UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA DE POSGRADO

DESARROLLO Y EVALUACIÓN DE MODELOS DE MACHINE LEARNING EN AMBIENTES DE COMPUTACIÓN EN LA NUBE ENFOCADOS A LA GENERACIÓN HIDROELÉCTRICA

PRESENTADO POR:

JOSÉ ERNESTO DE PAZ ÁLVAREZ

SALVADOR AMARU FLORES FLORES

PARA OPTAR AL TÍTULO DE:

MAESTRO EN INGENIERÍA PARA LA INDUSTRIA CON ESPECIALIZACIÓN EN TELECOMUNICACIONES.

CIUDAD UNIVERSITARIA, FEBRERO DE 2026.

UNIVERSIDAD DE EL SALVADOR

RECTOR:

MSc. JUAN ROSA QUINTANILLA

SECRETARIO GENERAL:

LCDO. PEDRO ROSALÍO ESCOBAR CASTANEDA

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

MSc. LUIS SALVADOR BARRERA MANCÍA

SECRETARIO:

ARQ. RAÚL ALEXANDER FABIÁN ORELLANA

ESCUELA DE POSGRADO

DIRECTOR:

MSc. ELMER ARTURO CARBALLO RUÍZ

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA DE POSGRADO

Trabajo de Graduación previo a la opción al Grado de:

MAESTRO EN INGENIERÍA PARA LA INDUSTRIA CON ESPECIALIZACIÓN EN TELECOMUNICACIONES.

Título:

DESARROLLO Y EVALUACIÓN DE MODELOS DE MACHINE LEARNING EN AMBIENTES DE COMPUTACIÓN EN LA NUBE ENFOCADOS A LA GENERACIÓN HIDROELÉCTRICA

Presentado por:

JOSÉ ERNESTO DE PAZ ÁLVAREZ

SALVADOR AMARU FLORES FLORES

Trabajo de Graduación Aprobado por:

Docente Asesor(a):

DR. CARLOS EUGENIO MARTÍNEZ CRUZ

SAN SALVADOR, FEBRERO DE 2026

Trabajo de Graduación Aprobado por:

Docente Asesor:

DR. CARLOS EUGENIO MARTÍNEZ CRUZ

Agradecimiento 1.

Agradezco profundamente a mi esposa y a mis padres que me han brindado apoyo incondicional en todo este proceso. Mi padre es un hombre luchador que me ha animado y ha mantenido en alto mi autoestima desde que tengo memoria, hasta el día de la presentación de esta investigación y su entrega me hace saber que estará siempre conmigo. Mi madre es una devota cristiana que ha dedicado sus oraciones por mí, todo el tiempo y su apoyo hasta este día ha sido incondicional en todo aspecto. Mi esposa ha sido mi apoyo fundamental desde el inicio de este proceso y está allí presente en cada iniciativa que se emprende. Estoy eternamente agradecido con ellos y este logro les pertenece.

Agradezco profundamente al Doctor Carlos Martínez por asesorar a este equipo en este trabajo de graduación y por su paciencia en los momentos difíciles de la realización del mismo y a mi compañero Salvador Flores por acompañarme en el desarrollo de esta investigación.

Así también agradezco al Ingeniero Rodolfo de Jesús Caceres en su cargo como gerente de producción de la Comisión Ejecutiva Hidroeléctrica del Río Lempa, quien posteriormente a la defensa de este trabajo de graduación se reunió con nosotros para para ofrecernos su apoyo e invitarnos a conocer las instalaciones de la represa 5 de noviembre en una visita técnica.

Agradezco también a los profesionales que laboran en la central hidroeléctrica 5 de noviembre que nos recibieron con mucha calidez, apertura y entusiasmo en la visita técnica desarrollada.

A todos mi más profunda gratitud.

José Ernesto De Paz Álvarez

Agradecimiento 2.

Incapaz de expresar plenamente mi gratitud hacia todas las personas que hicieron este camino posible, manifiesto mi más profundo agradecimiento a mi familia. A mi madre, Adela; a mi hermana, María; y a mi hermano, José, por ser la roca que soporta mis esfuerzos. De manera especial, dedico este logro a la memoria de mi querido padre, quien, aunque ya no nos acompaña físicamente, sigue siendo una de mis más grandes inspiraciones y el motor de mis sueños.

A mi compañero de maestría, José de Paz, por compartir las cargas de este proceso.

Y a mis amigos, por ser el refugio a las incontables tempestades que se manifiestan en la vida

Salvador Amaru Flores Flores

Índice general

Capítulo 1: Introducción

1.1. Abstract

The 5 de Noviembre Hydroelectric Power Plant is a strategic asset in El Salvador's energy matrix; however, the absence of disaggregated public data on the individual operation of its seven generating units prevents a detailed analysis of its performance and dispatch strategy. This research addresses this limitation through the development and application of Machine Learning models to estimate energy injection and characterize the functional role of each unit.

Using total plant injection data for the 2021-2023 period and a set of disaggregated data from 2024 as a training baseline, two complementary approaches were implemented in a cloud computing environment. The first, a multi-class classification model (Random Forest), identified the combination of active units in each hourly interval with high precision (average accuracy of 86.96%). The second, a Multiple Linear Regression model, estimated the continuous power injection (MW) per unit, achieving exceptional fidelity when compared to real data (coefficient of determination, R2, of 0.9998).

The analysis of the results made it possible to calculate the individual plant factors for the first time, revealing a clear operational hierarchy with base-load (U1, U3), secondary (U2, U5), and peak/reserve units (U4, U6, U7). A key finding is the reassessment of the performance of the expansion units (U6 and U7), demonstrating that their low plant factor does not indicate inefficiency but rather confirms their valuable strategic role as peaking generators, aligned with the original objectives of the expansion project.

In conclusion, this study validates a robust and scalable methodology that transforms aggregated energy data into detailed operational intelligence, effectively solving the problem of information scarcity and providing a tool for asset optimization and energy planning.

1.2. Resumen

La Central Hidroeléctrica 5 de Noviembre es un activo estratégico en la matriz energética de El Salvador; sin embargo, la ausencia de datos públicos desagregados sobre la operación individual de sus siete unidades generadoras impide un análisis detallado de su rendimiento y estrategia de despacho. Esta investigación aborda dicha limitación mediante el desarrollo y la aplicación de modelos de Machine Learning para estimar la inyección de energía y caracterizar el rol funcional de cada unidad.

Utilizando datos de inyección total de la planta para el período 2021-2023 y un conjunto de datos desagregados de 2024 como base de entrenamiento, se implementaron dos enfoques complementarios en un entorno de computación en la nube. El primero, un modelo de clasificación multiclase (Random Forest), identificó con alta precisión (exactitud promedio del 86.96%) la combinación de unidades activas en cada intervalo horario. El segundo, un modelo de Regresión Lineal Múltiple, estimó la inyección de potencia continua (MW) por unidad, logrando una fidelidad excepcional al ser comparado con los datos reales (coeficiente de determinación, R2, de 0.9998).

El análisis de los resultados permitió calcular por primera vez los factores de planta individuales, revelando una clara jerarquía operativa con unidades de carga base (U1, U3), secundarias (U2, U5) y de pico/reserva (U4, U6, U7). Un hallazgo clave es la reevaluación del desempeño de las unidades de expansión (U6 y U7), demostrando que su bajo factor de planta no indica ineficiencia, sino que confirma su valioso rol estratégico como generadoras de pico, alineado con los objetivos originales del proyecto de expansión.

En conclusión, este estudio valida una metodología robusta y escalable que transforma datos energéticos agregados en inteligencia operacional detallada, resolviendo eficazmente el problema de la escasez de información y proporcionando una herramienta para la optimización de activos y la planificación energética.

1.3. Introducción

La Central Hidroeléctrica 5 de Noviembre es un pilar fundamental en la matriz energética de El Salvador. Sin embargo, la evaluación precisa de su eficiencia operativa y el análisis de su estrategia de despacho se han visto históricamente limitados por una barrera crítica: la ausencia de datos públicos desagregados sobre la operación individual de sus siete unidades generadoras. Esta carencia de información granular no solo impide el cálculo de métricas de rendimiento clave, como el factor de planta por unidad, sino que también deja en la ambigüedad el verdadero impacto y rol estratégico de la expansión de 80 MW completada en 2016.

La presente tesis aborda este desafío directamente, proponiendo una metodología innovadora que utiliza algoritmos de Machine Learning para transformar datos agregados en inteligencia operacional detallada. Mediante el desarrollo de dos modelos predictivos complementarios —uno de clasificación para identificar qué unidades están activas y otro de regresión para estimar su inyección de potencia—, este estudio reconstruye el comportamiento histórico de cada generador. El enfoque se apoya en un conjunto limitado de datos desagregados del año 2024, utilizados como "verdad de terreno" para entrenar modelos capaces de interpretar los registros de inyección total de la planta desde 2021 hasta 2023.

El resultado es un análisis sin precedentes que no sólo valida la alta precisión de los modelos, sino que revela la lógica operativa subyacente de la central. Por primera vez, se caracterizan cuantitativamente los roles funcionales de cada unidad, diferenciando entre generadoras de carga base, secundarias y de pico. De manera crucial, esta investigación reevalúa el desempeño de las unidades de expansión, demostrando que su operación intermitente no es un signo de bajo rendimiento, sino la confirmación de su éxito en cumplir un rol estratégico como activos de reserva y respuesta rápida, vital para la estabilidad del sistema eléctrico nacional.

A través de esta investigación, se demuestra cómo la inteligencia artificial y la computación en la nube ofrecen una solución escalable y de bajo costo para superar la escasez de datos en infraestructuras críticas, sentando un precedente metodológico para la optimización de activos y la planificación energética en El Salvador.

Este trabajo desarrolla un conjunto de análisis de machine learning utilizando datos que son publicados en los registros públicos de la unidad de transacciones. Dicho ente, publica solo datos totales de generación hidroeléctrica en el país. Con la publicación de [1] se identificó la necesidad de proponer una metodología de desagregación virtual de unidades generadoras usando ML y variables hidrológicas de la central hidroeléctrica 5 de noviembre para evaluar la operación del proyecto de expansión realizado en el año 2016.

1.4. Planteamiento del problema

La Central Hidroeléctrica 5 de Noviembre, con una capacidad instalada de 179.4 MW, se erige como un activo de generación indispensable para la estabilidad y sostenibilidad de la matriz energética de El Salvador. A pesar de su importancia estratégica, existe una barrera fundamental que impide un análisis riguroso de su rendimiento y una comprensión profunda de su estrategia operativa: la indisponibilidad de datos públicos desagregados sobre la operación individual de sus siete unidades generadoras. Los registros oficiales de la Unidad de Transacciones (UT) reportan únicamente la inyección de potencia total de la planta, consolidando la producción de todas las unidades en una sola cifra horaria.

Esta falta de granularidad en los datos crea un significativo "punto ciego analítico" con múltiples consecuencias adversas. En primer lugar, imposibilita el cálculo de indicadores como el factor de planta individual. Sin esta métrica, es inviable evaluar la eficiencia con la que se utiliza cada generador. En este escenario, la planta, en su conjunto, opera como una "caja negra", ocultando la dinámica interna que define su contribución real al sistema eléctrico nacional.

El problema se agudiza al considerar la importante expansión de la central finalizada en 2016, que añadió dos nuevas turbinas (U6 y U7) de 40 MW cada una, con una inversión multimillonaria destinada a incrementar la oferta de energía limpia y desplazar la generación térmica. Un estudio preliminar ha sugerido que la eficiencia operativa de estas nuevas unidades es drásticamente inferior a la proyectada, con un factor de capacidad por debajo del 10%, en lugar del 18% esperado. Esta discrepancia plantea una pregunta crítica y sin respuesta: ¿están estos activos estratégicos siendo subutilizados, o cumplen un rol especializado que no se refleja en las métricas de energía tradicionales? La ausencia de datos específicos impide confirmar, refutar o recontextualizar este hallazgo, dejando en la incertidumbre la rentabilidad y el éxito estratégico de una de las inversiones más importantes del sector en la última década.

Más allá de la evaluación de activos individuales, esta carencia de información genera una ambigüedad sobre la estrategia de despacho global de la central. Se desconoce la jerarquía operativa entre las unidades más antiguas y las nuevas, o cómo se distribuye la carga para satisfacer la demanda y gestionar el recurso hídrico. ¿Operan todas las unidades como generadoras de base, o existe una especialización funcional donde algunas actúan como unidades de pico o de reserva para proveer servicios auxiliares a la red? Esta opacidad operativa limita la capacidad de los entes reguladores y planificadores para modelar con precisión el comportamiento de un activo que representa una porción sustancial de la capacidad hidroeléctrica del país.

En resumen, la escasez de datos desagregados no es una simple limitación técnica, sino un obstáculo fundamental que impide la optimización de recursos, la evaluación de inversiones estratégicas y la planificación energética informada. Se vuelve imperativo, por tanto, desarrollar una metodología capaz de superar esta barrera de información, permitiendo "desagregar" virtualmente la operación de la planta para revelar el comportamiento individual de sus componentes y, con ello, responder a las interrogantes críticas sobre su eficiencia, estrategia y valor real para el sistema eléctrico de El Salvador.

1.5. Formulación del problema

Dada la crítica falta de datos públicos desagregados que impide un análisis operativo detallado de la Central Hidroeléctrica 5 de Noviembre, la presente investigación se centra en responder la siguiente pregunta fundamental:

> ¿Es factible desarrollar y validar un modelo basado en Machine Learning que, utilizando datos de inyección total de la planta y un conjunto limitado de datos de operación por unidad como base de entrenamiento, pueda reconstruir con alta fidelidad el perfil de operación individual (estado y potencia inyectada) de cada una de las siete unidades generadoras para períodos históricos donde esta información no está disponible?

Esta pregunta principal se desglosa en las siguientes sub-preguntas de investigación, que guían el alcance del estudio:

1. ¿Puede una metodología de "desagregación virtual" alcanzar la precisión necesaria para permitir el cálculo confiable de métricas de rendimiento clave, como el factor de planta individual, superando así la principal barrera analítica existente?

2. ¿Permitirá el análisis de los perfiles reconstruidos caracterizar de manera inequívoca el rol estratégico de cada unidad —particularmente las de la expansión (U6 y U7)— y resolver la ambigüedad sobre su supuesto "bajo rendimiento"?

3. ¿Es posible, a través de esta reconstrucción, inferir la lógica de despacho subyacente de la central, identificando la jerarquía y la función especializada (base, secundaria o pico) de cada generador dentro del sistema?

1.6. Justificación

La presente investigación se justifica por su capacidad para resolver un problema de información crítico con implicaciones directas en la gestión operativa, la planificación estratégica y la evaluación de inversiones en el sector energético de El Salvador. La metodología propuesta no es un mero ejercicio académico, sino una solución pragmática y de alto impacto a la falta de transparencia en la operación de la Central Hidroeléctrica 5 de Noviembre, uno de los activos de generación más importantes del país.

Desde una perspectiva práctica y operativa, este estudio es fundamental porque transforma datos agregados, de limitado valor analítico, en inteligencia operacional detallada y accionable. Al reconstruir el perfil de funcionamiento de cada unidad generadora, se habilita por primera vez el cálculo de métricas de rendimiento individuales, como el factor de planta. Este conocimiento es indispensable para optimizar el uso del recurso hídrico, mejorar las estrategias de despacho diario, planificar mantenimientos predictivos basados en el desgaste real de los equipos y, en última instancia, maximizar la eficiencia y la vida útil de la central.

Estratégicamente, la investigación aporta una claridad sin precedentes sobre el valor de la expansión de 80 MW completada en 2016. Al refutar con evidencia empírica la noción de "bajo rendimiento" de las unidades U6 y U7, el estudio redefine su contribución, demostrando su rol crucial como unidades de pico y reserva. Esta reevaluación es vital para justificar la inversión multimillonaria realizada y para informar futuras decisiones sobre la expansión de la capacidad de generación renovable en el país. Proporciona a los tomadores de decisiones y a los planificadores energéticos una base fáctica para entender que el valor de un activo de generación no siempre reside en la energía constante que produce (GWh), sino también en la potencia flexible y la estabilidad que aporta al sistema (MW).

Metodológicamente, el trabajo se justifica por su enfoque innovador al aplicar algoritmos de Machine Learning como una herramienta de "desagregación virtual". Se demuestra que es posible superar las barreras impuestas por la escasez de datos mediante técnicas de ciencia de datos, estableciendo un precedente robusto y validado. La alta fidelidad de los modelos (con coeficientes de determinación superiores a 0.99) valida este enfoque como una alternativa viable y de bajo costo frente a la instalación de complejos sistemas de monitoreo individual.

Finalmente, la **relevancia de esta investigación trasciende a la Central 5 de Noviembre.** La metodología desarrollada es inherentemente escalable y replicable, ofreciendo una plantilla que puede ser adaptada para analizar otras infraestructuras de generación en El Salvador y la región que enfrenten limitaciones de datos similares. Por lo tanto, este estudio no solo resuelve un problema específico, sino que también contribuye con una herramienta poderosa para fortalecer la transparencia, la eficiencia y la planificación informada en todo el sector energético nacional.

1.7. Hipótesis

Se postula que la aplicación de un modelo dual de Machine Learning, entrenado a partir de datos de inyección total y un conjunto limitado de datos desagregados de muestra (año 2024), permitirá reconstruir con un alto grado de precisión el perfil de operación individual (estado y potencia) de cada una de las siete unidades de la Central Hidroeléctrica 5 de Noviembre para el período 2021-2023.

Esta reconstrucción no solo validará la viabilidad de la metodología de "desagregación virtual", sino que también permitirá caracterizar cuantitativamente el rol funcional de cada generador. Se anticipa que este análisis demostrará que la operación de las unidades de expansión (U6 y U7) responde a una lógica de despacho estratégica como unidades de pico, refutando así la noción de bajo rendimiento operativo y confirmando su valor para la flexibilidad y estabilidad del sistema eléctrico nacional.

1.8. Objetivos

1.8.1. Objetivo general

- Simular modelos de Machine Learning en entornos de computación en la nube para la industria 4.0, enfocado a la generación de Energía Hidroeléctrica en El Salvador.

1.8.2. Objetivos específicos

- Realizar un análisis exhaustivo del estado del arte para identificar y sintetizar las técnicas, herramientas y desafíos actuales en la simulación de modelos de Machine Learning aplicados a la computación en la nube, con un enfoque particular en su aplicación para la generación de energía Energía Hidroeléctrica en El Salvador.

- Desarrollar un modelo de Machine Learning personalizado que se adapte específicamente a las necesidades y características de la generación de Energía Hidroeléctrica en El Salvador, considerando variables locales y especificaciones técnicas del sector.

- Implementar y ajustar modelos de Machine Learning en un entorno de computación en la nube para simular y predecir comportamientos en la generación de Energía Hidroeléctrica, utilizando datos históricos y actuales de operaciones energéticas en El Salvador.

- Realizar un análisis comparativo entre los resultados obtenidos de los modelos de Machine Learning y los datos operacionales reales para identificar la precisión, eficacia y áreas de mejora en las simulaciones de generación de Energía Hidroeléctrica.

- Elaborar un conjunto de recomendaciones y directrices basadas en los hallazgos del estudio, orientadas a optimizar la implementación y operación de simulaciones de Machine Learning en la nube para aplicaciones industriales, con un enfoque específico en la industria de generación de energía Energía Hidroeléctrica.

Capítulo 2: Marco Teórico

El presente capítulo establece el marco contextual que sustenta esta investigación, describiendo la evolución y estructura del sistema energético salvadoreño, con especial énfasis en la relevancia de la generación hidroeléctrica dentro de la matriz nacional. Se analizan las políticas públicas, el marco institucional y regulatorio que rigen la producción y distribución de energía en el país, así como el papel histórico de la Comisión Ejecutiva Hidroeléctrica del Río Lempa (CEL) en el desarrollo de la infraestructura eléctrica nacional.

Asimismo, se incorpora una revisión de los fundamentos tecnológicos relacionados con el uso de Machine Learning y computación en la nube, herramientas que constituyen la base técnica del enfoque predictivo adoptado en esta tesis. De esta manera, el capítulo no solo contextualiza el estudio en el marco del sistema energético salvadoreño, sino que también establece la justificación científica y tecnológica para el desarrollo de modelos predictivos aplicados a la Central Hidroeléctrica 5 de Noviembre.

2.1. Contexto Energético en El Salvador.

La matriz energética de El Salvador ha experimentado una diversificación significativa en las últimas décadas, impulsada por políticas energéticas y por la introducción de nuevas fuentes de generación que han transformado el perfil energético nacional. Actualmente, el sistema se compone de una mezcla de energías renovables y de transición, incluyendo siete fuentes principales: hídrica, geotérmica, solar, eólica, biomasa, térmica y gas natural.

La hidroeléctrica, junto con otras fuentes renovables, constituye una parte importante de la matriz energética del país, contribuyendo al 59.4% de la generación total en 2024. Este avance en energías renovables ha fortalecido la sostenibilidad y la resiliencia del sistema energético salvadoreño, disminuyendo la dependencia de fuentes no renovables y mejorando la estabilidad del suministro energético. La planta hidroeléctrica 5 de Noviembre, en particular, desempeña un papel fundamental en esta estructura, al contribuir de manera continua y confiable a la red nacional.

Para comprender mejor la estructura de la matriz energética, es importante analizar la capacidad instalada de cada fuente. La siguiente tabla presenta la capacidad instalada por tipo de recurso en El Salvador, según datos de la "Guía sectorial Energía 2023":

Tipo de Recurso	Capacidad Instalada (MW)	Porcentaje del Total (%)
Hidroeléctrica	837	34
Geotérmica	566	23
Biomasa	221	9
Solar Fotovoltaica	209	8.5
GNL	396	16.1
Búnker y Diésel	194	7.9
Eólica	52	2.1

Total	2475	100
Fuente: Guía sectorial Energía 2023, Invest in El Salvador (https://investinelsalvador.gob.sv/wp-content/uploads/2023/12/Guia-Sectorial-Energia-2023.pdf)		

Tabla 1: Proyección de la demanda energética en GWh desde el 2011 hasta el 2030.

Como se observa en la tabla, la hidroeléctrica representa la mayor proporción de la capacidad instalada (34%), seguida por la geotérmica (23%) y el GNL (16.1%). Esta distribución de la capacidad instalada influye directamente en la generación de energía y en la composición de la matriz energética.[2]

El gas natural, que se incorporó en 2022, ha emergido rápidamente como un componente clave, representando el 32.1% de la generación total en el último año. La planta de Energía del Pacífico ha permitido una disminución significativa de la dependencia en fuentes térmicas basadas en búnker y diésel, sustituyéndolas por una fuente menos volátil en términos de costos y más amigable con el ambiente.

Este cambio hacia una matriz más diversificada también ha promovido un sistema eléctrico resiliente y autosuficiente, que evita riesgos de escasez energética y estabiliza los precios en el mercado eléctrico. La implementación de proyectos fotovoltaicos, impulsada en gran medida por empresas como AES, ha incrementado la generación solar, alcanzando un 7.21% de la matriz energética en 2024. Este crecimiento en energías limpias complementa el sistema y contribuye a los objetivos de sostenibilidad establecidos en la política energética nacional.

Es importante considerar que, a pesar de la diversificación de la matriz, la demanda energética del país continúa en aumento. Como se observa en la Gráfica 1 (Guía sectorial Energía, 2023), se proyecta un crecimiento constante de la demanda desde 2011 hasta 2030, superando los 8,000 GWh. Este crecimiento en la demanda subraya la necesidad de continuar invirtiendo en la expansión y modernización de la infraestructura de generación, transmisión y distribución de energía, así como en la optimización del uso de los recursos existentes. La confiabilidad de fuentes como la hidroeléctrica, y en particular la planta 5 de Noviembre, se vuelve crucial para satisfacer esta creciente demanda.[2]

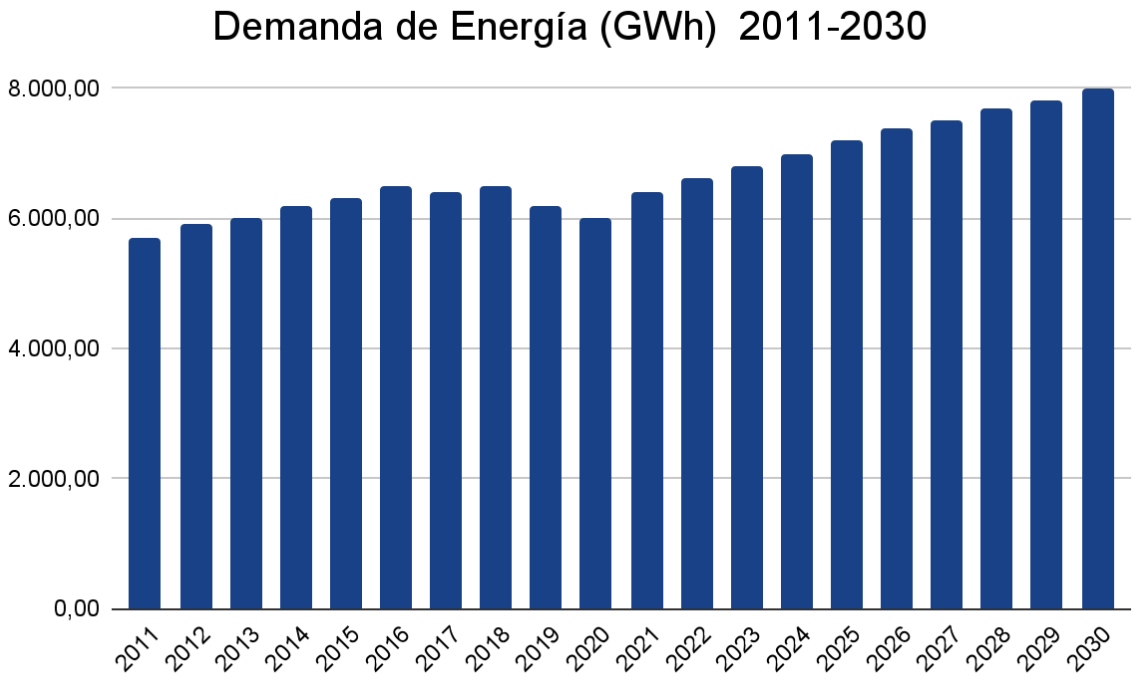


Figura 1: Proyección de la demanda energética en GWh desde el 2011 hasta el 2030.

La matriz energética de El Salvador ha evolucionado hacia un modelo diversificado que incorpora una combinación equilibrada de energías renovables, de transición y convencionales. Este enfoque contribuye no solo a la estabilidad de precios y al fortalecimiento de la seguridad energética del país, sino también a la reducción de su huella ambiental, promoviendo un futuro energético más sostenible y autosuficiente. Sin embargo, el crecimiento proyectado de la demanda energética, como se muestra en la Gráfica 1, plantea desafíos que requieren una planificación estratégica continua y una gestión eficiente de los recursos energéticos.

2.2. Rendimiento y Relevancia de la Generación Hidroeléctrica en la Matriz Energética de El Salvador

Durante el año 2024, la generación hidroeléctrica ha desempeñado un papel importante en el abastecimiento energético de El Salvador, siendo uno de los principales proveedores de energía en los meses de alta demanda. En julio, la generación hidroeléctrica alcanzó el 42.7% de la demanda energética nacional, superando a otras tecnologías en capacidad de suministro, en parte debido al aumento de lluvias en esta temporada. Las condiciones climáticas permitieron que las plantas hidroeléctricas del país, operadas por la Comisión Ejecutiva Hidroeléctrica del Río Lempa (CEL), como las centrales de Guajoyo, Cerrón Grande, 15 de Septiembre, 5 de Noviembre y 3 de Febrero, incrementaron su aporte [3].

La demanda energética total en julio de 2024 fue de 610 gigavatios hora (GWh), de los cuales el 73% correspondió a fuentes renovables, mientras que las tecnologías térmicas y las importaciones cubrieron el 24% y el 3% de la demanda, respectivamente. Según datos del Ministerio de Medio Ambiente y Recursos Naturales (MARN), julio fue el noveno mes

más lluvioso de los últimos 54 años, con precipitaciones que superaron en un 133% el promedio histórico, lo que favoreció el incremento en la generación hidroeléctrica [3].

El desempeño de la generación hidroeléctrica en períodos de alta demanda y condiciones climáticas favorables subraya la relevancia de esta fuente en la estabilidad y resiliencia del sistema energético salvadoreño. Para esta tesis, que busca modelar y predecir la inyección de energía por generador a partir de datos históricos, estos patrones de generación hidroeléctrica brindan un contexto valioso. La capacidad de la generación hidroeléctrica para ajustarse a la disponibilidad de recursos hídricos y su variabilidad en función del clima representan factores que influyen en el desarrollo y calibración de los modelos de *Machine Learning* propuestos para el análisis de inyección de energía. [4]

2.3 Comisión Ejecutiva Hidroeléctrica del Río Lempa (CEL): Marco Histórico

La Comisión Ejecutiva Hidroeléctrica del Río Lempa (CEL) fue concebida en un momento crucial para el desarrollo de El Salvador. Su establecimiento se formalizó mediante Decreto Ejecutivo el 3 de octubre de 1945, bajo la administración del presidente General Salvador Castaneda Castro, y su creación fue promulgada oficialmente en el Diario Oficial el 8 de octubre de ese mismo año. Sin embargo, el origen de CEL como la conocemos hoy fue un proceso de construcción institucional en dos etapas. Inicialmente, la entidad fue constituida como una comisión presidencial con un mandato específico y exploratorio: "hacer los estudios que determinen las posibilidades y las bases del desarrollo del Río Lempa para realizar la obra de electrificación nacional". Su propósito era evaluar la viabilidad técnica y económica de aprovechar el principal recurso hídrico del país para superar el déficit energético que limitaba el progreso industrial y social.

El paso decisivo hacia su consolidación ocurrió tres años después. Tras la finalización de los estudios preliminares, que confirmaron el vasto potencial del río, el proyecto fue elevado de una iniciativa presidencial a una política de Estado. Mediante el Decreto Legislativo N.º 137 [5], emitido el 18 de septiembre de 1948, la comisión fue transformada en la "Comisión Ejecutiva Hidroeléctrica del Río Lempa", una "institución autónoma de servicio público, sin fines lucrativos". Esta nueva ley le confirió personalidad jurídica, autonomía administrativa y las facultades necesarias para ejecutar obras de gran envergadura y gestionar sus propias finanzas, incluyendo la capacidad de emitir y colocar bonos en mercados nacionales e internacionales para financiar sus proyectos. Este cambio jurídico fue fundamental, pues dotó a CEL de la robustez institucional requerida para emprender la monumental tarea de la electrificación nacional.

2.4. Marco Regulatorio de las Hidroeléctricas en El Salvador

El funcionamiento y desarrollo de las hidroeléctricas en El Salvador están regidos por un conjunto de leyes y regulaciones que buscan promover la eficiencia, sostenibilidad y seguridad en la generación de energía. Entre las normativas más destacadas se encuentran:

2.4.1. Ley General de Electricidad

La Ley General de Electricidad de El Salvador fue promulgada mediante el Decreto Legislativo No. 843 el 10 de octubre de 1996 y publicada en el Diario Oficial el 25 de octubre de ese mismo año. Esta ley constituye el pilar regulador para las actividades de generación, transmisión, distribución y comercialización de energía eléctrica en el país. Su promulgación responde a la necesidad de un marco normativo que incentive la inversión en el sector eléctrico y fomente un mercado competitivo, debido a que la anterior Ley de Servicios Eléctricos de 1936 ya no cumplía con las demandas contemporáneas [6][7].

La ley busca promover el desarrollo económico y social mediante la ampliación de la producción energética, incrementando tanto la eficiencia como la competitividad del sector. Entre sus objetivos, destacan:

1. **Desarrollo de un mercado competitivo:** Asegura que la generación, transmisión, distribución y comercialización de electricidad se realicen en un entorno de competencia.

2. **Acceso a la red de transmisión y distribución:** Permite que las entidades generadoras accedan libremente a las instalaciones necesarias para el transporte de energía, con limitaciones únicamente establecidas por la ley.

3. **Uso racional y eficiente de los recursos energéticos:** Promueve prácticas sostenibles que optimicen los recursos y la infraestructura del sector.

4. **Accesibilidad de la energía eléctrica para todos los sectores:** Garantiza que la población tenga acceso al suministro eléctrico como un servicio esencial, beneficiando a sectores tanto urbanos como rurales.

5. **Protección de los derechos de usuarios y operadores:** Establece la defensa de los derechos de todas las partes involucradas en el sector energético, incluyendo consumidores y operadores de servicio

La ley designa a la **Superintendencia General de Electricidad y Telecomunicaciones (SIGET)** como la autoridad encargada de aplicar y supervisar las disposiciones de esta normativa. Las facultades otorgadas a la SIGET incluyen:

- **Supervisión de la competencia:** SIGET actúa como el organismo regulador que supervisa el cumplimiento de condiciones de competencia en el mercado eléctrico.

- **Gestión de concesiones y permisos:** La SIGET administra la emisión y el control de concesiones para la explotación de recursos energéticos como los hidráulicos y geotérmicos.

- **Resolución de conflictos y sanciones:** La superintendencia tiene la autoridad para resolver conflictos entre operadores y aplicar sanciones a quienes infrinjan la ley.

- **Control de precios y tarifas:** También se encarga de la aprobación de cargos por uso de redes de transmisión y distribución, asegurando que estos se basen en costos reales y protejan los intereses de los usuarios.

Este marco regulador fue diseñado para fortalecer el sector eléctrico de El Salvador, garantizar un suministro confiable de energía y fomentar la inversión en fuentes renovables y tecnologías eficientes.

2.4.2. Reglamento de la Ley General de Electricidad.

El Reglamento de la Ley General de Electricidad complementa y detalla los lineamientos establecidos por la Ley General de Electricidad, proporcionando un marco regulatorio que promueve la eficiencia, transparencia y competitividad en el sector eléctrico salvadoreño. Este reglamento, administrado y supervisado por la Superintendencia General de Electricidad y Telecomunicaciones (SIGET), asegura el cumplimiento de los principios fundamentales de la ley, facilitando un entorno adecuado para la inversión y operación dentro del sector [7][8]. A continuación, se describen los elementos clave del reglamento:

1.Disposiciones Generales

Las disposiciones iniciales establecen el propósito del reglamento y la autoridad de la SIGET para supervisar y hacer cumplir sus estipulaciones. En esta sección, también se establece la obligación de los operadores de cumplir con otras normativas nacionales aplicables, proporcionando así un marco legal integral y coherente para la regulación del sector eléctrico.

2.Concesiones para Generación Eléctrica

El reglamento detalla los procedimientos y requisitos necesarios para la concesión de permisos de generación eléctrica, especialmente en el caso de proyectos hidroeléctricos y geotérmicos. Este capítulo asegura que los concesionarios cumplan con normas técnicas y ambientales, y regula la duración y condiciones de renovación de las concesiones. Además, se establecen mecanismos para la modificación o terminación de las concesiones, con el objetivo de garantizar la continuidad y confiabilidad en el suministro eléctrico.

3.Mercado Mayorista de Electricidad

En este apartado se definen las reglas de operación para el mercado mayorista de electricidad, donde los generadores, distribuidores y comercializadores realizan transacciones de compra y venta de energía. La Unidad de Transacciones actúa como entidad administradora de este mercado, asegurando la competencia y regulando la participación de los agentes en un marco de transparencia y eficiencia operativa.

4.Contratos de Transmisión y Distribución

El reglamento regula las condiciones para la formalización de contratos entre generadores, transmisores y distribuidores. Esto incluye el acceso a la red de transmisión y distribución, así como los términos y tarifas de uso de infraestructura. La normativa fomenta una relación contractual justa entre las partes, protegiendo tanto los derechos de los operadores como los de los usuarios finales.

5.Tarifas y Precios

En este apartado, el reglamento establece los procedimientos para la fijación y ajuste de tarifas y precios en el sector eléctrico. La SIGET es responsable de aprobar los cargos por el uso de las redes de transmisión y distribución, con base en costos reales y en un marco de equidad. Esta regulación de precios tiene como objetivo evitar abusos y asegurar que el costo de la electricidad sea accesible y justo para todos los usuarios.

6.Calidad del Servicio Eléctrico

El reglamento establece los estándares mínimos de calidad que deben cumplir los operadores de generación, transmisión y distribución de electricidad. La normativa asegura que la prestación del servicio eléctrico sea continua, confiable y cumpla con parámetros de calidad que protejan los derechos de los consumidores. Además, establece medidas de compensación para los usuarios en caso de interrupciones o incumplimientos en el servicio.

7.Régimen Sancionatorio

La sección de sanciones detalla las infracciones y las penalidades correspondientes en caso de incumplimiento de la ley y el reglamento. Este régimen sancionador, administrado por la SIGET, busca disuadir conductas que puedan afectar la eficiencia y competitividad del sector, promoviendo el cumplimiento de las normas y el respeto a los derechos de los usuarios y operadores.

8.Disposiciones Transitorias y Finales

Las disposiciones transitorias y finales establecen los plazos para la entrada en vigencia del reglamento y la adaptación de los operadores a los nuevos requerimientos. Este apartado también incluye orientaciones administrativas que facilitan la implementación y adaptación de la normativa en el sector.

2.4.3. Política Energética Nacional.

El Salvador ha desarrollado un marco de políticas públicas para el sector energético que busca no solo satisfacer la demanda nacional de energía, sino también avanzar hacia un sistema más sostenible, seguro y diversificado. A través de la Política Energética Nacional 2010-2024 y la Política Energética Nacional 2020-2050, el país ha establecido metas y estrategias para fortalecer la matriz energética, adaptarse a los cambios tecnológicos y promover el uso de energías renovables [8].

La Política Energética Nacional 2010-2024 sentó las bases para una transición hacia un sistema energético más resiliente y diversificado. Este documento surgió como respuesta a la dependencia de combustibles fósiles y a la necesidad de incorporar energías renovables en la matriz. Entre sus principales objetivos estuvieron la diversificación de las fuentes de energía, la mejora en la eficiencia energética y la reducción de las emisiones de carbono. Durante este período, se realizaron inversiones en infraestructura y se promovieron proyectos de generación renovable, como la energía hidroeléctrica, geotérmica y solar,

contribuyendo al desarrollo de un sector más sostenible y menos dependiente de las fuentes no renovables [8].

Por su parte, la Política Energética Nacional 2020-2050 representa una visión a largo plazo que apunta a consolidar los avances de la década anterior y a enfrentar los nuevos desafíos del sector energético. Este documento enfatiza la sostenibilidad, la innovación y la seguridad energética, con un enfoque en la modernización del sistema eléctrico y el uso de tecnologías de bajo impacto ambiental. Entre sus ejes estratégicos destacan el impulso a la energía renovable, la eficiencia en el uso de los recursos, la integración regional y la creación de un entorno que favorezca la inversión en tecnologías limpias.

Ambas políticas han sido fundamentales para el desarrollo del sector eléctrico salvadoreño. La transición hacia un sistema de energía renovable ha permitido reducir la dependencia de fuentes fósiles y mitigar el impacto ambiental, mientras que la visión hacia 2050 consolida un marco de crecimiento sostenible e inclusivo. Este enfoque integrado no solo asegura un suministro de energía estable y asequible para la población, sino que también prepara a El Salvador para enfrentar los retos asociados al cambio climático y la transición energética global.

En conclusión, el marco de políticas energéticas de El Salvador refleja una evolución desde el fortalecimiento y diversificación del sistema en la última década, hasta una planificación sostenible y modernizada a largo plazo. Estos documentos son una referencia clave para orientar el desarrollo y adaptación del sector energético a las necesidades futuras, apoyando la investigación y proyectos que promuevan la eficiencia y el uso de recursos renovables, alineándose con las metas de sostenibilidad que son fundamentales para el país.

2.5. Importancia de la Planta Hidroeléctrica 5 de Noviembre y su Expansión

La **Central Hidroeléctrica 5 de Noviembre**, situada en “La Chorrera del Guayabo” sobre el río Lempa, en los departamentos de Cabañas y Chalatenango, representa un hito en la historia energética de El Salvador. Fue inaugurada el **21 de junio de 1954** con una capacidad inicial de **30 MW**, correspondiente a dos unidades generadoras de 15 MW cada una. Entre 1957 y 1966 se incorporaron tres unidades más (dos de 15 MW y una de 21.4 MW), elevando la capacidad instalada a **81.4 MW** [9].

La planta se constituye como una infraestructura hidráulica robusta y estratégica con una presa de gravedad de concreto de **65 metros de altura**, un vertedero de **7 compuertas**, y una casa de máquinas subterránea. Tiene un embalse que abarca **16 km²**, con un volumen total de **320 millones de m³** y útil de **87 millones de m³**, generando en promedio **457 GWh** anuales [9].

Con el paso del tiempo, CEL promovió la modernización del parque generador. Un avance clave ocurrió en 2013 cuando se lanzó el ambicioso **proyecto de expansión**, que consistía en la instalación de una nueva casa de máquinas con **dos turbinas Francis de 40 MW** cada una, lo que elevaría la capacidad total a **179.4 MW**. Este proyecto fue financiado con **US$57.5 millones** del **BCIE**, igual monto del **KfW (Banco Alemán de Desarrollo)** y una

donación de **€6 millones** de la **Unión Europea**, sumando aproximadamente **US$178.5 millones** en inversión total [10].

Según la **Memoria Anual del BCIE 2010**, el financiamiento de US$57.5 millones será clave para incrementar la capacidad instalada de la planta de 99.4 MW a **179.4 MW**, además de contribuir a la reducción estimada de **93,454 toneladas de CO_2 por año**, mediante programas de manejo ambiental como reforestación y conservación de cuencas [11].

En **noviembre de 2016**, CEL confirmó la conclusión de la ampliación, lo que permitió que la planta ampliada estuviera lista para operar con mayor generación. La nueva potencia de **179.4 MW** significó un aumento crítico en la oferta de energía limpia y renovable del país [12].

Figura 2: fotografía aérea de la Central hidroeléctrica 5 de Noviembre.

2.6. Fundamentos de Machine Learning

2.6.1. Contexto de Machine Learning

El **Machine Learning (ML)**, o aprendizaje automático, es una disciplina dentro de la inteligencia artificial que permite a los sistemas informáticos **aprender y mejorar su desempeño a partir de datos**, sin requerir programación explícita para cada tarea. Su propósito principal es identificar patrones en los datos y utilizarlos para generar predicciones o tomar decisiones informadas. De acuerdo con la Organización Internacional de Normalización (ISO), el ML constituye uno de los componentes clave de la inteligencia artificial moderna, pues posibilita que las máquinas aprendan directamente de la experiencia y adapten su comportamiento a diferentes contextos [13].

Los avances de mediados del siglo XX sentaron las bases para este campo. En 1959, **Arthur Samuel** acuñó el término *machine learning* al desarrollar un programa capaz de aprender a jugar damas, demostrando que un algoritmo podría mejorar progresivamente su rendimiento mediante la experiencia acumulada [13]. Años antes, **Alan Turing** ya había planteado la célebre pregunta “¿Pueden pensar las máquinas?” y diseñado el *Test de Turing* como criterio para evaluar la inteligencia de los sistemas computacionales [15].

Posteriormente, en 1958, **Frank Rosenblatt** presentó el *Perceptrón*, considerado precursor de las redes neuronales y del aprendizaje profundo, abriendo el camino para el desarrollo de arquitecturas más complejas [14]. No obstante, el progreso en este campo atravesó períodos de estancamiento conocidos como *inviernos de la IA* durante las décadas de 1970 y 1980, debido a limitaciones en capacidad computacional y expectativas no cumplidas. El resurgimiento llegó en los años 90, impulsado por el incremento en la potencia de cálculo y la disponibilidad de grandes volúmenes de datos. Desde entonces, el ML ha transformado múltiples industrias mediante aplicaciones en reconocimiento de patrones, clasificación, predicción y optimización [13].

En la actualidad, el Machine Learning se integra en un **ecosistema más amplio de la ciencia de datos**. Este ecosistema se apoya en la ciencia de la computación para proveer infraestructura tecnológica, en la estadística y matemáticas para el análisis de datos, y en la inteligencia artificial para simular procesos cognitivos humanos. Dentro de este marco, el ML ocupa un lugar central, destacándose por su capacidad de autoaprendizaje y su versatilidad para manejar grandes volúmenes de información en entornos dinámicos.

Existen tres categorías principales de Machine Learning:

- **Aprendizaje Supervisado:** Se basa en conjuntos de datos etiquetados donde cada entrada está asociada a una salida conocida. El modelo aprende esas relaciones para predecir resultados en datos nuevos. Se utiliza en problemas de **clasificación** (p. ej. detección de spam) y **regresión** (p. ej. predicción de precios o demanda energética).

- **Aprendizaje No Supervisado:** Trabaja con datos sin etiquetas, buscando patrones ocultos o estructuras subyacentes. Sus técnicas más comunes son el **clustering** (segmentación de clientes, agrupación de datos) y la **reducción de dimensionalidad** (simplificación de grandes volúmenes de información). Es útil en áreas como la detección de fraudes, análisis de mercados y biología.

- **3. Aprendizaje por Refuerzo:** Un agente aprende a tomar decisiones mediante la interacción con su entorno, recibiendo recompensas o penalizaciones según sus acciones. Este enfoque se aplica en la **robótica**, los **vehículos autónomos**, la **optimización energética** y en entornos complejos como los videojuegos.

Estos enfoques han permitido el desarrollo de aplicaciones en campos tan diversos como el diagnóstico médico, la detección de fraudes financieros, la automatización industrial, el procesamiento de lenguaje natural y los sistemas autónomos, consolidando al ML como una herramienta fundamental en la transformación digital contemporánea.

2.6.2. Aplicaciones del Machine Learning en el Sector Energético

En el sector energético, especialmente en la región de América Latina, estas tres categorías han mostrado ser de gran utilidad. Por ejemplo, el aprendizaje supervisado ha permitido mejorar significativamente la predicción de generación energética en plantas hidroeléctricas y eólicas mediante modelos que incorporan variables meteorológicas y operativas. El aprendizaje no supervisado ha facilitado la identificación de patrones de consumo anómalos

que podrían indicar pérdidas técnicas o fraudes eléctricos, mejorando así la eficiencia y la sostenibilidad financiera de los sistemas eléctricos. Por último, el aprendizaje por refuerzo está comenzando a utilizarse para optimizar la operación de redes eléctricas inteligentes, maximizando la eficiencia en el despacho de energía y minimizando costos operativos y ambientales.

En resumen, el Machine Learning presenta un potencial extraordinario para transformar y optimizar procesos en diversas industrias, siendo especialmente relevante en el contexto energético actual. La implementación efectiva de estos métodos puede generar importantes beneficios económicos, operativos y ambientales, posicionándose como un pilar fundamental en la evolución tecnológica hacia sistemas energéticos más sostenibles y eficientes.

2.7 Computación en la Nube

La computación en la nube ha transformado radicalmente la manera en que las organizaciones acceden y gestionan recursos informáticos. Este modelo permite el acceso bajo demanda a una variedad de servicios y recursos a través de internet, eliminando la necesidad de mantener infraestructuras físicas locales. Su mayor aporte es la flexibilidad, escalabilidad y eficiencia que ofrece, permitiendo a las empresas ajustar sus recursos tecnológicos según la demanda sin grandes inversiones iniciales en hardware o software [17].

Dentro de este modelo existen diferentes niveles de servicio que responden a distintas necesidades de gestión y control. Entre los más importantes se encuentran la Infraestructura como Servicio (IaaS), la Plataforma como Servicio (PaaS) y el Software como Servicio (SaaS). Cada uno representa un grado distinto de responsabilidad compartida entre el proveedor de la nube y el usuario, lo que define la forma en que se administran los recursos [18].

La **Infraestructura como Servicio (IaaS)** proporciona acceso bajo demanda a recursos informáticos esenciales como servidores, redes y almacenamiento, los cuales se pueden aprovisionar de manera flexible y escalable. Este modelo resulta ideal para empresas que necesitan mantener un mayor control sobre la infraestructura tecnológica, sin los costos asociados al hardware físico [18].

La **Plataforma como Servicio (PaaS)** se centra en ofrecer un entorno de desarrollo y despliegue en la nube que incluye herramientas, bibliotecas y servicios para crear aplicaciones. Bajo este esquema, el proveedor gestiona la infraestructura subyacente, lo que permite a los desarrolladores enfocarse en la lógica de negocio y en el código, acelerando así los procesos de innovación y colaboración sin preocuparse por la administración del hardware [18].

El **Software como Servicio (SaaS)** corresponde al nivel más accesible de la nube, ya que proporciona aplicaciones completas listas para usarse directamente desde un navegador o aplicación. En este caso, los usuarios finales no necesitan instalar, actualizar ni mantener software en sus dispositivos, puesto que el proveedor se encarga de todo el ciclo de gestión. Este modelo destaca por su accesibilidad desde cualquier dispositivo conectado a internet,

la reducción de costos en licencias y la capacidad de escalar rápidamente según la demanda de los usuarios [18].

Finalmente, el mercado de la computación en la nube está liderado por proveedores que ofrecen soluciones integrales a nivel global. **Amazon Web Services (AWS)** es uno de los pioneros, reconocido por su escalabilidad y amplia gama de servicios. **Microsoft Azure** se distingue por su integración con el ecosistema de productos de Microsoft y sus soluciones híbridas. **Google Cloud Platform (GCP)** ha ganado relevancia por sus capacidades en análisis de datos e inteligencia artificial. En Asia, **Alibaba Cloud** se ha posicionado como el mayor proveedor, con una fuerte presencia internacional en servicios de big data, seguridad y cómputo elástico. A su vez, **Tencent Cloud** ofrece soluciones especializadas en redes, bases de datos y servicios digitales, expandiendo su cobertura en regiones estratégicas de Asia, América y Europa [19].

En síntesis, en el presente capítulo se estableció el marco sectorial, institucional y tecnológico que contextualiza esta investigación: la evolución y diversificación de la **matriz energética salvadoreña**, la relevancia operativa de la **generación hidroeléctrica** —con énfasis en la Central **5 de Noviembre** y su expansión—, el **marco regulatorio** que rige al sector y los **fundamentos de Machine Learning y computación en la nube** que habilitan el enfoque propuesto. Con estos elementos, se dispone de una base suficiente para pasar del *qué* (contexto, políticas, capacidades y limitaciones) al *cómo* (procedimientos concretos para desagregar y estimar generación por unidad). En el **Capítulo 3** se detalla el **diseño metodológico**: fuentes y tratamiento de datos (UT, CEL y registros operativos), análisis exploratorio, definición de variables y construcción de **dos modelos complementarios** (clasificación multietiqueta y regresión continua) desplegados en entorno de **nube**. Esta transición marca el paso desde la comprensión del sistema eléctrico nacional y el papel de 5 de Noviembre hacia la **implementación técnica** de los modelos que permiten reconstruir patrones de operación y cuantificar la inyección por unidad en el período de estudio.

Capítulo 3: Diseño Metodológico

Este capítulo describe el enfoque metodológico empleado para el desarrollo de la investigación, estructurado bajo un paradigma cuantitativo explicativo–predictivo. Se detallan las fuentes de datos utilizadas, los procesos de tratamiento, integración y validación de la información y las fases de diseño de los modelos de Machine Learning aplicados.

El capítulo introduce la estrategia utilizada para abordar la falta de datos desagregados por unidad generadora en la Central 5 de Noviembre, proponiendo un enfoque mixto compuesto por dos modelos complementarios: un modelo de clasificación multietiqueta para determinar los estados operativos y un modelo de regresión continua para estimar la potencia inyectada. Finalmente, se justifica la implementación en entornos de computación en la nube, asegurando escalabilidad, trazabilidad y eficiencia en el procesamiento de datos, elementos esenciales para garantizar la validez técnica y científica de la investigación.

3.1. Enfoque de la Investigación

La presente investigación se enmarca en un enfoque cuantitativo de tipo explicativo–predictivo, apoyándose en el análisis de datos históricos y en la construcción de modelos de *Machine Learning* montados en servidores en la nube. dada la información obtenida y las condiciones de estudio, el objetivo metodológico es doble:

1. Identificar patrones de operación por unidad generadora en la central hidroeléctrica 5 de Noviembre y;
2. Estimar la inyección continua de cada unidad a partir de datos agregados y variables de operación.

Para ello, la secuencia metodológica contempló las siguientes fases:

3.1.1. Análisis exploratorio

La primera fase consistió en evaluar la generación global de la planta y calcular el factor de planta en el período de estudio. Para ello, se recopilaron y contrastaron distintas fuentes de datos:

- **Datos públicos de la Unidad de Transacciones (UT):** disponibles en su portal web, donde se publican reportes diarios de inyección de energía. Estos datos, aunque valiosos, se encontraban en formato general para la planta 5 de Noviembre y no permitían identificar el aporte específico por unidad generadora.

- **Datos proporcionados por el equipo de control de la central hidroeléctrica 5 de Noviembre:** obtenidos a través de reuniones de trabajo, los cuales ofrecieron un mayor nivel de detalle sobre el comportamiento de las unidades. Esta información

complementaria permitió robustecer el análisis y disponer de insumos más precisos para la construcción de los modelos predictivos.

A partir de esta integración, se diferenciaron los **datos de inyección real de 2024** separados en las 7 unidades (usados como base de entrenamiento) y los **reportes históricos de 2021–2023** de inyección total), conformando el conjunto de información que sustentó la creación de los modelos de predicción y estimación de inyección por unidad.

3.2. Fuentes y tratamiento de datos

3.2.1 Fuente de los Datos

Para cumplir los objetivos, se emplearon diversas fuentes de datos proporcionadas principalmente por la Unidad de Transacciones (UT), operador del mercado eléctrico salvadoreño, y complementados con información compartida por el equipo de control de la Central Hidroeléctrica 5 de Noviembre [1].

Fuente de datos	Descripción	Periodo	Resolución temporal	Tratamiento/Formato
Despacho real horario (UT)	Generación real total inyectada por la planta 5 de Noviembre en la red, por hora e inyección completa de la planta (en MW).	01/01/2012 – 31/12/2024	Horaria (24 valores/día)	Datos obtenidos del registro público de la UT.
Nivel del embalse / Recursos hídricos (CEL)	Datos del nivel de agua del embalse de 5 de Noviembre y/o caudal disponible para generación.	01/01/2012 – 31/12/2024	Horaria (24 valores/día)	Datos obtenidos del registro público de la UT.
Vertido	Datos del vertido de agua del embalse de 5 de Noviembre.	01/01/2012 – 31/12/2024	Horaria (24 valores/día)	Datos obtenidos del registro público de la UT.
Datos de Inyección real por unidad (UT)	Datos de inyección separados de las unidades 1 a la 7.	06/01/2024 - 31/12/2024	Horaria (24 valores/día)	Datos proporcionados mediante correo electrónico por autoridades de la UT.
Predespacho (UT)	Datos de programaciónes diarias de inyección	02/08/2020-31/ 12/2024	Horaria (24 valores/dia)	Datos del registro público de la UT.

Tabla 2: Fuentes de datos utilizadas en la metodología, con su periodicidad y tratamiento.

3.2.1 Tratamiento de los datos:

Cada una de estas fuentes fue sometida a un proceso de preparación y validación antes de su uso en los modelos:

1. **Integración:** los datos de inyección total real, nivel del embalse, vertido y las inyecciones individuales de las unidades (2024) se unificaron en un repositorio único, asegurando consistencia temporal (alineación de horas, fechas y zonas horarias).

2. **Limpieza:** se manejaron valores faltantes o anómalos (ej. inyecciones por encima de la capacidad real de las generadoras). En casos específicos se excluyeron datos del análisis.

3. **Cálculo del factor de planta global:** se estimó como la fracción entre la energía realmente generada y la energía máxima posible si todas las unidades hubiesen operado a plena capacidad. Este indicador permite caracterizar el desempeño general de la planta.

4. **Limitación de datos individuales:** los registros públicos de la UT reportan la generación de la central de forma general o agregada. Para solventar esta limitación, se gestionó ante la UT el acceso a datos desagregados por unidad durante 2024. Este conjunto, que abarcó desde el 6 de enero al 31 de diciembre, fue clave para entrenar los modelos y emplearlo como patrón de comportamiento de las siete unidades.

5. **Aplicación a períodos históricos:** con base en los datos de 2024, los modelos se entrenaron para desagregar y estimar la inyección horaria de las unidades en los años 2021–2023, permitiendo el cálculo individualizado del factor de planta para cada unidad, incluidas las dos más recientes de la expansión unidades 6 y 7.

3.3. Modelos predictivos

Se desarrollaron dos enfoques complementarios con el propósito de desagregar y estimar la generación individual de las siete unidades de la central hidroeléctrica 5 de Noviembre.

3.3.1. Modelo multietiqueta de clasificación.

1. **Codificación:** cada hora se representó mediante un vector binario de siete bits, en el cual cada posición corresponde a una unidad generadora (1 = operativa, 0 = inactiva).

2. **Objetivo:** identificar qué unidades estaban activas en cada intervalo horario, a partir de los datos agregados de inyección total, junto con variables contextuales (nivel de embalse y vertido).

3. **Algoritmos probados:** se evaluaron técnicas de aprendizaje supervisado como Decision Trees (DT), Logistic Regression (LoR) y K-Nearest Neighbors (KNN). La mejor precisión se obtuvo con KNN (k=5), alcanzando resultados superiores al 90% de exactitud de predicción en algunas unidades.

4. **Resultado esperado:** La clasificación permitió reconstruir patrones de operación unitarios y detectar con mayor claridad el rol de las unidades en funcionamiento.

5. **Dato representativo por unidad:** Contando con dicho patrón binario, se procedió a calcular los cuartiles y mediana de cada unidad de operación (buscando el el valor más representativo para cada unidad) en las estimaciones. A continuación se muestra el diagrama de caja y bigotes que ilustra los datos

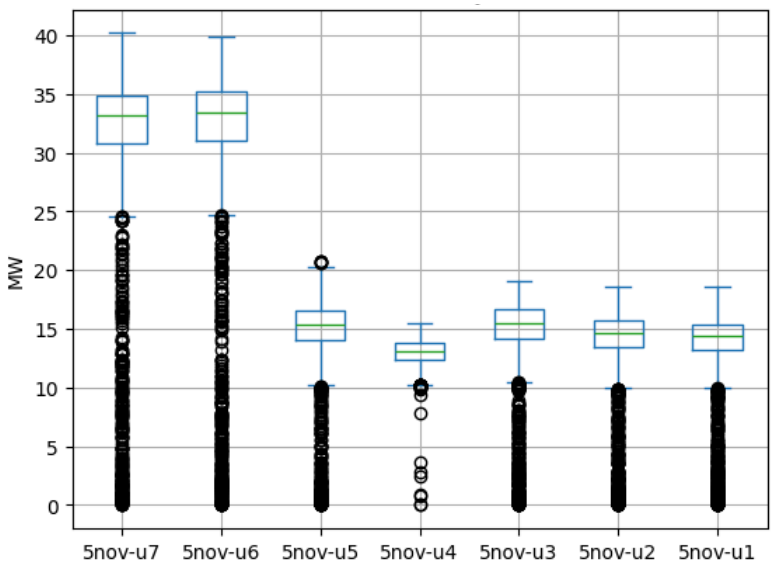


Figura 3: Diagrama de Caja y bigotes representando cuartiles y mediana de generación energética de la Hidroeléctrica 5 de noviembre.

6. Mediante prueba y error se logró determinar que el valor más representativo para para utilizar era la mediana de cada unidad. Esta mediana, se multiplica por el patrón binario unidad por unidad y posteriormente se realiza la suma de los datos para obtener la gráfica estimada de consumo la cual, al compararse con la de las inyección reales se observa de manera detallada en el apartado de Codificación de Estados Operativos en modelo multietiqueta de clasificación en el capítulo 4.

3.3.2. Modelo de predicción continua

1. **Entrenamiento:** se utilizaron como insumo los datos de inyección desagregada de 2024, junto con el nivel del embalse, el vertido y la inyección total del mismo año. Estos registros, al contar con información diferenciada por unidad, constituyeron la base de calibración del modelo.

2. **Algoritmos empleados:** se entrenaron dos tipos de modelos:

a. **Regresión lineal múltiple (RLM):** para capturar relaciones lineales entre las variables de entrada y la inyección individual de cada unidad.

b. **Redes neuronales artificiales (RNA):** para modelar relaciones no lineales y dinámicas complejas en la operación de la planta.

3. **Validación cruzada temporal:** la validación se realizó de forma interna dentro del conjunto de 2024 (hold-out y validación cruzada), a fin de medir el desempeño y evitar sobreajuste.

4. **Aplicación a periodos históricos:** una vez calibrados y validados los modelos con 2024, se aplicaron a los registros agregados de 2021–2023 para **estimar la generación horaria desagregada por unidad**, permitiendo reconstruir el comportamiento operativo en años donde no existía inyección individual reportada.

5. **Resultado esperado:** la predicción continua permitió obtener estimaciones horarias de potencia para cada unidad, evaluar el comportamiento histórico con base en patrones recientes y aproximar el factor de planta individual.

6. **Comparación con datos agregados:** las predicciones de generación individual obtenidas por cada modelo fueron sumadas y contrastadas con la serie de inyección total real publicada por la UT. Esta comparación permitió validar la coherencia de las estimaciones y confirmar que los modelos reproducen adecuadamente la dinámica global de la central. Estos resultados pueden visualizarse en el apartado: modelo de predicción continua por unidad en el capítulo 4.

En conjunto, ambos modelos proporcionan una visión integral: mientras el modelo multietiqueta de clasificación, resolvía el problema de identificación de operación por unidad, el continuo permitió cuantificar la magnitud de la potencia inyectada de manera directa. Esta complementariedad fortaleció la robustez metodológica y aumentó la capacidad explicativa del análisis.

3.4. Implementación en la nube

Los modelos fueron entrenados y desplegados en un entorno de computación en la nube, aprovechando su escalabilidad y capacidad de procesamiento para manejar el volumen de datos horarios. Tras evaluar diferentes opciones, se decidió utilizar Microsoft Azure como plataforma de pruebas y desarrollo, considerando varios factores:

1. **Rendimiento y costos:** Azure presentó una relación costo–beneficio competitiva en comparación con otras plataformas (Google Cloud, AWS, Alibaba Cloud), ofreciendo planes de facturación flexibles que se ajustaban al alcance de esta investigación.

2. **Ecosistema y familiaridad:** el equipo ya poseía experiencia previa con el ecosistema de Azure, lo que facilitó la curva de aprendizaje y permitió una implementación más rápida y oportuna.

3. **Particularidad en el área energética:** Azure cuenta con módulos especializados en **procesamiento de datos de series temporales e integración con IoT e infraestructura SCADA**, lo que lo convierte en una opción especialmente atractiva para proyectos vinculados a la gestión de energía y operación de plantas. Estas herramientas permiten manejar de manera eficiente grandes volúmenes de datos horarios, alineándose con las necesidades de análisis de despacho e inyección en sistemas eléctricos.

4. **Compatibilidad tecnológica:** la plataforma ofrece compatibilidad nativa con **Python, TensorFlow y Scikit-learn**, librerías utilizadas en el desarrollo de los modelos de clasificación y predicción continua, además de servicios de *Machine Learning* automatizado que facilitaron la experimentación con diferentes algoritmos.

5. **Seguridad y respaldo:** Azure proporciona entornos con altos estándares de seguridad y disponibilidad, aspectos cruciales al trabajar con datos del sector eléctrico, considerados estratégicos para el país.

Si bien las demás plataformas ofrecen funcionalidades similares, la combinación de **costos, ecosistema de trabajo ya conocido y servicios orientados al análisis de datos energéticos** hizo que Azure resultara la opción más viable y adecuada para el desarrollo de esta investigación.

Validación y análisis:
La validación de los modelos se realizó en varias etapas, con el fin de garantizar tanto su precisión técnica como su utilidad en el contexto operativo de la central hidroeléctrica 5 de Noviembre.

1. **Comparación con datos oficiales**
   Los resultados estimados por los modelos fueron contrastados con la curva oficial de inyección real publicada por la UT para los años 2021–2023. Esta comparación permitió evaluar la cercanía general entre las predicciones y los registros históricos, asegurando que el sistema reprodujera de forma adecuada la operación global de la planta.

2. **Métricas de desempeño**

    1. Modelo multietiqueta de clasificación.: se calcularon métricas de clasificación supervisada como exactitud (accuracy), recall y F1-score, con el fin de medir la capacidad del modelo de identificar correctamente la combinación de unidades activas en cada intervalo horario.

    2. Modelo de predicción continua: se utilizaron métricas de error estándar, entre ellas: MAE (Mean Absolute Error): midió la desviación promedio entre las inyecciones predichas y las reales y RMSE (Root Mean Square Error): penalizó con mayor peso los errores grandes, evaluando la fidelidad del modelo frente a eventos extremos. Estas métricas se calcularon inicialmente sobre los datos desagregados de 2024 (set de entrenamiento y validación principal) y posteriormente se aplicaron como criterio de consistencia al

extrapolar las predicciones a los años 2021–2023.

3. **Coeficiente de determinación R^2**
   Para complementar las métricas anteriores, se empleó el coeficiente de determinación, que mide el grado de ajuste entre la curva estimada y la real. Este indicador se interpreta en un rango de 0 a 1, donde valores cercanos a 1 implican que las curvas son prácticamente idénticas en tendencia y magnitud.

4. **Escenarios de validación**
   Con el fin de analizar la robustez de los modelos bajo diferentes niveles de exigencia, se establecieron tres escenarios de tolerancia para la comparación entre valores estimados y reales: ±20%, ±15% y ±5%. Esto permitió evaluar en qué medida los modelos se mantenían confiables al incrementar el rigor de la validación.

5. **Análisis de utilidad práctica**
   Finalmente, se evaluó la capacidad explicativa de los modelos más allá de las métricas numéricas. En particular, se analizó su aporte para comprender el comportamiento histórico de la planta, incluyendo la estimación del factor de planta individual de cada unidad y la identificación de patrones de operación que no son evidentes en los reportes agregados de la UT. Este enfoque permitió validar que los modelos no solo alcanzaran precisión matemática, sino también relevancia práctica para el análisis y la toma de decisiones en el ámbito energético.

En conjunto, este proceso de validación demostró que los modelos no solo alcanzaron niveles de precisión superiores al 90% en escenarios de clasificación y predicción, sino que además ofrecieron una **herramienta práctica para la gestión operativa** y para la toma de decisiones sobre el mantenimiento y la planificación energética en el país.

En el capítulo 3 se estableció la base metodológica de esta investigación, definiendo el enfoque cuantitativo explicativo-predictivo, las fuentes de datos utilizadas, los procesos de limpieza e integración, y la formulación de los dos modelos centrales —el de **clasificación multietiqueta** y el de **predicción continua**— implementados en entornos de **computación en la nube**. A partir de esta estructura, el siguiente capítulo profundiza en la **caracterización de los datos** y en la descripción detallada de los **algoritmos aplicados**, explicando los procedimientos de codificación, entrenamiento, validación y comparación de resultados. En este sentido, el **Capítulo 4** representa la transición del diseño conceptual hacia la **ejecución práctica de los modelos de Machine Learning**, mostrando cómo los datos tratados se transforman en insumos analíticos capaces de reconstruir y estimar la operación histórica de la central hidroeléctrica 5 de Noviembre [1].

Capítulo 4: Datos y Algoritmos

4.1 Descripción de los Datos

Este capítulo presenta la estructura y preparación de los datos utilizados en la investigación, abarcando registros provenientes de la Unidad de Transacciones (UT) y de la Comisión Ejecutiva Hidroeléctrica del Río Lempa (CEL). Los datos analizados comprenden el período 2012–2024, con especial énfasis en el año 2024, que contiene los registros desagregados por unidad generadora (U1–U7) empleados para el entrenamiento de los modelos.

Se describe el proceso de integración, limpieza y alineación temporal de series horarias —inyección total, nivel del embalse, vertido y predespacho— y se detalla la construcción de la codificación binaria de estados operativos, utilizada para identificar combinaciones de unidades activas. A partir de estos datos, se desarrollan dos modelos predictivos complementarios:

1. Un modelo de clasificación multiclase, basado en algoritmos de árboles de decisión y Random Forest, para inferir qué unidades estaban operando por hora.

2. Un modelo de regresión continua, implementado mediante Regresión Lineal Múltiple (RLM) y Redes Neuronales Artificiales (RNA), para estimar la potencia inyectada individualmente.

Los modelos fueron implementados y validados en la plataforma Microsoft Azure, utilizando herramientas de análisis en Python (pandas, scikit-learn, NumPy, Matplotlib). Este capítulo constituye el eje técnico que transforma los datos históricos en una base cuantitativa sólida para el análisis de desempeño presentado en el siguiente capítulo.

4.1.1 Fuentes de los datos

Para llevar a cabo este estudio, se emplearon cuatro fuentes principales de información, todas ellas relacionadas con la operación de la central 5 de Noviembre:

Datos de registros públicos:

- **Despacho real horario (UT):** Es el registro de la generación efectiva inyectada a la red por la central, medido hora a hora en MW. Estos datos históricos también provienen de los registros públicos de la UT y abarcan un rango más amplio (desde el 1 de enero de 2012 hasta el 31 de diciembre de 2024, con valores horarios). Representan la producción real de energía de la planta en cada intervalo horario, permitiendo comparar lo planificado vs. lo realmente generado.

- **Datos de inyección real por unidad (UT):** Consisten en registros desagregados de la inyección de cada una de las siete unidades generadoras que conforman la

central. Este conjunto de datos abarca desde el 6 de enero de 2024 hasta el 31 de diciembre del mismo año, con resolución horaria (24 valores por día). Fueron proporcionados mediante comunicación directa con las autoridades de la UT y resultan de particular relevancia para esta investigación, ya que constituyen la base para entrenar y validar los modelos predictivos, al ofrecer por primera vez información diferenciada de la operación de cada unidad.

- **Nivel del embalse (CEL):** Son datos referentes al recurso hídrico disponible, en particular el nivel de agua del embalse de la central 5 de Noviembre (y, de ser pertinente, el caudal afluente o vertido). Esta información fue obtenida también de los registros públicos de la UT. Se tienen datos desde el 1 de enero de 2012 hasta el 31 de diciembre de 2024 con mediciones horarias. Estos datos permiten incorporar una perspectiva ambiental y de disponibilidad de agua al análisis, complementando los datos puramente eléctricos.

- **Vertido (UT):** Se refiere al volumen de agua liberado desde el embalse de la central, sin aprovechamiento para generación eléctrica. Los datos son también obtenidos de registros públicos de la UT y cubren el periodo del 1 de enero de 2012 al 31 de diciembre de 2024, con resolución horaria. Al provenir de registros públicos de la UT, esta información se incorpora como una variable de contexto clave, pues permite analizar las pérdidas hídricas y su relación con la eficiencia operativa de la central.

- **Programaciones de predespacho diario (UT):** Corresponden a la generación planificada para cada hora del día siguiente en la central 5 de Noviembre, expresada en megavatios (MW). Estos datos, son también obtenidos en registros públicos de la Unidad de Transacciones (UT) del sistema eléctrico, cubren el periodo desde el 2 de agosto de 2021 hasta el 31 de diciembre de 2024, con resolución horaria (24 valores por día). En otras palabras, cada informe diario de predespacho contiene 24 puntos que indican cuánta potencia se programó inyectar en cada hora de ese día por dicha central.

Estas fuentes brindan una visión completa de la operación de la central, abarcando lo planificado (predespacho), la energía real inyectada y las condiciones ambientales (nivel del embalse y vertido) de la generadora 5 de Noviembre.

4.1.2 Integración y alineación temporal

Antes de proceder con el análisis cuantitativo, fue necesario realizar un exhaustivo proceso de limpieza, transformación e integración de los datos obtenidos de las distintas fuentes. A continuación, se describen los principales pasos de este pre-procesamiento:

Dado que los datos provienen de fuentes heterogéneas y con diferentes formatos de publicación, se inició el proceso unificando el formato de fechas y horas, y alineando temporalmente todas las series. En particular, los registros de **despacho real total y los datos proporcionados por la UT (los datos por separado de las 7 unidades de 2024)** fueron obtenidos a partir de informes diarios generados de manera automática en la página oficial de la UT, los cuales presentan la programación y la inyección efectiva en bloques horarios de 24 valores por día.

Para garantizar la coherencia del conjunto, fue necesario **seleccionar únicamente la información pertinente** y organizarla en función de un índice temporal común (*timestamp*). A cada registro horario se le asoció tanto la generación planificada (MW) como la generación efectiva inyectada (MW), construyendo así una **serie de tiempo horaria continua para el período 2021–2024**.

Durante esta fusión se verificó la consistencia temporal, ajustando los formatos cuando fue necesario y asegurando que los cortes horarios (por ejemplo, las 00:00 horas de cada día) coincidieran exactamente entre despacho real, nivel de embalse, Vertido y los datos separados correspondientes del año 2024. De esta forma, se obtuvo una base de datos cronológicamente ordenada y lista para el análisis comparativo y el entrenamiento de los modelos predictivos.

En el proceso de unificación se detectaron brechas y valores atípicos que podían comprometer la calidad del análisis. Algunas horas carecían de registros de generación real, ya fuera por interrupciones en los informes de la UT posiblemente porque no hubo inyección o simplemente los datos no se registraron. Por otro lado, en las bases de datos de las unidades separadas de 2024, los datos de los primeros 5 dias de enero eran inconsistentes en su totalidad con los datos de inyección real totales de la otra base de datos por lo tanto fueron eliminados para evitar que los datos erróneos produjeran resultados erróneos en el análisis. Por otro lado de manera general:

1. Si se confirmó que en cierta hora la planta no generó, se imputó un valor de 0 MW en el despacho real (inyección nula).

2. Si la ausencia correspondía a un error de registro pero era razonable esperar un valor distinto de cero, se dejó marcado para análisis posterior o se interpoló suavemente en cálculos agregados.

Asimismo, se identificaron outliers obvios que no eran físicamente plausibles y que distorsionan la serie. Un ejemplo claro fueron algunos registros automáticos obtenidos de la UT que reportaban valores de inyección muy superiores a la capacidad nominal total de los generadores de la central. Dichos valores fueron suprimidos, ya que no representaban un evento real de operación sino un error de digitación o de consolidación en los informes automáticos.

Otro caso, era en horas continuas que el nivel de embalse pasaba en una hora por ejemplo en un valor de 180 m s. n. m., a la hora siguiente dicho registro era de 0 y a la hora siguiente volvía a estar en 180 m s. n. m. En casos como ese, es obvio que el nivel del embalse no se puede vaciar instantáneamente. Tal caso indica que en esa hora particular se pasó por alto el registro del dato y lo que se hizo en esos casos es promediar los registros superiores e inferiores para colocar un valor promedio en ese espacio.

Con estas medidas de limpieza se depuró el conjunto de datos, eliminando ruido e inconsistencias y garantizando que la serie final fuera coherente y adecuada para el entrenamiento de los modelos predictivos.

4.1.3. Codificación Binaria de Estados Operativos

Para representar el comportamiento de la central a nivel individual, se construyó una codificación binaria de siete bits, en la que cada posición corresponde a una de las unidades generadoras de la planta (cinco originales y dos de expansión). El valor **1** indica que la unidad estuvo en operación (inyectando potencia) en una hora determinada, mientras que el valor **0** indica que permaneció apagada o sin inyección.

Esta codificación se generó a partir de los **datos de inyección real por unidad del año 2024**, proporcionados directamente por la UT. Dichos registros constituyen una fuente confiable y precisa, pues reflejan de manera directa el estado operativo de cada unidad, sin necesidad de recurrir a inferencias a partir de indisponibilidades o programaciones. De este modo, cada observación horaria quedó asociada a un vector binario que resume qué unidades estuvieron activas.

El orden de precedencia del binario indica que la unidad **más significante es la unidad 7** Por ejemplo, un registro horario con vector **1110000** indica que durante esa hora operaron las unidades 7, 6 y 5. mientras las cuatro restantes permanecieron inactivas. Esta representación binaria resultó fundamental para el entrenamiento del modelo, permitiendo capturar patrones de operación por unidad y aportar granularidad al análisis predictivo.

4.1.4. Procesamiento para Modelo de Predicción Continua

Se utilizaron para entrenamiento los datos íntegros separados del año 2024 posterior al tratamiento. Adicional a ello, se agregó también los datos específicos del año 2024 respecto de inyección total, nivel de embalse y vertido. Estos datos contienen el esqueleto correspondiente del comportamiento de las unidades con el cual se pretende obtener la separación de los datos por unidad compartiendo nada más los datos de inyecciones totales, nivel de embalse y vertido.

4.1.5. Herramientas Utilizadas

La gestión y preparación de datos se realizó con Python y librerías de análisis de datos. Se usó pandas intensivamente para leer CSVs, filtrar por fechas, combinar tablas y manipular series temporales. Operaciones como agrupar por día/mes/año, sumar horas a totales diarios y calcular estadísticas (media, máx., mín.) se facilitaron con pandas y NumPy.

La flexibilidad de Python permitió integrar fácilmente pasos adicionales, como la aplicación de funciones *lambda* personalizadas para generar las etiquetas binarias de estado por unidad. Toda esta preparación se realizó en un entorno interactivo tipo Jupyter Notebook, que facilitó documentar cada paso y visualizar resultados intermedios.

Para las visualizaciones (gráficos de líneas, barras, histogramas, etc.) y para el desarrollo de los modelos de *machine learning* se emplearon librerías como Matplotlib, Seaborn y scikit-learn.

Adicionalmente, la implementación y entrenamiento de los modelos se realizó en la plataforma de Microsoft Azure, aprovechando su ecosistema de servicios en la nube. Azure permitió disponer de máquinas virtuales escalables con capacidad de cómputo suficiente

para procesar grandes volúmenes de datos horarios, además de integrar de forma nativa bibliotecas y entornos de Python. Entre las ventajas más relevantes se destacan:

1. **Escalabilidad y rendimiento**: posibilidad de ejecutar cargas de trabajo intensivas en paralelo, reduciendo tiempos de entrenamiento.

2. **Azure Machine Learning (AML):** servicio especializado para la gestión de experimentos de aprendizaje automático, con herramientas de control de versiones, gestión de modelos e implementación rápida de pipelines.

3. **Compatibilidad con notebooks y librerías estándar:** integración directa con Jupyter y soporte para frameworks de *machine learning* como scikit-learn, TensorFlow y PyTorch.

4. **Seguridad y disponibilidad:** altos estándares de ciberseguridad y respaldo, fundamentales para proyectos que trabajan con datos estratégicos del sector energético.

5. **Servicios orientados a datos energéticos e IoT:** módulos preparados para manejar series temporales y flujos de datos industriales, lo que facilitó la adaptación al contexto eléctrico.

La combinación del ecosistema Python con la infraestructura de Azure permitió contar con un entorno robusto, reproducible y seguro para el desarrollo del proyecto, asegurando tanto la eficiencia computacional como la trazabilidad de los experimentos.

4.2 Descripción de los Algoritmos

Con el conjunto de datos consolidado, el siguiente paso fue realizar un análisis exploratorio para comprender tendencias, relaciones y características importantes de la operación de la central. Este análisis preliminar permitió extraer intuiciones y guiar el desarrollo de los modelos posteriores.

Posterior al tratamiento de los datos, se procedió a calcular el total horario de la generación de los datos separados para poder validar que dicha totalidad se correspondiera con los totales publicados en los registros públicos de la UT. Se pudo validar que los datos separados de las 7 unidades eran inconsistentes en las 24 horas de los primeros 5 días (del 1 al 5 de enero) dado que el total obtenido allí y el total publicado por la UT era diferente por mucho y al validar el dato individual en ocasiones duplicaba la capacidad nominal de cada unidad por lo cual se concluyó que los datos de esos días no eran consistentes y podían ser perjudiciales en el análisis por lo cual fueron eliminados y se trabajó precisamente con los datos desde el 6 de enero hasta el 31 de diciembre.

Otro aspecto explorado fue la influencia directa de las condiciones hídricas sobre la capacidad de generación de la central. Para ello, se analizaron conjuntamente los datos de nivel del embalse y las inyecciones horarias.

El análisis de las gráficas permitió identificar patrones claros asociados a la disponibilidad del recurso hídrico:

- En la **temporada seca**, los niveles del embalse tienden a descender y con ello la generación de la central se reduce de forma evidente.

- En la **temporada lluviosa**, los niveles del embalse se elevan, lo que permite un mayor margen de generación y, en algunos casos, la necesidad de realizar **vertidos controlados** cuando se alcanzan cotas críticas de almacenamiento.

A continuación se observa una gráfica que contiene los niveles de embalse promedios diarios a través del tiempo desde el año 2021 hasta el 2024 considerando la central de Guajoyo, Cerrón Grande, 5 de noviembre y 15 de septiembre para que se pueda observar las diferencias en el nivel en cada una. A manera de resumen, tanto en Guajoyo como en Cerrón Grande, la temporada seca y la temporada lluviosa puede notarse claramente.

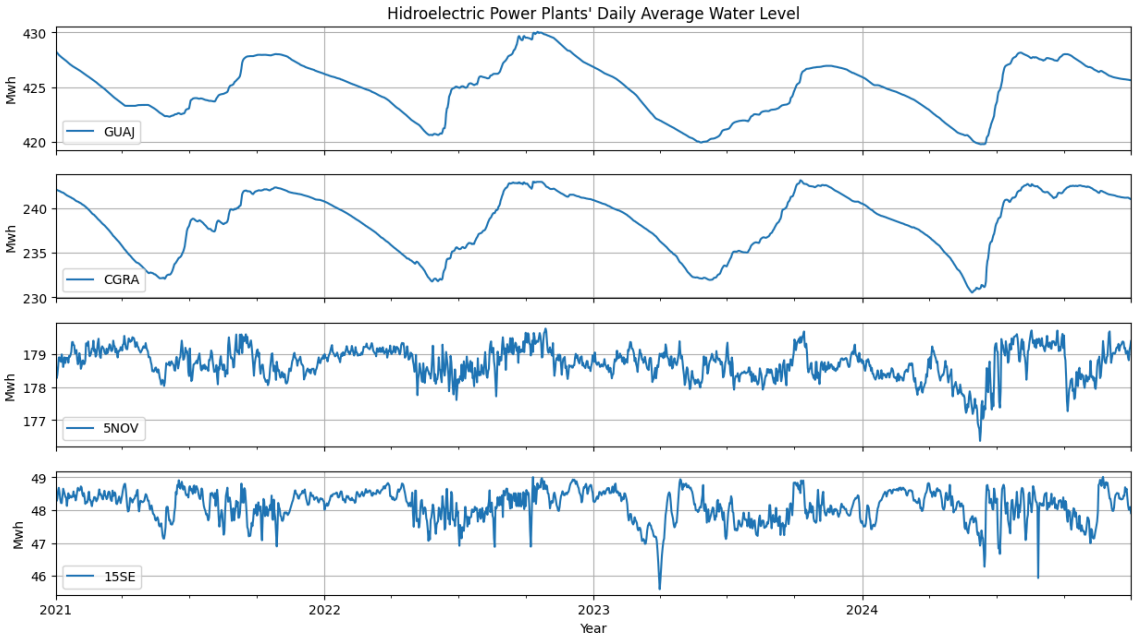


Figura 4: Niveles de embalse de las diferentes hidroeléctricas a nivel nacional en los años 2021-2024.

En la central 5 de noviembre es menos evidente debido a que dicho embalse es pequeño y depende de diversidad de variables. Entre ellas, la más importante, es mantener el embalse de 5 de noviembre alto debido a su servicio de reserva secundaria (AGC).

Adicionalmente, la revisión de las series de tiempo y diagramas de dispersión mostró una correlación positiva general entre el nivel del embalse y la potencia generada, aunque no lineal. En particular, se observó que al superar ciertos niveles de embalse, la planta no necesariamente incrementa más su generación, sino que opta por realizar descargas para mantener la seguridad de la infraestructura.

Estos hallazgos confirman que el nivel del embalse es una variable fundamental en la operación de la central, tanto para explicar las fluctuaciones en la generación real como para justificar la inclusión del **vertido** como un insumo relevante en los modelos predictivos.

4.2.1. Modelo de Clasificación de Estados Operativos

A continuación, en la Figura 2 se presenta el flujo de trabajo seguido para el modelo de clasificación multiclase, donde se ilustran las etapas de codificación, entrenamiento, predicción y postprocesamiento.

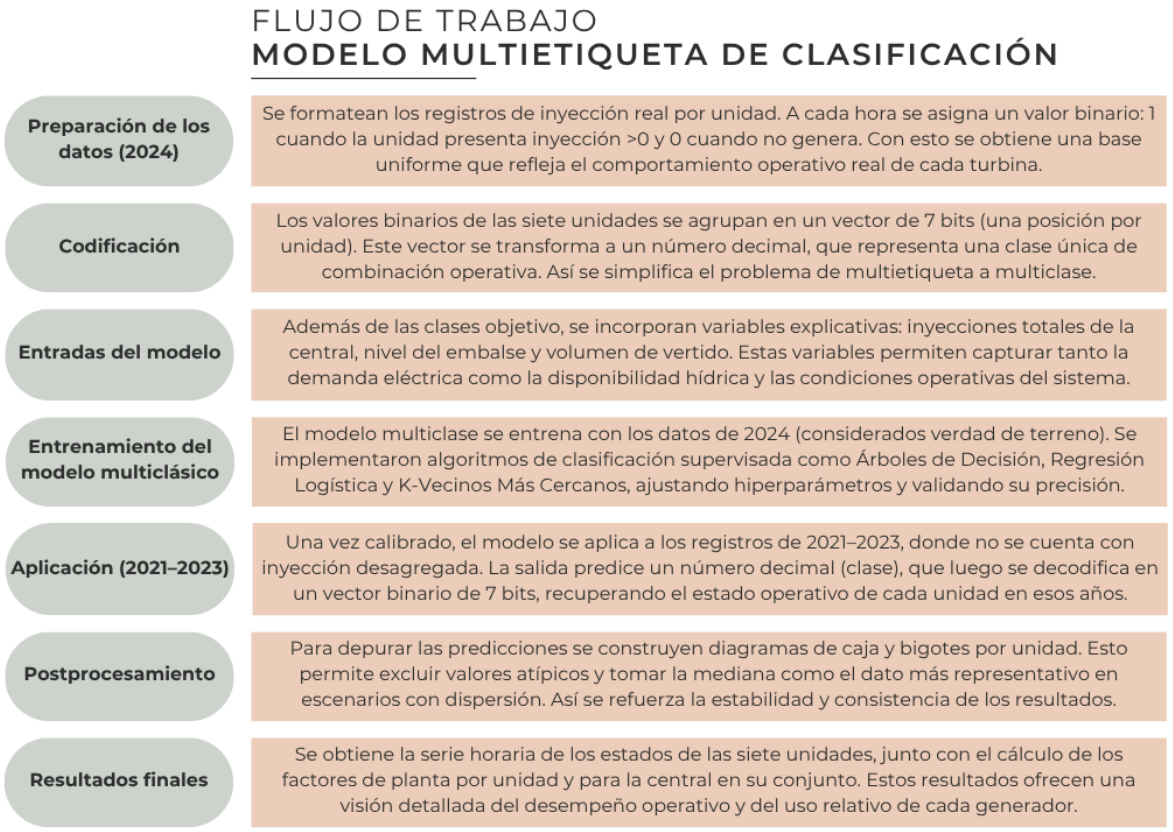


Figura 5: Flujo de trabajo del modelo multietiqueta de clasificación.

Para el desarrollo del modelo predictivo fue necesario establecer un esquema de representación uniforme del estado de operación de las siete unidades generadoras de la central 5 de Noviembre. El objetivo de esta codificación es transformar los registros horarios en un formato que permita a los algoritmos de aprendizaje automático identificar patrones y combinaciones de funcionamiento de manera eficiente.

El procedimiento seguido fue el siguiente:

1. Para cada hora del registro histórico, se construyó un vector de 7 bits en el que cada posición corresponde a una unidad generadora.
2. El valor **1** indica que la unidad estuvo operativa (inyectando energía a la red en esa hora).
3. El valor **0** indica que la unidad permaneció apagada o no generó.
4. Ejemplo: el vector `1110000` indica que las unidades 7, 6 y 5 estuvieron activas mientras que las cuatro restantes permanecieron apagadas.
5. Una vez obtenido el vector binario, este se transformó en un número decimal que representa de forma compacta la combinación de unidades operativas en cada hora.

6. Por ejemplo, el vector binario `1110000` se convierte en el número decimal 112, que se corresponde con que la unidad 7, 6 y 5 estuvieron activas a esa hora mientras que `0000111` se convierte en 7, indicando que las unidades 3, 2 y 1 estuvieron activas a esa hora.
7. De esta manera, cada posible combinación de unidades corresponde a una clase distinta dentro de un problema de clasificación multiclase.

4.2.1.1. Entrenamiento y predicción multiclase

1. Los algoritmos de clasificación (Bosques aleatorios, árboles de decisión, regresión logística y k-vecinos más cercanos) fueron entrenados para predecir directamente el número decimal asociado a cada hora.
2. Generado el modelo, se introducen datos de inyección, nivel de embalse y vertido de un rango que no se corresponde con los datos de entrenamiento, en este caso valores de los años 2021 a 2023 respectivamente y la salida obtenida es ese número decimal que posteriormente se reconvirtió al vector binario correspondiente, obteniendo así el estado de cada unidad generadora. De esta manera obtenemos la estimación del registro de unidades en operación en la hora específica para los rangos en los que no se tiene la inyección separada.
3. Con el fin de depurar las salidas y obtener una representación comparativa con los datos reales de inyección total, se desarrolló un diagramas de caja y bigotes en los datos separados reales de 2024 para conocer el resumen estadistico en cuanto a los cuartiles y mediana para elegir el mas representativo para cada unidad.
4. Se realizó pruebas del cuartil 1, 3 y la mediana multiplicando el dato correspondiente de cada unidad por su respectivo binario de 7 bits. Con esto se obtuvo la tabla completa de generaciones estimadas por separado. Se procedió a realizar la sumatoria de estos datos estimados para compararlos con el valor total de inyecciones reales para visualizar las curvas generadas y se obtuvo el siguiente resultado.

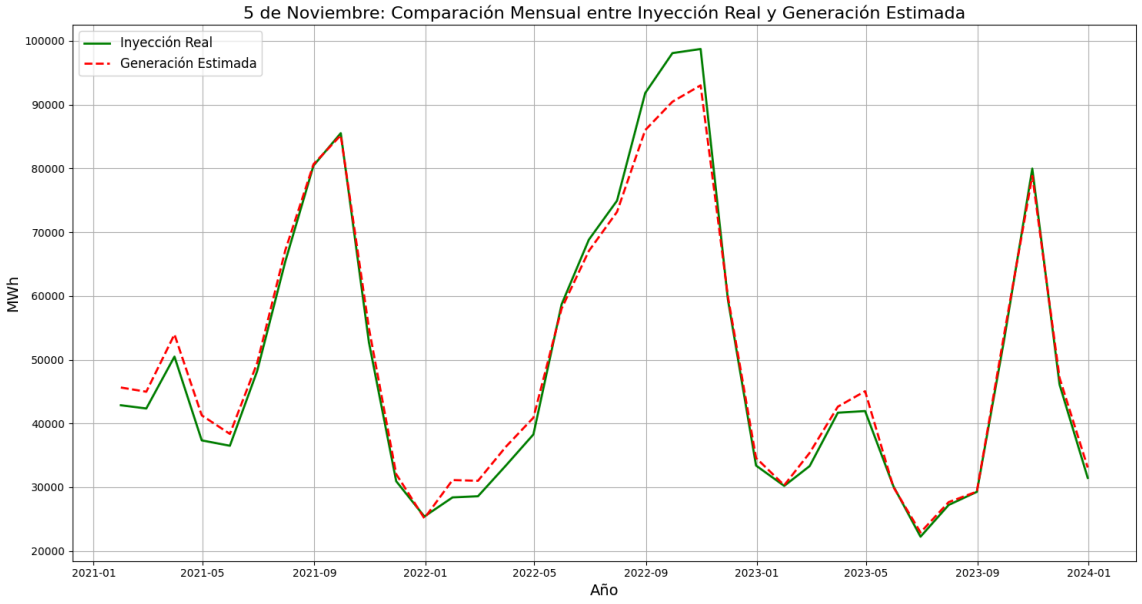


Figura 6: Inyección real Vs.Generación estimada en el modelo multietiqueta de clasificación.

Como se puede observar, la gráfica verde se corresponde con el valor real de inyecciones promedio y la curva roja se corresponde con la generación estimada con el modelo multietiqueta de clasificación. Haciendo uso del coeficiente de determinación, La gráfica de Generación estimada, se acerca a la de Inyección real en un **98.61%** y el mejor resultado, se correspondió con el algoritmo “Bosques aleatorios”.

A continuación se muestran todos los parámetros que demuestran la precisión del algoritmo “bosques aleatorios” el cual mostró un resultado más cercano en rendimiento respecto de regresión logística, El vecino más cercano y Árbol de decisiones.

Unidad	Accuracy	Precision	Recall	F1_score
5nov-u7	0.9179	0.9161	0.9096	0.9126
5nov-u6	0.8982	0.8992	0.8828	0.8897
5nov-u5	0.8213	0.7901	0.7666	0.7765
5nov-u4	0.8109	0.7343	0.7141	0.7230
5nov-u3	0.8618	0.8297	0.8047	0.8158
5nov-u2	0.8918	0.8767	0.8602	0.8678
5nov-u1	0.8959	0.8803	0.8747	0.8774

Tabla 3: Métricas de confiabilidad del algoritmo “bosques aleatorios”.

Este enfoque de **codificación binaria–decimal** permitió transformar el problema original de múltiples etiquetas simultáneas en un problema de clasificación multiclase más manejable, garantizando además que la predicción refleje de manera fiel las combinaciones reales de operación de la central hidroeléctrica.

4.2.2. Modelo Predictivo de Inyección Continua

En la Figura 3 se muestra el diagrama de flujo correspondiente al Modelo de predicción continua por unidad, que sintetiza el proceso de entrenamiento con los datos de 2024 y su posterior aplicación a los años 2021–2023 para estimar la inyección horaria por unidad.

Además del modelo de clasificación multiclase, se desarrolló un modelo de regresión continua con el propósito de estimar directamente la potencia horaria de cada unidad generadora. A diferencia del enfoque discreto basado en combinaciones de estados, este modelo trabaja con valores continuos en megavatios (MW), lo que permite aproximar la contribución individual de cada turbina de manera más precisa y detallada y de la misma manera, desglosada de manera horaria para los años 2021 a 2023. Este esquema es fundamental para reconstruir series históricas de inyección por unidad en periodos donde no se dispone de registros desagregados, manteniendo coherencia con las condiciones hídricas y la planificación operativa.

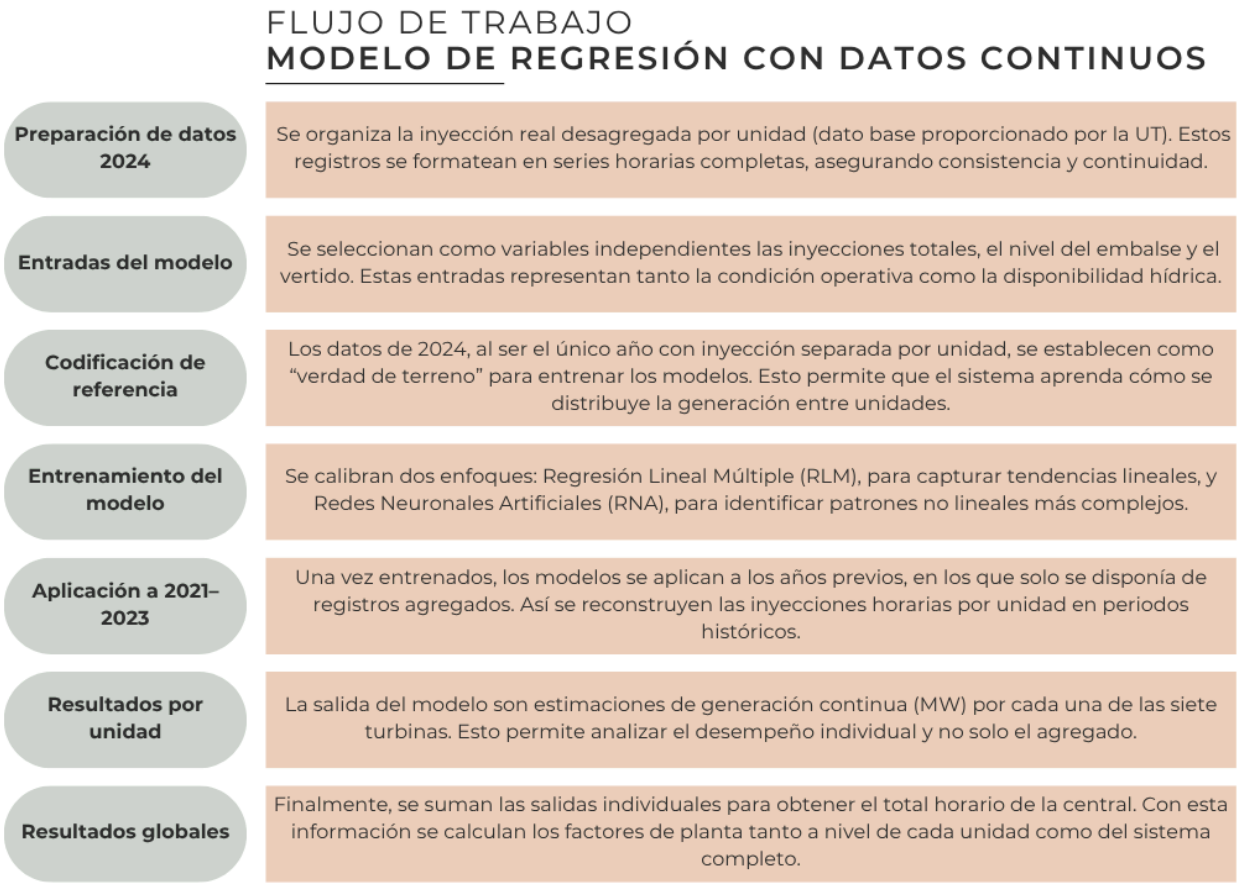


Figura 7: Flujo de trabajo del modelo de regresión con datos Continuos.

El objetivo de este modelo es de igual manera estimar directamente la potencia horaria por unidad (MW), sin discretizar combinaciones de estados. En este caso se incluyen directamente los datos de inyección total horaria, nivel del embalse y vertido y la inyección desagregada por unidad de 2024 se usó como verdad de terreno para entrenar.

Los tipos de algoritmos utilizados fueron regresión lineal múltiple (RLM) como línea base y redes neuronales artificiales (RNA) para relaciones no lineales. El modelo se entrenó con todos los datos antes mencionados en la fracción correspondiente al año 2024, posteriormente se aplicaron los datos disponibles de 2021–2023 como entradas los cuales son Inyecciones reales totales, Nivel de embalse y Vertido para reconstruir la inyección horaria por unidad.

Las métricas correspondientes Métricas. MAE, RMSE y coeficiente de determinación (R^2), calculados sobre particiones de 2024 y empleados luego como criterio de consistencia al extrapolar a 2021–2023.

Al obtener los resultados separados de las unidades, se procedió a realizar la sumatoria de las inyecciones estimadas en ambos modelos para poder compararlas con las inyecciones reales y el resultado fue el siguiente.

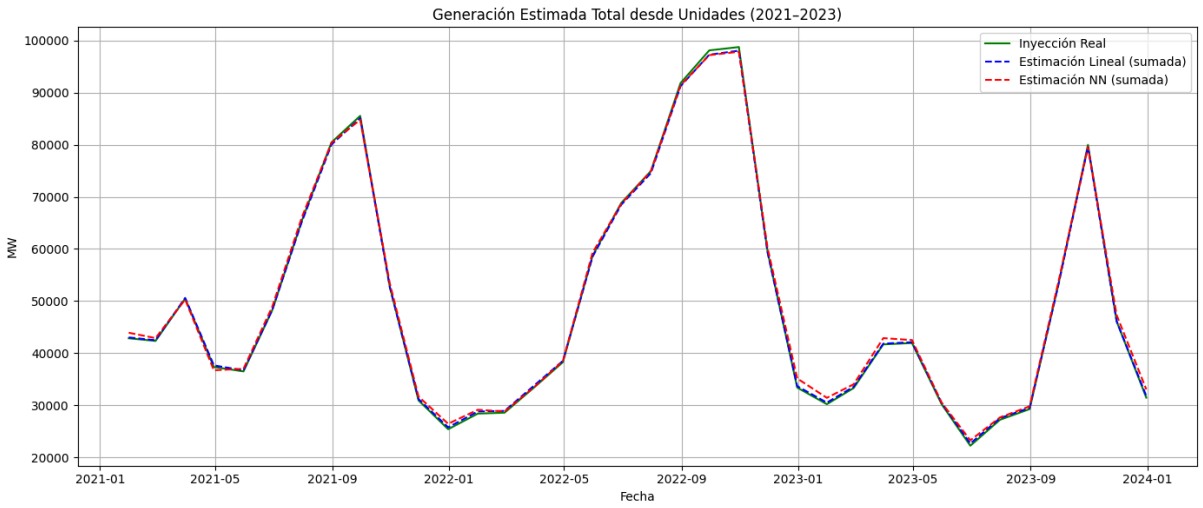


Figura 8: Inyecciones reales vs sumatoria de estimaciones totales obtenidas tanto con el algoritmo de redes neuronales como en el de regresión lineal múltiple.

Al igual que en el modelo anterior, las gráficas tanto de la totalidad de las estimaciones con regresión lineal como las de Redes neuronales, se comparan con el total real de inyecciones. La gráfica verde se corresponde con el valor real de inyecciones promedio y la curva roja se corresponde con la generación estimada con el redes neuronales, por otro lado la gráfica azul se corresponde con la estimación de regresión lineal múltiple.

Al comparar las curvas con la de Inyecciones reales, utilizando también el coeficiente de determinación, la curva obtenida con el algoritmo de redes neuronales esta se corresponde en un **99.87%** sin embargo la curva de regresión lineal múltiple, se corresponde a la de inyecciones reales en un **99.98%**.

En resumen, el mejor algoritmo por su cercanía en los resultados a la curva real de inyecciones es la de Regresión Lineal Múltiple.

En la tabla 4 se muestran los resultados de las métricas de los resultados correspondientes (MSE, RMSE y MAE) tanto al análisis utilizando redes neuronales como regresión lineal múltiple en ambos casos se realiza una valoración por unidad de cuál fue el mejor modelo basado en todos los parámetros y principalmente en el coeficiente de determinación. Si este anda entre 0 y 1, los valores son favorables en esa escala (0 inferior y 1 superior).

Unidad	Modelo	MSE	RMSE	MAE
5nov-u1	Lineal	58.97	7.68	4.05
5nov-u1	Neuronal	17.47	4.18	2.26
5nov-u2	Lineal	44.94	6.70	4.39
5nov-u2	Neuronal	368.17	19.19	3.09
5nov-u3	Lineal	66.46	8.15	4.72

5nov-u3	Neuronal	31.02	5.57	3.05
5nov-u4	Lineal	133.20	11.54	4.29
5nov-u4	Neuronal	392.43	19.81	3.78
5nov-u5	Lineal	44.18	6.65	4.98
5nov-u5	Neuronal	700.87	26.47	4.43
5nov-u6	Lineal	300.14	17.32	7.74
5nov-u6	Neuronal	510.46	22.59	4.77
5nov-u7	Lineal	353.33	18.80	7.29
5nov-u7	Neuronal	103.16	10.16	3.97

Tabla 4: Métricas de confiabilidad de los algoritmos de redes neuronales y regresión lineal múltiple.

El RMSE representa el error promedio en las mismas unidades que la variable analizada, facilitando su interpretación operativa. En este análisis, se observaron valores de RMSE inferiores a 5 en las unidades 5nov-u1 y 5nov-u3 bajo el modelo neuronal, lo cual indica predicciones relativamente precisas.

En contraste, las unidades 5nov-u4, 5nov-u5 y 5nov-u6 presentaron valores de RMSE superiores a 19, alcanzando incluso valores mayores a 26 en la unidad 5nov-u5, lo que implica desviaciones significativas entre los valores reales y los estimados.

Estos altos valores de RMSE reflejan una limitada capacidad del modelo para representar adecuadamente el comportamiento dinámico de estas unidades, probablemente asociada a factores externos no incluidos en el conjunto de variables de entrada, como mantenimientos, restricciones operativas o decisiones de despacho.

4.3. Validación y Análisis Comparativo

La validación es clave para comprobar la precisión y la capacidad de generalización de los modelos. En esta investigación se aplicó un esquema de validación temporal: los modelos se entrenaron con los datos desagregados de 2024 (considerados la “verdad de terreno”) y luego se aplicaron a los años 2021–2023, donde solo se disponía de registros agregados.s

Para el modelo de clasificación multiclase se utilizó la métricas de la exactitud global por unidad y el coeficiente de determinación. En el modelo de predicción continua se emplearon indicadores de error como el MAE, el RMSE y el coeficiente de determinación R^2. Este enfoque permitió evaluar tanto la precisión puntual como la coherencia de las curvas estimadas frente a las reales, asegurando resultados consistentes y comparables entre ambos modelos.

4.3.1. Validación del Modelo de Clasificación Multiclase

El modelo se entrenó con los datos desagregados de 2024 y se aplicó de forma retrospectiva a los años 2021–2023. Para evaluar su desempeño se utilizaron métricas estándar en problemas multiclase: exactitud global y el coeficiente de determinación.

Los árboles de decisión y los bosques aleatorios demostraron ser los modelos más efectivos, logrando un equilibrio óptimo entre precisión y facilidad de interpretación. Por otro lado, la regresión logística y el KNN, aunque útiles para comparaciones, mostraron limitaciones al intentar modelar relaciones no lineales. La claridad inherente a los árboles de decisión facilitó la extracción de reglas operativas concretas, como la activación de las unidades de expansión (U6 y U7) en situaciones de alta demanda o cuando los niveles del embalse excedían ciertos umbrales.

4.3.2. Validación del Modelo de Predicción Continua

El modelo se entrenó con los datos de 2024 y se aplicó posteriormente a los registros de 2021–2023, utilizando validación cruzada temporal para garantizar consistencia. El desempeño se evaluó mediante RMSE, MAE y el coeficiente de determinación R^2.

Los resultados mostraron un ajuste adecuado entre las curvas reales y predichas, con errores bajos en la mayoría de horas y discrepancias más notorias en eventos extremos de generación. La regresión lineal múltiple (RLM) ofreció una línea base sólida, mientras que las redes neuronales artificiales (RNA) capturaron mejor las relaciones no lineales y presentaron mayor capacidad de generalización.

4.3.3. Comparación de Algoritmos

En la clasificación multiclase, los métodos basados en árboles (árboles de decisión y Random Forest) ofrecieron un mejor balance entre precisión y capacidad explicativa, superando a la regresión logística y a KNN. Los árboles, además, permitieron identificar reglas operativas interpretables, lo que resulta valioso en un contexto energético.

En el modelo continuo, las redes neuronales artificiales (RNA) mostraron un desempeño superior en escenarios no lineales, capturando relaciones complejas entre variables. Sin embargo, la regresión lineal múltiple (RLM) se mantuvo como una línea base sólida y fácilmente interpretable, lo que la hace útil para análisis comparativos y validaciones rápidas.

En conjunto, la elección del algoritmo depende del objetivo: maximizar la precisión con RNA y Random Forest, o priorizar la interpretabilidad con RLM y árboles de decisión. Este equilibrio entre exactitud y explicabilidad fue clave para la propuesta metodológica de este trabajo.

4.3.4. Integración y Coherencia de Resultados

Para garantizar la validez práctica de las predicciones, se verificó la coherencia entre los dos modelos desarrollados: el de clasificación multiclase (estados de las unidades) y el continuo (MW generados). La integración se basó en reglas de consistencia: si el modelo

continuo estimaba una generación superior a la capacidad disponible según las unidades activas, la predicción se ajustaba a la baja. De forma inversa, cuando el modelo multiclase anticipa la entrada de más unidades, se redistribuye la generación total estimada para reflejar esa condición.

De esta manera, las series finales de potencia por unidad resultaron físicamente plausibles, coherentes con la capacidad instalada y representativas del comportamiento operativo real de la central.

4.3.5. Análisis de Utilidad Práctica

Más allá de su desempeño técnico, los modelos desarrollados aportan valor en la interpretación operativa de la central hidroeléctrica 5 de Noviembre. La capacidad de estimar tanto el estado horario de las unidades como la inyección continua de potencia permite identificar patrones de utilización de cada generador y comprender cómo responden a cambios en la demanda o en las condiciones hídricas.

Este nivel de detalle ofrece una conexión directa con métricas del sector eléctrico, en particular con el factor de planta, indicador clave para evaluar el aprovechamiento de la infraestructura. El análisis de factores por unidad, derivado de los modelos, evidencia el papel de las unidades base frente a las de respaldo y ayuda a planificar estrategias de operación más eficientes.

Finalmente, los resultados no solo describen el comportamiento histórico, sino que también brindan insumos para la planificación y gestión de recursos hídricos y energéticos, favoreciendo decisiones orientadas a mejorar la confiabilidad y eficiencia del sistema eléctrico nacional.

El Capítulo 4 delimitó el universo de datos, documentó su limpieza e integración temporal y presentó los algoritmos empleados —clasificación multiclase y predicción continua— junto con sus criterios de entrenamiento y validación. A partir de esta base técnica, el Capítulo 5 traslada el foco desde el cómo se modela hacia el qué significan los resultados: primero, evalúa rigurosamente el desempeño de los modelos (exactitud, MAE, RMSE, R2R^2R2); luego, interpreta operativamente las predicciones para caracterizar el rol funcional de cada unidad (base, secundaria, pico/reserva) y su factor de planta; finalmente, explora proyecciones y líneas de investigación futura orientadas a optimización hídrica y escenarios de despacho. Esta transición marca el paso de la construcción metodológica a la extracción de conocimiento accionable para la gestión de la Central 5 de Noviembre.

Capítulo 5. Conclusiones y líneas futuras.

El presente capítulo expone los resultados derivados de los modelos predictivos desarrollados, evaluando su precisión técnica y coherencia operativa frente a los datos reales de la Central Hidroeléctrica 5 de Noviembre. Los análisis se centran en el período 2021–2023, utilizando como referencia los registros verificados de 2024 para validar las predicciones.

Se analizan las métricas de desempeño obtenidas —exactitud promedio del 86.9% para el modelo de clasificación multiclase y un coeficiente de determinación de $R^2 = 0.9998$ para el modelo de regresión lineal—, demostrando la fidelidad de los resultados. A partir de estas predicciones se reconstruye el comportamiento histórico de las siete unidades generadoras, permitiendo calcular por primera vez su factor de planta individual y rol operativo (base, secundaria o de pico).

Finalmente, el capítulo introduce proyecciones exploratorias orientadas a la optimización de la eficiencia hídrica, la reducción de vertidos y la simulación de escenarios de despacho alternativo. De esta forma, se consolidan los hallazgos del estudio y se abren nuevas líneas de investigación aplicables a la gestión energética nacional.

5.1. Evaluación del Desempeño Predictivo y Fidelidad de los Modelos

La credibilidad de cualquier análisis derivado de modelos predictivos depende fundamentalmente de su capacidad para replicar con precisión la realidad. Esta sección se dedica a una evaluación rigurosa del desempeño de los modelos de clasificación y regresión, no sólo presentando métricas de rendimiento, sino también interpretando sus implicaciones sobre la naturaleza de la operación de la planta.

El primer modelo se diseñó para resolver una pregunta fundamental: en una hora determinada, ¿qué combinación de unidades generadoras estaba activa? Para ello, se entrenó un modelo de clasificación multiclase utilizando los datos desagregados de 2024 como verdad de terreno. Entre los algoritmos evaluados, los métodos basados en árboles –en particular *Random Forest* (Bosque Aleatorio)– demostraron una capacidad superior para capturar la lógica de despacho de la central.

5.1.1 Precisión en la Identificación de Estados Operativos (Modelo de Clasificación Multiclase)

El primer modelo se diseñó para resolver una pregunta fundamental: Para un momento determinado, ¿qué combinación de unidades generadoras estaba activa? Para ello, se entrenó un modelo de clasificación multiclase utilizando los datos desagregados de 2024

como verdad de terreno. Entre los algoritmos evaluados, los métodos basados en árboles –en particular *Random Forest* (Bosque Aleatorio)– demostraron una capacidad superior para capturar la lógica de despacho de la central.

La **Tabla 5** presenta un resumen consolidado del desempeño de varios algoritmos de clasificación. El modelo *Random Forest* alcanzó una exactitud promedio del **86.96%** en la predicción del estado (activo/inactivo) de todas las unidades, superando a los demás algoritmos probados. Este alto nivel de precisión general confirma la viabilidad de utilizar variables agregadas (inyección total, nivel de embalse y vertido) para inferir el estado operativo a detalle.

Unidad	Decision Tree	Random Forest	Regresión Logística	KNN
U1	0.8652	0.8976	0.8861	0.8866
U2	0.8635	0.8936	0.8930	0.8832
U3	0.8479	0.8623	0.8329	0.8577
U4	0.7825	0.8057	0.7611	0.8068
U5	0.8022	0.8149	0.7901	0.8097
U6	0.8566	0.8971	0.9115	0.8982
U7	0.8809	0.9161	0.9121	0.9092
Promedio	**0.8427**	**0.8696**	**0.8624**	**0.8646**

Tabla 5: Exactitud (accuracy) por unidad y modelo de clasificación supervisada (validación con datos de 2024).

Un análisis detallado de los resultados por unidad revela patrones significativos en la exactitud, sirviendo como indicador de la regularidad operativa. Las Unidades 7 y 1, con una precisión del 91.61% y 89.76% respectivamente, demuestran roles operativos altamente consistentes y fuertemente correlacionados con las variables de entrada. Esto sugiere que ambas unidades siguen reglas de despacho claras y bien definidas, operando continuamente (carga base) o activándose bajo condiciones específicas y predecibles (por ejemplo, superando un umbral crítico de nivel del embalse).

En contraste, la Unidad 4 presenta la exactitud más baja (80.57%). Este resultado no indica un fallo del modelo, sino que su operación es más compleja, estocástica o dependiente de factores externos a las variables de entrada. Su activación podría responder a necesidades dinámicas del sistema eléctrico, como la regulación secundaria de frecuencia (AGC) o requerimientos de despacho no programados, que son menos predecibles a partir de datos de inyección total y niveles hídricos. De este modo, la métrica de exactitud se convierte en un indicador de la complejidad y regularidad del rol operativo de cada activo.

La superioridad de los métodos basados en árboles sobre modelos lineales, como la Regresión Logística, confirma la naturaleza no lineal de la lógica de despacho de la Central

5 de Noviembre. Existen interacciones complejas y umbrales no lineales que determinan la activación de las unidades, como la dependencia de una combinación no lineal de la demanda y la tasa de cambio del nivel del embalse para el encendido de una unidad de pico. El modelo Random Forest, al ser un ensamblaje de árboles de decisión, es capaz de modelar estas relaciones complejas, lo que explica su robusto desempeño.

Por otro lado, en la tabla 3, podemos validar las características propias del modelo random forest en las cuales pueden notarse los siguientes parámetros.

Para la unidad 7, tanto el rendimiento como el puntaje del modelo supera el 91.3%. lo que indica que para esta unidad si se predice correctamente el encendido, es estable y el patrón operativo es claro.

Para la unidad 6, tanto el rendimiento como el puntaje del modelo supera el 88.9%. lo que indica que para esta unidad si se predice correctamente el encendido, es estable y el patrón operativo es claro a pesar que es inferior que en la unidad 7.

Para la unidad 5, tanto el rendimiento como el puntaje del modelo supera el 87.7%. lo que indica que es bastante estable y que tiene un buen comportamiento.

Para la unidad 4, apenas es de 72.2% lo cual indica que los patrones de esta unidad son erráticos y en ellos existe mayor incertidumbre.

Para la unidad 3, tanto el rendimiento como el puntaje del modelo supera el 81.5%. lo que indica que es un poco más variable que las unidades mayores (7, 6 y 5) y esto indica que requiere más datos para un mejor abordaje.

Para la unidad 2, tanto el rendimiento como el puntaje del modelo supera el 86.7%. lo que indica que es más estable que las 3 y 4 lo que indica que es aún usable pero posee un leve margen de error.

Para la unidad 1, tanto el rendimiento como el puntaje del modelo supera el 87.7%. lo que indica que es bastante estable y tiene un buen comportamiento al igual que la unidad 5.

5.1.2. Fidelidad en la Estimación de Inyección Continua (Modelo de Regresión)

El segundo modelo abordó el problema desde una perspectiva continua, buscando estimar la inyección de potencia (en MW) de cada unidad para cada hora. Se probaron dos enfoques: Regresión Lineal Múltiple (RLM) y Redes Neuronales Artificiales (RNA). Los resultados, visualizados en la Gráfica 4 del capítulo anterior, fueron excepcionales y revelaron una característica fundamental de la operación de la planta.

El modelo de RLM, el más simple de los dos, logró un coeficiente de determinación R^2 de 0.9998 al comparar la suma de sus predicciones individuales con la inyección total real de la planta para el período 2021-2023. La RNA obtuvo un resultado ligeramente inferior, con un

R^2 de 0.9987. Este hallazgo –donde un modelo más simple supera a uno más complejo– es profundamente revelador.

La predicción del modelo lineal sugiere que la lógica de despacho agregada de la planta es, altamente lineal y aditiva. La inyección total de la central en un momento dado puede ser explicada como una suma ponderada de las variables de entrada (inyección total programada, nivel del embalse, vertido) sin necesidad de modelar interacciones no lineales complejas a nivel agregado. Esto implica que no existen efectos sinérgicos o supresores significativos entre las unidades que alteren drásticamente la relación lineal; la producción total es, en esencia, la suma de sus partes. El rendimiento ligeramente inferior de la RNA podría atribuirse a un leve sobreajuste a los patrones específicos de los datos de entrenamiento de 2024, mientras que la RLM capturó la relación fundamental, más simple y generalizable, que gobierna el sistema.

Esta conclusión tiene implicaciones que van más allá de la simple validación del modelo. Sugiere que el modelo de RLM no es solo una herramienta predictiva, sino que funciona como un "gemelo digital" de la lógica de despacho de los operadores de la planta. Los coeficientes del modelo lineal para cada unidad pueden ser interpretados como una aproximación cuantitativa de las reglas operativas: revelan el peso relativo que tienen la demanda total y la disponibilidad hídrica en la decisión de cuánta potencia despachar desde cada generador. Esto abre una vía directa e interpretable para el análisis y la optimización de la estrategia operativa, sugiriendo que las mejoras pueden provenir del análisis de estas reglas lineales fundamentales en lugar de la implementación de sistemas de control de inteligencia artificial más complejos.

5.1.3. Análisis de Coherencia Global y Magnitud del Error

Para contextualizar el desempeño de los modelos en términos prácticos, es crucial analizar la magnitud del error absoluto. Se calcularon el Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE) para las predicciones agregadas de ambos modelos frente a los datos reales de inyección total para el período 2021-2023 los cuales se detallan en la tabla 4.

El modelo de RLM, consistente con su mayor R^2, mostró un MAE consistentemente bajo, indicando que, en promedio, sus predicciones horarias se desvían muy poco del valor real. Es fundamental contextualizar este error en relación con la capacidad total de la planta, que es de 179.4 MW [20]. Por ejemplo, un MAE de 3 MW representaría una desviación promedio de sólo el 1.67% de la capacidad total, un nivel de error que es altamente aceptable para la mayoría de las aplicaciones de planificación y análisis operativo.

El análisis del RMSE proporciona una perspectiva complementaria, ya que penaliza más fuertemente los errores grandes. Una comparación entre el MAE y el RMSE permite evaluar la capacidad del modelo para predecir eventos extremos. Si el RMSE es significativamente mayor que el MAE, indicaría la presencia de errores grandes ocasionales, sugiriendo que el modelo puede tener dificultades para predecir con precisión picos de generación muy agudos o caídas repentinas. Este análisis es vital para entender los límites de fiabilidad del modelo en condiciones operativas atípicas y para informar a los usuarios finales sobre los escenarios en los que las predicciones deben ser tratadas con mayor cautela. En conjunto,

el bajo error promedio y un RMSE controlado confirman la robustez y la utilidad práctica de los modelos para reconstruir el historial operativo de la planta con un alto grado de confianza.

En cuanto al rendimiento propiamente del modelo para cada unidad, en la tabla 4, pueden valorarse los elementos para cada parámetro por unidad los cuales se detallan a continuación.

5.2. Determinación de Roles Operativos a partir del Factor de Planta Estimado

El factor de planta, definido como la relación entre la energía realmente generada y la energía máxima que podría haberse generado operando a plena capacidad, es un indicador clave para evaluar la utilización de un activo de generación. Al aplicar los modelos para estimar la generación individual de cada unidad durante el período 2021-2023, fue posible calcular por primera vez sus factores de planta individuales.

La Tabla 6 consolida estos cálculos y asigna un rol operativo a cada unidad basado en su patrón de utilización. Los resultados de ambos modelos (regresión y clasificación) son notablemente consistentes y pintan un cuadro claro de una estrategia de despacho diferenciada.

Unidad	Capacidad Nominal (MW)	Factor de Planta Promedio (RLM, 2021-23)	Factor de Planta Promedio (Clasificación, 2021-23)	Rol Operativo Designado
U1	15.0	0.620	0.643	Base
U2	15.0	0.443	0.484	Secundaria
U3	21.4	0.721	0.823	Base (Primaria)
U4	15.0	0.159	0.135	Pico / Reserva
U5	15.0	0.448	0.537	Secundaria
U6	40.0	0.272	0.218	Pico / Reserva
U7	40.0	0.286	0.254	Pico / Reserva

Tabla 6: Caracterización funcional y factor de planta promedio por unidad (Estimación 2021-2023).

La clasificación revela una jerarquía operativa clara:

5.2.1. Unidades Base

Las unidades **U1** y, especialmente, **U3** muestran factores de planta consistentemente altos. La U3, según ambos modelos, opera por encima del 72% del tiempo, llegando a un

impresionante ~94.8% en 2022 según el modelo de clasificación. Por lo cual podríamos inferir que estas son las "Las generadoras más importantes" de la central, responsables de la generación continua y de mantener el flujo mínimo de agua turbinada.

5.2.2. Unidades Secundarias

Las unidades **U2** y **U5** presentan factores de planta moderados, típicamente en el rango de 0.40 a 0.60. Su rol es complementar a las unidades base, activándose para satisfacer aumentos sostenidos de la demanda o cuando las condiciones hídricas permiten una mayor generación general.

5.2.3. Unidades de Pico y Reserva

Las unidades **U4**, **U6** y **U7** muestran factores de planta consistentemente bajos, a menudo por debajo de 0.30. Este es el hallazgo más significativo, ya que identifica su función como activos de uso intermitente, diseñados no para la generación constante de energía (GWh), sino para proporcionar potencia (MW) rápidamente cuando es más necesario.

Este análisis permite unificar dos descripciones aparentemente contradictorias de la planta. Por un lado, información técnica oficial describe la operación de la central como de "hilo de agua" (run-of-the-river), lo que implica una generación constante para turbinar el caudal entrante del río[9]. De hecho, la central tiene un embalse relativamente pequeño, de apenas 87 millones de m 3 útiles[9], y durante décadas se consideró principalmente una planta de operación base. Por otro lado, los datos de los modelos demuestran inequívocamente la existencia de unidades de pico operando de forma intermitente. La resolución de esta paradoja reside en una estrategia operativa híbrida y sofisticada: la planta como sistema cumple con el requisito de "hilo de agua" a través de la operación continua de sus unidades base (U1 y U3) –manteniendo siempre un mínimo de generación base–. Simultáneamente, utiliza la limitada capacidad de almacenamiento de su embalse para modular la producción de sus unidades de pico (U4, U6, U7), proporcionando flexibilidad, capacidad de reserva y servicios auxiliares valiosos para la estabilidad de la red eléctrica nacional.

En otras palabras, la central 5 de Noviembre opera con una estrategia dual: la mayor parte del tiempo actúa como central de base (aprovechando su caudal como una planta de filo de agua), pero cuando las condiciones lo requieren, despacha potencia adicional a través de las unidades de pico. Cabe recordar que tras la expansión de 2017, la capacidad instalada total aumentó de 99.4 MW a 179.4 MW[20] con la incorporación de dos unidades nuevas, lo que prácticamente duplicó el potencial de generación de la central. Esta investigación proporciona la primera evidencia cuantitativa de esta estrategia operativa dual dentro de la Central 5 de Noviembre.

5.3. Análisis Específico de las Unidades de Expansión (U6 y U7)

El planteamiento del problema de esta tesis menciona un estudio previo que sugería una eficiencia operativa de las nuevas unidades (U6 y U7) muy por debajo de lo proyectado, con un factor de capacidad promedio inferior al 10% en lugar del ~18% esperado. Vale la pena

aclarar que ese 18% esperado proviene de la meta de generación anual (~130 GWh) planteada para el proyecto de expansión [21] . Los resultados de este trabajo, aunque confirman el bajo factor de planta en términos absolutos (alrededor del 25-28% en promedio durante 2021-2023 para cada unidad, según la Tabla 5), permiten reinterpretar completamente este hallazgo.
El error radica en evaluar estas unidades con una métrica de energía (factor de planta) cuando su verdadero valor está en la potencia y la flexibilidad que aportan. El proyecto de expansión, que añadió 80 MW de capacidad (dos turbinas de 40 MW cada una), fue financiado con el objetivo explícito de incrementar la oferta de energía limpia y desplazar la generación térmica en picos de demanda[21].

En El Salvador, como en muchos sistemas eléctricos, la generación térmica (principalmente plantas de combustibles fósiles) se utiliza a menudo para cubrir la demanda pico. Por lo tanto, el hecho de que las unidades U6 y U7 operen con un bajo factor de planta no indica un fracaso, sino que confirma su éxito en cumplir un rol estratégico como unidades de pico, exactamente el tipo de función necesaria para sustituir a las plantas térmicas en horas de mayor demanda.

Para visualizar este comportamiento, se propone un análisis correlacional. Imaginemos un diagrama de dispersión que grafique la activación predicha de U6 y U7 (según el modelo de clasificación) contra dos ejes: la inyección total de la planta (como proxy de la demanda del sistema) y el nivel del embalse. Se esperaría que la activación de estas unidades ocurra predominantemente en dos escenarios:

1. **Demanda del sistema muy alta:** Cuando la inyección total de la planta es elevada, indicando que se necesitan U6/U7 para satisfacer la demanda pico nacional (por ejemplo, en horas pico diarias o durante eventos de demanda récord).

2. **Embalse en niveles máximos:** Cuando el nivel del embalse alcanza cotas cercanas al vertedero. En tal caso, U6 y U7 se activan para evitar el vertido (spillage) y convertir en energía útil el agua excedente que de otro modo se habría desperdiciado. De hecho, la propia justificación de la expansión señalaba que con las nuevas turbinas se aprovecharían los excedentes de agua en época lluviosa que antes eran descargados por el vertedero [22].

Este análisis permite reformular la narrativa del proyecto de expansión. No se trató de una inversión de bajo rendimiento, sino de una adición estratégica exitosa que dotó a la red salvadoreña de 80 MW de capacidad de pico limpia y renovable, mejorando la flexibilidad operativa y la seguridad del sistema. La contribución de esta tesis es proporcionar la evidencia basada en datos para sustentar esta reevaluación estratégica, demostrando que U6 y U7 cumplen con creces el objetivo original de su instalación (aportar energía renovable en picos de demanda y reducir la dependencia de generación térmica[21]).

5.4. Síntesis de Hallazgos y Contribución a la Resolución del Problema

La presente investigación propuso determinar si era posible utilizar algoritmos de *Machine Learning* para predecir con una precisión confiable la inyección real de energía por

generador en la Central Hidroeléctrica 5 de Noviembre, utilizando únicamente datos agregados y un conjunto limitado de datos desagregados para entrenamiento. Los resultados presentados en este capítulo confirman de manera concluyente la hipótesis. Los modelos desarrollados no solo lograron predecir el comportamiento individual de las unidades con un alto grado de exactitud cuantificable, sino que también permitieron desvelar la lógica operativa subyacente de la central.

A continuación, se sintetizan los hallazgos clave y las contribuciones principales de este análisis:

5.4.1. Caracterización Funcional de Activos

Por primera vez, se ha caracterizado cuantitativamente el rol operativo de cada una de las siete unidades generadoras de la planta. Se identificó una clara jerarquía compuesta por unidades de carga base (U1, U3), unidades secundarias de seguimiento de carga (U2, U5) y unidades de pico y reserva (U4, U6, U7). Esta clasificación, derivada empíricamente de los datos, proporciona una comprensión sin precedentes de la estrategia de despacho interno de la central, la cual combina operación *base-load* con capacidad de respuesta rápida.

5.4.2. Reevaluación Estratégica de la Expansión

El análisis refuta la noción previa de que las unidades de expansión (U6 y U7) son activos de "bajo rendimiento". Por el contrario, la evidencia demuestra que cumplen un rol estratégico y valioso como unidades de pico, proporcionando flexibilidad y capacidad de reserva que son cruciales para la estabilidad del sistema eléctrico y para el desplazamiento de la generación térmica. Esto está en línea con los objetivos originales del proyecto de expansión de 2017, el cual buscaba aumentar 80 MW de capacidad para desplazar generación térmica y atender el crecimiento de la demanda[21]. Nuestros hallazgos confirman que dichos objetivos se han logrado en la práctica, al menos en cuanto a la funcionalidad aportada por U6 y U7.

5.4.3.Descubrimiento de la Lógica de Despacho

Si bien las decisiones de activación de unidades individuales responden a dinámicas no lineales complejas y a veces difíciles de predecir (especialmente en unidades pico como U4), la operación agregada de la planta sigue un patrón lineal y predecible. La estrategia global de despacho parece basarse en un conjunto de reglas aditivas bien definidas (por ejemplo, “X metros cúbicos de agua disponibles resultan aproximadamente en Y MW distribuidos entre las unidades base, más Z MW si cierta condición se cumple para activar una unidad pico”). Este comportamiento lineal agregado ha sido capturado eficazmente por el modelo de regresión lineal, que actuó como un espejo matemático de la planta. La implicación de este hallazgo es que las mejoras en la operación global podrían alcanzarse optimizando esas reglas lineales fundamentales, antes de considerar intervenciones más complejas.

En última instancia, la contribución fundamental de esta tesis es el desarrollo y la validación de una metodología que transforma exitosamente datos energéticos agregados y públicamente disponibles en inteligencia operacional a detalle y accionable. Este enfoque proporciona una solución escalable y de bajo costo a un problema común en el sector

energético: la falta de transparencia en la operación de activos individuales. Los modelos y análisis aquí presentados no solo resuelven el problema específico de inferir la operación interna de la Central 5 de Noviembre, sino que establecen un precedente metodológico aplicable a otras infraestructuras de generación en El Salvador y en la región.

Al reconstruir virtualmente el comportamiento histórico de cada unidad generadora y revelar la lógica oculta tras el despacho hidroeléctrico, esta investigación aporta herramientas nuevas para la gestión optimizada de recursos hídricos y la planificación energética nacional. La Central 5 de Noviembre, con sus 179.4 MW de capacidad instalada[20] y su doble propósito de generación base y pico, puede ahora ser mejor comprendida y aprovechada en el contexto de la matriz eléctrica salvadoreña, cerrando así la brecha de información que motivó inicialmente este estudio.

5.5. Líneas de Investigación Futura (Proyecciones)

El desarrollo de los modelos presentados en esta investigación abre la posibilidad de realizar análisis complementarios orientados a la optimización operativa y al aprovechamiento integral del recurso hídrico. Si bien los resultados actuales se limitan a la reconstrucción histórica y validación del comportamiento de la central, la solidez de los modelos permite proyectar **escenarios exploratorios** en los que podrían evaluarse nuevas métricas y estrategias.

Estas líneas no constituyen resultados comprobados, sino **hipótesis derivadas** del desempeño observado en los modelos, que apuntan a ampliar el potencial analítico del enfoque propuesto y a generar futuras aplicaciones en planificación energética, eficiencia hídrica y gestión predictiva.

5.5.1. Optimización de la Eficiencia Hídrica y Reducción de Pérdidas

Los modelos permiten cuantificar el **costo de oportunidad del agua vertida**, es decir, el ingreso perdido cada vez que se libera agua sin pasar por las turbinas. La metodología consiste en:

- **Identificar eventos de vertido** en los datos históricos.
- Utilizar el modelo de clasificación para saber **qué combinación de generadores estaba operando** en esos momentos.
- **Cuantificar la energía no generada** y su valor económico según el precio del mercado en esa hora.

Este análisis transforma el modelo en una herramienta de gestión económica. Permite responder preguntas estratégicas como: "¿Cuántos ingresos adicionales se habrían generado si otras unidades hubieran estado activas para aprovechar el exceso de agua?". Así, se pueden tomar decisiones informadas para minimizar el vertido, por ejemplo, justificando la operación de unidades de pico durante la temporada lluviosa aunque la demanda no lo exija estrictamente.

5.5.2. Simulación de Escenarios para un Despacho Óptimo

Los modelos pueden funcionar como un **motor de simulación para explorar escenarios operativos alternativos**. En lugar de solo predecir lo que ocurrió, se puede investigar qué habría pasado con otras estrategias. Por ejemplo:

- Se toma un día histórico y se fija el perfil de generación total que se cumplió.
- Se utilizan los modelos para simular **diferentes combinaciones de unidades** que podrían haber logrado esa misma generación total.
- Se evalúa cada escenario bajo criterios como: **reducción del vertido, equilibrio en el desgaste de las unidades o mejora en la capacidad de reserva** del sistema.

Este enfoque permite pasar de la predicción a la prescripción, sentando las bases para desarrollar un **sistema de recomendación de predespacho** que sugiera a los operadores la combinación óptima de unidades en tiempo real para alcanzar objetivos más amplios que solo cumplir con la demanda.

Bibliografía

[1] C. E. Martínez-Cruz, "Analysis of the expansion of the 5 de Noviembre Hydroelectric Power Plant using ML algorithms," *Escuela de Ingeniería Eléctrica*, Universidad de El Salvador, 2024. [En línea]. Disponible en: https://www.elsalvador.edu.

[2] Invest in El Salvador, *Guía Sectorial Energía 2023*, Invest in El Salvador, 2023. [En línea]. Disponible en: https://investinelsalvador.gob.sv/wp-content/uploads/2023/12/Guia-Sectorial-Energia-2023.pdf

[3] K. Molina, "(2024, sep. 27). El 59% de la energía generada en El Salvador en 2024 es renovable," *La Prensa Gráfica*. [En línea]. Disponible en: https://www.laprensagrafica.com/economia/El-59-de-la-energia-generada-en-El-Salvador-en-2024-es-renovable-20240927-0092.html.

[4] U. Alemán, "(2024, ago. 22). La hidroeléctrica fue el principal generador en julio, con 42% de la demanda de energía," *Diario El Mundo*. [En línea]. Disponible en: https://diario.elmundo.sv/economia/la-hidroelectrica-fue-el-principal-generador-en-julio-con-42-de-la-demanda-de-energia.

[5] Comisión Ejecutiva Hidroeléctrica del Río Lempa (CEL), *Guía del Sistema Institucional de Gestión Documental y Archivos*, CEL, 2008. [En línea]. Disponible en: https://transparencia.gob.sv/descarga_archivo.php?id=NTY4NzA4

[6] Asamblea Legislativa de la República de El Salvador, *Ley General de Electricidad*, Decreto Legislativo No. 843, 10 de octubre de 1996, Diario Oficial No. 204, Tomo 333, 25-oct-1996. [En línea]. Disponible en: https://www.siget.gob.sv/download/ley-general-de-electricidad-2/

[7] Consejo Nacional de Energía, *Política Energética Nacional de El Salvador 2010-2024*, San Salvador, El Salvador, 2010. [En línea]. Disponible en: https://biblioteca.olade.org/opac-tmpl/Documentos/cg01016.pdf

[8] Consejo Nacional de Energía, *Política Energética Nacional 2020-2050: Construyendo un futuro energético sostenible*. San Salvador: CNE, 2020. [En línea]. Disponible en: https://cdn.climatepolicyradar.org/navigator/SLV/2020/2020-2050-national-energy-policy_0b3b3fff88e35984f0ab231c09730595.pdf

[9] EcuRed, *Central Hidroeléctrica 5 de Noviembre*, [En línea]. Disponible en: https://www.ecured.cu/Central_Hidroel%C3%A9ctrica_5_de_Noviembre

[10] BCIE, “Con $57.5 millones el BCIE financia ampliación de central hidroeléctrica en El Salvador,” 28 sep. 2011. [En línea]. Disponible en: https://www.bcie.org/novedades/eventos/evento/con-575-millones-el-bcie-financia-ampliacion-de-central-hidroelectrica-en-el-salvador

[11] Banco Centroamericano de Integración Económica, *Memoria Anual 2010*, p.43. [En línea]. Disponible en: https://www.bcie.org/fileadmin/bcie/espanol/archivos/novedades/publicaciones/memorias_anuales/00-27-Memoria_Anual_2010.pdf

[12] ÚltimaHora.sv, “Ampliación de presa 5 de Noviembre lista para iniciar operaciones,” 4 nov. 2016. [En línea]. Disponible en: https://ultimahora.sv/ampliacion-de-presa-5-de-noviembre-lista-para-iniciar-operaciones-2/

[13] ISO, “Machine learning (ML) is a type of artificial intelligence that allows machines to learn from data without being explicitly programmed...,” What is ML?, accessed 2025. [En línea]. Disponible en: https://www.iso.org/artificial-intelligence/machine-learning

[14] A. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959.

[15] A. M. Turing, “Computing Machinery and Intelligence,” *Mind*, vol. 59, no. 236, pp. 433–460, Oct. 1950.

[16] Dartmouth College, “Artificial Intelligence (AI) Coined at Dartmouth,” Dartmouth College, 2016. [En línea]. Disponible en: https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth

[17] Instituto Nacional de Estándares y Tecnología (NIST), *La definición del NIST sobre la computación en la nube*. Departamento de Comercio de EE. UU., Publicación Especial 800-145, Gaithersburg, MD, 2011. [En línea]. Disponible en: https://doi.org/10.6028/NIST.SP.800-145

[18] Microsoft Azure, “¿Qué es la computación en la nube? Una guía para principiantes,” *Microsoft Learn*, 2023. [En línea]. Disponible en: https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/overview/what-is-cloud-computing

[19] Gartner, “Participación de mercado: Servicios de TI, a nivel mundial 2023,” *Gartner Research*, abr. 2024. [En línea]. Disponible en: https://www.gartner.com

[20] LAIF - Facilidad de Inversión en América Latina, "Extensión de la Central Hidroeléctrica '5 de Noviembre'," LAIF, 2019. [En línea]. Disponible en: https://www.eulaif.eu/es/proyectos/extension-de-la-central-hidroelectrica-5-de-noviembre

[21] Banco Centroamericano de Integración Económica (BCIE), "Dossier de Energía de El Salvador," BCIE, 2022. [En línea]. Disponible en: https://www.bcie.org/fileadmin/user_upload/v2DossierEnergiaElSalvador_02Junio2022.pdf

[22] Noticias de Energía, "Ampliarán capacidad generadora en Central 5 de Noviembre," Noticias de Energía, 2011. [En línea]. Disponible en: https://eeyer.wordpress.com/2011/01/28/ampliaran-capacidad-generadora-en-central-5-de-noviembre/

[23] Reportes estadísticos de la Unidad de Transacciones para la obtención de datos públicos de inyección totales, Nivel de embalse y Vertido.

https://www.ut.com.sv/reportes?p_p_id=MenuReportesEstadisticosPublicReports_WAR_PublicReports&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&_MenuReportesEstadisticosPublicReports_WAR_PublicReports_reportName=07nivelesembalse

.[24] Datos desagregados de las unidades 1 a la 7 de la central hidroeléctrica 5 de noviembre para el año 2024.

https://github.com/JoseDePaz94/5_Nov_TBE_Git/blob/main/Datos_UT_Procesados.xlsx

Anexos

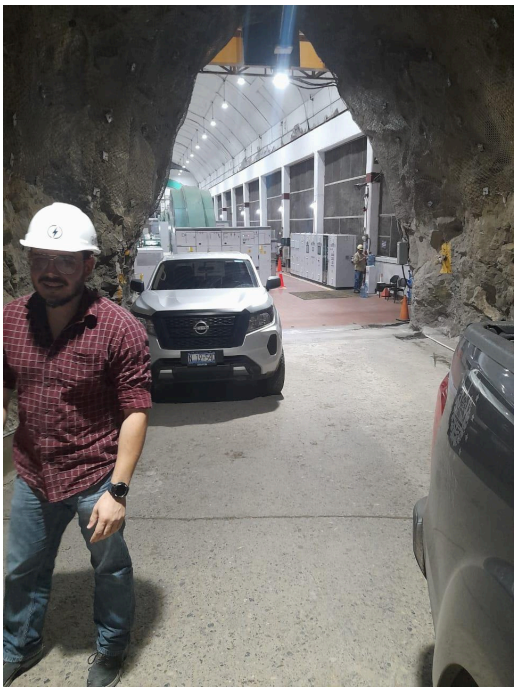
Anexo A: Realización de visita técnica a la central hidroeléctrica 5 de noviembre.

Posterior a la realización de la defensa final (desarrollada el día 22 de noviembre de 2025, fuimos invitados a el desarrollo de una visita técnica a las instalaciones de la central hidroeléctrica 5 de noviembre. Dicha visita se desarrolló los días 11 y 12 de diciembre de 2025, en la cual se brindó apertura a mostrar las máquinas, el funcionamiento y se brindó un recorrido completo de las unidades u1 a u5 y a su vez del proyecto de expansión unidades u6 y u7. En esta visita se nos permitió visualizar el funcionamiento de toda la tecnología así como corroborar ciertos elementos involucrados en nuestra investigación los cuales se detallan en este apartado.

En la siguiente figura se observan los monumentos correspondientes en uno de los espacios principales de la central hidroeléctrica. La placa y una turbina conmemorativa.

Se registró la visualización del embalse desde diferentes puntos y las imágenes de referencia se muestran a continuación.


Se visitó la “cueva” en la cual se encuentran inmersas las máquinas del proyecto primario. Máquinas u1 a u5 las cuales entran en operación según el predespacho diario. Se permitió tomar algunas fotos ilustrativas del lugar sin embargo no de equipo específico detallado por razones empresariales.

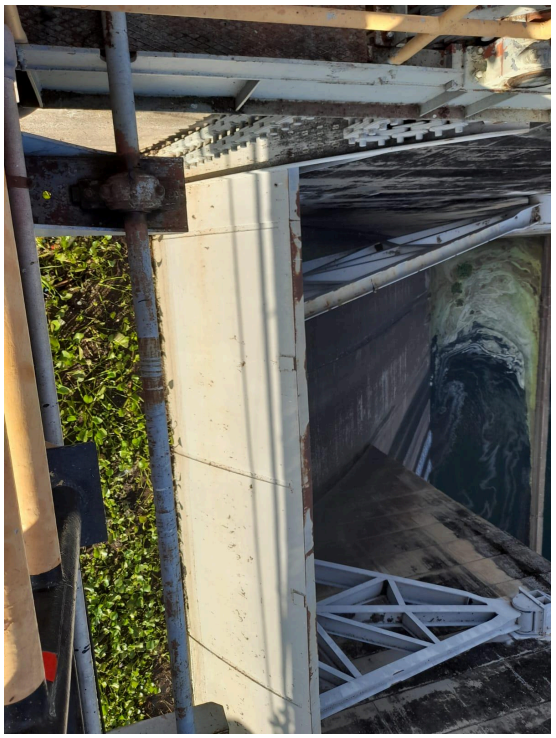
Se visitó también el edificio del proyecto de expansión y se visualizó la gama de elementos de tecnología de punta existente tanto en el proyecto de expansión como en el proyecto principal. Se logró registrar fotografías de las tuberías que llevan el agua a las máquinas u6 y u7 y la subestación asociada a dichos generadores.

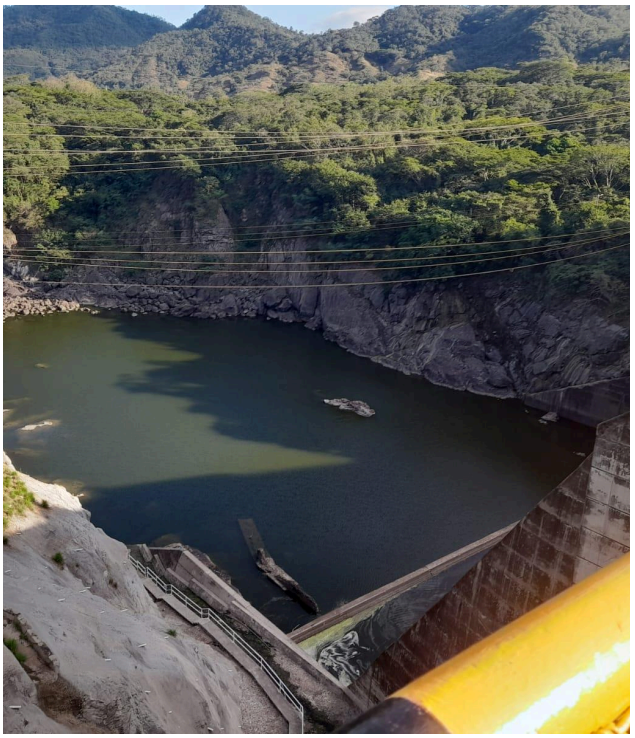
En camino a la “cueva” se visualizó también el lugar donde sale el agua producto de la generación y producto del vertido. En la siguiente figura se observa el espacio por donde brota el agua que sale de las maquinas u1 a u5.

En las siguientes figuras se observa el espacio en que se unen tanto los afluentes de agua de las unidades u1 a u5 como el vertido.

También se logró visualizar la subestación que toma la energía de las unidades u1 a u5. Dichas imágenes se muestran a continuación.

Por último, también se logró registrar el vertedero de manera cercana. Este puede verse a continuación.





Anexo B: Comparación de datos compartidos post defensa final del trabajo de graduación vs generados por los modelos de ML.

Posterior a la visita técnica, con los contactos realizados se solicitó el factor de utilización (Factor de planta) por separado de la central para hacer una comparación de dichos resultados contra los datos resultado de las simulaciones de machine learning en ambos modelos. La CEL compartió el conjunto de factores de planta separados de 2020 a 2023 de los dos proyectos llamados G9 y G10. Siendo G9 el proyecto de expansión y G10 el proyecto principal. Para efectos de este trabajo de graduación, los datos más importantes son los de G9 dado que el objeto de estudio ha sido separar la generación del conjunto de las unidades.
La comparación de los resultados de factor de planta del proyecto de expansión (unidades 6 y 7) se muestra a continuación.

Año	FP CEL	FP Multiclasico (Bosques aleatorios)	FP Continuo (Lineal múltiple).
2021	33.33%	23.76%	28.56%
2022	31.76%	31.63%	37.48%
2023	22.16%	15.29%	17.70%

Es importante aclarar que el cálculo de los factores de planta estimados se realizó utilizando como base el año 2024. La estimación realizada es un espejo del año 2024 tomando como referencia los datos totales de 2021 a 2023. Para que los datos puedan acercarse más, es imprescindible alimentar el modelo con datos de varios años.

Una observación importante realizada por autoridades de la hidroeléctrica al realizar una presentación, fue que la unidad 4 era una unidad poco utilizada, pero esto se debe a que en el año 2024 esta unidad pasó en mantenimiento la mayor parte del tiempo pero en realidad al estar en operación es realmente equiparable a las unidades 1, 2, 3 y 5 que son las que entran en funcionamiento de manera más constante.

Los datos estimados bajos respecto de la unidad 4 para 2021-2023, se debieron también a que dicha estimación se realizó en base al año 2024, año en el cual dicha unidad se encontró en mantenimiento. Por otro lado se explicó que los datos por separado no se publican por situaciones empresariales. Dicha información es sensible dado que al poseerse información específica de cada unidad, y teniendo el dato correspondiente del pliego tarifario, podría estimarse el ingreso completo de la central, lo cual implica información sensible de la central hidroeléctrica.

Anexo C: Código Fuente y Visualizaciones

En este apartado se encuentra el primer bloque de programas para analizar las tendencias y verificar los comportamientos generales de los datos del documento de embalses (donde se encuentran los totales de nivel, inyecciones y vertido).

En el programa 1 se desarrolla el ingreso de las bibliotecas necesarias para ejecutar las directivas necesarias para correr los programas y realizar las gráficas, sumatorias, procedimientos necesarios, etc.

```
#Programa 1

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

#machine Learning

import sklearn as sk

from sklearn.linear_model import LogisticRegression

from sklearn import svm
```

En el programa 2 se desarrolla la carga del archivo que contiene los datos totales de inyección, nivel y vertido de todas las presas hidroeléctricas del país. El registro generado en este trabajo de investigación posee datos desde 2012 hasta 2024 de manera horaria.

```
#Programa 2

#Definir la ruta de los datos y leer datos de un unico archivo con los datos UT

#main_dir    ='/mnt/data/'

fname       ='embalses12-24.csv'

df = pd.read_csv(fname, parse_dates=['HoraFecha'], dayfirst=True)

print(df.head())
```

En el programa 3 se delimitan las fechas en un rango de 4 años 2021-2024 el cual es el rango de fechas para el cual se desarrolla el estudio.

```
#Programa 3
```

```
#fechas

inicio= '2021-1-1'

final= '2024-12-31'

#Seleccionar entre dos fechas

data = df.loc[(df['HoraFecha']>=inicio)&(df['HoraFecha']<=final)]

print(data.tail)
```

En el programa 4 separan los datos de inyección de todas las hidroeléctricas en un solo dataframe llamado iny, todos los de nivel en un dataframe llamado niv, y el vertido en un solo dataframe llamado vert.

```
#Programa 4

#separar actividad segun reporte UT: Inyección (MW), nivel(msm) y vertido

iny     = data [['HoraFecha','15se_iny','3feb_iny','5nov_iny','cgra_iny','guaj_iny']]

niv     = data [['HoraFecha','15se_niv','3feb_niv','5nov_niv','cgra_niv','guaj_niv']]

ver     = data [['HoraFecha','15se_vert','3feb_vert','5nov_vert','cgra_vert','guaj_vert']]

print(iny)

print(niv)

print(ver)
```

En el programa 5 se agrega como columna principal la columna de HoraFecha la cual posee el dato horario en la fecha correspondiente.

```
#Programa 5

#Set Index

iny.set_index('HoraFecha',inplace=True)

niv.set_index('HoraFecha',inplace=True)

ver.set_index('HoraFecha',inplace=True)
```

En el programa 6 se realizan agrupaciones diarias, mensuales y anuales. Se realizan las sumatorias correspondientes y se guardan en dataframe iny_D, iny_M e iny_Y para cada uno. Es importante agregar que los datos agrupados para este análisis solo es el de inyecciones.

```
#Programa 6

#inyecciones

#Se suman los valores de inyecciones (MWh) mediante agrupamiento diario (D), Mensual (M), Anual(Y)

iny_D   =iny.groupby(pd.Grouper(freq='D')).sum()

iny_M   =iny.groupby(pd.Grouper(freq='M')).sum()

iny_Y   =iny.groupby(pd.Grouper(freq='Y')).sum()

#Forma porcentual

iny_M_N =iny_M.divide(iny_M.sum(axis=1),axis=0)

print(iny_D)

print(iny_M)

print(iny_Y)
```

En el programa 7 se realizan los cálculos generales de factor de planta para cada una de las generadoras (Ojo. Para fines del análisis, el único dato relevante es el factor de planta de la generadora 5 de noviembre). También es importante entender que el objetivo del análisis es obtener los factores de planta del proyecto de expansión de la hidroeléctrica 5 de noviembre para los años 2021-2024). Posterior al código también se encuentra una captura de pantalla de los resultados de los factores de planta generales para el mismo período.

```
#Programa 7

#Factor de planta: anual y mensual

#La capacidad (MW) para cada central utiliza datos nominales de potencia en MW

hours      = 24

hours_year  = 24*365

capacidad   = np.array([184,66,180,173,20])

#Anual

fpyear     = iny_Y.iloc[:,:]/(capacidad*hours_year)

print(fpyear)

#Mensual
```

```
hours_month = hours*iny_M.index.daysinmonth

fpmonth     = iny_M.loc[:,:].div(hours_month,axis=0)

fpmonth     = fpmonth.loc[:,:].div(capacidad)

print('**********')

print(fpmonth.mean(0))
```

```
            15se_iny  3feb_iny  5nov_iny  cgra_iny  guaj_iny
HoraFecha
2021-12-31  0.365471  0.000000  0.379447  0.332798  0.285309
2022-12-31  0.452664  0.000000  0.451997  0.415956  0.489407
2023-12-31  0.312830  0.151166  0.296596  0.262189  0.220193
2024-12-31  0.430547  0.349996  0.382993  0.371817  0.344478
**********
15se_iny    0.389329
3feb_iny    0.124765
5nov_iny    0.377072
cgra_iny    0.345244
guaj_iny    0.334792
dtype: float64
```

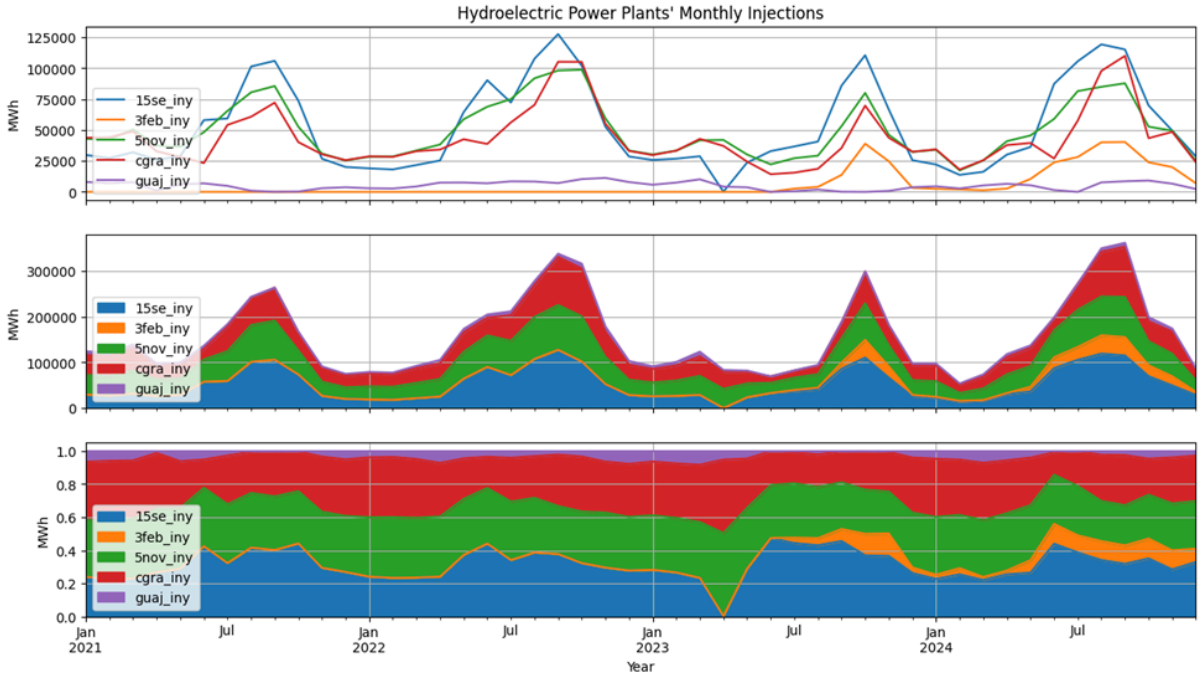
En el programa 8 se realiza el agrupamiento de los valores del nivel de agua de manera diaria en la variable Niv_D para poder graficar los datos

```
# Programa 8

#Niveles

#Se promedian los valores de nivel de agua (msm) mediante: Agrupamiento diario (D), Semanal (W)

Niv_D       = niv.groupby(pd.Grouper(freq='D')).mean()

print(Niv_D.head())
```

En el programa 9 se grafican las inyecciones en diferentes gráficos el primer grafico tiene las gráficas superpuestas que muestran la cantidad inyectada en el periodo correspondiente. En el segundo grafico se observan los datos montados de generadora sobre generadora para apreciar el acumulado en el tiempo. En el tercer grafico se muestra el mismo segundo, pero en forma porcentual para validar el porcentaje que brindó cada distribuidora en el periodo correspondiente.

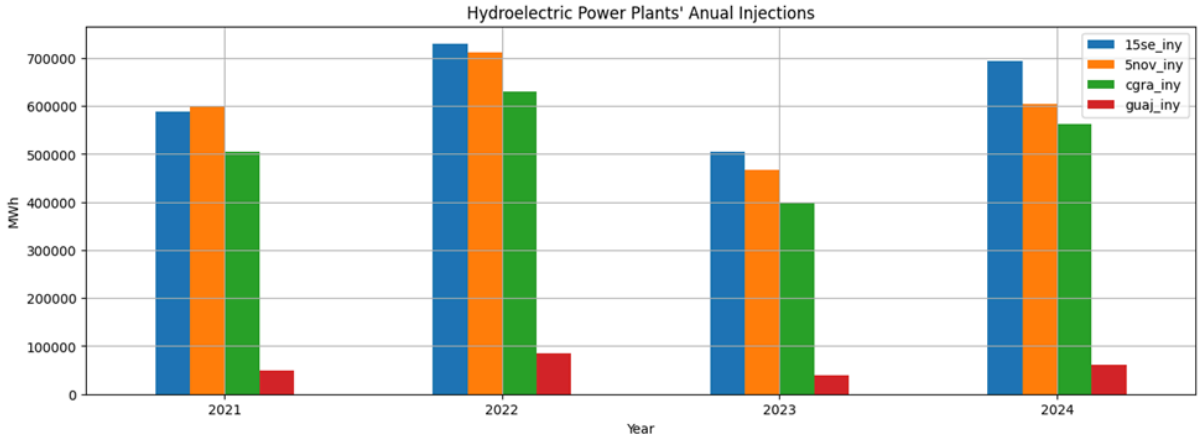
```
#Programa 9

#plot Inyecciones. Tres graficos diferentes representando lo mismo: Inyecciones mensuales.

fig,axs=plt.subplots(3,sharex='all')
```

```
iny_M.plot(ax=axs[0],figsize=(15,8))

iny_M.plot.area(ax=axs[1],figsize=(15,8))

iny_M_N.plot.area(ax=axs[2],figsize=(15,8))

axs[0].set_title('Hydroelectric Power Plants\' Monthly Injections')

axs[2].set_xlabel('Year')

axs[0].set_ylabel('MWh')

axs[1].set_ylabel('MWh')

axs[2].set_ylabel('MWh')

axs[0].grid()

axs[1].grid()

axs[2].grid()

axs[0].legend(loc='lower left')

axs[1].legend(loc='lower left')

axs[2].legend(loc='lower left')
```



En el programa 10 se grafican las inyecciones anuales en gráficos de barras. Se presentan sólo las 4 generadoras más importantes una al lado de la otra en los años 2021-2024.

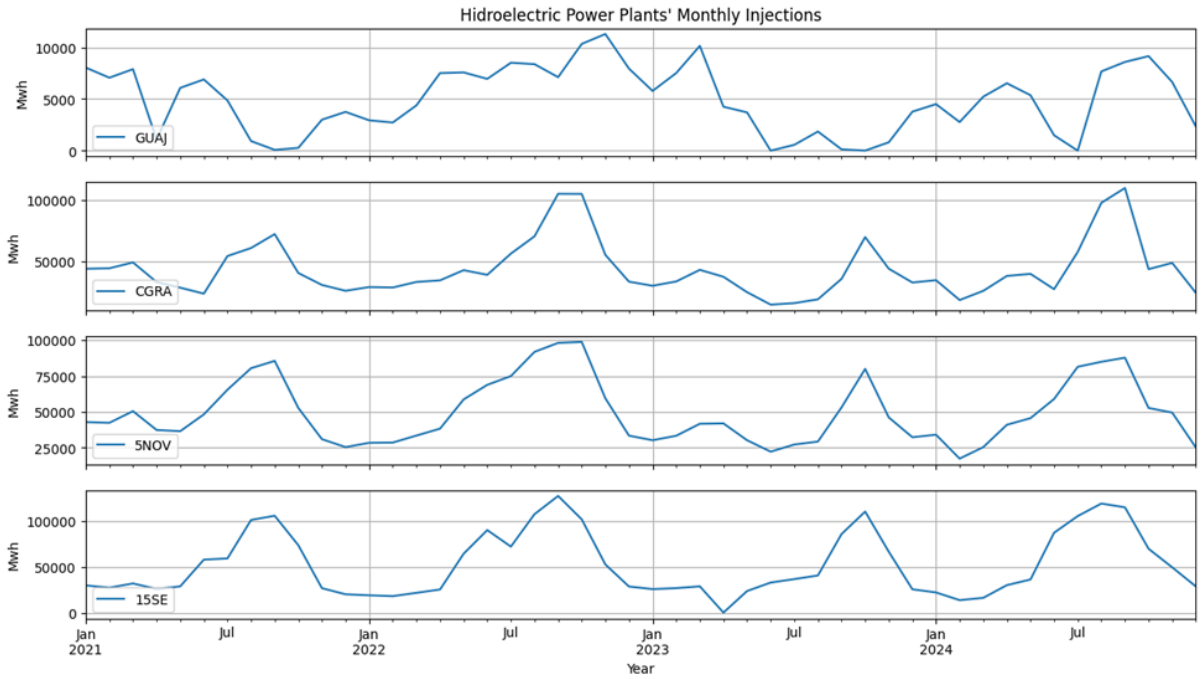
```
#Programa 10

iny_barras            = iny_Y.drop(['3feb_iny'],axis=1)

iny_barras.index      = iny_barras.index.year

iny_barras.plot.bar(rot = 0, figsize=(15,5))

plt.title('Hydroelectric Power Plants\' Anual Injections')

plt.xlabel('Year')

plt.ylabel('MWh')

plt.grid()

plt.legend(loc='upper right')
```



En el programa 11 se grafican las inyecciones mensuales por generadora en el periodo de 2021 a 2024. Como puede observarse en la gráfica, en las temporadas lluviosas un mayor índice de generación, se anuales en gráficos de barras.

```
#Programa 11

fig, axs = plt.subplots(4, sharex='all', figsize=[15, 8])

 iny_M['guaj_iny'].plot(ax=axs[0], label='GUAJ')

iny_M['cgra_iny'].plot(ax=axs[1], label='CGRA')

iny_M['5nov_iny'].plot(ax=axs[2], label='5NOV')

iny_M['15se_iny'].plot(ax=axs[3], label='15SE')
```

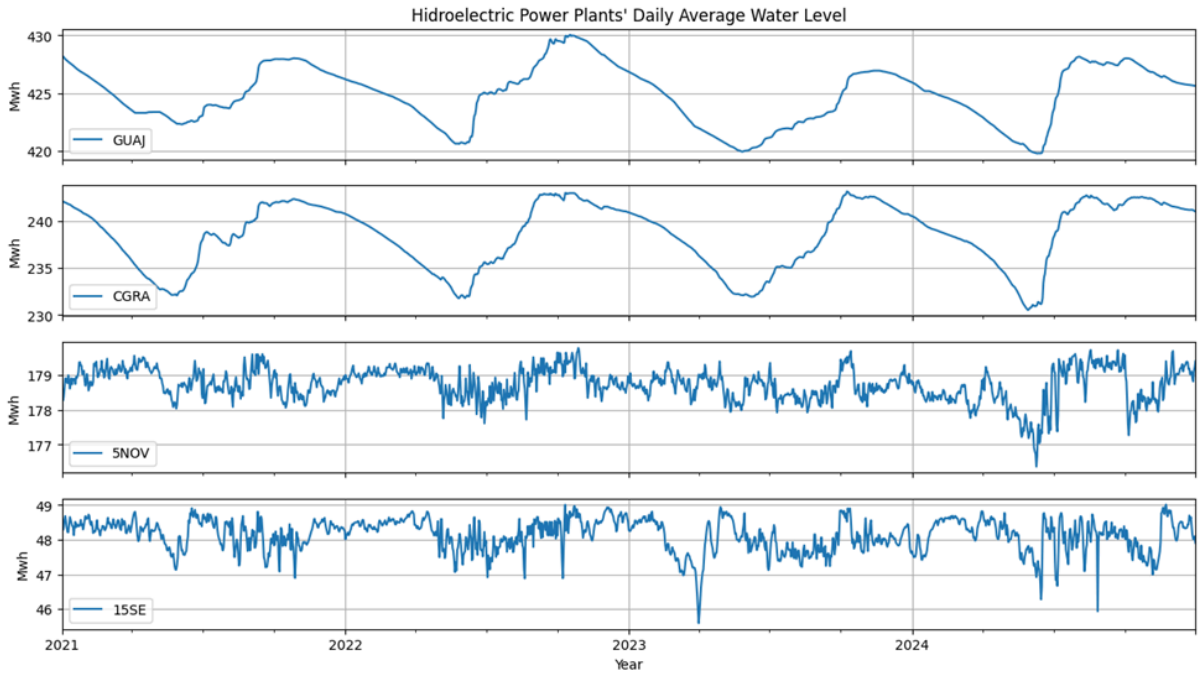
```
axs[0].set_title("Hidroelectric Power Plants' Monthly Injections")

axs[3].set_xlabel('Year')

axs[0].set_ylabel('Mwh')

axs[1].set_ylabel('Mwh')

axs[2].set_ylabel('Mwh')

axs[3].set_ylabel('Mwh')

for ax in axs:

    ax.grid()

    ax.legend(loc='lower left')

plt.show()
```



En el programa 12 se grafica el nivel de los embalses de las 4 generadoras principales. Como puede verse, la caída en los niveles de embalse es evidente en guajoyo y cerrón grande en las temporadas secas y el llenado en las temporadas de lluvia. En 5 de noviembre el efecto es menos evidente al igual que 15 de septiembre.

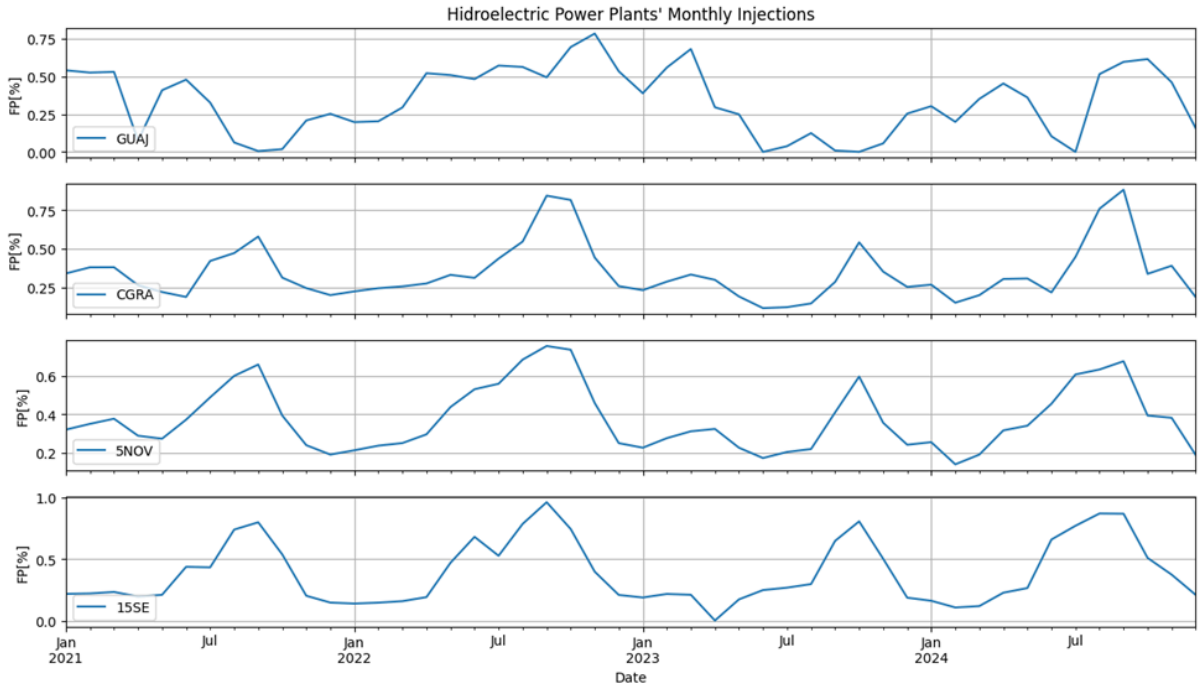
```
#Programa 12

#plot niveles

fig, axs = plt.subplots(4, sharex='all', figsize=[15, 8])
```

```
Niv_D['guaj_niv'].plot(ax=axs[0], label='GUAJ')

Niv_D['cgra_niv'].plot(ax=axs[1], label='CGRA')

Niv_D['5nov_niv'].plot(ax=axs[2], label='5NOV')

Niv_D['15se_niv'].plot(ax=axs[3], label='15SE')

axs[0].set_title("Hidroelectric Power Plants' Daily Average Water Level")

axs[3].set_xlabel('Year')

axs[0].set_ylabel('Mwh')

axs[1].set_ylabel('Mwh')

axs[2].set_ylabel('Mwh')

axs[3].set_ylabel('Mwh')

for ax in axs:

    ax.grid()

    ax.legend(loc='lower left')

plt.show()
```



En el programa 13 se grafica el factor de planta de las 4 generadoras. Dicho factor de planta se agrupa de manera mensual y en las siguientes graficas se observa la variación de este de manera general por hidroeléctrica.

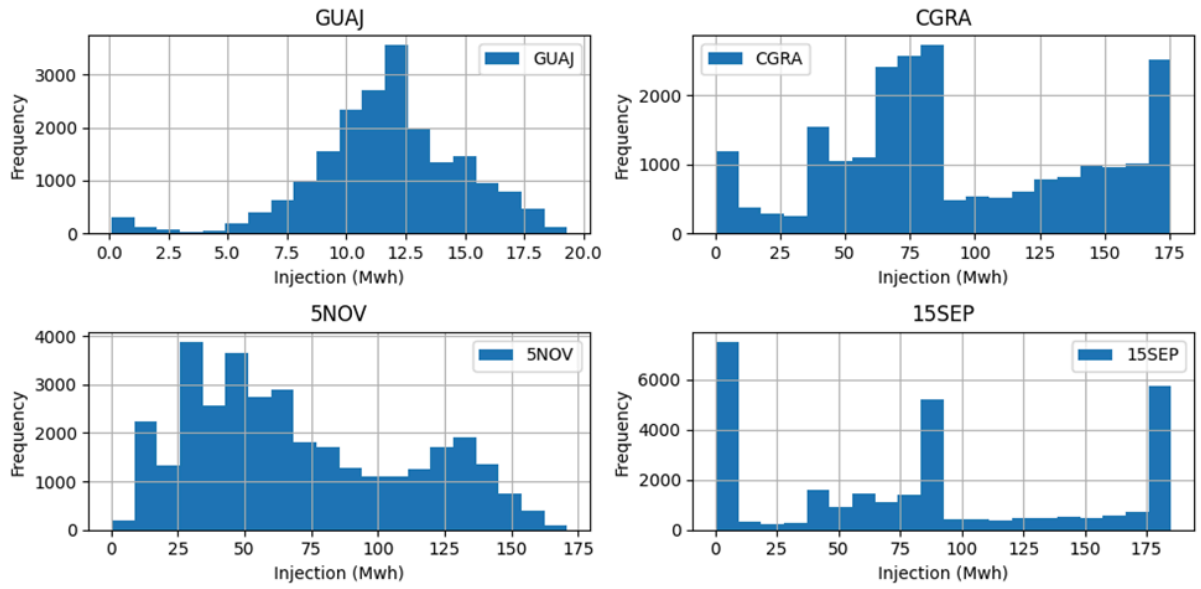
```
#Programa 13

#plot mensual y anual

fig, axs = plt.subplots(4, sharex='all', figsize=[15, 8])

fpmonth['guaj_iny'].plot(ax=axs[0], label='GUAJ')

fpmonth['cgra_iny'].plot(ax=axs[1], label='CGRA')

fpmonth['5nov_iny'].plot(ax=axs[2], label='5NOV')

fpmonth['15se_iny'].plot(ax=axs[3], label='15SE')

axs[0].set_title("Hidroelectric Power Plants' Monthly Injections")

axs[3].set_xlabel('Date')

axs[0].set_ylabel('FP[%]')

axs[1].set_ylabel('FP[%]')

axs[2].set_ylabel('FP[%]')

axs[3].set_ylabel('FP[%]')

for ax in axs:

    ax.grid()

    ax.legend(loc='lower left')

plt.show()
```



En el programa 14 se grafican 4 histogramas en las que se observa la tendencia de entradas de los generadores en el tiempo.

```
#Programa 14

# Filtrar datos

guaj = iny.loc[iny['guaj_iny'] != 0]['guaj_iny']

cgra = iny.loc[iny['cgra_iny'] != 0]['cgra_iny']

nov = iny.loc[iny['5nov_iny'] != 0]['5nov_iny']

sep = iny.loc[iny['15se_iny'] != 0]['15se_iny']

# Crear subplots

fig, axs = plt.subplots(2, 2)

 # Graficar histogramas

guaj.hist(ax=axs[0, 0], bins=20, label='GUAJ',figsize=(10, 5))

cgra.hist(ax=axs[0, 1], bins=20, label='CGRA',figsize=(10, 5))

nov.hist(ax=axs[1, 0], bins=20, label='5NOV',figsize=(10, 5))

sep.hist(ax=axs[1, 1], bins=20, label='15SEP',figsize=(10, 5))

# Añadir leyendas
```

```
axs[0, 0].legend()

axs[0, 1].legend()

axs[1, 0].legend()

axs[1, 1].legend()

# Añadir títulos

axs[0, 0].set_title('GUAJ')

axs[0, 1].set_title('CGRA')

axs[1, 0].set_title('5NOV')

axs[1, 1].set_title('15SEP')

# Añadir etiquetas de ejes

for ax in axs.flat:

    ax.set_xlabel('Injection (Mwh)')

    ax.set_ylabel('Frequency')

plt.tight_layout()

plt.show()
```



Como se puede observar en la primera grafica guajoyo parece contar con una sola unidad que opera mayoritariamente (alrededor de 3500 horas) a una potencia aproximada de 12

MWh) , por otro lado hay un máximo cercano a 19 MWh y un minimo cercano a cero (pero no cero) con menos de 500 horas cada uno.

En el caso de Cerron grande, parece haber dos unidades. Superando las 2000 horas, la central entra en promedio a 75 MWh y a 175MWh también superando las 2000 horas. Claramente esos dos rangos implican que existen 2 unidades en operación. Al estar cercano a cero, no hay ninguna operando, al estar en 75 hay una operando y al estar en valores cercanos a 175, las dos se encuentran operando.

En el caso de 15 de septiembre ocurre lo mismo. En un espacio cercano a 85 MWh (aproximadamente 5000 horas) implica que hay una sola unidad operando, de la misma manera (cercano a 6000 horas) entra en operación a 175 MWh lo cual indica que las dos unidades están operando.

Como puede observarse en 5 de noviembre, al existir 7 unidades, puede observarse un espectro más amplio sin poderse determinar cuántas o cuales de las 7 están en operación. Es de allí donde sale el cuestionamiento de si es posible determinar tanto las joras a las que entró en operación cada unidad como los valores con los que entró en operación.

Para esto, se realizó 2 análisis de Machine learning con el objetivo de determinar dichos parámetros. El primero es un analisis multiclasico en el que se intenta calcular las horas a las que entró en operación cada unidad en el período de 2021 a 2024.

Se consiguió datos reales separados por unidades del año 2024 por parte de autoridades de la UT con los cuales se generan los modelos y en base a esto se realizan las estimaciones.

El modelo principal (modelo multiclasico) busca formar un binario de 7 bits cada bit se corresponde con una unidad precisa entre la 1 y la 7 de la generadora 5 de noviembre.

Se usan datos de entrenamiento solo del año 2024 para realizar el modelo. Estos datos son inyecciones totales, nivel de embalse, vertido y el conjunto separado de las unidades de 2024 para las 7 unidades.

Posteriormente se calcula la mediana de las unidades separadas de 2024 para utilizar ese dato de cada unidad como dato representativo.

A continuación, se presentan los programas del primer análisis de Machine learning (Análisis Multiclásico).

A.1. Gráficos y Tendencias

En el programa 1 se colocan todas las bibliotecas que sirven tanto para tratamiento de datos como para análisis de machine learning multiclasico (binario). En este se pueden destacar los modelos de Árbol de decisiones, vecino mas cercano, Regresión Logística y RandomForest.

```
#programa 1
```

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

#machine Learning

import sklearn as sk

from sklearn.linear_model import LogisticRegression

from sklearn import svm

from sklearn.model_selection import train_test_split

from sklearn import tree

from sklearn.svm import SVC

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.preprocessing import StandardScaler

from sklearn.pipeline import make_pipeline

import seaborn as sns

from sklearn.metrics import (accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix,classification_report,r2_score)

from sklearn.ensemble import RandomForestClassifier

from sklearn.multioutput import MultiOutputClassifier
```

En el programa 2 se toman los datos proporcionados por la UT (Datos separados por unidades del año 2024) y se agregan al dataframe fname para su procesamiento posterior en df2. Se procede a llenar todas las celdas vacías con ceros y a ajustar los formatos de las fechas utilizadas.

```
#programa 2

fname = 'Datos_UT_procesados_2024_completo.csv'

df2 = pd.read_csv(fname, parse_dates=['Fecha'], dayfirst=True, encoding='ISO-8859-1', thousands=',')

# Llenar valores faltantes con 0
```

```
df2 = df2.fillna(0)

# Asegurarse de que la columna 'Fecha' esté en formato datetime

df2['Fecha'] = pd.to_datetime(df2['Fecha'], errors='coerce')

# Imprimir las primeras filas del DataFrame para verificar la carga correcta

print(df2.head())# %% [markdown]

# # Programa 2: Carga y Exploración de Datos

# Carga del archivo CSV y primera exploración
```

En el programa 3 se define el rango de fechas para considerar respecto del análisis para generar el modelo. Se toma en consideración desde el 6 de enero hasta el 31 de diciembre de 2024. (se eliminan los primeros 5 días de enero dado que eran datos que se encontraban alterados y que se correspondían con los de inyecciones del nivel de embalses en sus totalidades.

```
#programa 3

inicio = pd.to_datetime('2024-01-6')

final = pd.to_datetime('2024-12-31')

df2 = df2.loc[(df2['Fecha']>=inicio)&(df2['Fecha']<=final)]

print(df2.tail())
```

En el programa 4 se toman los datos totales correspondientes al documento de embalses (inyecciones totales, nivel y vertido) en fname para posteriormente pasar a df y procesarlos desde alli

```
#Programa 4

#Definir la ruta de los datos y leer datos de un unico archivo con los datos UT

#main_dir    ='/mnt/data/'

fname       ='embalses12-24.csv'

df  =  pd.read_csv(fname,  parse_dates=['HoraFecha'],  dayfirst=True,  encoding='ISO-8859-1',
thousands=',')

print(df.head())
```

En el programa 5 se definen los mismos límites de fechas que en el documento anterior, Del 6 de enero hasta el 31 de diciembre de 2024.

```
#programa 5

#fechas

inicio= '2024-1-6'

final= '2024-12-31'

#Seleccionar entre dos fechas

data = df.loc[(df['HoraFecha']>=inicio)&(df['HoraFecha']<=final)]

print(data.tail)
```

En el programa 6 se agregan los elementos X_2024 y Y_2024 los cuales servirán para entrenar los modelos de clasificación correspondientes. Partiendo de Y_2024, los datos se transforman en binario para formar el modelo multiclasico. Para estos casos se usarán RandomForest, Arbol de decisiones, el vecino más cercano y Regresión logística. El programa 6 se corresponde con el modelo RandomForest

```
#Programa 6

# Analisis usando RandomForestClassifier

# Entradas y salidas

X_2024 = data[['5nov_iny', '5nov_niv', '5nov_vert']]

Y_2024 = df2[['5nov-u7', '5nov-u6', '5nov-u5', '5nov-u4', '5nov-u3', '5nov-u2', '5nov-u1']]

Y_2024_binario = (Y_2024 > 0).astype(int)

# División de datos

X_train, X_test, y_train, y_test = train_test_split(X_2024, Y_2024_binario, test_size=0.2, random_state=42)

# Entrenamiento del modelo

modelo_base = RandomForestClassifier()

modelo = MultiOutputClassifier(modelo_base)

modelo.fit(X_train, y_train)

# Predicciones

y_pred = modelo.predict(X_test)

# Evaluación general (accuracy promedio)
```

```
accuracy_promedio = accuracy_score(y_test, y_pred)

print(f"Accuracy promedio (todas las unidades): {accuracy_promedio:.4f}")

# Accuracy por unidad

print("\nAccuracy por unidad:")

for i, col in enumerate(y_test.columns):

    acc = accuracy_score(y_test[col], y_pred[:, i])

    print(f"{col}: {acc:.4f}")
```

Accuracy promedio (todas las unidades): 0.8697

Accuracy por unidad:

5nov-u7: 0.9179

5nov-u6: 0.8971

5nov-u5: 0.8149

5nov-u4: 0.8074

5nov-u3: 0.8623

5nov-u2: 0.8918

5nov-u1: 0.8965

En el programa 6 V2 se realiza el mismo análisis del programa 6 pero utilizando el modelo Árbol de decisiones.

```
#Programa 6 V2

# Analisis usando DecisionTreeClassifier

# Entradas y salidas

X_2024 = datos[['5nov_iny', '5nov_niv', '5nov_vert']]

Y_2024 = df2[['5nov-u7', '5nov-u6', '5nov-u5', '5nov-u4', '5nov-u3', '5nov-u2', '5nov-u1']]

Y_2024_binario = (Y_2024 > 0).astype(int)

# División de datos

X_train, X_test, y_train, y_test = train_test_split(X_2024, Y_2024_binario, test_size=0.2, random_state=42)
```

```
# Entrenamiento con Decision Tree

modelo_base = DecisionTreeClassifier(random_state=42)

modelo = MultiOutputClassifier(modelo_base)

modelo.fit(X_train, y_train)

# Predicciones

y_pred = modelo.predict(X_test)

# Evaluación general

accuracy_promedio = accuracy_score(y_test, y_pred)

print(f"\nAccuracy promedio (Decision Tree): {accuracy_promedio:.4f}")

# Accuracy por unidad

print("\nAccuracy por unidad:")

for i, col in enumerate(y_test.columns):

    acc = accuracy_score(y_test[col], y_pred[:, i])

    print(f"{col}: {acc:.4f}")
```

Accuracy promedio (Decision Tree): 0.8427

Accuracy por unidad:

5nov-u7: 0.8809

5nov-u6: 0.8566

5nov-u5: 0.8022

5nov-u4: 0.7825

5nov-u3: 0.8479

5nov-u2: 0.8635

5nov-u1: 0.8652

En el programa 6 V3 se realiza el mismo análisis del programa 6 pero utilizando el modelo de Regresión Logística.

```
#Programa 6 V3

# Analisis usando Regresión Logistica
```

```
# Entradas y salidas

X_2024 = datos[['5nov_iny', '5nov_niv', '5nov_vert']]

Y_2024 = df2[['5nov-u7', '5nov-u6', '5nov-u5', '5nov-u4', '5nov-u3', '5nov-u2', '5nov-u1']]

Y_2024_binario = (Y_2024 > 0).astype(int)

# División de datos

X_train, X_test, y_train, y_test = train_test_split(X_2024, Y_2024_binario, test_size=0.2,
random_state=42)

# Modelo base: Regresión logística

modelo_base = LogisticRegression(max_iter=1000)

modelo = MultiOutputClassifier(modelo_base)

modelo.fit(X_train, y_train)

# Predicción

y_pred = modelo.predict(X_test)

# Evaluación general

accuracy_promedio = accuracy_score(y_test, y_pred)

print(f"\nAccuracy promedio (Regresión Logística): {accuracy_promedio:.4f}")

# Accuracy por unidad

print("\nAccuracy por unidad:")

for i, col in enumerate(y_test.columns):

    acc = accuracy_score(y_test[col], y_pred[:, i])

    print(f"{col}: {acc:.4f}")
```

Accuracy promedio (Regresión Logística): 0.8553

Accuracy por unidad:

5nov-u7: 0.9121

5nov-u6: 0.9115

5nov-u5: 0.7901

5nov-u4: 0.7611

5nov-u3: 0.8329

5nov-u2: 0.8930

5nov-u1: 0.8861

En el programa 6 V4 se realiza el mismo análisis del programa 6 pero utilizando el modelo Vecino más cercano (KNN).

```
#Programa 6 V4

# Entradas y salidas

X_2024 = datos[['5nov_iny', '5nov_niv', '5nov_vert']]

Y_2024 = df2[['5nov-u7', '5nov-u6', '5nov-u5', '5nov-u4', '5nov-u3', '5nov-u2', '5nov-u1']]

Y_2024_binario = (Y_2024 > 0).astype(int)

# División de datos

X_train, X_test, y_train, y_test = train_test_split(X_2024, Y_2024_binario, test_size=0.2, random_state=42)

# Modelo base: KNN

modelo_base = KNeighborsClassifier(n_neighbors=5)  # Puedes ajustar el valor de k

modelo = MultiOutputClassifier(modelo_base)

modelo.fit(X_train, y_train)

# Predicción

y_pred = modelo.predict(X_test)

# Accuracy global

accuracy_promedio = accuracy_score(y_test, y_pred)

print(f"\nAccuracy promedio (KNN): {accuracy_promedio:.4f}")

# Accuracy por unidad

print("\nAccuracy por unidad:")

for i, col in enumerate(y_test.columns):

    acc = accuracy_score(y_test[col], y_pred[:, i])
```

```
    print(f"{col}: {acc:.4f}")
```

Accuracy promedio (KNN): 0.8652

Accuracy por unidad:

5nov-u7: 0.9092

5nov-u6: 0.8982

5nov-u5: 0.8097

5nov-u4: 0.8068

5nov-u3: 0.8577

5nov-u2: 0.8832

5nov-u1: 0.8866

En el programa 7, luego de haber generado los modelos (debe elegirse el modelo particular) se inicia con el brindado de datos al modelo para hacer las predicciones. Se colocan como límite principal el 1 de enero de 2021 y como límite final el 31 de diciembre de 2023.

```
#Programa 7

#fechas

inicio= '2021-1-1'

final= '2023-12-31'

#Seleccionar entre dos fechas

data = df.loc[(df['HoraFecha']>=inicio)&(df['HoraFecha']<=final)]

print(data.tail)
```

En el programa 8, se brindan los datos de inyección, nivel y vertido de los años 2021 a 2023 (correspondiente al documento de los embalses), se prueba el modelo y al generarse la predicción de datos, esta se guarda en un archivo csv llamado predicción_2021-2023.csv

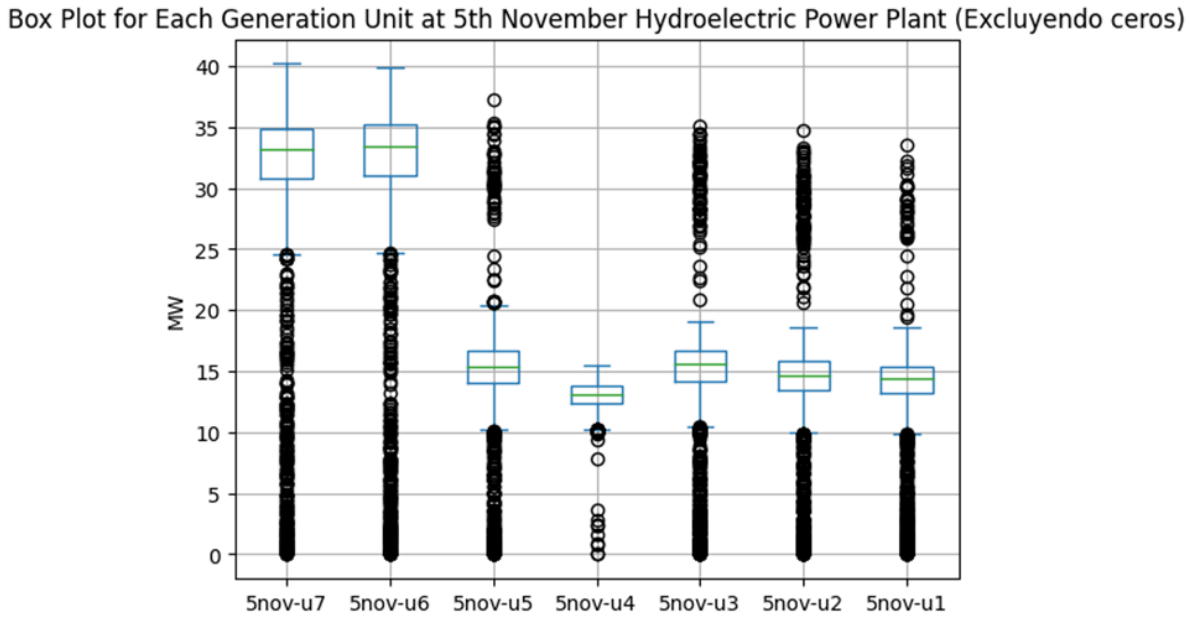
```
#Programa 8

# 1. Selección de columnas de entrada para predicción

X_pred = data[['5nov_iny', '5nov_niv', '5nov_vert']]

# 2. Predicción binaria por unidad (multi-output)

predicciones_binarias = modelo.predict(X_pred)
```

```
# 3. Convertimos cada fila a una cadena binaria tipo '1011001'

binarios_resultantes = [''.join(map(str, fila)) for fila in predicciones_binarias]

# 4. Añadimos las predicciones al DataFrame original

data['Pred_binario'] = binarios_resultantes

# 5. Seleccionamos columnas de interés: inyección, nivel, vertido y binario predicho

df_interes = data[['HoraFecha','5nov_iny', '5nov_niv', '5nov_vert', 'Pred_binario']]

# 6. Guardamos en archivo CSV

df_interes.to_csv('predicción 2021-2023.csv', index=False)

print("Archivo 'predicción 2021-2023.csv' generado exitosamente.")
```

En el programa 9, se analizan las columnas del archivo de los datos reales brindados por la UT correspondientes a las unidades separadas para el año 2024. Se desarrolla un diagrama de caja y bigotes para observar los valores de cuartiles y mediana para considerar cuales son los valores más representativos para generar la gráfica.

```
#Programa 9

#Programa para hacer diagrama de caja y bigotes sin contar los ceros y mostrando cuartiles y mediana.

# Seleccionar solo las columnas de las unidades 5nov-u1 a 5nov-u7

columnas_unidades = ['5nov-u7', '5nov-u6', '5nov-u5', '5nov-u4', '5nov-u3', '5nov-u2', '5nov-u1']

df_unidades = df2[columnas_unidades]

# Filtrar para excluir ceros - creamos una copia para no modificar el original

df_unidades_sin_ceros = df_unidades.copy()

# Reemplazar ceros con NaN (que pandas/matplotlib ignorará automáticamente)

df_unidades_sin_ceros[df_unidades_sin_ceros == 0] = None

# Alternativa: Filtrar filas donde todas las unidades son cero

# df_unidades_sin_ceros = df_unidades[(df_unidades != 0).any(axis=1)]

# Crear boxplot sin ceros

plt.figure(figsize=(20, 5))
```

```
df_unidades_sin_ceros.plot(kind='box',sharex=False,sharey=False,title='Box Plot for Each Generation Unit at 5th November Hydroelectric Power Plant (Excluyendo ceros)')

plt.ylabel('MW')

plt.grid(True)

plt.show()

# Obtener resumen estadístico excluyendo ceros

resumen = df_unidades_sin_ceros.describe(percentiles=[0.25, 0.5, 0.75])

# Mostrar solo los datos relevantes para el diagrama de caja y bigotes

resumen_filtrado = resumen.loc[['min', '25%', '50%', '75%', 'max']]

print("Resumen estadístico por unidad (Boxplot - excluyendo ceros):")

print(resumen_filtrado)
```



En el programa 10, se realiza la separación del binario de 7 bits en 7 columnas diferentes, se toman los datos de la mediana del programa anterior y se multiplican las columnas por cada mediana. Los datos actualizados se guardan en el archivo predicción_actualizada.csv.

```
#Programa 10

# Extraer mediana (50%) desde el resumen estadístico excluyendo ceros

medianas = resumen.loc['50%']  # Esto devuelve una Serie con los valores medianos por unidad

# Ordenar las medianas en el orden correspondiente de los bits en 'Pred_binario'
```

```
factores = [medianas['5nov-u7'],medianas['5nov-u6'],medianas['5nov-u5'],medianas['5nov-u4'],medianas['5nov-u3'],medianas['5nov-u2'],medianas['5nov-u1']]

print("Factores (medianas) aplicados:", factores)

# Paso 1: Separar los bits de Pred_binario

bits_separados = data['Pred_binario'].apply(lambda x: [int(bit) for bit in list(x)]).apply(pd.Series)

bits_separados.columns = ['u7', 'u6', 'u5', 'u4', 'u3', 'u2', 'u1']

for i, col in enumerate(['u7', 'u6', 'u5', 'u4', 'u3', 'u2', 'u1']):

    data[f'Estim_{col}'] = bits_separados[col] * factores[i]

# Paso 2: Calcular la suma total (opcional)

data['Generación_estim'] = data[[f'Estim_{col}' for col in ['u7', 'u6', 'u5', 'u4', 'u3', 'u2', 'u1']]].sum(axis=1)

# Paso 3: Verificación

print(data[['Pred_binario'] + [f'Estim_{col}' for col in ['u7', 'u6', 'u5', 'u4', 'u3', 'u2', 'u1']] + ['Generación_estim']].head())

# Paso 4: Guardar en CSV (opcional)

data.to_csv('predicción_actualizada.csv', index=False)

print("Archivo 'predicción_actualizada.csv' generado exitosamente con los 7 valores estimados por bit.")
```

En el programa 11, se realiza la sumatoria de las columnas para obtener el consolidado de los estimados de las unidades 1 a la 7. La nueva columna con el total se guarda en el archivo predicción_actualizada.csv.

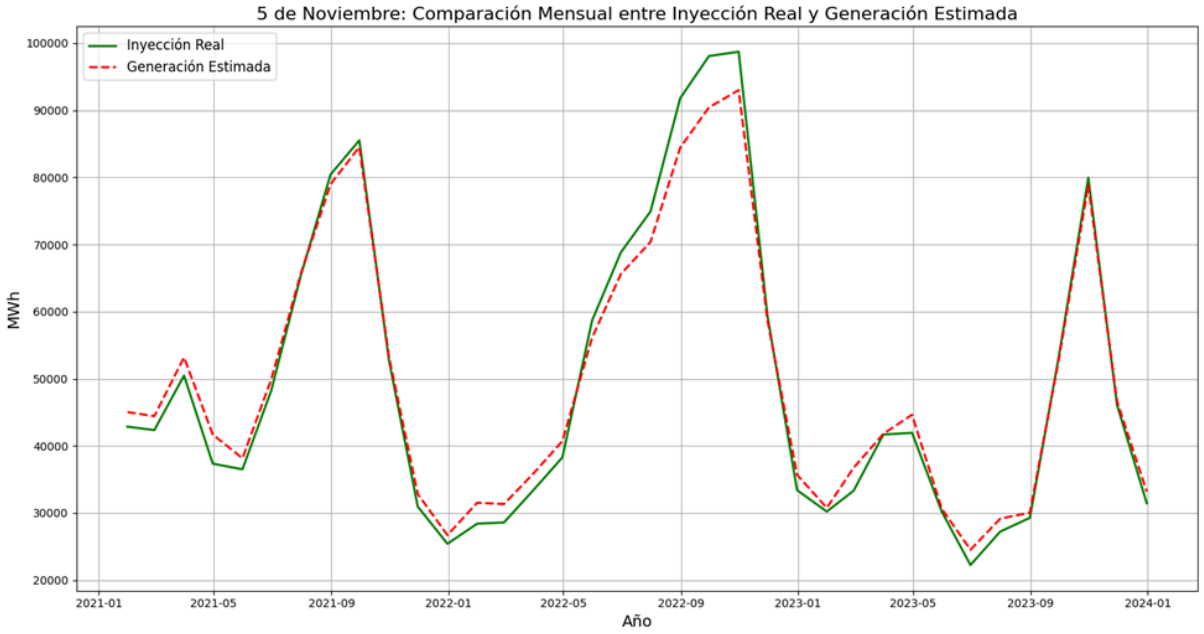
```
#Programa 11

# Paso 1: Separar los bits de Pred_binario

bits_separados = data['Pred_binario'].apply(lambda x: [int(bit) for bit in list(x)]).apply(pd.Series)

bits_separados.columns = ['u7', 'u6', 'u5', 'u4', 'u3', 'u2', 'u1']

# Paso 2: Aplicar las medianas como factores individualmente y guardarlos en columnas nuevas

for i, col in enumerate(['u7', 'u6', 'u5', 'u4', 'u3', 'u2', 'u1']):

    data[f'Estim_{col}'] = bits_separados[col] * factores[i]
```

```
# Paso 3: Calcular la suma total

data['Generación_estim'] = data[[f'Estim_{col}' for col in ['u7', 'u6', 'u5', 'u4', 'u3', 'u2', 'u1']]].sum(axis=1)

# Paso 4: Verificación con HoraFecha incluida

columnas_ver = ['HoraFecha', 'Pred_binario'] + [f'Estim_{col}' for col in ['u7', 'u6', 'u5', 'u4', 'u3', 'u2', 'u1']] + ['Generación_estim']

print(data[columnas_ver].head())

# Paso 5: Guardar en CSV con HoraFecha

data.to_csv('predicción_actualizada.csv', index=False)

print("Archivo 'predicción_actualizada.csv' generado exitosamente con HoraFecha y los 7 valores estimados por bit.")
```

En el programa 12, se realiza la comparativa del consolidado del programa anterior y el consolidado de inyecciones totales de 2021 a 2023 del documento de los embalses. El resultado de la comparación se puede observar en la gráfica que se encuentra después de este programa.

```
#Programa 12

# # Cerrar cualquier figura existente

plt.close('all')

# Asegurarse de que las fechas estén en formato datetime

data['HoraFecha'] = pd.to_datetime(data['HoraFecha'])

# Agrupar por frecuencia mensual

df_mensual = data.groupby(pd.Grouper(key='HoraFecha', freq='M')).sum()

# Crear una nueva figura

plt.figure(figsize=(15, 8))

plt.plot(df_mensual.index, df_mensual['5nov_iny'], label='Inyección Real', color='g', linewidth=2)

plt.plot(df_mensual.index, df_mensual['Generación_estim'], label='Generación Estimada', color='r', linestyle='--', linewidth=2)

plt.title('5 de Noviembre: Comparación Mensual entre Inyección Real y Generación Estimada', fontsize=16)
```

```
plt.xlabel('Año', fontsize=14)

plt.ylabel('MWh', fontsize=14)

plt.grid(True)

plt.legend(loc='upper left', fontsize=12)

plt.tight_layout()

plt.show()
```



En el programa 13, se realiza la correlación del parecido de las dos graficas en base al coeficiente de determinación. Según la ejecución del modelo la generación estimada se aproxima en un 98.25% a la gráfica de inyecciones reales.

```
#Programa 13

#Extraer las series de datos

Y_real = df_mensual['5nov_iny']

Y1 = df_mensual['Generación_estim']

# Calcular solo R²

r2_1 = r2_score(Y_real, Y1)

# Mostrar los resultados

print(f"R² para 'Generación_Estim': {r2_1:.4f}")
```

```
# Determinar la curva más cercana a 'Inyecc_real'

print("************************************")
```

```
R² para 'Generación_Estim': 0.9825
**********************************
```

En el programa 14, se procesan las columnas con los valores estimados manipulándolos para poder calcular los factores de planta individuales y en conjunto (tanto el conjunto de unidades 6 y 7 como el conjunto total).

```
# Programa 14

# Separar actividad según reporte UT: Inyección por unidades (MW)

iny5N = data[['HoraFecha', 'Estim_u1', 'Estim_u2', 'Estim_u3', 'Estim_u4', 'Estim_u5', 'Estim_u6',
'Estim_u7']]

# Sumar u6 + u7 en una columna nueva

iny5N['Estim_u7+u6'] = iny5N['Estim_u6'] + iny5N['Estim_u7']

# Agregar columna con suma total de Estim_u1 a Estim_u7

iny5N['Estim_u1_to_u7'] = iny5N[['Estim_u1', 'Estim_u2', 'Estim_u3', 'Estim_u4', 'Estim_u5',
'Estim_u6', 'Estim_u7']].sum(axis=1)

# Mostrar el resultado

print(iny5N)
```

En el programa 15, se realizan los acumulados de las columnas estimadas del programa anterior tanto de la 1 a la 7, 6 y 7 y totales para poder con esto procesar los factores de planta.

```
#programa 15

#inyecciones

#Se suman los valores de inyecciones (MWh) mediante agrupamiento diario (D), Mensual (M),
Anual(Y)

iny5N.set_index('HoraFecha', inplace=True)

iny_D  =iny5N.groupby(pd.Grouper(freq='D')).sum()

iny_M  =iny5N.groupby(pd.Grouper(freq='M')).sum()

iny_Y  =iny5N.groupby(pd.Grouper(freq='Y')).sum()
```

```
#Forma porcentual

iny_M_N =iny_M.divide(iny_M.sum(axis=1),axis=0)

#print(iny_D)

#print(iny_M)

print(iny_Y)
```

En el programa 16, Se calcula los factores de planta de las unidades individuales utilizando el número de horas anual, las capacidades nominales y los valores estimados del modelo. Posterior al programa se encuentra la tabla de resultados de los factores de planta de los años 2021, 2022 y 2023 de las unidades por separado, del conjunto de las unidades 6 y 7 y de la totalidad de las unidades.

```
#Programa 16

#Factor de planta: anual y mensual

#La capacidad (MW) para cada central utiliza datos nominales de potencia en MW

hours      = 24

hours_year  = 24*365

capacidad   = np.array([15,22.2,15,15,23.8,44.19,44.19,88.38,179.4])

#Anual

fpyear     = iny_Y.iloc[:,:]/(capacidad*hours_year)

print(fpyear)

#Mensual

hours_month = hours*iny_M.index.daysinmonth

fpmonth     = iny_M.loc[:,:].div(hours_month,axis=0)

fpmonth     = fpmonth.loc[:,:].div(capacidad)

print('**********')

print(fpmonth.mean(0))
```

```
            Estim_u1  Estim_u2  Estim_u3  Estim_u4  Estim_u5  Estim_u6  \
HoraFecha
2021-12-31  0.687552  0.519229  0.858114  0.137728  0.560503  0.201530
2022-12-31  0.718924  0.545568  0.954565  0.094975  0.573360  0.292703
2023-12-31  0.572122  0.409798  0.701308  0.169719  0.475534  0.136629

            Estim_u7  Estim_u7+u6  Estim_u1_to_u7
HoraFecha
2021-12-31  0.251604     0.226567        0.390980
2022-12-31  0.317053     0.304878        0.441637
2023-12-31  0.150945     0.143787        0.305298
**********
Estim_u1          0.658926
Estim_u2          0.491514
Estim_u3          0.839156
```

Adicional al modelo multiclasico, se realizó un análisis con datos continuos. Los modelos utilizados fueron Regresión lineal y Redes neuronales. A continuación, se comparten los programas y los resultados así como los gráficos correspondientes.

Análisis de modelo de predicción continua por unidad. Redes neuronales y regresión lineal.

En el programa 1 se desarrolla el ingreso de las bibliotecas necesarias para ejecutar las directivas necesarias para correr los programas y realizar los modelos, gráficas, sumatorias, procedimientos necesarios, etc en la predicción de datos continuos.

```
# ================================

# Programa 1: Importaciones

# ================================

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.linear_model import LinearRegression

from sklearn.neural_network import MLPRegressor

from sklearn.model_selection import train_test_split

from sklearn.metrics import r2_score, mean_squared_error

from sklearn.preprocessing import StandardScaler
```

```
from sklearn.pipeline import Pipeline
```

En el programa 2 se toman los datos proporcionados por la UT (Datos separados por unidades del año 2024) y se agregan al dataframe fname para su procesamiento posterior en df_2024. Se procede a llenar todas las celdas vacías con ceros y a ajustar los formatos de las fechas utilizadas.

```
# ================================

# Programa 2: Cargar datos 2024

# ================================

fname = 'Datos_UT_procesados_2024_completo.csv'

df_2024 = pd.read_csv(fname, parse_dates=['Fecha'], dayfirst=True, encoding='ISO-8859-1', thousands=',')

df_2024 = df_2024.fillna(0)

df_2024['Fecha'] = pd.to_datetime(df_2024['Fecha'], errors='coerce')

print("Datos 2024:")

print(df_2024.head())
```

En el programa 3 se realiza la lectura del archivo de embalses en el que se encuentran los datos de inyecciones totales, nivel de embalse y vertido. Especificando la carga de datos desde el 1 de enero de 2021 hasta el 31 de diciembre de 2023.

```
# ================================

# Programa 3: Cargar datos 2021–2023

# ================================

fname = 'embalses12-24.csv'

df_hist = pd.read_csv(fname, parse_dates=['HoraFecha'], dayfirst=True, encoding='ISO-8859-1', thousands=',')

df_hist = df_hist.loc[(df_hist['HoraFecha'] >= '2021-01-01') & (df_hist['HoraFecha'] <= '2023-12-31')]

#print(df_hist.head())
```

En el programa 4 se realizan las definiciones de los modelos tanto el de regresión lineal como el de redes neuronales brindándoles los datos de ‘”X” correspondientes a inyección real, nivel y vertido y “Y” las unidades individuales, todos correspondientes al año 2024.

```
# ====================================================

# Programa 4: Predicción por unidad (Lineal y NN)

# ====================================================

# Entrenamiento por unidad (2024)

df_hist = df_hist.rename(columns={'5nov_iny': 'Inyecc_real'})

X_hist = df_hist[['Inyecc_real', '5nov_niv', '5nov_vert']]

y_unidades = df_2024[['5nov-u1', '5nov-u2', '5nov-u3', '5nov-u4', '5nov-u5', '5nov-u6', '5nov-u7']]

X_train_u, X_test_u, y_train_u, y_test_u = train_test_split(X, y_unidades, test_size=0.2, random_state=42)

# Modelo 1: Regresión lineal

modelo_lineal_unit = LinearRegression()

modelo_lineal_unit.fit(X_train_u, y_train_u)

y_pred_lin_unit = modelo_lineal_unit.predict(X_test_u)

# Modelo 2: Red neuronal

modelo_nn_unit = Pipeline([

    ('scaler', StandardScaler()),

        ('nn', MLPRegressor(hidden_layer_sizes=(100, 50), activation='relu', solver='adam', max_iter=1000, random_state=42))

])

modelo_nn_unit.fit(X_train_u, y_train_u)

y_pred_nn_unit = modelo_nn_unit.predict(X_test_u)
```

En el programa 5 se prueban los modelos entregándole los datos provenientes del programa 3 correspondientes a inyecciones, nivel y vertido de los años 2021 al 2023 para realizar las predicciones en ambos modelos en conjunto. Los datos resultantes producto de la predicción se guardan en el archivo prediccion_unidades_2021_2023.csv en los apartados de generación_lineal y Generación_nn. Por último se hacen las sumatorias de los estimados de cada modelo para obtener los totales y compararlos de manera gráfica con la gráfica de inyecciones reales del mismo período.

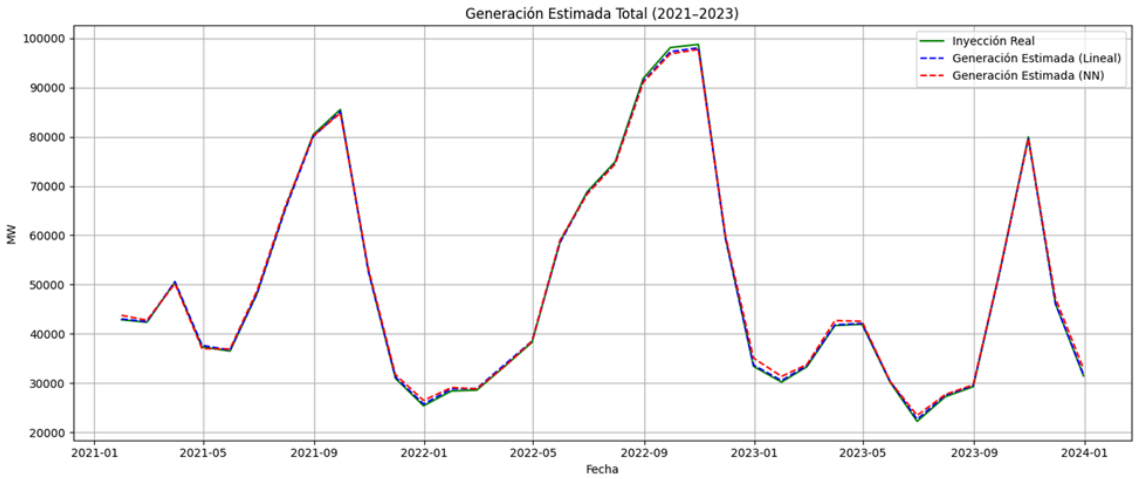
```
# ====================================================
```

```
# Programa 5: Predicción por unidad (2021–2023)

# ====================================================

pred_lin_unit = modelo_lineal_unit.predict(X_hist)

pred_nn_unit = modelo_nn_unit.predict(X_hist)

cols_units = ['5nov-u1', '5nov-u2', '5nov-u3', '5nov-u4', '5nov-u5', '5nov-u6', '5nov-u7']

df_lin_units = pd.DataFrame(pred_lin_unit, columns=[f'Estim_Lin_{c}' for c in cols_units])

df_nn_units = pd.DataFrame(pred_nn_unit, columns=[f'Estim_NN_{c}' for c in cols_units])

# Concatenar

df_hist = pd.concat([df_hist.reset_index(drop=True), df_lin_units, df_nn_units], axis=1)

# Sumar las unidades

df_hist['Generacion_lineal_sumada'] = df_lin_units.sum(axis=1)

df_hist['Generacion_nn_sumada'] = df_nn_units.sum(axis=1)

# Guardar

df_hist.to_csv('prediccion_unidades_2021_2023.csv', index=False)

print("\nArchivo 'prediccion_unidades_2021_2023.csv' generado exitosamente.")
```

En el programa 6 se pueden confrontar tanto las gráficas de generación estimada (Regresión Lineal y Redes Neuronales) como la de inyecciones reales y se pueden corroborar las similitudes entre las mismas.

```
# ====================================================

# Programa 6: Visualización mensual comparativa

# ====================================================

df_mensual_unidades = df_hist.groupby(pd.Grouper(key='HoraFecha', freq='M')).sum()

plt.figure(figsize=(14, 6))

plt.plot(df_mensual_unidades.index, df_mensual_unidades['Inyecc_real'], label='Inyección Real',
color='green')

plt.plot(df_mensual_unidades.index, df_mensual_unidades['Generacion_lineal_sumada'],
label='Estimación Lineal (sumada)', linestyle='--', color='blue')
```

```
plt.plot(df_mensual_unidades.index, df_mensual_unidades['Generacion_nn_sumada'], label='Estimación NN (sumada)', linestyle='--', color='red')

plt.title('Generación Estimada Total desde Unidades (2021–2023)')

plt.xlabel('Fecha')

plt.ylabel('MW')

plt.grid(True)

plt.legend()

plt.tight_layout()

plt.show()
```



En el programa 7 se hace la comparación numérica en el coeficiente de correlación entre las 3 graficas. Comparando las generaciones estimadas con la inyección real. Al revisar los resultados, se puede corroborar en el análisis realizado que la curva de generación estimada lineal se corresponde en un 99.98% a la curva de inyección real. Por otro lado, la curva estimada con Redes neuronales, se corresponde en un 99.87 % con la de Inyección real.

```
#Programa 7

#Extraer las series de datos

Y_real = df_mensual_unidades['Inyecc_real']

Y1 = df_mensual_unidades['Generacion_lineal_sumada']

Y2 = df_mensual_unidades['Generacion_nn_sumada']

# Calcular solo R²
```

```
r2_1 = r2_score(Y_real, Y1)

r2_2 = r2_score(Y_real, Y2)

# Mostrar los resultados

print(f"R² para 'Generación lineal sumada': {r2_1:.4f}")

print(f"R² para 'Generacion red_neu sumada': {r2_2:.4f}")

# Determinar la curva más cercana a 'Inyecc_real'

print("************************************")
```

```
R² para 'Generación lineal sumada': 0.9998
R² para 'Generacion red_neu sumada': 0.9987
************************************
```

En el programa 8 se realiza la manipulación de los datos del modelo que brinda el mejor resultado (Regresión lineal) para hacer el análisis de factores de planta con los valores estimados. Se preparan las columnas para obtener los factores de planta de las unidades estimadas de la 1 a la 7, el conjunto de las unidades 6 y 7 y el conjunto de las unidades 1 a la 7 (la totalidad).

```
#Programa 8

#separar actividad segun reporte UT: Inyección por unidades (MW)

#iny5N     = df [['Fecha','5nov-u1','5nov-u2','5nov-u3','5nov-u4','5nov-u5','5nov-u6','5nov-u7']]

df_mensual_unidades = df_mensual_unidades.reset_index()

iny5N = df_mensual_unidades[['HoraFecha','Estim_Lin_5nov-u1','Estim_Lin_5nov-u2','Estim_Lin_5nov-u3','Estim_Lin_5nov-u4','Estim_Lin_5nov-u5','Estim_Lin_5nov-u6','Estim_Lin_5nov-u7']]

iny5N['Estim_Lin_5nov-u6-u7'] = iny5N['Estim_Lin_5nov-u6'] + iny5N['Estim_Lin_5nov-u7']

print(iny5N)
```

En el programa 9 se realizan las sumatorias diaria, mensual y anual de las estimaciones de las inyecciones correspondientes a la regresión lineal.

```
#programa 9

#inyecciones
```

```
#Se suman los valores de inyecciones (MWh) mediante agrupamiento diario (D), Mensual (M), Anual(Y)

iny5N.set_index('HoraFecha', inplace=True)

iny_D   =iny5N.groupby(pd.Grouper(freq='D')).sum()

iny_M   =iny5N.groupby(pd.Grouper(freq='M')).sum()

iny_Y   =iny5N.groupby(pd.Grouper(freq='Y')).sum()

#Forma porcentual

iny_M_N =iny_M.divide(iny_M.sum(axis=1),axis=0)

#print(iny_D)

#print(iny_M)

print(iny_Y)
```

En el programa 10 se realizan los cálculos de los factores de planta brindando los datos de potencia nominal, los valores horarios por año, y los consolidados de estimaciones de inyección correspondientes a regresión lineal.

```
#Programa 10

#Factor de planta: anual y mensual

#La capacidad (MW) para cada central utiliza datos nominales de potencia en MW

hours      = 24

hours_year  = 24*365

capacidad   = np.array([15,22.2,15,15,23.8,44.19,44.19,88.38,179.4])

#Anual

fpyear      = iny_Y.iloc[:,:]/(capacidad*hours_year)

print(fpyear)

#Mensual

hours_month = hours*iny_M.index.daysinmonth

fpmonth     = iny_M.loc[:,:].div(hours_month,axis=0)

fpmonth     = fpmonth.loc[:,:].div(capacidad)
```

```
print('**********')

print(fpmonth.mean(0))
```

```
            Estim_Lin_Snov-u1  Estim_Lin_Snov-u2  Estim_Lin_Snov-u3  \
HoraFecha
2021-12-31           0.623982           0.445983           0.723506
2022-12-31           0.718199           0.507749           0.807360
2023-12-31           0.519049           0.375040           0.633398

            Estim_Lin_Snov-u4  Estim_Lin_Snov-u5  Estim_Lin_Snov-u6  \
HoraFecha
2021-12-31           0.152399           0.450012           0.277772
2022-12-31           0.134249           0.499773           0.366728
2023-12-31           0.192408           0.393738           0.171687

            Estim_Lin_Snov-u7  Estim_Lin_Snov-u6-u7  Estim_Lin_Snov-u1-a-u7
HoraFecha
2021-12-31           0.293545              0.285658               0.381025
2022-12-31           0.382989              0.374858               0.452585
2023-12-31           0.182373              0.177030               0.298303
**********
```

```
Estim_Lin_Snov-u1         0.619616
Estim_Lin_Snov-u2         0.442444
Estim_Lin_Snov-u3         0.720695
Estim_Lin_Snov-u4         0.159368
Estim_Lin_Snov-u5         0.447492
Estim_Lin_Snov-u6         0.271793
Estim_Lin_Snov-u7         0.286045
Estim_Lin_Snov-u6-u7      0.278919
Estim_Lin_Snov-u1-a-u7    0.376915
dtype: float64
```

Se presenta en las siguientes imágenes el resultado de las estimaciones del factor de planta producto del modelo de regresión lineal (directamente con datos continuos).

En síntesis, se presenta en la siguiente tabla la comparativa de los resultados de factores de planta generales. Se compara el factor de planta general calculado en el primer análisis en conjunto con el calculado en el modelo multiclasico a su vez comparado con el modelo de regresión lineal y redes neuronales.

En conclusión, los valores para los 3 años son bastante cercanos y se puede inferir que los modelos son representativos. Ahora se debe concluir mostrando los datos estimados de factores de planta del conjunto de las unidades 6 y 7 los cuales se ven en la siguiente tabla, como se puede observar, los datos son diferentes para cada modelo, pero se puede inferir que con respecto a la estimación hecha por la CEL de que el proyecto de expansión operaría en un factor de planta del 18%, se ha cumplido.

Año	FP Real Total (embalses)	FP Modelo Multiclasico	FP Regresión Lineal y Redes Neuronales
2021	0.379447	0.39098	0.381025
2022	0.451997	0.441637	0.452585
2023	0.296596	0.305298	0.298303

Año	FP Unidades 6 y 7 Multiclasico	FP unidades 6 y 7 regresión Lineal y Redes Neuronales
2021	0.226567	0.285658
2022	0.304878	0.374858
2023	0.143787	0.17703

Anexo D: Repositorio de programación de los desarrollos realizados en este trabajo.

https://github.com/JoseDePaz94/5_Nov_TBE_Git