatoriana, sin descender a niveles de distribución local. El estudio abarca un periodo de estudio de más de 16 años desde julio de 2009 hasta diciembre de 2025; la granularidad de los datos es diaria permitiendo así captar variaciones estacionales y eventos críticos. Se pretende identificar regímenes operativos mediante aprendizaje no supervisado (clustering) para segmentar el comportamiento del sistema, modelado explicativo de los cuatro flujos energéticos del SNI (generación hidroeléctrica, generación térmica, importación y exportación), cálculo de explicabilidad cuantitativa a través de valores SHAP permitiendo identificar el peso de factores que intervienen en el comportamiento.

Para mantener la viabilidad y el enfoque quedan fuera del alcance la predicción y los pronósticos de demanda o generación para fechas futuras, no se busca optimizar el despacho económico de carga ni proponer mejoras en la eficiencia de las plantas, no se

incluirá el modelado individual de plantas específicas, no se realizará estudios de flujos de potencia, estabilidad de voltaje o análisis de contingencias en las líneas de transmisión. El sistema propuesto es una herramienta para el análisis pos-operativo y de planificación estratégica.

1.4. Objetivos

1.4.1. Pregunta de Investigación

¿Qué factores hidrológicos, climáticos y demográficos explican los regímenes operativos del Sistema Nacional Interconectado del Ecuador y cómo se cuantifica su contribución a los flujos de generación e intercambio energético en el período 2009-2025?

De esta pregunta principal se desprenden dos preguntas secundarias:

a) ¿Es posible identificar regímenes operativos recurrentes del sistema a partir de sus propios datos históricos, sin conocimiento previo de eventos de racionamiento?

b) ¿En qué medida las variables externas al operador (caudales, ONI, precipitación, población) permiten reproducir o caracterizar esos regímenes?

1.4.2. Objetivo general

Analizar los factores determinantes en la generación e intercambio energético del Sistema Nacional Interconectado del Ecuador mediante técnicas de aprendizaje automático, utilizando datos históricos del período 2009 – 2025.

1.4.3. Hipótesis

Los estados operativos del Sistema Nacional Interconectado del Ecuador conforman un número reducido de patrones recurrentes identificables desde sus propios datos históricos, y las variables externas al operador (caudales, ONI, precipitación y demografía) aportan información suficiente para caracterizar estos patrones y explicar buena parte del comportamiento de los flujos de generación e intercambio del sistema, aunque no lo cubren en su totalidad.

1.4.4. Objetivos específicos

Construir un conjunto de datos integrado del Sistema Nacional Interconectado con variables endógenas (generación, demanda, intercambio) y exógenas (caudales, índice ENSO, precipitación, población) a partir de fuentes públicas, con granularidad diaria para el período 2009-2025.

Identificar los regímenes operativos del sistema eléctrico mediante técnicas de aprendizaje no supervisado, validando la coherencia de los grupos obtenidos contra los periodos de racionamiento documentados oficialmente, considerando como criterio de validación la proporción de meses con racionamiento que el algoritmo asigna a regímenes de severidad alta.

Desarrollar modelos que expliquen los flujos de generación (hidroeléctrica y térmica), importación y exportación de energía, cuantificando la contribución de cada variable al resultado mediante métodos de explicabilidad mediante técnicas de atribución de importancia por variable.

Evaluar si las variables exógenas disponibles (caudales, ONI, precipitación, población), sin incluir información de generación, son suficientes para caracterizar los regímenes operativos identificados, mediante clasificación supervisada, cuyo desempeño se medirá con métricas apropiadas para problemas multiclase, interpretando el valor obtenido como evidencia parcial o total de explicabilidad física.

Desarrollar un análisis de escenarios basado en patrones históricos que, a partir de un régimen operativo identificado, describa el comportamiento del sistema en períodos previos con condiciones similares y permita consultar precedentes específicos mediante técnicas de recuperación por similitud.

Presentar los resultados del análisis explicativo y los escenarios históricos identificados mediante visualizaciones que ayuden a interpretar los hallazgos del proyecto.

Capítulo 2

2. Revisión de Literatura

2.1. Estado del Arte

El análisis de los sistemas energéticos ha ganado peso en la última década por el aumento de la demanda, la integración de fuentes renovables y la creciente complejidad operativa del sector eléctrico. Distintos trabajos han abordado este comportamiento aplicando técnicas de análisis de datos, inteligencia artificial y modelos estadísticos, con el fin de entender la relación entre demanda, generación e intercambio.

El empleo de las técnicas de aprendizaje automático para el estudio de sistemas energéticos ha crecido notablemente en la última década. En su revisión sistemática sobre las publicaciones en el ámbito de la predicción de variables en sistemas de ingeniería eléctrica, basada en la inteligencia artificial Sánchez y Coto, (2022) concluyeron que aquellos ámbitos con mayor nivel de trascendencia son la predicción del consumo y la producción de energía eléctrica, en donde destacan las redes neuronales artificiales y las máquinas de soporte vectorial como soluciones principales. No obstante, los autores citados afirman que el enfoque explicativo, el que se centra en la explicación de las causas del comportamiento del sistema, más allá de la simple predicción de este, queda poco considerado.

Clustering en sistemas eléctricos. La detección de patrones operativos a partir de métodos no supervisados se ha establecido como una de las técnicas para analizar redes eléctricas. Miraftabzadeh et al. (2023) muestran una revisión de más de 440 artículos sobre K-Means y técnicas de agrupación alternativas en la red eléctrica moderna, poniendo de

manifiesto su uso en la segmentación de perfiles de consumo, categorización de los estados operativos y reducción de la dimensionalidad para la planificación. Aksan et al. (2021) aplican clustering al estudio de la calidad de la energía en plantas virtuales, comparando K-Means con algoritmos aglomerativos para clasificar las mediciones del sistema. Wang et al. (2019) revisan la aplicación de clustering en datos de medidores inteligentes para programas de respuesta a la demanda, mostrando las mejores prácticas a seguir al agrupar consumidores eléctricos. El algoritmo K-Means, presentado por MacQueen (1967), es el más conocido de los algoritmos de agrupación debido a su sencillez computacional y a su capacidad para descubrir grupos naturales en datos multivariados.

Modelos de gradient boosting en energía. XGBoost, que fue planteado en el trabajo de Chen y Guestrin (2016, p. 785), es hoy uno de los algoritmos más usados para el modelado de series energéticas, en su aplicación a datos tabulares, siendo capaz de resistir el sobreajuste de los datos, por medio de las regularizaciones L1 y L2. Hong et al. (2020) presentan una revisión integral de métodos de predicción energética estadísticos, de aprendizaje automático e híbridos, donde los modelos de gradient boosting destacan por su balance entre rendimiento y eficiencia computacional. En el contexto latinoamericano, Martínez (2025), a partir de datos provenientes del sistema ecuatoriano y haciendo uso de técnicas de aprendizaje automático, presenta un trabajo de pronósticos de la demanda eléctrica aplicadas en el sector residencial, evidenciando así la posibilidad de aplicar estas técnicas.

Explicabilidad en modelos de energía. La interpretabilidad de los modelos de aprendizaje automático aplicados a sistemas energéticos se ha vuelto un requisito para tomar decisiones informadas. Machlev et al. (2022) presentan la primera revisión comprensiva que mapea técnicas de inteligencia artificial explicable (XAI) a aplicaciones en sistemas de potencia, cubriendo SHAP, LIME y mecanismos de atención para pronóstico de carga, diagnóstico de fallas y gestión energética. Alsaigh et al. (2023) abordan la explicabilidad y gobernanza de IA en sistemas energéticos inteligentes, indicando que el método preferido es el de SHAP debido a su fundamentación en teoría de juegos cooperativos. El método SHAP fue publicado por Lundberg y Lee (2017) y se presenta como una forma unificada para interpretar las predicciones de los modelos basados en los valores de Shapley. Molnar (2019) presenta la teoría y la práctica de estos métodos en la referencia más completa sobre aprendizaje automático interpretable.

Sistema eléctrico ecuatoriano. Ecuador enfrenta desafíos particulares por su alta dependencia hidroeléctrica, y la literatura reciente ha abordado el problema desde distintos enfoques metodológicos. A continuación, se revisan cinco estudios representativos del contexto local, discutiendo en cada caso sus alcances y las limitaciones que dejan abiertas frente al análisis diario de los flujos del SNI propuesto en este trabajo.

Naranjo Silva et al. (2026) estudian la crisis energética 2023-2024 a través del modelado LEAP (Long-range Energy Alternatives Planning), una herramienta de planificación energética que simula escenarios de oferta y demanda a partir de supuestos macro. Trabajan con datos agregados anuales del sector energético ecuatoriano y

proyectan escenarios hasta 2050, encontrando que la dependencia hidroeléctrica del 67,4% es insostenible bajo escenarios de sequía recurrente y atribuyendo a la crisis un impacto económico aproximado de 2.000 millones de dólares. Su enfoque opera a nivel agregado y prospectivo, por lo que no captura la dinámica diaria de los flujos del SNI ni cuantifica la contribución específica de cada variable sobre cada flujo del sistema.

Peña et al. (2025) realizan una evaluación de confiabilidad del sistema eléctrico ecuatoriano utilizando métricas estándar de la ingeniería eléctrica, específicamente “Loss of Load Expectation” (LOLE), la cual estima cuantos días del año es posible que la capacidad de generación disponible no abastezca la demanda, para el año 2024 se estimó en 91 días/año y el 2023 fue de 75 días/año. Construyen una curva diaria a partir de los picos de demanda mensual publicados por el sistema (CENCE). La idea de este estudio es cuantificar la vulnerabilidad del sistema bajo distintos escenarios. Los resultados de este trabajo son indicadores de confiabilidad y no se explica que variables específicas influyen en el comportamiento.

En la publicación *“La crisis del sector eléctrico: vino para quedarse”* publicado por la USFQ, describen al problema como estructural es decir que aunque se presenten mejores condiciones climáticas, las fallas de inversión y planificación acumuladas siguen presentes. Se presenta un déficit de 600MW ya que actualmente se tiene una demanda pico de 4,900 MW frente a una oferta disponible de 4,300 MW. Este aporte es cualitativo y la distancia con el presente trabajo es de método. Ya que Vela(2024) se queda en el diagnóstico macro y no examina variables técnicas.

En la línea del aprendizaje automático aplicado al caso ecuatoriano se ubica el trabajo de Guamán et al. (2025), que proyecta el consumo eléctrico nacional hasta 2050 combinando redes neuronales MLP con redes autorregresivas no lineales NARNET. Los NARNET alimentan al MLP con variables socioeconómicas proyectadas a futuro y los autores muestran que esta combinación supera en precisión a modelos de Support Vector Machines y regresión lineal robusta, un aporte directo para la planificación del país. La distancia con el presente trabajo es de propósito y de granularidad. El autor pronostica demanda anual a escala nacional, mientras que aquí se analiza el comportamiento diario de los cuatro flujos del SNI.

Gallo Cruz (2020) en la Escuela Politécnica Nacional, propone un análisis predictivo basado en minería de datos para proyectar la demanda de potencia a corto plazo en el sistema ecuatoriano. Trabaja con datos diarios del operador del sistema y modelos clásicos de minería de datos. Su contribución radica en la aplicación de estas técnicas al contexto local, pero el enfoque sigue siendo predictivo y limitado a la demanda, sin abordar la composición de la generación ni los flujos de intercambio energético del sistema.

Sistemas de soporte a decisiones basados en patrones históricos. Más allá del modelado explicativo, existe una línea de la literatura que propone usar el comportamiento documentado del sistema bajo condiciones similares como base para el soporte a decisiones, en lugar de recurrir a forecasting. El paradigma del razonamiento basado en casos (Case-Based Reasoning, CBR), formalizado por Aamodt y Plaza (1994), plantea que un problema nuevo puede resolverse recuperando experiencias anteriores con características similares y adaptando sus soluciones, lo cual resulta particularmente útil

cuando la dinámica del sistema está influida por factores no observables como decisiones humanas, contratos comerciales o restricciones políticas que limitan la capacidad predictiva de los modelos puramente físicos. El algoritmo K-Nearest Neighbors (KNN), propuesto originalmente por Cover y Hart (1967), maneja un concepto basado en la clasificación o predicción de nuevos datos utilizando la similitud y la cercanía con sus vecinos más próximos, tomando la función principal de un mecanismo de búsqueda por similitud.

Brecha identificada. De lo ya antes analizado, el Sistema Nacional Interconectado (SNI) ha enfocado en tres puntos específicos:

i. Planificación del sistema eléctrico.
ii. Proyecciones de la demanda eléctrica.
iii. Análisis de confiabilidad del sistema.

La brecha más clara de esta operación y enfoque del SNI, es que los puntos específicos no abordan de manera conjunta el análisis del comportamiento diario para los principales flujos operativos de generación eléctrica, térmica, para importación y exportación de energía.

Considerando lo antes ya mencionado para este trabajo se realizará un análisis de más de 16 años de operación diaria del SNI, utilizando técnicas de aprendizaje automático. Aunque, existen metodologías de agrupamiento cuya base principal son las técnicas de agrupamiento orientados a los sistema eléctricos aplicarlos directamente a la operación

del SNI ecuatoriano convierte estos modelos en limitados. Es por ello que para este análisis se desarrollará el análisis utilizando los siguientes mecanismos:

i. K-Means para identificar regímenes operativos.
ii. Random Forest para validar su clasificación.
iii. XGBoost y SHAP para determinar las variables que explican cada comportamiento.
iv. K-Nearest Neighbors para identificar días históricos con condiciones operativas similares.

De esta manera, se evidencia la necesidad de investigaciones que integren múltiples variables del sistema energético y generen conocimiento estratégico a partir de datos históricos. Por lo que el presente proyecto se propone el desarrollo de un sistema inteligente orientado al análisis explicativo del sistema energético del Ecuador, diferenciándose de los enfoques tradicionales al centrarse en la comprensión del comportamiento energético y no en la predicción.

2.2. Marco teórico

El marco teórico del trabajo se organiza en seis subsecciones que cubren los pilares conceptuales del proyecto: el Sistema Nacional Interconectado del Ecuador como objeto de estudio, los algoritmos de aprendizaje no supervisado y supervisado, la explicabilidad de modelos mediante valores de Shapley, el razonamiento basado en casos como complemento al modelado predictivo, y las métricas de evaluación que sustentan los resultados reportados.

2.2.1. El Sistema Nacional Interconectado del Ecuador

El Sistema Nacional Interconectado (SNI) es la infraestructura eléctrica que articula la generación, la transmisión y la distribución de energía a nivel nacional en Ecuador. Su operación se rige por la Ley Orgánica del Servicio Público de Energía Eléctrica (LOSPEE; República del Ecuador, 2015) y se encuentra bajo la regulación de la Agencia de Regulación y Control de Energía y Recursos Naturales No Renovables (ARCERNNR, 2020).

El SNI se compone de tres niveles funcionales. La generación, donde participan empresas públicas como CELEC EP y generadoras privadas, integra cuatro tipos de fuentes: hidroeléctrica, termoeléctrica, no convencional (solar, eólica y biomasa) y las interconexiones internacionales con Colombia y Perú. La transmisión cuya responsabilidad recae sobre la empresa estatal CELEC EP - Transelectric, que opera la red de alta tensión a nivel nacional. Finalmente, la distribución de la energía hasta los consumidores es responsabilidad de las compañías eléctricas regionales.

El Centro Nacional de Control de Energía (CENACE), es la entidad pública responsable de coordinar la operación del sistema y determina cuales son los registros que se utilizarán para el análisis del comportamiento del SNI.

Para este trabajo se estudiarán los datos de generación hidroeléctrica, térmica e intercambios mediante compra y venta de energía eléctrica extranjera. Al realizar un análisis detallado de este tipo de generación se logrará estudiar el comportamiento y como el SNI cubre diariamente la demanda de consumo. Se realizará un énfasis en la generación hidroeléctrica cuya concentración de generación representa el 67,4% del total demandado,

ya que es el sistema de generación mayormente expuesto a problemas por el cambio climático (Naranjo Silva et al., 2026).

2.2.2. Aprendizaje no supervisado: K-Means

Dentro del aprendizaje automático se encuentra el aprendizaje no supervisado, la idea de estos algoritmos es descubrir estructura en los datos sin necesidad de indicarle que es lo que está describiendo es decir no son necesarias la etiquetas. Al finalizar se obtienen varios grupos que contienen observaciones similares entre sí.

El algoritmo de clustering K-means formulado por MacQueen (1967), su funcionamiento se centra en elegir el parámetro *K,* que indica el número de grupos que se quieren encontrar y estos puntos *K* se colocan al azar dentro del conjunto de datos. El algoritmo realiza su proceso iterativo de asignar puntos al centroide más cercano usando la distancia euclidiana y se recalcula el centroide como el promedio de todos los puntos. Este proceso iterativo lo que busca es minimizar la distancia total entre cada punto y su centroide, a lo que se conoce como inercia o suma de errores cuadráticos (SSE). Formalmente se conoce como (Within-Cluster Sum of Squares) y se la define a continuación:

$$\mathrm{WCSS} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{x_i \in S_j} \left\| x_i - u_j \right\|^2 \tag{2.1}$$

Donde k es el número de clusters o grupos, s_j es el conjunto de todos los datos asignados al cluster j y $u_j$es el centroide del clúster j es decir el punto promedio que representa al grupo, x_j representa cada punto de dato individual dentro de ese cluster.

La selección de *k* es compleja si se realiza solo por intuición, para definir este parámetro k se tiene el método del codo el cual gráfica el SSE para distintos valores de *k* y se busca el punto donde agregar más clusters deja de reducir el error significativamente. La fundamentación matemática parte de los criterios de selección de grupos (Thorndike, 1953), o el índice de Calinski-Harabasz, desarrollado por Caliński y Harabasz (1974), el cual compara la dispersión entre clústeres con la dispersión dentro de cada clúster. A diferencia de otros métodos de clustering en este algoritmo cada centroide es un valor en el mismo espacio de datos, lo cual facilita interpretar el cluster y darle un nombre significativo.

2.2.3. Aprendizaje supervisado: Random Forest y XGBoost

El aprendizaje supervisado se diferencia del no supervisado en que cada observación incluye una etiqueta o valor objetivo; en problemas de clasificación, una categoría, y en regresión, un valor numérico (Russell & Norvig, 2020). El algoritmo aprende la relación entre las variables predictoras y la variable objetivo a partir de un conjunto de entrenamiento, y la aplica para inferir resultados sobre observaciones nuevas.

Los árboles de decisión constituyen la base conceptual de los dos algoritmos supervisados que utiliza este trabajo. Un árbol particiona el espacio de variables predictoras en regiones rectangulares, asignando a cada región un valor predicho. La construcción del árbol busca maximizar la pureza dentro de cada partición usando criterios

como la impureza de Gini o la ganancia de información para clasificación, y el error cuadrático para regresión.

Random Forest, propuesto por Breiman (2001), cuyo principio es, agrupar o construir varios árboles de decisión a partir de algunas muestras aleatorias del conjunto de datos. Cuando construye cada árbol también selecciona variables para decidir las divisiones, generando modelos que trabajan en paralelo o conjunto. La predicción final es el promedio (en regresión) o el voto mayoritario (en clasificación) de los árboles individuales. Este enfoque, conocido como bagging, reduce la varianza del modelo y mejora la generalización. Una propiedad útil de Random Forest es la estimación out-of-bag (OOB), que permite estimar el error sin necesidad de un conjunto de validación separado.

XGBoost (Extreme Gradient Boosting), introducido por Chen & Guestrin (2016, p. 788), es una implementación optimizada de gradient boosting. A diferencia de Random Forest, los árboles se entrenan secuencialmente y cada nuevo árbol se ajusta para corregir los errores residuales del modelo acumulado hasta ese punto. La predicción final es la suma ponderada de las predicciones de todos los árboles. *XGBoost incorpora regularización L_1 (Lasso) y L_2 (Ridge) en su función objetivo:*

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.2}$$

donde $l(\cdot)$ es la función de pérdida (típicamente la raíz del error cuadrático medio [RMSE] en regresión o *log-loss* en clasificación) y $\Omega(\cdot)$ es el término de regularización aplicado a cada árbol f_k, que penaliza los pesos grandes y reduce el riesgo de sobreajuste.

La capacidad de manejar datos tabulares heterogéneos, valores faltantes y la velocidad de entrenamiento son las razones por las que XGBoost ha dominado los benchmarks de aprendizaje supervisado sobre datos tabulares durante la última década.

2.2.4. Explicabilidad de modelos: valores de Shapley y SHAP

La explicabilidad de modelos de aprendizaje automático es un campo que busca describir, de manera comprensible para humanos, cómo se generan las predicciones de un modelo. La distinción central es entre explicabilidad global, que caracteriza el comportamiento promedio del modelo, y explicabilidad local, que se enfoca en una predicción individual y descompone su origen variable por variable.

Los valores de Shapley, originados en la teoría de juegos cooperativos, proporcionan una manera matemáticamente fundamentada de distribuir la ganancia total de un juego entre sus jugadores (Shapley, 1953). Aplicados a aprendizaje automático, los jugadores son las variables predictoras y la ganancia es la diferencia entre la predicción del modelo para una observación específica y un valor de referencia (típicamente la predicción promedio sobre el conjunto). El valor de Shapley para la variable i se define formalmente como:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \cdot [v(S \cup \{i\}) - v(S)] \tag{2.3}$$

Donde *N* es el conjunto de variables, *S* un subconjunto que no incluye a la variable *i*, y *v(S)* la predicción del modelo usando solo las variables en *S*. La suma considera todas las posibles coaliciones de variables. Los valores de Shapley cumplen cuatro axiomas: (a)

eficiencia, donde la suma de las contribuciones de todas las variables es igual a la diferencia entre la predicción y el valor de referencia; (b) simetría, si dos variables aportan lo mismo en todas las combinaciones reciben la misma contribución; (c) variable nula, una variable que no aporta nada en ninguna combinación recibe contribución cero; y (d) aditividad, donde las contribuciones se suman cuando se combinan modelos.

SHAP (Shapley Additive Explanations), propuesto por Lundberg y Lee (2017), es una implementación práctica de los valores de Shapley para modelos de aprendizaje automático. Su versión TreeExplainer aprovecha la estructura interna de los modelos basados en árboles (Random Forest, XGBoost) para calcular valores de Shapley exactos de manera eficiente, con complejidad polinomial respecto al tamaño del modelo (número de árboles, hojas y profundidad), en lugar de la aproximación combinatoria que requeriría calcularlos directamente sobre todas las posibles combinaciones de variables.

Los valores SHAP permiten responder preguntas concretas para cualquier observación específica: cuánto aportó cada variable a la predicción, en qué dirección (aumentó o redujo el valor predicho), y qué tan grande fue su efecto en relación con las demás variables. Esta granularidad es lo que permite hacer análisis caso por caso, no solo informes agregados. Machlev et al. (2022), en su revisión sobre XAI en sistemas de potencia, ubican a SHAP como el método de mayor adopción en el campo.

2.2.5. Razonamiento basado en casos: ciclo CBR y K-Nearest Neighbors

El razonamiento basado en casos (Case-Based Reasoning, CBR) es un paradigma de resolución de problemas formulado por Aamodt y Plaza (1994) que se distingue del

aprendizaje supervisado tradicional. Mientras que un modelo supervisado aprende una función general que mapea entradas a salidas, el CBR aborda cada problema nuevo recuperando experiencias anteriores con condiciones similares y adaptando sus soluciones.

El ciclo CBR canónico se compone de cuatro fases (Aamodt & Plaza, 1994). La primera es *Retrieve*, que dado un problema nuevo busca en la base de casos aquellos con condiciones más parecidas. La segunda es *Reuse*, donde se toman las soluciones de los casos recuperados como punto de partida. La tercera es *Revise*, que adapta la solución reutilizada al contexto específico considerando las diferencias. La cuarta es *Retain*, donde el nuevo caso resuelto se incorpora a la base, enriqueciendo el sistema para futuras consultas.

K-Nearest Neighbors (KNN), formulado originalmente por Cover y Hart (1967), es una técnica clásica de aprendizaje basado en instancias que se utiliza naturalmente como mecanismo de *retrieval* en sistemas CBR. KNN no construye un modelo explícito durante el entrenamiento; almacena el conjunto de datos completo y, cuando llega una consulta, calcula la distancia entre esta y todas las instancias almacenadas, devolviendo las *K* más cercanas.

La función de distancia más utilizada es la euclidiana, que mide la separación geométrica entre puntos en el espacio de variables:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

(2.4)

donde *x* e *y* son dos observaciones cualesquiera dentro del espacio de variables y *n* es el número de variables consideradas; cada término bajo la raíz compara ambas observaciones, variable por variable antes de sumarlas.

Cuando las variables tienen escalas distintas, es necesario normalizar antes del cálculo de distancia para evitar que las variables con rangos grandes dominen el cálculo. La normalización estándar (*z-score*) transforma cada variable de la siguiente manera:

$$x' = \frac{(x - \mu)}{\sigma}$$

(2.5)

donde μ es la media de la variable y σ su desviación estándar. La elección de k es un equilibrio entre la estabilidad y la especificidad puesto que números más bajos dan mejores soluciones para cada caso pero también son más afectados por el ruido mientras que números más altos suavizan más casos pero poseen mayor estabilidad.

De las cuatro fases del ciclo CBR, este trabajo implementa principalmente la primera (*Retrieve*) mediante *K*-NN. Las fases de *Reuse* y *Revise* quedan a discreción del analista que consulte el sistema, dado que adaptar la información del precedente al contexto actual depende del juicio humano sobre las diferencias entre situaciones. La fase

de *Retain* no se implementa en este alcance, pero podría incorporarse como trabajo futuro mediante la actualización dinámica de la base de casos.

2.2.6. Métricas de evaluación

Para evaluar modelos de aprendizaje automático es necesario utilizar métricas específicas según el tipo de tarea. En el presente trabajo se emplean tres familias: métricas de clustering, de clasificación y de regresión.

Para evaluar la calidad de los regímenes operativos identificados por K-Means se utilizan tres índices internos. El silhouette score mide qué tan bien cada observación encaja en su clúster comparado con los demás clústeres; sus valores van de -1 a 1, donde los valores positivos cercanos a 1 indican agrupaciones bien separadas. El índice de Calinski-Harabasz compara la dispersión entre clústeres con la dispersión dentro de cada cluster, donde valores altos indican clústeres compactos y bien separados. El índice de Davies-Bouldin mide la similitud promedio entre cada clúster y el más cercano, de modo que valores más bajos indican una mejor separación (Davies & Bouldin, 1979). La selección del número óptimo de clústeres utiliza estos índices de forma conjunta con el método del codo.

Para evaluar el clasificador Random Forest empleado en la validación de regímenes se utilizan métricas estándar de clasificación (Kuhn y Johnson, 2013). La precisión mide qué proporción de las predicciones positivas son correctas, y el recall, también llamado sensibilidad, mide qué proporción de los casos positivos reales fueron identificados. El F1-score es la media armónica entre precisión y recall, una métrica especialmente útil cuando

ambas dimensiones tienen una importancia equivalente o ante la presencia de conjuntos de datos desbalanceados, calculándose de la siguiente manera:

$$F_1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}} \tag{2.6}$$

Donde *Precisión* es la proporción de predicciones positivas correctas y *Recall* la sensibilidad del modelo.

Esta media armónica castiga con más fuerza los casos en que una de las dos métricas es baja, aunque la otra sea alta, evitando que un buen recall disimule una precisión pobre o viceversa.

La accuracy mide la proporción de predicciones correctas en total y es muy informativa cuando las clases están balanceadas (Kuhn y Johnson, 2013). Dado que el problema involucra cinco regímenes potencialmente desbalanceados, por lo que se reporta el F1 macro, que promedia los F1-score por clase sin ponderar por frecuencia, de modo que cada categoría de régimen sea igualmente importante.

Para valorar los modelos XGBoost de los flujos del sistema se emplean métricas de regresión. El coeficiente de determinación (R^2) mide la proporción de varianza de la variable objetivo que el modelo es capaz de explicar; valores cercanos a 1 indican un mejor ajuste y valores cercanos a 0 indican que el modelo no aporta información más allá del promedio histórico:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(2.7)

donde cada y_i es el valor observado, cada $\hat{y}_i$ el valor predicho por el modelo, y $\bar{y}$ el promedio de los valores observados en el conjunto; el cociente compara así el error del modelo frente al que cometería una predicción ingenua basada solo en ese promedio.

El error absoluto medio (MAE) informa la desviación promedio en las unidades originales de la variable objetivo:

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

(2.8)

Donde n es el número total de observaciones y $|y_i - \hat{y}_i|$ representa el error absoluto de cada predicción.

A diferencia del RMSE, el MAE pondera todos los errores por igual sin importar su magnitud, por lo que resulta menos sensible a valores atípicos.

La raíz del error cuadrático medio (RMSE) mide la magnitud del error cuadrático medio y, a diferencia del MAE, penaliza en mayor medida los errores grandes debido a la elevación al cuadrado, lo que la vuelve una métrica idónea para la detección de valores atípicos u outliers:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

(2.9)

Por esta razón, cuando el RMSE es sustancialmente mayor que el MAE, es señal de que existen errores puntuales grandes en las predicciones del modelo.

La validación se efectúa a través de una separación temporal estricta del conjunto de datos: entrenamiento en el periodo 2009-2022 y evaluación en el periodo 2023-2025, sin recurrir a una división aleatoria (Hyndman & Athanasopoulos, 2018). Esta estrategia evita la fuga de información (data leakage) del futuro hacia el pasado y respeta la naturaleza temporal intrínseca del problema. Para Random Forest se usa validación cruzada estratificada de 5 pliegues sin barajar los meses, para no romper el orden temporal de los datos.

2.2.7. Cierre del marco teórico

Los seis pilares conceptuales antes mencionados se materializan en el pipeline metodológico que se detalla en el Capítulo 3. El algoritmo K-Means se utilizará para discernir los regímenes operativos del Sistema Nacional Interconectado (SNI) a partir de los datos diarios. Posteriormente el algoritmo Random Forest examinará en qué medida esos regímenes pueden ser explicados por las variables exógenas físicas. Por su parte el algoritmo XGBoost modelará los flujos de generación e intercambio del sistema, mientras que el enfoque de SHAP descompondrá cada predicción en contribuciones individuales por

variable para garantizar la interpretabilidad del modelo. Finalmente, KNN recuperará los precedentes históricos con condiciones similares para el análisis de escenarios. Las métricas previstas en la sección 2.2.6 serán utilizadas para valorar el rendimiento de cada componente, usando la división temporal indicada para respetar la naturaleza secuencial y cronológica de los datos.

Capítulo 3

3. Desarrollo

3.1. Enfoque metodológico y diseño de la investigación

Para sustentar el enfoque metodológico se debe considerar que la investigación no generó datos nuevos, si no, tomó información histórica ya existente del SNI y fue analizada utilizando métodos de aprendizaje automático y estadísticos. Está investigación está basada en un estudio cuantitativo que parte de mediciones ya registradas, patrones y relaciones. Para ello es importante notar que para dicha investigación no se realizaron encuestas, entrevistas ni recolección propia; todos los datos provienen de fuentes públicas.

La investigación utiliza los datos bajo los cuales el SIN operó entre 2009 y 2025, cuyo carácter es longitudinal, sin intervenir ni modificar las variables utilizadas durante la operación. Es decir dicha investigación se categorizará como no experimental al no manipular los datos ya existentes.

La investigación es categorizada como exploratoria y explicativa.

i. Exploratoria porque no encontramos antecedentes locales que apliquen al SNI ecuatoriano el mismo conjunto de técnicas que se emplean en el presente trabajo.
ii. Explicativa porque no solo se enfoca en describir el comportamiento del sistema, sino que busca cuantificar cuánto pesan los factores externos al

operador tales como caudales, índice ONI, precipitación y demografía sobre la generación y el intercambio energético.

El objeto de investigación es el comportamiento global del SNI, más no sus componentes individuales, tales como, plantas distribuidoras por cantones o ciudades. Los datos referentes a la generación se recolectan día por día, mientras que también existen patrones dentro de la operación que pueden ser mensuales por la velocidad de respuesta o toma de datos debido a factores climáticos, tales como el índice ONI y la precipitación acumulada.

Las fuentes empleadas son todas públicas: SMEC del CENACE para los datos operativos, GloFAS/Copernicus para caudales, NOAA para el índice ENSO, ERA5 para precipitación en zonas de embalses e INEC para proyecciones poblacionales. El detalle de la integración se aborda en las secciones siguientes.

3.2. Construcción del conjunto de datos

La fuente principal de datos es el archivo histórico del Sistema SMEC del CENACE (Centro Nacional de Control de Energía del Ecuador), que registra la operación diaria del Sistema Nacional Interconectado. Se seleccionó esta fuente por su granularidad diaria, cobertura temporal (2009-2025), y por incluir las variables clave del sistema: demanda, generación por tipo, importación, exportación y pérdidas.

Los datos de CENACE del período 2009-2025 se complementaron con cuatro fuentes exógenas:

• GloFAS (Copernicus/UE): Caudales diarios de los ríos Paute, Coca, Pastaza y Napo, medidos sobre las coordenadas de las represas principales del país (Harrigan et al., 2020).

• NOAA PSL: Índice Oceánico El Niño (ONI), que mide el fenómeno ENSO y afecta los patrones de lluvia en Ecuador (NOAA, 2025).

• Open-Meteo / ERA5 (ECMWF): Precipitación diaria en las zonas de embalses hidroeléctricos (Hersbach et al., 2020).

• INEC: Proyecciones de población nacional 1990-2035, interpoladas a resolución diaria (INEC, 2024).

3.2.1. Alineación temporal entre fuentes heterogéneas

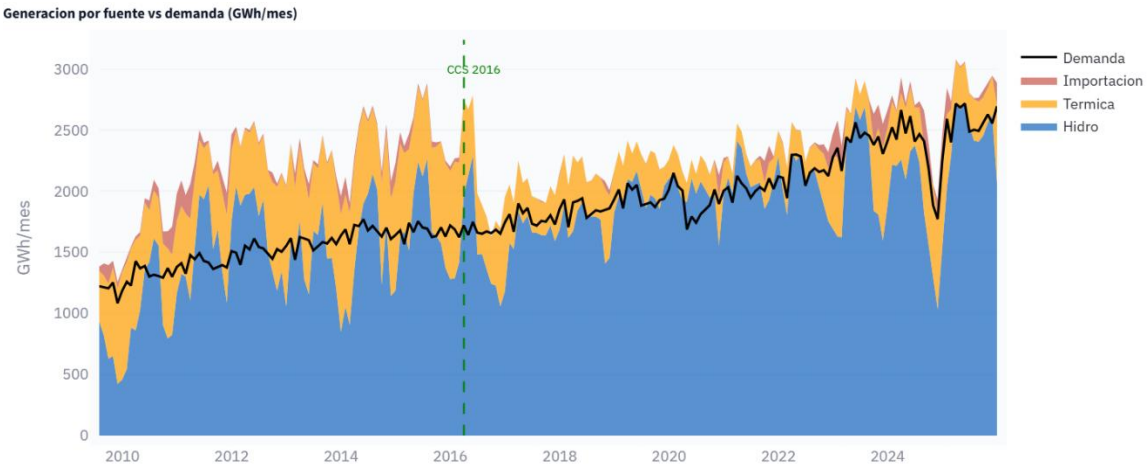
No todas las fuentes vienen con la misma granularidad temporal, por lo que fue necesario llevarlas a un único nivel de detalle: diario. Las que ya están a ese nivel (CENACE, GloFAS-ERA5 y ERA5/Open-Meteo) se integraron directamente usando la fecha como llave común. Para el índice ONI, que se publica de forma mensual, se replicó el valor del mes en cada uno de sus días correspondientes; esto tiene sentido porque el ONI describe una condición climática que se mantiene durante todo el mes y no cambia día a día. En el caso de las proyecciones de población del INEC, que vienen en valores anuales, se aplicó interpolación lineal entre años consecutivos para obtener un valor estimado por día. Considerando que los datos de crecimiento poblacional no son diarios se realizó una estimación de población, donde se asumió una variación entre los datos de dos años consecutivos, dando como resultado, una serie diaria, continua e integrada con el resto de las fuentes de información.

3.2.2. Descripción del conjunto de datos

El dataset resultante de la integración de las fuentes descritas en la sección 3.2, está organizado con una resolución diaria y comprende el periodo entre el 1 de julio de 2009 y el 31 de diciembre de 2025, lo que equivale a aproximadamente 6,000 registros por columna. Cada fila representa un día de operación del Sistema Nacional Interconectado y agrupa, en formato tabular, las variables endógenas del sistema (demanda, generación por tipo, importación, exportación y pérdidas reportadas por el CENACE), las variables exógenas físicas (caudales fluviales, índice ENSO, precipitación sobre las zonas de embalses y proyecciones poblacionales) y las variables derivadas calculadas a partir de las anteriores (porcentaje de generación hidroeléctrica sobre el total, dependencia de la importación y excedente de exportación).

Figura 1

Composición mensual de la generación eléctrica del SNI frente a la demanda nacional (2009–2025)



Nota. Elaboración propia. Áreas apiladas para hidroeléctrica, térmica e importación; línea negra para la demanda. La línea vertical verde marca la entrada en operación de Coca Codo Sinclair en abril de 2016, cuando la participación hidroeléctrica dio un salto estructural.

El conjunto de datos se almacena en formato CSV con la fecha como clave de unión y constituye el insumo único para todos los modelos del pipeline metodológico descrito en las siguientes secciones.

Tabla 1

Descripción general del conjunto de datos

Características	Valor
Nombre del archivo	sni_diario_2009_2025.csv
Período	1 julio 2009 - 31 diciembre 2025
Granularidad	Diaria
Filas	6.028
Columnas	16 (variables efectivamente utilizadas en el modelado)
Valores Nulos	0
Gaps Temporales	0
Fechas duplicadas	0

Nota. Se describen las características generales del conjunto de datos construido donde se puede visualizar que el total de registros es de 6.028 (equivalente a 96.448 celdas por 16 columnas).

A continuación, se detallan las variables que componen el conjunto de datos.

Tabla 2

Descripción de las variables

#	Variable	Tipo	Fuente	Rol en el modelado	Descripción
1	fecha	Date	CENACE	Identificador temporal	Fecha del registro
2	tipo_dia	Texto	Legislación EC	Feature XGBoost	Laboral, sábado, domingo o Festivo
3	demanda_gwh	Float	CENACE	Feature XGBoost	Demanda eléctrica nacional diaria en GWh
4	gen_hidraulica_gwh	Float	CENACE	Variable objetivo XGBoost (gen. hidro)	Generación hidroeléctrica diaria en GWh
5	gen_termica_total_gwh	Float	Derivada	Variable objetivo XGBoost (gen. térmica)	Suma de las siete subclases de generación térmica
6	total_generacion_gwh	Float	CENACE	Auxiliar de cálculo	Suma total de generación hidráulica y térmica
7	total_importacion_gwh	Float	CENACE	Variable objetivo XGBoost (importación)	Importación neta desde Colombia y Perú
8	total_exportacion_gwh	Float	CENACE	Variable objetivo XGBoost (exportación)	Exportación neta hacia países vecinos
9	total_perdidas_transporte_gwh	Float	CENACE	Auxiliar de cálculo	Pérdidas de energía en la red de transmisión
10	pct_hidro	Float	Derivada	Feature XGBoost / var. estado K-Means	Fracción hidroeléctrica sobre la generación total
11	dependencia_importacion	Float	Derivada	Variable de estado K-Means	Cociente entre importación y demanda
12	excedente_exportacion	Float	Derivada	Variable de estado K-Means	Cociente entre exportación y demanda

13	caudal_ponderado	Float	Derivada de GloFAS	Feature RF / KNN / XGBoost	Promedio ponderado de los caudales de los ríos Paute, Coca, Pastaza y Napo según capacidad hidroeléctrica
14	oni_mensual	Float	NOAA	Feature RF / KNN / XGBoost	Oceanic Niño Index (ENSO), publicado mensualmente
15	precip_embalses	Float	ERA5 (ECMWF)	Feature RF / KNN / XGBoost	Precipitación acumulada en las zonas de embalses hidroeléctricos
16	poblacion	Int	INEC	Feature RF / XGBoost	Proyección de población nacional interpolada a resolución diaria

Nota. Elaboración Propia. Se explica que representa cada una de las 16 columnas o features que componen el conjunto de datos, se detalla el tipo de dato y su unidad en caso de ser necesario.

3.2.3. Fundamento teórico de selección de variables

La selección de variables mantiene una relación física, operativa o documental vinculado al SNI. Bajo este criterio se definieron las variables que forman parte del análisis y se excluyeron aquellas cuya disponibilidad y resolución resultan limitados.

Variables endógenas del sistema. Estas variables son elementos básicos en el análisis y modelado de los sistemas eléctricos ya que mantienen el equilibrio entre el consumo y la generación (Hong et al., 2020). Las variables que permiten definir el comportamiento del SIN a nivel general son:

i. Demanda diaria en GWh.

ii. Generación desagregada por fuentes hidroeléctricas y térmicas.

iii. Importación y exportación de energía.

La importación y exportación se incluyen como variables objetivo del análisis ya que estas alteran la operación normal del SNI cuando existen cambios internos en la generación, tal como quedó documentado en el análisis de la crisis 2023-2024 en Ecuador (Naranjo et al., 2026).

La variable hidroeléctrica (pct_hidro) representa la participación de generación hidroeléctrica respecto a la generación total del SNI.

Se incorpora como indicador de estado y no como variable objetivo ya que responde a la alta participación de la hidro generación del Ecuador. La razón es que la dependencia del SNI frente a la hidroelectricidad es del 67% de la generación según los datos del 2024, fecha donde ocurrió la crisis energética. Si esta proporción tiende a disminuir es por que suele estar acompañada de mayor consumo en la utilización de energía térmica o por importaciones de energía. Es por ello que esta variable permite determinar balances energéticos.

Variables exógenas hidro climáticas. Con la finalidad de representar la disponibilidad hídrica se trabajop con la variable de caudal sobre los ríos Paute, Coca, Pastaza y Napo, los cuales alimentan a los complejos y centrales hidroeléctricas más grandes del Ecuador. Los datos se obtuvieron de GloFAS-ERA5 (Harrigan et al., 2020), que ofrece cobertura continua diaria sobre los caudales principales manteniendo una cobertura temporal continua que permite mantener una serie homogénea de datos durante el análisis de caudal.

Además, se trabajó con la precipitación, la misma que se consideró solo en zonas específicas asociadas a embalses y cuencas principales. Esto permite relacionar

directamente la lluvia con al disponibilidad de los recursos hídricos, para ello se utilizó la fuente ERA5 (Hersbach et al., 2020).

El comportamiento del fenómeno ENSO se incorporó utilizando el “Oceanic Niño Index” (ONI) publicado por la NOAA. Se optó por el ONI porque su publicación mensual continua permite alinear la serie con la granularidad del resto del dataset.

Variables exógenas estructurales y calendáricas. Se agregó la población como variable en el contexto socioeconómico, es una de las variables utilizadas para la proyección más eficientes ya que tienen una relación directa con la evolución de la demanda eléctrica. Para la recolección de data se utilizaron las proyecciones del INEC, ya que permite representar cambios a largo plazo.

El tipo de día (laboral, sábado, domingo, festivo) ya que permite determinar las variaciones semanales de demanda eléctrica, debido a su variación por los comportamientos socioeconómicos de acuerdo con cada tipo de jornada.

Variables descartadas. Para la construcción del conjunto de datos se evaluaron algunas variables que no fueron incorporadas tales como:

i. Temperatura ambiente promedio ya que odia ocultar diferencias regionales importantes o a su vez podría aportar información limitada al modelo.
ii. Precios spot del mercado regional debido a una falta de una serie pública continua y homogénea.
iii. Demanda desagregada por los sectores residenciales, comerciales debido a que su resolución temporal no era compatible con los análisis diarios.

Estas exclusiones son consideradas como limitaciones para la presente investigación y estudio y se las retomas en las recomendaciones para próximas investigaciones.

3.2.4. Selección de variables según el rol de análisis

Se seleccionaron únicamente las variables que aportan información relevante para el análisis energético y para el entrenamiento de los diferentes modelos de inteligencia artificial. La selección respondió a criterios de relevancia técnica y disponibilidad histórica de la información.

Variables predictoras de los modelos. El modelo XGBoost utiliza siete variables: demanda_gwh, pct_hidro, caudal_ponderado, oni_mensual, precip_embalses, tipo_dia (codificado numéricamente como tipo_dia_num antes del entrenamiento) y población. El clasificador Random Forest, empleado para validar los regímenes, recibe solo las cuatro variables exógenas físicas del conjunto anterior: caudal_ponderado, oni_mensual, precip_embalses y población. Las variables endógenas del sistema se dejan fuera a propósito, porque el objetivo es evaluar si los regímenes pueden recuperarse a partir de condiciones externas al operador.

Variable objetivo principal (total_importacion_gwh): Es el flujo más sensible al estado del sistema y el que más se activa durante las crisis, por lo que concentra el mayor valor explicativo del análisis.

Variables objetivo secundarias (total_exportacion_gwh, gen_hidraulica_gwh, gen_termica_total_gwh): Estas se modelan como variables objetivo de modelos

complementarios para tener una visión completa de los flujos del sistema, no como entregable principal.

Variables descriptivas (fecha, total_generacion_gwh, total_perdidas_transporte_gwh): No se utilizan en los modelos.

Variables derivadas (pct_hidro, dependencia_importacion, excedente_exportacion): Son indicadores calculados en función de las variables base.

Limitación reconocida sobre pct_hidro como predictora del modelo de importación. La variable pct_hidro merece una observación aparte. Se calcula como gen_hidraulica/total_generacion, y total_generacion entra en la ecuación de balance del SNI junto con la importación. Esto significa que pct_hidro arrastra información de la variable que el modelo de importación intenta predecir, lo que abre un riesgo de fuga de información (data leakage). Es una limitación que conviene admitir desde ya: en la fase de desarrollo del trabajo se entrena una segunda versión del modelo de importación sin pct_hidro, para ver qué tanto cambian los R^2 y los SHAP, y así verificar si las explicaciones siguen siendo robustas.

3.3. Volumen y cobertura

El dataset cubre más de 16 años de operación del SNI ecuatoriano (julio 2009 - diciembre 2025), incluyendo: Períodos de crisis con racionamiento oficial: 2009 (Decreto Ejecutivo No. 124) y 2023-2024; El quiebre estructural por la entrada en operación de Coca Codo Sinclair en 2016, la central hidroeléctrica más grande del país con 1,500 MW de capacidad instalada; Eventos de El Niño y La Niña registrados en el índice ONI; La pandemia

COVID-19 y su efecto en la demanda eléctrica (2020-2021); Variabilidad estacional completa (16 ciclos anuales de datos).

3.4. Pipeline metodológico y arquitectura del sistema

El desarrollo de la oslución se organizó en cinco etapas principales, las cuales abarcan desde la recopilación y preparación de los datos hasta la construcción y evaluación de sistema.

Figura 2

Metodología implementada

Nota. Elaboración propia. Diagrama del pipeline metodológico de cinco fases desde la ingesta de datos hasta la entrega del recomendador.

Fase 1: Adquisición y consolidación de datos. Se recopilaron datos históricos de múltiples plataformas institucionales, cubriendo el período del 1 de julio de 2009 al 31 de diciembre de 2025.

Fase 2: Integración y limpieza. Se evaluó la calidad analítica del conjunto de datos y se calcularon métricas de consistencia temporal para descartar duplicados y discontinuidades en los historiales.

Fase 3: Análisis exploratorio. Se generaron gráficas ilustrativas para entender la composición y el comportamiento de los datos antes del entrenamiento.

Fase 4: Modelado y entrenamiento. Se implementaron en paralelo arquitecturas de aprendizaje supervisado y no supervisado para caracterizar la operación del sistema energético y aislar sus factores condicionantes.

Fase 5: Construcción de la recomendación. No constituye otro modelo de IA, sino una etapa que orquesta las salidas de los componentes anteriores. Para un mes consultado, el sistema detecta el régimen con K-Means, revisa las estadísticas históricas de ese régimen, busca los cinco meses más parecidos con KNN y genera una recomendación.

3.5. Evaluación comparativa de modelos

La selección de los algoritmos del pipeline no se da por asumida. Cada una de las tres tareas que aborda el proyecto (clustering, clasificación y regresión), se trabaja con un modelo principal escogido por su alineamiento con el alcance, pero se compara contra al menos dos alternativas. La intención es validar empíricamente que la elección es defendible y no producto de inercia metodológica.

El núcleo metodológico del sistema se desarrolló a partir de un enfoque analítico explicativo, el cual combina cuatro técnicas de Machine Learning:

1. Identificación de Regímenes Operativos (Aprendizaje No Supervisado)

Para la clasificación de los comportamientos del SNI se aplicó el algoritmo K-Means Clustering. Dada la característica de inercia mensual que presentan las dinámicas de la planificación eléctrica, se consideró la construcción y normalización de un subconjunto de datos agregados mensualmente. Se normalizaron las variables mediante StandardScaler, con el objetivo de eliminar el efecto de las diferentes escalas de medición y evitar que una variable tuviera mayor influencia que otra durante el proceso de agrupamiento. Las métricas de decisión son el silhouette score, el índice de Calinski-Harabasz y el de Davies-Bouldin, evaluadas sobre el mismo conjunto y con el mismo preprocesamiento para las tres opciones.

2. Validación de Regímenes desde Variables Exógenas

Una vez identificados los regímenes operativos ocultos, se entrenó un clasificador basado en Random Forest que se ocupa de predecir la pertenencia al régimen de acuerdo con variables no controladas por la operadora eléctrica (meteorológicas y demográficas). Se compara contra una máquina de soporte vectorial (SVM) con kernel RBF, que captura fronteras no lineales, pero exige un escalado cuidadoso, y contra una regresión logística multinomial como referencia interpretable de baja complejidad. La métrica de decisión es el F1 macro evaluado mediante validación cruzada estratificada de 5 folds, dado que los regímenes pueden estar desbalanceados y el F1 macro pondera todas las clases por igual.

3. Modelado Explicativo del Intercambio Energético (Regresión y SHAP)

Para el análisis del intercambio energético se seleccionó como el modelo principal a XGBoost, escogido por su buen desempeño en problemas con datos tabulares y su compatibilidad con herramientas de interpretabilidad como SHAP. Sus resultados se comparan con LightGBM y Random Forest Regressor con la finalidad de contar con modelos de referencia y poder así evaluar las diferencias del desempeño predictivo.

La evaluación se realizó mediante el coeficiente de determinación R^2 y la raíz del error cuadrático medio (RMSE), ambos evaluados sobre el conjunto de test temporal 2023-2025.

4. Recuperación de Precedentes Históricos (Aprendizaje Basado en Instancias)

Se emplea K-Nearest Neighbors (KNN) con distancia euclidiana normalizada sobre cuatro variables mensuales (caudal ponderado, ONI, precipitación en embalses y demanda). El algoritmo recupera los cinco meses históricos semejantes a un mes de consulta, sin construir un modelo predictivo explícito. No se compara contra alternativas porque el propósito no es maximizar precisión predictiva sino permitir el análisis de escenarios por analogía histórica, un rol que KNN cumple de forma directa e interpretable.

División temporal en la comparación: Para los modelos de regresión, las tres opciones se evalúan con la misma división temporal 2009-2022 para entrenamiento y 2023-2025 para test, sin recurrir a validación cruzada con barajado. Esta restricción evita que los resultados se vean inflados por información futura filtrada al entrenamiento y asegura que la comparación represente el escenario real de uso del sistema. Para clustering y

clasificación se aplica el criterio estándar de cada técnica, dado que la estructura del problema en esos casos admite barajado de las observaciones.

Capítulo 4

4. Resultados

Este capítulo presenta los resultados de la implementación descrita en el Capítulo 3, y cierra con una validación técnica de la aplicación desplegada, no solo de los modelos en el notebook, sino del sistema tal como lo usaría un tomador de decisiones. El análisis parte de un dataset diario de 6.028 observaciones (julio 2009–diciembre 2025), agregado también a nivel mensual (198 meses).

4.1. Descripción final del conjunto de datos

El dataset base, resultado de la integración descrita en el Capítulo 3 (CENACE + GloFAS + NOAA + ERA5 + INEC), tiene 6.028 filas y 16 columnas: fecha, tipo de día, demanda, generación desagregada por fuente (hidráulica y térmica total), totales de generación/importación/exportación/pérdidas, los cuatro indicadores de estado (fracción hidro, dependencia de importación, excedente de exportación, caudal ponderado), las variables exógenas (ONI, precipitación en embalses, población) y el flag de racionamiento oficial. Confirmando la ausencia de valores nulos, registros de fechas duplicadas y sin discontinuidades temporales.

También se realizó una versión extendida de 43 columnas en total, la idea de agregar estas variables derivadas es poder utilizarlas en el tablero de visualización, las cuales se detallan a continuación: codificación cíclica de calendario (seno/coseno de día de semana y día del año), clima diario detallado (temperatura, humedad, precipitación, viento), caudales individualizados por represa (Paute, Coca, Pastaza, Napo) y un indicador de

período COVID. Las 16 columnas del archivo base están contenidas dentro de esas 43 columnas ampliadas, con los mismos valores. Por lo tanto no hay una segunda fuente de verdad solo una extensión posterior para fines de visualización.

4.2. Pruebas de Concepto

El primer paso para probar el diseño es verificar que la arquitectura completa funcione, el flujo que se sigue es el siguiente: Se aplica K-means sobre los 198 meses del conjunto de datos y entrega grupos con centroides bien delimitados; Después se toman los grupos como etiquetas asociadas a las variables exógenas para entrenar un Random Forest; los cuatro regresores de XGBoost se ajustaron sobre la participación temporal de datos diarios y generaron predicciones; se calcula la contribución por variable mediante SHAP; y KNN recuperó vecinos válidos para meses de prueba.

4.3. K-Means: regímenes operativos del sistema eléctrico

Para implementar este modelo se realiza una normalización con Z-score a las cuatro variables de estados, se evaluaron valores de k entre 2 y 8, se seleccionó el más adecuado según el índice Calinski-Harabasz que mide la separación de los clusters contra que tan compacto están sus datos internamente. La configuración con la que se entrenó el modelo definitivo (n_init=50, max_iter=1000, random_state=42), seleccionando k=5 el índice es de 113,5.

El modelo final obtuvo un Silhouette Score de 0,314 queda dentro de la banda de estructura débil y un índice Davies-Bouldin de 0,983 está dentro de la separación

razonable. Estos valores son esperables dado que los regímenes de un sistema eléctrico no son categorías discretas sino puntos en un espectro continuo de estrés hídrico.

Tabla 3

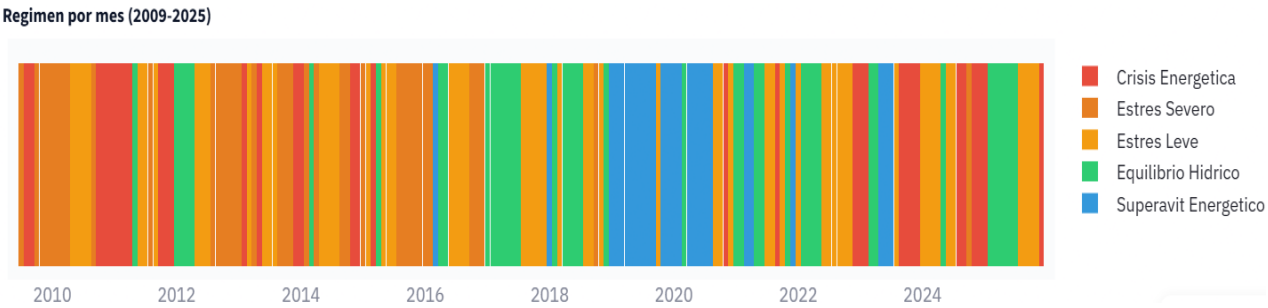
Regímenes operativos identificados mediante K-Means (k=5)

Régimen	Meses	% del total	Hidro promedio	Importación promedio	Caudal promedio
Superávit Energético	26	13%	89%	1,2 GWh/mes	430 m^3/s
Equilibrio Hídrico	41	21%	85%	12,6 GWh/mes	515 m^3/s
Estrés Leve	60	30%	82%	31,6 GWh/mes	326 m^3/s
Estrés Severo	37	19%	60%	35,8 GWh/mes	279 m^3/s
Crisis Energética	34	17%	67%	180,5 GWh/mes	339 m^3/s

Nota. Elaboración propia. Silhouette=0,314; Davies-Bouldin=0,983; Calinski-Harabasz=113,5. Promedios sobre 198 meses (jul. 2009–dic. 2025).

Figura 3

Regímenes operativos del SNI ecuatoriano mes a mes (julio 2009 – diciembre 2025)



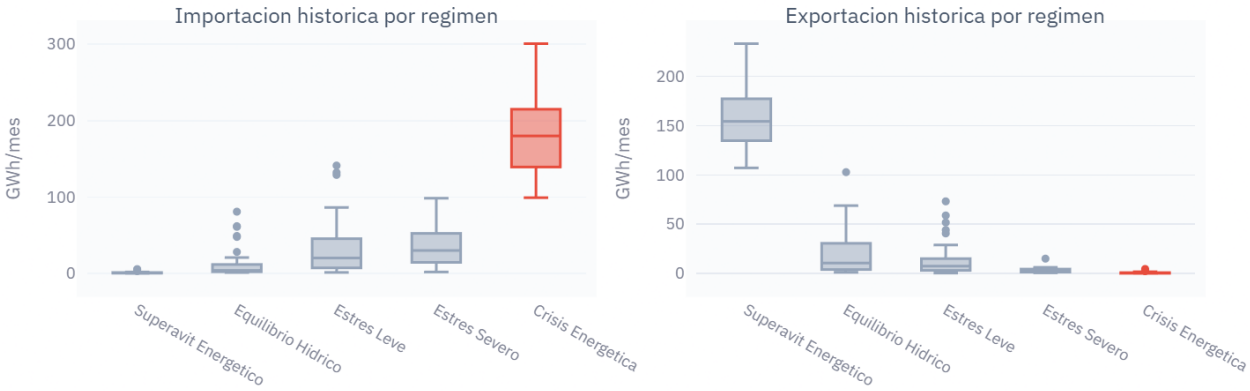
Nota. Elaboración propia. Cada barra corresponde a un mes coloreado según el régimen asignado por K-Means. Los períodos oficiales de racionamiento (noviembre 2009 – febrero

2010 y varios episodios de 2023–2024) coinciden visualmente con las franjas rojas y naranjas, sin que esa información se haya usado en el entrenamiento.

El salto en importación entre Estrés Severo y Crisis Energética (de 35,8 a 180,5 GWh/mes, más de cinco veces) mientras la fracción hidro en realidad sube unos puntos, sugiere que Crisis Energética no es simplemente "menos agua", sino un estado cualitativamente distinto donde el sistema recurre masivamente a la importación aun con condiciones hídricas favorables.

Figura 4

Distribución de importación y exportación mensual por régimen operativo



Nota. Cada caja muestra el rango de valores mensuales de cada régimen. Se ve el salto de importación al pasar a Crisis Energética, y que la exportación se da casi solo cuando hay superávit.

4.3.1. Validación contra racionamientos oficiales

K-Means no recibió ninguna información sobre racionamientos; esa variable no participa del clustering. Al cruzar los 15 meses con racionamiento oficial documentado

(noviembre 2009–febrero 2010, y los tres episodios de 2023-2024) contra los regímenes detectados, 11 de esos 15 meses (73%) cayeron en Crisis Energética o Estrés Severo. Los meses no detectados como graves tenían fracción hidro entre 65% y 89% e importación muy por debajo del centroide de Crisis Energética (180 GWh/mes); en esos casos el racionamiento tuvo carácter preventivo o respondió a limitaciones externas de importación por ejemplo, en octubre 2024 Colombia atravesaba su propia crisis hídrica y restringió las exportaciones al Ecuador (CENACE, 2025), y no al patrón cuantitativo del sistema que define Crisis. El 73% de coincidencia, obtenido sin supervisión, es la evidencia más fuerte de que los regímenes descubiertos capturan algo real.

4.4. Random Forest: ¿explican las variables exógenas los regímenes?

Con los regímenes de K-Means como etiqueta, se entrenó un Random Forest usando solo caudal ponderado, índice ONI, precipitación en embalses y población; variables que un operador conoce sin ver la generación del día. La validación se hizo con StratifiedKFold de 5 pliegues sin mezclar el orden temporal.

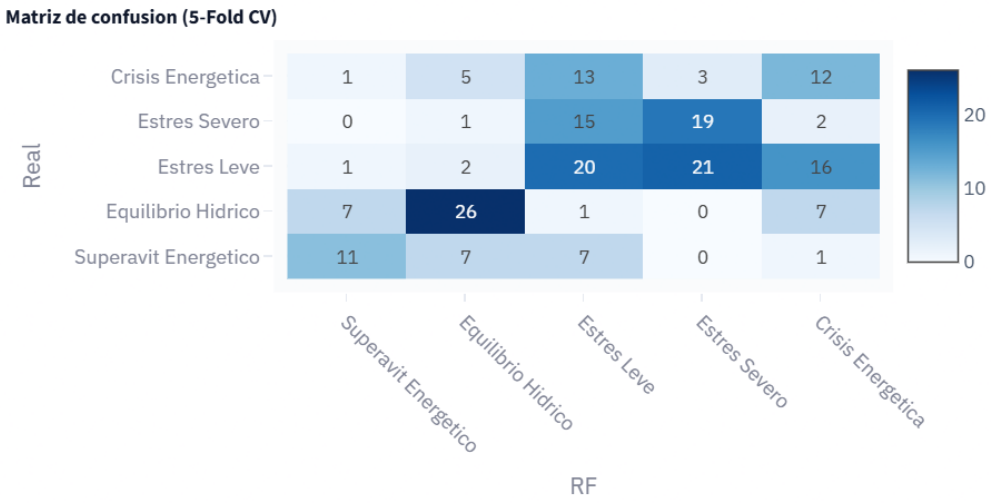
El resultado fue un F1-macro de 0,396, valor que duplica el desempeño esperado del azar (0,225) y multiplica por 4,3 el de una predicción por clase mayoritaria (0,093). Es un desempeño moderado: no significa que el modelo falle, sino que las condiciones físicas externas explican el estado del sistema solo parcialmente. El resto de la varianza probablemente responde a factores que estas cuatro variables no capturan; la entrada de Coca Codo Sinclair en 2016, decisiones operativas de CENACE, acuerdos de intercambio con Colombia y Perú. Un F1 bajo es en sí mismo un hallazgo: el estado del SNI no se reduce a "cuánta agua hay".

La validación se hace sin barajar los meses porque meses consecutivos suelen compartir régimen; barajarlos podría inflar el F1 al mezclar información temporal cercana entre pliegues.

La matriz de confusión muestra errores concentrados entre regímenes adyacentes, no entre extremos consistentes con un espectro continuo más que categorías bien separadas.

Figura 5

Matriz de confusión del Random Forest



Nota. Elaboración propia. Matriz obtenida por validación cruzada estratificada de 5 pliegues sobre los 198 meses del dataset (F1-macro global = 0,396). La mayoría de los errores caen entre regímenes contiguos, no entre extremos.

4.5. XGBoost + SHAP: modelado explicativo diario

Los valores SHAP que a continuación se presentan descomponen cada predicción del modelo en la contribución que cada variable aporta a dicha predicción. Esta contribución sirve como descriptor del peso que tiene dentro del modelo, y no es una relación causal directa del sistema físico: en este sentido, dos variables pueden aparecer con una alta contribución porque están correlacionadas con la salida durante el periodo de entrenamiento, sin tener mecanismos causales entre ellas. Uno de los casos más claros se observa con la variable población: su alta contribución al modelo de exportación no responde a que más personas sean la causa del incremento en exportaciones, sino a que esta variable marca temporalmente un antes y después que trajo consigo Coca Codo Sinclair en 2016. Bajo esta perspectiva se interpretan las descomposiciones subsiguientes.

Los cuatro modelos se entrenaron sobre 4.932 días (jul. 2009–dic. 2022) y se evaluaron sobre 1.096 días de prueba (ene. 2023–dic. 2025), respetando el orden cronológico.

Tabla 4

Desempeño de los modelos XGBoost por flujo energético

Modelo	R^2 train	R^2 test	MAE test
Generación térmica	0,989	0,947	1,14 GWh
Generación hidroeléctrica	0,980	0,680	7,11 GWh
Importación	0,860	0,384	2,52 GWh
Exportación	0,737	0,216	0,81 GWh

Nota. Elaboración propia a partir de metadata_diario.joblib. Entrenamiento: 4.932 días (jul. 2009–dic. 2022); prueba: 1.096 días (ene. 2023–dic. 2025). MAE en GWh.

El patrón es decreciente: cuanto más "físico" es el fenómeno, mejor lo explica el modelo. La generación térmica responde casi mecánicamente a la falta de hidro (R^2=0,947), y SHAP confirma que la fracción hidroeléctrica es, por lejos, el factor dominante. La hidro (R^2=0,680) depende de caudal y demanda observados, así que se explica razonablemente bien, aunque no del todo.

En generación térmica el gap entre R^2 de entrenamiento y prueba es casi nulo; en hidro es amplio (0,30 puntos), aunque menor que en importación y exportación, donde alcanza valores superiores a 0,45. Ese margen refleja en parte un sobreajuste residual pese a la regularización que XGBoost aplica por diseño, pero sobre todo un techo estructural: las variables físicas disponibles no cubren la parte comercial y operativa del intercambio energético con Colombia y Perú, componente que este trabajo delimita como fuera de alcance. La prueba complementaria con rezagos temporales de la sección siguiente confirma esta lectura al recuperar buena parte del R^2 de importación sin agregar variables no físicas.

4.6. Explicación de importación y exportación

SHAP muestra que la importación sube cuando la fracción hidro es baja y la demanda alta, lo que tiene sentido físico claro. El modelo de exportación es más difícil de interpretar: la variable población aparece como la más influyente no porque más personas causen más exportaciones, sino porque actúa como proxy temporal, distinguiendo la era antes y después de 2016.

Figura 6

Contribución SHAP de cada variable a la importación diaria; ejemplo del 15 de octubre de 2024

Detalle por variable

Variable	Valor del día	Promedio histórico	Estado	Contribución SHAP	Efecto
Poblacion	17771857.00	16449052.79	↑ +8% vs promedio	+2.86 GWh	aumentó
Fracc. hidro	0.65	0.77	↓ -16% vs promedio	+0.88 GWh	aumentó
Caudal	182.74	371.89	↓ -51% vs promedio	+0.27 GWh	aumentó
ONI	-0.21	0.06	↓ -456% vs promedio	-0.24 GWh	redujo
Demanda	58.51	60.73	↓ -4% vs promedio	+0.09 GWh	aumentó

Factor principal: **Poblacion** incremento en 2.86 GWh.

Nota. Elaboración propia. Valores SHAP para la importación de un día del racionamiento de 2024. Las barras rojas suman al valor final, las azules restan. Población aparece primero, aunque no es una relación causal directa: actúa como marca temporal del cambio estructural que trajo Coca Codo Sinclair en 2016. La segunda variable, fracción hidroeléctrica, sí tiene lectura física. Para este día específico el modelo predice una importación mayor que la observada (5,26 GWh frente a 0,12 GWh reales). La diferencia es coherente con lo documentado en el capítulo: octubre 2024 coincidió con la crisis hídrica

de Colombia, país desde el cual el SNI recibe la mayor parte de sus importaciones, y esa restricción externa de origen comercial y no físico queda fuera del alcance de las variables del modelo (CENACE, 2025).

Impacto de pct_hidro sobre el modelo de importación. La variable pct_hidro se calcula como el cociente entre la generación hidráulica y la generación total. Dado que la generación total participa del balance del sistema junto con la importación, pct_hidro arrastra información indirecta de la propia variable objetivo del modelo de importación. Este es el riesgo de fuga que se documenta en la selección de variables.

Para saber cuánto pesa realmente ese sesgo, se entrenó una versión de control del modelo de importación quitando pct_hidro y dejando el resto de las variables igual. El modelo principal, con pct_hidro, alcanza un R^2 de test de 0,384. Al retirar la variable, el R^2 cae a 0,315. La diferencia es de 0,069 puntos: es real, pero no afecta significativamente el resultado. Se aplicó el mismo procedimiento al modelo de exportación: el R^2 de test cambia de 0,216 a 0,237 al remover pct_hidro (aquí el cambio es marginal e incluso ligeramente positivo, dado el bajo desempeño base del modelo).

Se acepta la limitación: el R^2 reportado en la Tabla 4 para importación está inflado en 0,069 puntos por el efecto de pct_hidro. Los resultados cuantitativos deben leerse con esa salvedad. Los resultados cualitativos quiénes son los factores explicativos y en qué orden no dependen de esa variable, como confirma el análisis SHAP más adelante.

Prueba complementaria con variables de rezago. Sobre los modelos de control sin pct_hidro presentados arriba (R^2 importación = 0,315; R^2 exportación = 0,237), se probó si el

bajo desempeño se debía a que XGBoost trata cada día como independiente. Al agregar rezagos de 1, 7 y 30 días junto con promedios móviles:

- Importación: el R^2 test subió de 0,315 a 0,649 (+0,334), lo que indica que la importación es mayormente explicable a partir de condiciones físicas acumuladas de una a tres semanas.

- Exportación: el R^2 test subió de 0,237 a 0,346 (+0,109), mejora modesta. Buena parte de lo que caracteriza la exportación no está en los datos físicos disponibles; responde a un componente comercial y contractual fuera del alcance de este proyecto.

Ese salto en importación es mucho mayor que el aporte de *pct_hidro* y muestra que la mayor parte de lo que ayuda a explicar la importación no es la variable contaminada, sino la memoria hidrológica del sistema en escalas de una a tres semanas.

Implicación práctica: los modelos físicos pueden anticipar razonablemente cuándo el sistema necesitará importar, pero no cuándo podrá exportar.

En cuanto a la interpretación con SHAP, el ordenamiento de las variables más influyentes no se invierte al quitar *pct_hidro*: el caudal ponderado, la demanda y el índice ONI aparecen entre los primeros lugares tanto en el modelo principal como en el de control. La conclusión conceptual del trabajo (que la importación está fuertemente ligada a las condiciones hidrológicas y a la demanda del sistema), se sostiene con o sin *pct_hidro*.

4.7. KNN: recuperación de precedentes históricos

Sobre las mismas cuatro variables mensuales, con distancia euclidiana normalizada, el modelo recupera los cinco meses históricos más parecidos a un mes de consulta. Para octubre de 2024 (mes de estrés operativo), los vecinos recuperados corresponden en su mayoría a meses de bajo caudal e importación elevada de años anteriores, la métrica agrupa correctamente episodios comparables, aunque estén separados por años.

4.8. El recomendador: validación contra eventos históricos conocidos

El recomendador es el único componente interactivo del sistema desplegado. Mientras el resto del análisis describe la estructura histórica del SNI a nivel agregado, el recomendador responde a una consulta puntual: dada una fecha específica, entrega al usuario el régimen operativo que corresponde a dicho mes, la caracterización estadística del régimen, los meses históricos con condiciones similares y una acción sugerida acompañada de evidencia cuantitativa.

Frente a una fecha, el recomendador emite una severidad en vocabulario operativo (oportunidad, normal, precaución, alerta, crítico), presenta la mediana, los cuartiles y el rango histórico de importación, exportación y generación para meses de condiciones equivalentes, y devuelve los cinco meses históricos con condiciones más parecidas al mes consultado, recuperados por distancia euclidiana normalizada sobre caudal ponderado, ONI, precipitación en embalses, demanda y mes del año.

4.8.1. Diseño de la evaluación

Para valorar la utilidad práctica del recomendador se seleccionaron ocho meses del período 2011-2022 que cubren los cinco regímenes del sistema. Para cada mes se registró el régimen asignado, la severidad emitida, los valores reales de importación y exportación y los promedios correspondientes al régimen según la caracterización estadística. La coherencia se juzgó por magnitud: si el valor real cae dentro del orden de magnitud típico del régimen, la recomendación que el sistema entregaría para ese mes es útil como referencia; si se aparta significativamente, el sistema estaría sugiriendo una lectura incorrecta.

Tabla 5

Evaluación del recomendador sobre ocho meses representativos de los cinco regímenes

Mes	Régimen K-means	Severidad	Importación (real / prom.)	Exportación (real / prom.)	Coherente
2011-11	Crisis Energética	crítico	183,3 / 180,5	0,4 / 0,5	Sí
2013-07	Estrés Leve	precaución	34,3 / 31,6	6,9 / 5,0	Sí
2014-04	Estrés Severo	alerta	30,0 / 35,8	2,1 / 2,0	Sí
2016-01	Estrés Severo	alerta	27,8 / 35,8	2,1 / 2,0	Sí
2019-09	Superávit Energético	oportunidad	0,8 / 1,2	146,5 /155,0	Sí
2020-01	Superávit Energético	oportunidad	1,2 / 1,2	127,7 / 155,0	Sí
2020-12	Estrés Leve	precaución	35,0 / 31,6	3,4 / 5,0	Sí

2022-02	Equilibrio Hídrico	normal	11,4 / 12,6	4,3 /15,0	Sí*

Nota. Elaboración propia. Valores en GWh. El promedio del régimen proviene de la caracterización de la Tabla 3. (*) 2022-02 muestra exportación de 4,3 GWh, por debajo del promedio del régimen Equilibrio Hídrico (15 GWh) pero dentro del rango habitual del clúster; se marca como coherente parcial.

4.8.2. Lectura de los resultados

En los ocho meses la importación real se mantuvo dentro del orden de magnitud esperado por el régimen. Los casos de Crisis y Superávit muestran los mayores contrastes de magnitud entre regímenes (183 GWh frente a 1 GWh de importación) y el recomendador los distingue con claridad. Los regímenes intermedios (Equilibrio, Estrés Leve y Estrés Severo) son más difíciles de separar por importación, pero la caracterización mantiene la coherencia con los valores reales en todos los casos revisados.

La exportación es menos estable. En el mes de Equilibrio Hídrico el valor real quedó por debajo del promedio del régimen, sin apartarse del rango del cluster. Esto es consistente con lo observado en el análisis XGBoost + SHAP: los flujos de exportación dependen en parte de acuerdos comerciales que las variables físicas del sistema no capturan directamente.

4.8.3. Nota metodológica

La severidad que emite el recomendador es una traducción determinista del régimen K-Means, no un cálculo independiente. Por eso esta evaluación no funciona como verificación del clustering (para eso se emplean los criterios internos y la validación externa

contra racionamientos discutidos en secciones previas), sino como medición de si la caracterización estadística que el recomendador entrega al usuario es una referencia útil de magnitud para una fecha específica.

4.9. Comparación contra modelos alternativos

Con el propósito de verificar que los tres modelos escogidos para este trabajo se sostienen empíricamente, cada modelo seleccionado fue entrenado junto a dos alternativas equivalentes sobre el mismo set de datos, misma partición temporal y sin ajustes finos de hiperparámetros. Para el proceso de selección, no solo se considera los niveles de rendimiento, sino criterios de explicabilidad.

4.9.1 Clustering (Regímenes): K-means, DBSCAN y clustering aglomerativo

Se compararon tres algoritmos sobre las cuatro variables de estado normalizadas (pct_hidro, dependencia_importacion, caudal_ponderado, y excedente_exportacion) con k=5.

Tabla 6

Comparación de algoritmos de clustering

Modelo	Silhouette	Davies-Bouldin	Calinski-Harabasz	Clústeres
K-Means (k=5)	0,314	0,983	113,5	5
DBSCAN (eps=0,4)	0,305	0,996	59,1	6
Agglomerative Ward (k=5)	0,244	1,159	86,3	5

Nota. Silhouette y Calinski-Harabasz: mayor es mejor. Davies-Bouldin: menor es mejor. Evaluados sobre 198 meses agregados del período 2009-2025.

K-Means destaca dentro de las tres métricas mientras que DBSCAN queda cerca de Silhouette sin embargo esta primera divide en seis grupos y su Calinski-Harabasz cae a la mitad. El clustering aglomerativo (método de Ward) produce grupos con menor separación. K-means se sostiene como la mejor selección para este problema, además de ofrecer centroides interpretables que se traducen directamente en la caracterización de los cinco regímenes reportados en este capítulo.

4.9.2 Clasificación: Random Forest, SVM y regresión logística multinomial

Los tres clasificadores se evaluaron con las mismas cuatro variables exógenas físicas (caudal ponderado, ONI, precipitación en embalses y población) sobre el conjunto de regímenes de K-Means, con validación cruzada estratificada con cinco subconjuntos.

Tabla 7

Comparación de clasificadores para los regímenes

Modelo	F1-macro (media)	F1-macro (std)
Random Forest (n=200)	0,557	0,120
SVM (RBF)	0,518	0,043
Regresión logística multinomial	0,472	0,082

Nota. Elaboración propia. Validación cruzada estratificada de 5 pliegues con barajado.

Random Forest obtiene el F1-macro más alto. SVM se ubica segundo con menor variabilidad entre pliegues; la regresión logística queda por debajo, coherente con la naturaleza no lineal del problema. Además del rendimiento, RF ofrece medidas de importancia de variables directamente calculables, lo que resulta clave para interpretar

cuáles condiciones exógenas explican los regímenes, análisis reportado antes en este capítulo.

Es importante hacer una aclaración metodológica. Esta comparación usa validación con barajado para que los tres modelos se enfrenten en igualdad de condiciones. El F1-macro de 0,396 reportado antes para el clasificador principal corresponde a una validación temporal sin barajado, más conservadora frente a la autocorrelación mensual entre regímenes. Los dos números responden a preguntas distintas: la comparación entre modelos y el desempeño del modelo elegido bajo la validación más exigente.

4.9.3 Regresión: XGBoost, LightGBM y Random Forest Regressor

Los tres regresores se entrenaron sobre las siete variables predictoras del modelo principal con la misma partición temporal (entrenamiento 2009-2022, prueba 2023-2025) para los cuatro flujos energéticos del sistema.

Tabla 8

Comparación de regresores por flujo energético

Flujo (variable objetivo)	XGBoost R^2	LightGBM R^2	RF Regressor R^2	Ganador
Importación	0,318	0,274	0,254	XGBoost
Exportación	0,067	0,204	0,115	LightGBM
Generación hidroeléctrica	0,648	0,630	0,714	RF Regressor
Generación térmica	0,911	0,923	0,902	LightGBM

Nota. Elaboración propia. R^2 sobre el conjunto de prueba (1.096 días). Los valores difieren de la Tabla 4 porque esta comparación usa hiperparámetros por defecto para las tres alternativas, sin ajustes finos, con el fin de mantener condiciones equivalentes entre modelos.

XGBoost supera claramente en importación, que es el flujo central del análisis por ser el más sensible al estado del sistema y el que activa las señales de crisis. En los otros tres flujos, los tres modelos se mueven en un rango estrecho: LightGBM y Random Forest Regressor superan a XGBoost por márgenes de 0,01 a 0,06 puntos de R^2. Ninguno domina en todos los flujos.

La elección de XGBoost como modelo principal no se sostiene solo en la ventaja en importación. La razón central es la compatibilidad con TreeSHAP, que calcula valores de Shapley exactos sobre árboles de decisión de forma eficiente y permite las descomposiciones día a día presentadas antes en este capítulo. LightGBM ofrece una implementación SHAP similar pero menos madura para la interpretación local que este trabajo requiere; Random Forest Regressor no ofrece SHAP nativo. Los tres modelos son competitivos en precisión, pero solo XGBoost habilita el nivel de explicabilidad que da sentido a las conclusiones del capítulo.

En clustering la elección se sostiene en métricas puras; en clasificación y en regresión la elección se apoya en un criterio adicional al rendimiento: la posibilidad de extraer explicaciones interpretables (importancia de variables en Random Forest, valores SHAP en XGBoost). Ese criterio es coherente con el enfoque del trabajo, que busca explicar el comportamiento del SNI y no solo predecirlo.

4.10. Validación técnica de la aplicación desplegada

Además de validar los modelos, se probó la aplicación (sni-ia-dashboard) como pieza de software ejecutándola con el framework oficial de pruebas de Streamlit (AppTest) en un entorno limpio con las versiones exactas fijadas en requirements.txt.

4.10.1. Módulos operativos verificados

Las secciones Evaluación y Sistema Inteligente cargan y ejecutan sin errores. Las cifras que la interfaz muestra (F1=0,396, los regímenes, las métricas de XGBoost) se recalcularon de forma independiente contra los archivos .joblib y coinciden con lo reportado en las secciones anteriores.

4.10.2. Incidencia detectada en el módulo de recomendación

Este problema ha sido identificado en 2 de los 8 casos en los que se ha realizado la ejecución controlada del Módulo Recomendador. Este problema existe debido a la falta de compatibilidad entre las bibliotecas dentro del entorno de ejecución y no se debe a ningún error en la lógica del programa. Se recomienda que la ejecución del módulo se valide en el entorno de despliegue después de cualquier actualización de las dependencias.

4.10.3. Lectura conjunta de los resultados

Tomados en conjunto, estos resultados dibujan una jerarquía clara de qué tan predecible es cada parte del sistema eléctrico ecuatoriano. Lo puramente físico se explica bien con las variables disponibles; lo que involucra negociación entre países y decisiones operativas se explica solo parcialmente, y esa limitación es un reflejo honesto de que el

sistema energético ecuatoriano tiene una capa contractual y política que ningún dataset climático va a poder sustituir.

Con el 73% de coincidencia entre los regímenes de K-Means y los racionamientos oficiales, junto con la evaluación del recomendador sobre los ocho meses de la sección anterior, la arquitectura de cinco componentes cumple lo propuesto en el Capítulo 1. La validación técnica de la aplicación confirma además que esos resultados no quedan encerrados en el proceso de desarrollo: son consultables a través de una interfaz, con las salvedades de robustez señaladas en las secciones anteriores, que quedan documentadas para su corrección antes de un despliegue en producción.

Capítulo 5

5. Conclusiones y Recomendaciones

5.1. Conclusiones

El objetivo general de este trabajo fue analizar los factores que influyen en la generación e intercambio energético del SNI usando aprendizaje automático sobre los datos históricos de 2009-2025. También con los resultados del capítulo 4 se comprueba, uno por uno, si se cumplieron los objetivos específicos y la hipótesis planteada al inicio, y en qué medida.

Se realizó la construcción del dataset en donde se integró una base diaria (julio 2009-diciembre 2025) unificando variables operativas de CENACE con caudales de GloFAS, el índice ONI de NOAA, precipitación de ERA5 y proyecciones poblacionales de INEC, sin valores faltantes ni discontinuidades. Entonces esto comprueba que es posible reunir fuentes públicas dispersas del sector eléctrico ecuatoriano en un solo conjunto analíticamente útil.

El algoritmo K-Means sin ningún conocimiento de los racionamientos identifico cinco regímenes operativos y la mayoría de los meses con racionamiento oficial pertenecientes a los años 2009 - 2010 y 2023 - 2024 fueron categorizados como Crisis Energética o Estrés Severo. Así esta coincidencia obtenida sin supervisión demuestra la primera parte de la hipótesis de que el sistema sí tiene regímenes operativos recurrentes detectables desde sus propios datos.

El clasificador Random Forest usando solamente las variables caudal, ONI, precipitación y población tuvo un desempeño moderado al clasificar los regímenes de K-Means. Por lo tanto, las condiciones físicas explican una parte estado del sistema, el comportamiento del SNI depende también de decisiones operativas y de la infraestructura instalada en cada momento, algo que ninguna variable hidrológica captura por sí sola.

La generación hidroeléctrica, la generación térmica, la importación y la exportación podrían modelarse con distintos niveles de éxito usando modelos de regresión y SHAP permitió dividir esas predicciones variable por variable. Además, la importación puede describirse en buena medida a través de condiciones físicas acumuladas mientras que la exportación tiene un componente comercial y contractual que no se encuentran en los datos.

En conjunto fue posible cuantificar el peso de cada factor externo sobre el sistema, pero ese peso no es constante en todos los casos, es alto en térmica e hidroeléctrica, moderado en importación, bajo en exportación, donde entran factores comerciales que los datos disponibles no capturan. Este segmento de la hipótesis se valida con matices.

El modelo KNN pudo recuperar los meses más similares a un escenario dado y el recomendador que integra K-Means, la caracterización de los regímenes y los precedentes de KNN mostró coherencia entre la caracterización emitida y los valores reales observados. Esto significa que a partir de patrones observados sin recurrir a pronósticos es posible describir escenarios análogos del sistema.

Cumpliendo el objetivo de traducir el análisis técnico en un formato consultable por un tomador de decisiones sin necesidad de leer código, los regímenes las métricas de cada modelo, los gráficos SHAP y el recomendador se integraron en una interfaz gráfica (sni-ia-dashboard).

En conjunto la arquitectura de cinco componentes incluyendo K-Means, Random Forest, XGBoost con SHAP, KNN y el módulo de recomendación cumplió con lo mencionado en el capítulo 1 para dar un paso de una descripción aislada del comportamiento energético a un marco que identifica regímenes, cuantifica causas y recupera precedentes útiles para la planificación, dentro del alcance definido, delimitado a omitir pronósticos de demanda, estimaciones de generación futura, optimización del despacho económico y modelado individual de centrales o líneas de transmisión.

El trabajo confirma los tres componentes de la hipótesis dentro de los márgenes que la propia formulación reconoce, donde el SNI presenta regímenes operativos recurrentes identificables desde sus datos históricos. Las variables exógenas (caudal, ONI, precipitación y población) caracterizan esos regímenes con un desempeño moderado, y explican buena parte de los flujos de generación e intercambio, con alta intensidad en la generación térmica e hidroeléctrica, moderada en la importación y limitada en la exportación. Lo que queda sin explicar corresponde a decisiones operativas y factores contractuales que quedan fuera de las fuentes públicas disponibles, coincidiendo con la salvedad "aunque no lo cubren en su totalidad" ya prevista en la hipótesis.

5.2. Limitaciones

Existen limitaciones metodológicas y técnicas del proyecto, así como límites en la disponibilidad de los datos los cuales deberían ser expresados.

En el recomendador se realizó una comparación cualitativa de ocho meses de datos históricos con la caracterización de su régimen. No existe una métrica que mida en números qué tan cerca estuvo la recomendación de la decisión que realmente tomó el operador, así que su desempeño no se puede medir con el mismo rigor que el de los modelos de XGBoost.

El algoritmo Random Forest que intenta reconocer el régimen del sistema basándose únicamente en variables externas se entrenó con cuatro datos físicos que incluyen caudal, índice ONI, precipitación y población. El desempeño de este algoritmo fue moderado ya que le faltó información acerca de la disponibilidad de las plantas y de las operaciones mensuales.

El análisis se mantuvo a nivel nacional agregado, sin variables específicas de las centrales de mayor capacidad como Coca Codo Sinclair y el complejo hidroeléctrico integrado por las centrales Paute, Mazar y Sopladora, ni datos de restricciones de transmisión. Asimismo, no se contó con variables comerciales de exportación tales como precios spot regionales, contratos bilaterales.

5.3. Recomendaciones

Los organismos de planificación energética pueden beneficiarse de este hallazgo que existe una coincidencia del 73% entre los regímenes no supervisados y racionamientos oficiales. También implica que los controles mensuales del régimen operativo que se está utilizando pueden indicar las señales de estrés incluso antes de que aparezcan como racionamiento. Además, los valores SHAP al ser introducidos en los comités técnicos pueden indicar con precisión bajo qué circunstancias se vuelve necesario recurrir a la costosa energía térmica y las importaciones.

La validación contra ocho eventos históricos hecha en el recomendador fue cualitativa. En estudios posteriores la validación se puede hacer en un periodo más extenso de meses con una métrica cuantitativa como la distancia entre la recomendación y la acción realmente tomada por el operador como una medida del desempeño similar a como se validó el desempeño del modelo XGBoost.

Los modelos ya han sido serializados usando joblib y el tablero está accediendo directamente a esos artefactos serializados en vez de entrenarlos de nuevo y por lo tanto el proceso de actualización de la línea de producción puede ser automatizado ya que CENACE actualiza sus datos diariamente. El proceso puede ser automatizado junto con el entrenamiento periódico de los modelos y la versión de los modelos para hacer de este proyecto una herramienta de monitoreo continuo.

Como línea de investigación futura, y fuera del alcance definido para este trabajo, valdría la pena explorar si los regímenes identificados aquí pueden servir como insumo para

un modelo de pronóstico de corto plazo (no reemplazando el enfoque explicativo desarrollado, sino como una capa adicional que combine la comprensión causal lograda en este proyecto con capacidad predictiva), algo que la literatura revisada en el Capítulo 2 trata de forma separada y que rara vez se integra en un mismo sistema para el caso ecuatoriano.

Referencias

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications, 7*(1), 39-59. https://content.iospress.com/articles/ai-communications/aic7-1-04

Agencia de Regulación y Control de Energía y Recursos Naturales No Renovables [ARCERNNR]. (2020). *Regulación de la infraestructura y operación del Sistema Nacional Interconectado.* ARCERNNR.

Aksan, F., Jasiński, M., Sikorski, T., Kaczorowska, D., Rezmer, J., Suresh, V., Leonowicz, Z., Kostyła, P., Szymańda, J., & Janik, P. (2021). Clustering methods for power quality measurements in virtual power plant. *Energies, 14*(18), 5902. https://doi.org/10.3390/en14185902

Alsaigh, R., Mehmood, R., & Katib, I. (2023). AI explainability and governance in smart energy systems: A review. *Frontiers in Energy Research, 11*(1071291). https://doi.org/10.3389/fenrg.2023.1071291

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1-27. https://doi.org/10.1080/03610927408827101

CENACE. (2025). *Rendición de cuentas 2024.* Retrieved 15 de 8 de 2026, from https://www.cenace.gob.ec/wp-content/uploads/downloads/2025/07/Informe-de-rendicion-de-cuentas-2024_CENACE_final-rev.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (págs. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27. https://ieeexplore.ieee.org/document/1053964

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224-227. https://doi.org/10.1109/TPAMI.1979.4766909

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. En *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (págs. 226-231). https://ojs.aaai.org/index.php/KDD/article/view/18198

Gallo Cruz, A. O. (2020). *Análisis predictivo para minería de datos y proyección a corto plazo de la demanda de potencia en el sistema eléctrico ecuatoriano.* [Trabajo de titulación, Escuela Politécnica Nacional], Repositorio Institucional EPN. https://bibdigital.epn.edu.ec/handle/15000/21303

Guamán , W., Benalcazar, P., Cordova Garcia, J., & Torres, M. (2025). Machine learning-based projections of long-term electricity consumption: The case study of Ecuador. En *Advanced research in technologies, information, innovation and sustainability (ARTIIS 2024).* Springer. https://doi.org/10.1007/978-3-031-83432-5_12

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., & Pappenberger, F. (2020). GloFAS-ERA5 operational global river discharge reanalysis 1979–present. *Earth System Science Data*. https://doi.org/https://doi.org/10.5194/essd-12-2043-2020

Hersbach, H., Bell, B., Berrisford, P., & al et. (2020). The ERA5 global reanalysis. *Q J R Meteorol Soc*. https://doi.org/https://doi.org/10.1002/qj.3803

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., & Zareipour, H. (2020). Energy forecasting: A review and outlook. *Open Access Journal of Power and Energy, XX*, 1-13. https://doi.org/10.1109/OAJPE.2020.3007915

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2.ª ed.). OTexts. https://otexts.com/fpp2/

Icaza, D., Jurado, F., Flores, C., & Reivan, G. (2023). Ecuadorian electrical system: Current status, renewable energy and projections. *Heliyon*. https://doi.org/10.1016/j.heliyon.2023.e16010

Instituto Nacional de Estadística y Censos. . (2024). *Ecuador en Cifras*. Proyecciones poblacionales del Ecuador 2010-2050.: https://www.ecuadorencifras.gob.ec/proyecciones- poblacionales/

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer. https://link.springer.com/book/10.1007/978-1-4614-6849-3

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (págs. 4765–4774). Curran Associates, Inc. https://arxiv.org/pdf/1705.07874

Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable artificial intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI, 9*(100169). https://www.sciencedirect.com/science/article/pii/S2666546822000246?via%3Dihub

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, págs. 281–297). University of California Press. https://digicoll.lib.berkeley.edu/record/113015/files/5_1_28.pdf

Martínez Garófalo, S. A. (2025). *Pronóstico de la demanda de energía eléctrica aplicando técnicas de Machine Learning para sistemas de gestión en el sector residencial.* [Trabajo de titulación, Universidad Politécnica Salesiana], Repositorio Institucional UPS. https://dspace.ups.edu.ec/bitstream/123456789/31508/1/TTS2348.pdf

Miraftabzadeh, S. M., Colombo, C. G., Longo, M., & Foiadelli, F. (2023). K-means and alternative clustering methods in modern power systems. *IEEE Access, 11*, 119596-119633. https://doi.org/10.1109/ACCESS.2023.3327640

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable (versión del 21 de febrero de 2019)* (1.ª ed.). Leanpub. https://christophm.github.io/interpretable-ml-book/

Naranjo Silva, S., Punina Guerrero , D. J., Jacome Dominguez, E. A., Escobar Segovia , K., & Laverde Albarracín, C. (2026). Energy planning under climate pressure in Ecuador: Insights from the 2023–2024 crisis using LEAP modeling. *Sustainability, 18*(4), 2112. https://doi.org/10.3390/su18042112

National Oceanic and Atmospheric Administration [NOAA]. (2025). *Oceanic Niño Index*. Climate Prediction Center: https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/enso/oni/v6/

Peña, D., Aguila Téllez, A., & Jurado, F. (2025). Reliability assessment of Ecuador's power system: Metrics, vulnerabilities, and strategic perspectives. *Energies, 18*(12), 3059. https://www.mdpi.com/1996-1073/18/12/3059

República del Ecuador. (2015). Ley Orgánica del Servicio Público de Energía Eléctrica [LOSPEE]. Registro Oficial Suplemento 418.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4.ª ed.). Pearson.

Sánchez Solís, J., & Coto Jiménez, M. (2022). Estado del arte de la predicción de variables en sistemas de ingeniería eléctrica basada en inteligencia artificial. *E-Ciencias De La Información, 12*(1). https://doi.org/10.15517/eci.v12i1.47628

Shapley, L. S. (1953). A value for n-person games. En *Contributions to the Theory of Games (AM-28), Volume II* (págs. 307-318). Princeton University Press.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika, 18*(4), 267-276.

Vela, M. d. (septiembre de 2024). La crisis del sector eléctrico: Vino para quedarse. *Koyuntura*(108), págs. 1-25. https://www.usfq.edu.ec/sites/default/files/2024-09/koyuntura-108.pdf

Wang, Y., Chen, Q., Hong, T., & Kang, C. (2019). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*. https://doi.org/10.1109/TSG.2018.2805

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244.