BALANCE ENERGÉTICO - PROYECCIÓN EN LA GENERACIÓN DE PLANTAS DE ENERGÍA.

Autor

Mauricio Duque Quintero

Informe de práctica como requisito para optar al título de:

Ingeniero Electrónico

Asesor Interno

Sebastián Isaza Ramírez

Profesor Vinculado Universidad de Antioquia, PhD

Universidad de Antioquia
Facultad de Ingeniería, Departamento de Ingeniería Electrónica
Medellín, Colombia
2022.

BALANCE ENERGÉTICO - PROYECCIÓN EN LA GENERACIÓN DE PLANTAS DE ENERGÍA.

Cita	Duque Quintero [1]
Referencia	[1] M. Duque Quintero, " Balance Energético – Proyección en la generación de plantas de energía.", Semestre de Industria, Ingeniería Electrónica, Universidad de Antioquia, Medellín, 2022.
Estilo IEEE (2020)	

Resumen

En el área de planeación de la empresa para la cual se desarrolló este proyecto, calculaban por medio de Excel el promedio de generación de energía a partir de los datos históricos en cada una de sus plantas hidroeléctricas y con ellos pronosticaban su generación para próximos años.

Por tanto, el propósito de este proyecto fue implementar una solución de analítica avanzada, a través de un ensamble de modelos estadísticos y Machine Learning que permitiera mejorar y automatizar el proceso de pronóstico en la generación de energía para el año 2022 con granularidad mensual, en 16 plantas hidroeléctricas, reduciendo los cálculos "manuales" de sus analistas, para aumentar la productividad y rentabilidad en su negocio. Los resultados del desempeño del modelo muestran que de forma satisfactoria en la mayoría de sus plantas se logra mejor el pronóstico de energía en hasta un 11% con respecto a los cálculos clásicos que ellos realizaban, dando un alto grado de confiabilidad en los datos obtenidos.

Tabla de Contenido

Índice de tablas

Índice de ilustraciones

1. Introducción

Con un 68% de participación en la generación de energía, las centrales hidroeléctricas son la mayor fuente de oferta energética del país [1]. Es por esto, que cada día diferentes compañías de energía invierten en diferentes alternativas de innovación, con fin de potenciar al máximo los recursos para aumentar su productividad y obtener una mayor rentabilidad en su negocio.

La "Empresa" [1] requiere pronosticar para los próximos 12 meses, la generación de energía para cada una de sus plantas hidroeléctricas. A la fecha, el departamento de operación y planeación de esta empresa, realizan los pronósticos de forma manual, basados en la experiencia de sus analistas y técnicos sobre los eventos hidrológicos, fallas y mantenimiento, analizando las estadísticas de los datos históricos de la generación de energía en Megavatios (MW) con medias estándar, desviaciones y percentiles.

Con el fin de sistematizar dicho proceso, se busca construir una data histórica del comportamiento de las plantas desde el año 2010, que luego de ser procesadas a través de Machine Learning, permitan a la "Empresa" tomar decisiones a futuro y mapear las necesidades de energía para llegar a las metas estimadas, minimizando la carga operativa actual del proceso, teniendo un control sobre los posibles eventos que se puedan presentar y desarrollar un crecimiento sostenible de la operación, que permitan mejorar a través de la analítica, el pronóstico de la generación de energía.

El proyecto balance energético, es una iniciativa de las diferentes estrategias de negocio para buscar soluciones, haciendo uso de las herramientas de la era digital como lo es el Machine Learning, bajo diseño de arquitectura en la nube, de código abierto.

El presente documento, contiene el diseño y la descripción de la solución analítica avanzada que se desarrollará en el proyecto, con el uso de algoritmos de series de tiempo convencionales y técnicas de Machine Learning, así como los elementos asociados entre los datos y el análisis del desempeño de los resultados obtenidos.

[1] A lo largo de este documento, se menciona el nombre "Empresa" haciendo referencia al cliente con el cual se está desarrollando la solución requerida, esto debido a que, la empresa contratante DataKnow S.A.S, se reserva el derecho de revelar su nombre y/o tipo de vinculación.

2. Objetivos

Objetivo General

Desarrollar una herramienta flexible e interactiva que pronostique la generación de energía en las diferentes centrales hidroeléctricas de la "Empresa", utilizando modelos de Machine Learning y servicios en la nube, con el fin de disminuir operatividad manual al realizar los cálculos y análisis en su área de planeación y operación de mercados energéticos, generando escenarios que permitan responder a las dinámicas de mercado.

Objetivos específicos

- Realizar un análisis exploratorio de los datos EDA, con el propósito de definir un umbral de decisión, en el cual los valores atípicos o faltantes no afecten el desarrollo de los modelos predictivos.

- Construir un data set, que permita definir las variables de interés de la "Empresa" y la granularidad según el tipo de planta y la generación de energía.

- Configurar modelos de series de tiempo, utilizando técnicas clásicas estadísticas como ARIMA, algoritmos de Machine Learning (ML) basados en árboles de regresión XGBoost, para evaluar la capacidad predictiva o medida de desempeño, de cada uno de los modelos en la generación de energía para los próximos doce meses.

- Realizar un ensamble de los modelos evaluados, con la media como estadístico de prueba, con el fin de capturar características o tendencias que alguno de los modelos de forma independiente no logre mapear.

- Implementar los diferentes modelos en los servicios en la nube de Azure Databricks, que permitan el aprendizaje automatizado en un entorno de producción.

- Diseñar y crear un tablero en Power BI, donde se presente el consolidado de los resultados obtenidos, con los diferentes indicadores que requiera el negocio.

3. Marco Teórico

3.1. Forecasting

Forecasting es un método para realizar predicciones a cerca de un conjunto de variables de interés, usando datos históricos o lo que se denomina series de tiempo como entrada principal para determinar el comportamiento de las tendencias futuras. [2] Ser capaz de predecir con alto grado de precisión tendencias eventos futuros es útil en muchos contextos porque se puede utilizar para:

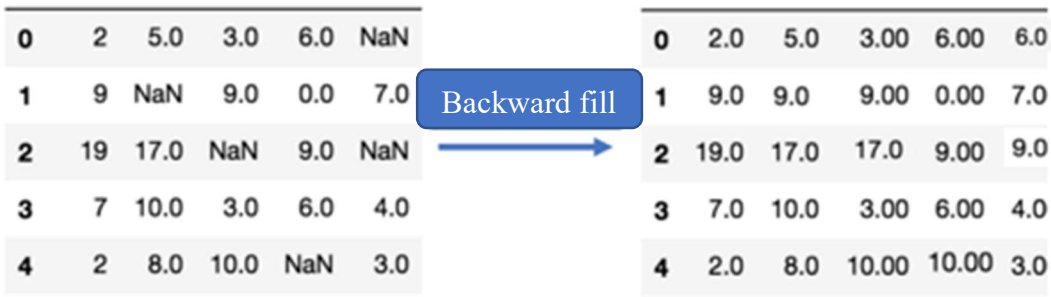
- Aumentar las probabilidades del éxito y la atención a las necesidades financieras de una empresa.
- Garantizar la consistencia operativa de la empresa, Formular planes efectivos para el futuro.
- Ayudar a los gerentes a tomar las decisiones correctas.

Antes de llevar a cabo un pronóstico, que en este caso será la generación de energía, se requiere que todos los elementos del sistema necesitan ser revisados y sus valores relativos analizados. Dependiendo del pronóstico requerido, esto puede implicar un análisis exploratorio en profundidad de cualquier elemento relevante del sistema de ventas.

3.2. Imputación de datos

La **imputación de datos** es el proceso de identificar y sustituir los registros sin información o que poseen valores ausentes y valores atípicos en una serie de tiempo. La estimación se puede hacer a partir de la información que contiene el conjunto de variables de interés. Usualmente los métodos de imputación se utilizan con variables métricas de intervalo o de razón. [3]

Para el tratamiento de los valores ausentes se utiliza el **método Backward fill** el cual consiste en estimar el dato faltante a partir de su valor anterior en una estructura temporal, como se muestra en la Figura 1. Por otra parte, para el manejo de los valores atípicos se hace uso del **filtro de Hampel**, a través de una ventana deslizante de ancho configurable para "suavizar" dichos valores, calculando la mediana y la desviación estándar de sus K vecinos, si el valor de esta sobrepasa una cantidad de desvíos, entonces es atípico y se reemplaza por su valor medio. Como el filtro usa una ventana deslizante, tiene más sentido usarlo con datos de series de tiempo, donde el orden de los datos se rige por el tiempo. [4]



0	2	5.0	3.0	6.0	NaN
1	9	NaN	9.0	0.0	7.0
2	19	17.0	NaN	9.0	NaN
3	7	10.0	3.0	6.0	4.0
4	2	8.0	10.0	NaN	3.0

0	2.0	5.0	3.00	6.00	6.0
1	9.0	9.0	9.00	0.00	7.0
2	19.0	17.0	17.0	9.00	9.0
3	7.0	10.0	3.00	6.00	4.0
4	2.0	8.0	10.00	10.00	3.0

Figura1. Método backward fill de imputación

3.3. Modelo ARIMAX

Autoregressive Integrated Moving Average (ARIMA) son modelos de serie de tiempo y análisis estadístico clásico, que tiene como objetivo interpretar los datos históricos y usar regresiones lineales para realizar predicciones futuras. Son modelos de series de tiempo que pueden identificar o caracterizar las series de acuerdo con su comportamiento, dinámicas y estacionariedad, mostrando tendencias, varianza no constante, o ciclos en las series; descomponiendo la serie de tiempo en sus elementos AR (Autorregresivos) de Raíces unitarias (I) de Medias Móviles (MA) e incorporando en algunos casos variables externas (X). Un modelo ARIMA se caracteriza por 3 parámetros (p, d, q) que les asignan valores enteros específicos que indican el tipo de modelo [5]:

$$ARIMA(p,d,q)(P,D,Q)[s] \qquad (1)$$

- **p** es el número de términos autorregresivos o el número de "observaciones de retraso" y determina el resultado del modelo al proporcionar puntos de datos rezagados.

- **d** es el número de diferencias necesarias para que la serie temporal sea estacionaria. Y si la serie temporal ya es estacionaria, entonces d = 0.

- **q** es el número de errores de pronóstico en el modelo y también se conoce como el tamaño de la ventana de media móvil.

- **s** Los procesos estacionales están correlacionados s, 2s, 3s, ... periodos o distancias temporales en múltiplos de s, con s par.

El modelo ARIMA se puede utilizar para pronosticar cantidades futuras basados en datos históricos. Por lo tanto, para que el modelo sea confiable, los datos deben ser veraces y mostrar un período de tiempo relativamente largo durante el cual se

han recopilado. Siendo así, una opción viable para pronosticar la generación de energía a partir de la data histórica con la que se cuenta.

3.4. Modelo XGBoost

XGBoost es un algoritmo de aprendizaje automático, que pertenece a una familia de algoritmos de impulso, basada en árboles de decisión o regresión que utiliza un marco de aumento de gradiente (Gradient Boosting Machines - GBM). El algoritmo XGBoost fue desarrollado como un proyecto de investigación en la Universidad de Washington por Tianqi Chen y Carlos Guestrin. [7]

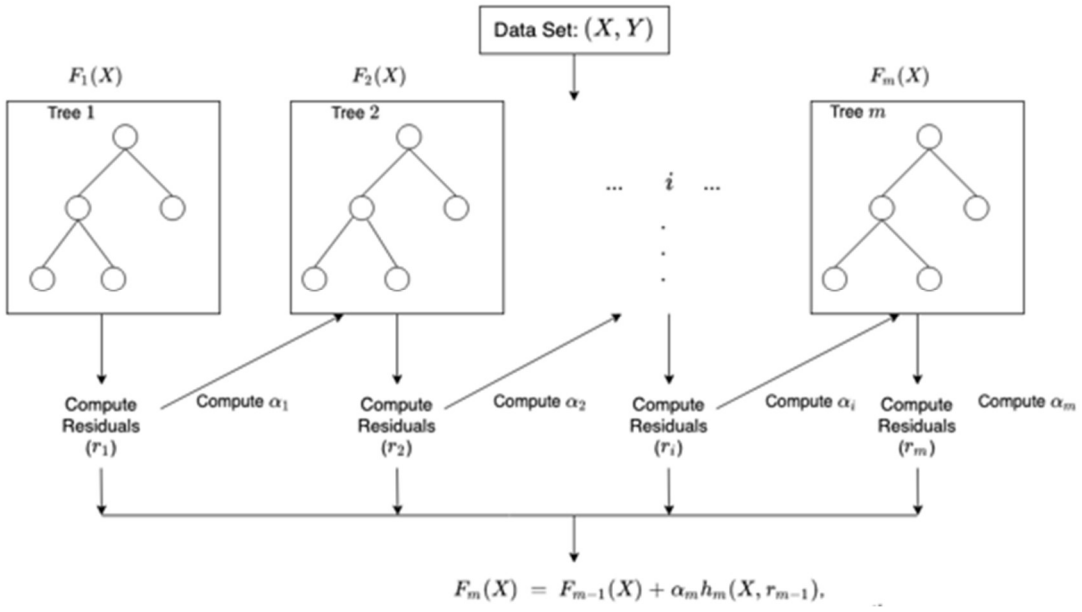


Figura 2. Arquitectura modelo XGBoost

XGBoost posee un conjunto de características que lo convierten en un algoritmo sobresaliente en comparación con otros algoritmos de forecasting:

- Brinda una paralelización en la implementación del proceso.
- Cada nuevo árbol minimiza los errores de los árboles anteriores.
- Regula los parámetros para evitar sobre ajustes.
- Permite hacer validación cruzada incorporado en cada iteración.
- Posee un gran rendimiento computacional y menor tiempo de procesamiento.
- Una amplia gama de aplicaciones al utilizarse para resolver problemas de regresión, clasificación y predicción definidos por el usuario.

Es por esto, que se convierte en una gran opción, para ser incluido en los modelos que se usaron en la realización del proyecto.

3.5. Microsoft Azure

La plataforma **Microsoft Azure** está compuesta por más de 200 productos y servicios en la nube diseñados para brindar soluciones que permitan resolver las dificultades actuales, al crear, ejecutar y administrar aplicaciones en varias nubes, en el entorno local y sus alrededores, con las herramientas y los marcos que se prefiera [8]. Para el desarrollo de este proyecto se hará uso de los siguientes servicios, que se trabajan bajo las suscripciones de la "Empresa":

- **Azure Databricks** es una plataforma de servicios en la nube de Azure utilizada para el análisis de datos, bajo un ambiente clusterizado que proporciona las últimas versiones de Apache Spark y permite la integración sin problemas con bibliotecas de código abierto.

- **Azure Blob Storage** es la solución de almacenamiento de objetos de Microsoft para la nube. Blob Storage está optimizado para el almacenamiento de cantidades masivas de datos. Los contenedores de blobs se pueden utilizar para almacenar los registros del servicio de puntuación de un modelo de Machine Learning y para recopilar tanto los datos de entrada como la predicción del modelo. Después de cierta transformación, los registros se pueden usar para el reentrenamiento de modelos.

Debido a que "La Empresa" cuenta con licenciamientos y cuentas de Microsoft Azure, es por esto que se utilizan cada uno de los servicios mencionados anteriormente.

3.6. Métricas de desempeño MAPE y RMSE

Error Porcentual Absoluto Medio (MAPE) es un indicador del desempeño utilizado para medir el tamaño del error absoluto en términos de porcentaje. El hecho que se estime una magnitud del error porcentual lo hace un indicador frecuentemente utilizado para el pronóstico de la demanda debido a su fácil interpretación.

$$MAPE = \left(\frac{1}{n}\sum_{1}^{n}\frac{\left|Vreal - \widehat{Vpronostico}\right|}{Vreal}\right) * 100 \qquad (2)$$

Error Cuadrático Medio (RMSE) mide la desviación de error que hay entre dos conjuntos de datos. RMSE compara un valor predicho con respecto a un valor conocido o histórico**.** Entre mayor el resultado mayor es el error y menos preciso el modelo. Es una medida altamente sensible a desviaciones grandes entre la demanda real y pronosticada. [10]

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}\left(Vreal - \widehat{Vpronostico}\right)^2} \quad (3)$$

3.7. Planta Hidroeléctrica

Una **central hidroeléctrica** es una instalación que permite el aprovechamiento de la hidrología que circulan por los ríos o embalses, para transformarlas en energía eléctrica, utilizando turbinas acopladas a generadores. Después de este proceso, el agua se devuelve al río en las condiciones en que se tomó, de modo que se puede volver a usar por otra central situada aguas abajo o para consumo [9].

Es importante tener un conocimiento previo de las variables que se van a tratar, para lograr un mejor entendimiento y análisis de los datos dados por la "empresa" y el pronóstico que arrojen los diferentes modelos.

4. Metodología

En este capítulo se aborda cada una de las etapas que se llevaron a cabo para lograr los objetivos propuestos del proyecto, donde se explica los procedimientos e implementación realizadas.

4.1. Arquitectura

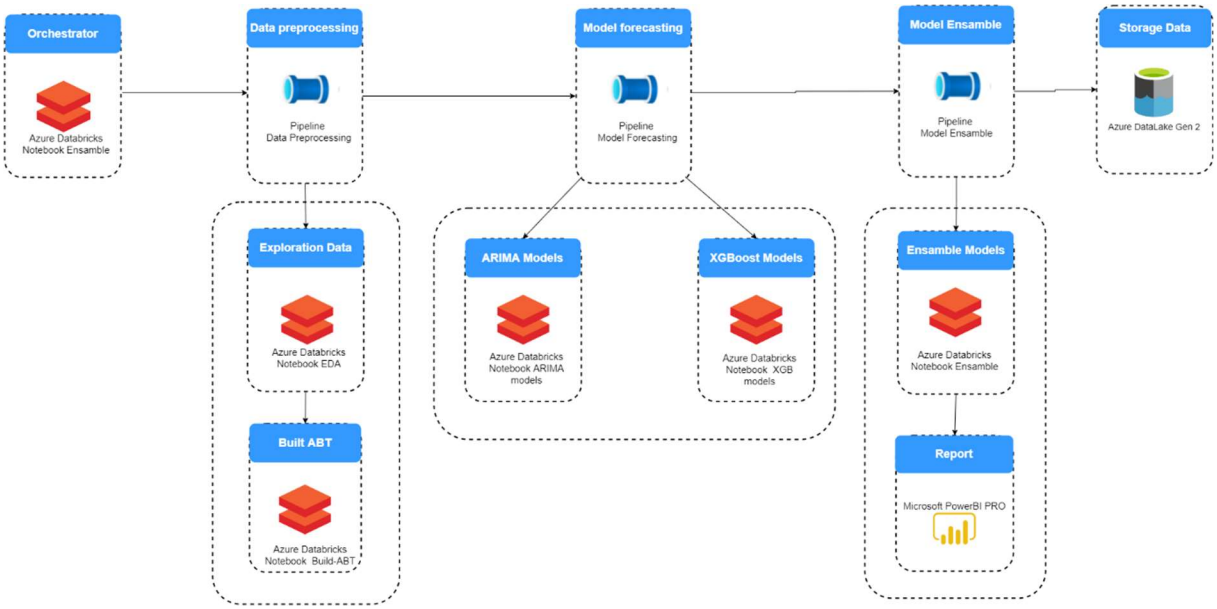


Figura 3. Arquitectura del proyecto

Para el desarrollo del proyecto, se implementó la arquitectura expuesta en la Figura 3. Que consta de las siguientes etapas, las cuales se especifican a continuación:

Orchestrator: el proceso se encuentra administrado bajo un orquestador encargado de ejecutar cada una de las etapas de forma automática en un ambiente productivo, a través, de flujos previamente definidos. Todo ello en función de asegurar entornos de interacción tanto a niveles de dispositivos (tableros de Power BI) como a niveles de servicios (herramientas de Microsoft).

Data preprocessing: una vez se cuenta con los datos históricos, se realiza un análisis exploratorio de los datos y se realiza la construcción de la base analítica (ABT) que permitirá preparar el data set que se consumirá en la etapa de modelamiento

Model Forecasting: encargado de realizar el desarrollo y entrenamiento de los modelos planteados ARIMA y XGBoost y todo el proceso de evaluación de desempeño.

Model Ensamble: Cuando se tiene el resultado del pronóstico en cada uno de los diferentes modelos se construye el ensamble de los modelos. Por último, una vez se tenga los resultados finales, se carga la información que será consumida por el tablero en Power BI.

4.2. Análisis exploratorio de los datos

Con el fin de validar y verificar la integridad de los datos históricos que se obtuvieron en la etapa de extracción, se realiza un análisis exploratorio de los datos (EDA), que permitirán preparar y transformar los datos que son el suministro de la etapa de modelación. El análisis EDA consiste en determinar las siguientes variables:

4.2.1. Faltantes y discontinuidades

Para verificar si la información histórica contenía datos faltantes para las series de tiempos a modelar, se realizó el conteo de días que habían transcurrido a partir de la fecha en que comienza a operar y los días que se tenían registro de la generación en cada una de las plantas.

4.2.2. Valores ceros

La presencia de valores ceros en las series de tiempo, pueden representar una caída en el pronóstico, el cual no se acerque al esperado. Por lo tanto, fue necesario realizar un análisis de dichos valores. Para este proceso, se dividieron los datos históricos en 4 franjas horarias: 1-6, 7-12, 13-18, 19-24, con el fin de determinar el número de días, donde la generación es cero. Luego, se creó una variable Booleana que permite verificar si para una fecha determinada el valor en cada hora es cero. Es decir, 0 si tiene un valor positivo o 1 si el valor es igual a 0. Por último, se agrupa de forma mensual y se suma la variable booleana para identificar cuando puede ocurrir un evento de importancia.

4.2.3. Valores Atípicos y negativos

Para el análisis de valores negativos se realizó a partir de un gráfico de cajas y bigotes, el cual permite identificar la distribución para la generación de energía de cada una de las plantas, así como su valor mínimo y máximo de generación, donde se evidenció que el valor mínimo de generación en todas las plantas es de cero, por lo que no existen valores negativos en las series de tiempo.

4.3. Construcción ABT (Analytical Base Table)

La construcción de la tabla de base analítica (ABT), se realizó con el fin de procesar los datos faltantes y atípicos resultantes del análisis exploratorio, expuesto en el numeral anterior, para garantizar un set de datos "limpio" y que contengan las variables necesarias, que permita un buen rendimiento en el modelamiento de las series de tiempo.

4.3.1. Imputación de valores ausentes

Para el tratamiento de los valores ausentes, en primera medida se realizó un recorte temporal en las plantas RPD1, RMR1, RFR1 y RFR2 que poseían una generación cero de energía en el periodo de 2008 - 2009 que según la "Empresa" eran plantas que aún se encontraban en proceso de adecuación. Para los valores faltantes de las demás plantas, se utilizó el método Backward fill expuesto en la Figura 1, que transforma el dato faltante en el valor previo de la serie. Este proceso se realiza con el fin de garantizar una continuidad en la serie de tiempo.

4.3.2. Filtrado y suavizado de valores atípicos

El manejo que se le dio a los valores atípicos fue la aplicación de filtros que permiten suavizar y controlar la distribución de la serie, disminuyendo la varianza del error. Para el desarrollo del proyecto se utiliza el filtro de Hampel, que, a partir de la información histórica de las series de tiempo, detecta y reemplaza valores atípicos en línea mientras preserva la demás información de los datos. Este método es eficiente para las series de tiempo no estacionarias, debido a su ventana deslizante de la desviación absoluta media.

4.3.3. Inclusión Eventos y fenómenos

De acuerdo con patrones establecidos por la "Empresa" fue necesario la incorporación de lógicas de mantenimiento y eventos climatológicos como la sequía y fenómenos del niño y niña, con el fin de detectar picos y tendencias inesperadas en la serie que los modelos no son capaces de detectar únicamente con una estructura temporal, que, para este caso, era la granularidad mensual de la generación de energía.

Para introducir estos eventos al modelado, se crearon variables dicotómicas independientes, en las cuales, si en la fecha correspondiente ocurrió un evento, entonces el valor de la variable era 1, de lo contrario era 0.

En el caso de los fenómenos del niño y la niña se partió de la información suministrada por "El Niño and La Niña Years and Intensities "que se muestra en la tabla 1 y se crea una lógica que permite definir en variables dicotómicas los periodos mensuales en que se presenta el fenómeno del niño y la niña a partir de la temperatura superficial del océano pacífico.

Tabla 1. Histórico de temperaturas superficiales - fenómeno del niño y la niña

Temperatura superficial del océano pacífico

WL	2008-2009	-0,4	-0,3	-0,3	-0,4	-0,6	-0,7	-0,8	-0,7	-0,5	-0,2	0,1	0,4
ME	2009-2010	0,5	0,5	0,7	1	1,3	1,6	1,5	1,3	0,9	0,4	-0,1	-0,6
SL	2010-2011	-1	-1,4	-1,6	-1,7	-1,7	-1,6	-1,4	-1,1	-0,8	-0,6	-0,5	-0,4
ML	2011-2012	-0,5	-0,7	-0,9	-1,1	-1,1	-1	-0,8	-0,6	-0,5	-0,4	-0,2	0,1
	2012-2013	0,3	0,3	0,3	0,2	0	-0,2	-0,4	-0,3	-0,2	-0,2	-0,3	-0,3
	2013-2014	-0,4	-0,4	-0,3	-0,2	-0,2	-0,3	-0,4	-0,4	-0,2	0,1	0,3	0,2
WE	2014-2015	0,1	0	0,2	0,4	0,6	0,7	0,6	0,6	0,6	0,8	1	1,2
VSE	2015-2016	1,5	1,9	2,2	2,4	2,6	2,6	2,5	2,1	1,6	0,9	0,4	-0,1
WL	2016-2017	-0,4	-0,5	-0,6	-0,7	-0,7	-0,6	-0,3	-0,2	0,1	0,2	0,3	0,3
WL	2017-2018	0,1	-0,1	-0,4	-0,7	-0,8	-1	-0,9	-0,9	-0,7	-0,5	-0,2	0
WE	2018-2019	0,1	0,2	0,5	0,8	0,9	0,8	0,8	0,7	0,7	0,7	0,5	0,5
	2019-2020	0,3	0,1	0,2	0,4	0,5	0,6	0,5	0,5	0,4	0,2	-0,1	-0,3
ML	2020-2021	-0,4	-0,6	-0,9	-1,2	-1,3	-1,2	-1,1	-0,9	-0,8	-0,7	-0,5	-0,4
	2021-2022	-0,4	-0,5	-0,7	-0,8	-1	-1	-1	-0,92	-0,7	-0,6	-0,44	-0,25
	2022-2029	-0,09	-0,03	-0,02	0,10	0,10							
ENSO Type	Season	JJA	JAS	ASO	SON	OND	NDJ	DJF	JFM	FMA	MAM	AMJ	MJJ

Para determinar los valores futuros de ambos fenómenos que permitan el pronóstico de la generación de energía, se toma el registro de posibles temperaturas de una serie de modelos proporcionados por "Climatic Predition Center".

En cuanto a los periodos de sequía, se obtuvieron de acuerdo con el comportamiento histórico de las series, donde se presentaba una disminución significativa en la generación de energía, analizados a través de un diagrama de cajas y bigotes para cada una de las plantas.

4.4. Modelamiento

Para realizar el pronóstico de la generación de energía para los próximos doce meses, se utilizaron dos enfoques de modelos, uno de análisis estadístico clásico ARIMAX y un enfoque de Machine Learning XGBoost, donde se realizaron diferentes versiones para cada uno de ellos midiendo y mejorando las medidas de desempeño en cada versión a través del MAPE y el RMSE.

Con el fin de mejorar el rendimiento de los modelos, se usan técnicas de ajuste que son exhaustivas y aleatorias de hiperparámetros que permiten obtener de manera automática las configuraciones que optimizan la exactitud de las estimaciones, a lo que se le denomina tuneo de hiperparámetros.

Por otra parte, se dividió el data set en 3 segmentos: entrenamiento, validación y test, donde para los dos últimos se tomó una franja temporal de 1 año.

Las etapas que se llevaron a cabo en el proceso de modelado fueron las siguientes;

- La primera versión consistió en entrenar un modelo ARIMA y un modelo XGBoost, predefiniendo los parámetros configurables de cada modelo, teniendo únicamente como referencia la serie temporal con granularidad mensual.

- Luego se entra de nuevo, un modelo ARIMAX y un XGBoost, a los cuales se les agrega una hiper-parametrización. Y se le anexa a la data set la variable foránea de mantenimientos.

- Con el fin de mejorar los modelos anteriores, para que tuvieran una mejor captura en la tendencia de la serie se agregan las variables de sequía, y fenómenos del niño y La Niña tanto al modelo ARIMAX como al XGBoost.

Estas versiones permiten generar un abanico de posibilidades sobre el comportamiento de las series con respecto a la inclusión de las variables foráneas, detectando así cambios en las estructuras de la serie y tener un concepto más amplio para definir los modelos que componen el ensamble.

4.5. Evaluación de desempeños

4.5.1. Cálculo de métricas de error

Para evaluar el desempeño de los modelos , se obtuvieron dos métricas: Error Porcentual Absoluto Medio (MAPE) y el Error Cuadrático Medio (RMSE), donde la primera de ellas es un error que premia los aciertos, es decir permite determinar qué tan preciso es el valor estimado frente al real y el segundo representa la sensibilidad a desviaciones grandes entre la generación real y la pronosticada; entre mayor el resultado mayor es el error y menos preciso el modelo.sus unidades se dan en la escala de la variable objetivo, en este caso la generación de energía medida en (MW).

4.5.2. Validación cruzada

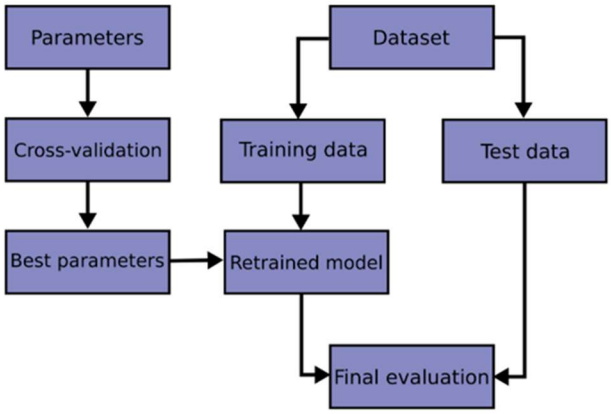


Figura 4. Proceso de validación cruzada

Con el fin de determinar los segmentos apropiados en los que se evaluaría el rendimiento de la serie, se hizo uso de las técnicas de validación cruzada o Cross Validation, las cuales consisten en definir diferentes submuestras de conjunto de datos para entrenamiento, test y validación. Es importante aclarar, que en el caso de series temporales no se puede definir los subconjuntos de forma aleatoria, sino que se debe respetar la dependencia temporal que posee la serie, debido a que no tiene sentido usar los valores del futuro para pronosticar valores en el pasado. En pocas palabras, se quiere evitar mirar hacia el futuro cuando entrenamos el modelo. [11] Existe una dependencia temporal entre las observaciones y se debía preservar esa relación durante las pruebas.

Para el desarrollo del proyecto, se probaron 3 diferentes ventanas de tiempo para el conjunto de prueba y validación:

- 6 meses de prueba – 6 meses de validación
- 6 meses de prueba – 8 meses de validación
- 1 año de prueba – 1 año de validación

4.5.3. Validación año 2020 y BackTesting

Se tomó como conjunto de validación el año 2020 según decisiones de la "Empresa", debido a que el año 2021 se consideraba atípico y podía generar cambios drásticos en el pronóstico para el año 2022. Otra alternativa que se llevó a cabo, para evaluar el desempeño de los resultados obtenidos del modelo final es comparar con un periodo (anual) que tengan una tendencia similar al año de validación, es decir, un año que tenga un comportamiento similar al 2020; como en este prevaleció fenómeno de la niña entonces el año que más se asemeja es

el 2017, donde también ocurrió de manera significativa dicho fenómeno. Este proceso es denominado BackTesting.

4.6. Ensamble de modelos

Cuando un solo modelo no logra capturar de buena manera las dinámicas de la serie, es necesario recurrir a un ensamble que permita agrupar un conjunto de modelos con diferentes características, para lograr obtener un modelo final que abarque la mayor cantidad de tendencias posibles que pueda tener la serie temporal.

De acuerdo con las métricas evaluadas (MAPE) en cada uno de los modelos, se tomó la decisión de escoger aquellos modelos que tengan el menor MAPE. Para realizar el ensamble del modelo, el primer paso que se llevó a cabo fue la asignación de pesos que debía tener cada uno de los modelos, para lograr minimizar el error de estimación. Este proceso se realiza a través de una optimización Bayesiana, el cual permite encontrar rápidamente los valores que ayuda a minimizar el error evaluado, a partir de procesos Gaussianos.

Una vez encontrados los pesos para cada modelo seleccionado, se realiza una sumatoria de la multiplicación del valor estimado de cada modelo por el peso parametrizado.

$$ensamble_j = \sum_{i=1}^{n} modelo_i \times w_i \qquad (4)$$

4.7. Cálculo de percentiles

Para tener diferentes escenarios de decisión en la generación de energía futura en la que se pueda garantizar en un 95% de los casos el valor en la generación real, se utilizan los percentiles 10, 25, media, 75 y 90, que se calculan a partir de la distribución de la generación histórica de energía horaria a nivel mensual. Los percentiles 10 y 90 definen los limites inferior y superior respectivamente, entre la cual se posiciona la generación histórica, los percentiles 25, media y 75 será medidas que según las variaciones del negocio en el tiempo ya sea comerciales o climatológicas, permitirán evaluar y tomar acciones, frente a su comportamiento en la tendencia de las series.

4.8. Visualización en Power BI

Una vez se cuenta con los resultados en formato tabular del ensamble de los modelos, se almacenó la información en Azure Blob Storage. Por consiguiente, se creó un tablero con la herramienta Power BI de Microsoft, en la que usuario puede interactuar de forma dinámica y gráfica con los datos obtenidos, ya que este será el insumo final que consumirá.

5. Resultados y análisis

5.1. Preprocesamiento de datos

Luego de calcular el porcentaje de valores nulos, para determinar los datos Faltantes y discontinuidades. Se obtuvo el resultado de la tabla 2:

Tabla 2. Días Faltantes en cada una de las plantas

PLANTA	FECHA_INICIO	FECHA_FIN	DIAS_GENERACION	DIAS_TRANSCURRIDOS	DIFF_DIAS	%NULOS
RPD1	1/01/2008	13/01/2022	4917	5126	209	4,08%
RMR1	1/01/2008	13/01/2022	4976	5126	150	2,93%
RFR1	1/01/2008	13/01/2022	4976	5126	150	2,93%
RFR2	1/01/2008	13/01/2022	4976	5126	150	2,93%
ALBG	1/01/2008	13/01/2022	5117	5126	9	0,18%
SLVJ	1/01/2008	13/01/2022	5117	5126	9	0,18%
PRD4	1/01/2008	13/01/2022	5117	5126	9	0,18%
PRDO	1/01/2008	13/01/2022	5117	5126	9	0,18%
RCL1	1/01/2008	13/01/2022	5117	5126	9	0,18%
CLMG	1/01/2008	13/01/2022	5117	5126	9	0,18%
NIM1	1/01/2008	13/01/2022	5117	5126	9	0,18%
ALTG	21/01/2012	13/01/2022	3638	3645	7	0,19%
HMO1	22/05/2012	13/01/2022	3517	3523	6	0,17%
2QV2	11/09/2014	13/01/2022	2675	2681	6	0,22%
CUC1	11/12/2014	13/01/2022	2584	2590	6	0,23%
AMA1	1/08/2010	8/01/2022	4172	4178	6	0,14%

Se evidenció según la Tabla 2 que para las plantas RPD1, RMR1, RFR1 y RFR2 no hay registro de información en la fuente de origen de los datos en los primeros meses de operación, generando esto un porcentaje de datos nulos mayor al 2%. Por lo que se tomó la decisión junto con el cliente de tomar como punto de partida de la data histórica desde el 2010-01-01, debido que para el periodo entre 2008 y 2009, dichas plantas aún se encontraban en proceso de adecuación y generaban intermitencias en la generación.

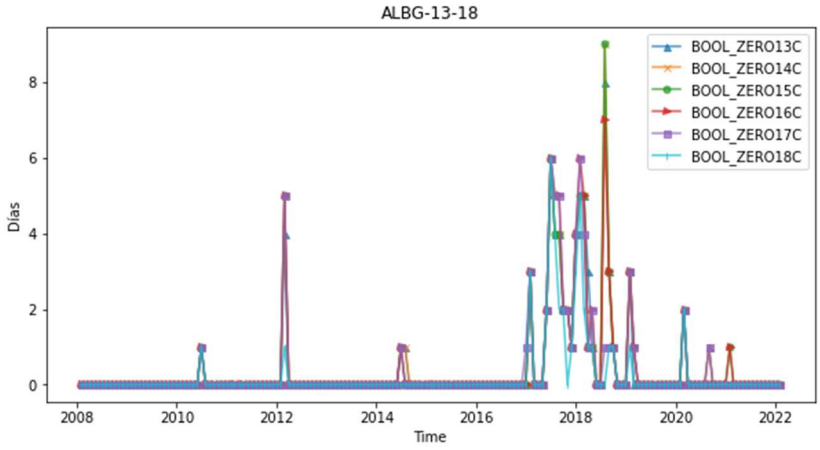


Figura 5. Análisis planta ALBG franja 13-18

Por otra parte, en la Figura 5, se logra ejemplificar el tratamiento dado a los valores ceros, en donde se evidencia, que para la planta ALBG en la franja horaria de 13-18, posee una mayor cantidad de valor cero de generación de energía al rededor del año 2018, en algunos casos llegando hasta 8 días. Esto debido a que, la planta se para el dicho año se encontraba en mantenimiento correctivo por fallas ocurridas.

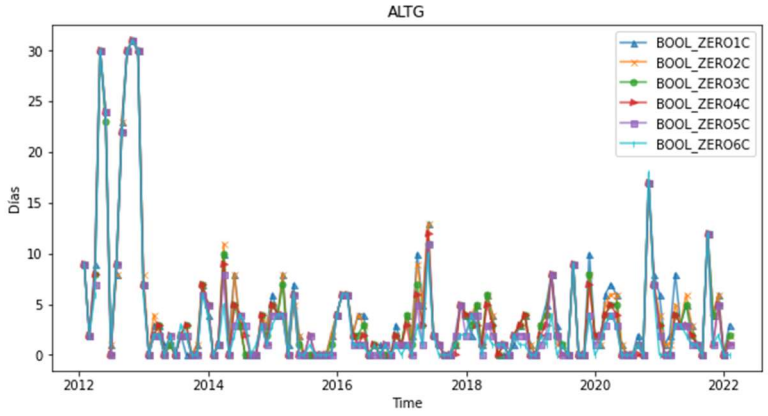


Figura 6. Análisis planta ALTG franja 1-6

En cambio, para la planta ALTG en la franja horaria de 1-6, según la Figura 6, se observa que a lo largo de los años posee varios picos en los que posee valor cero y esto es debido a que como es una central secundaria, la capacidad del embalse es mucho menor a que la planta ALBG, por lo que en dicho periodo horario la planta generalmente se encuentra apagada.

En las ilustraciones 7 y 8 que representa diagrama de cajas y bigotes, se tiene la distribución de la generación de energía en cada una de las plantas primarias y secundarias con el fin de determinar en cada una de ellas los valores atípicos:

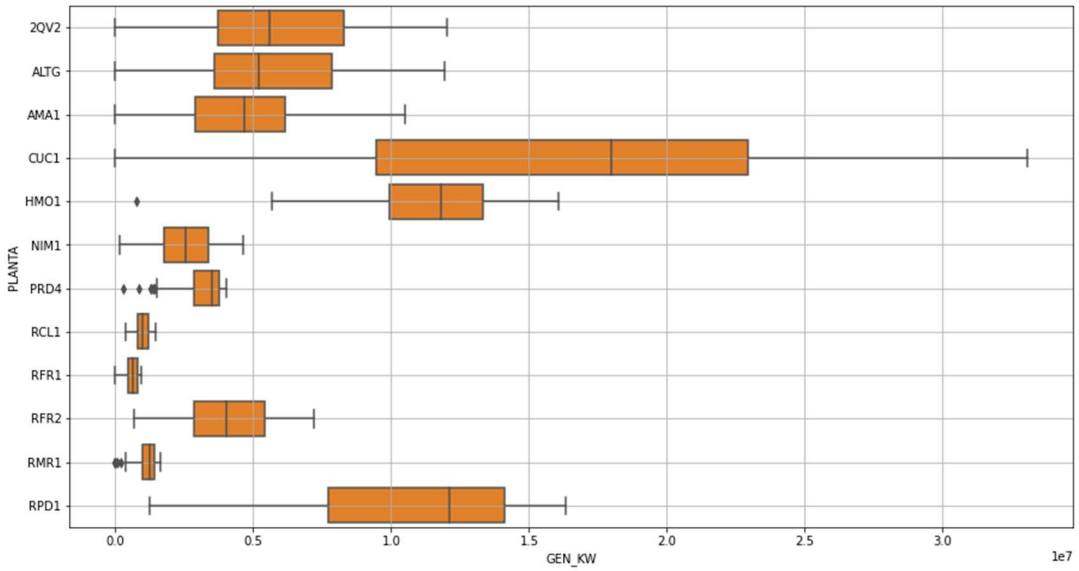


Figura 7. Cajas y bigotes para determinar datos atípicos en plantas secundarias

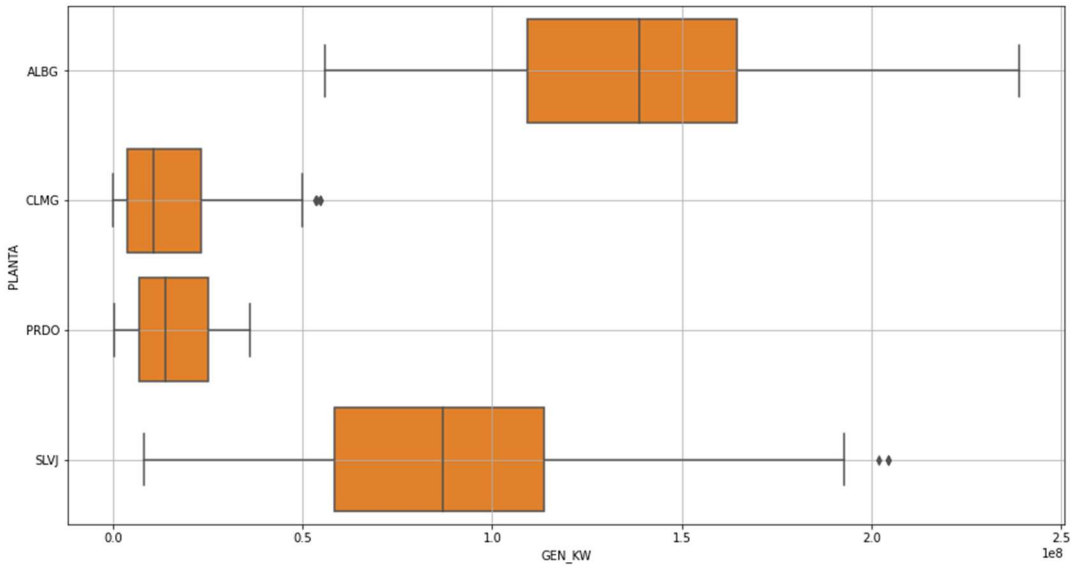


Figura 8. Cajas y bigotes para determinar datos atípicos en plantas principales

Los valores atípicos para cada una de las plantas, según la Figura 7 y 8, se presentan por debajo o por encima de los valores máximos o mínimos, más allá de la posición del primer o tercer cuartil definida por la distribución de los datos, por lo que para las plantas secundarias HMO1, PRD4, RMR1, estos valores se presentan por debajo, mientras que para las plantas principales CLMG y SLVJ los valores atípicos se encuentran por encima de los valores máximos según su distribución.

Uno de los resultados de la aplicación del filtro de Hampel, se observa en la siguiente Figura (9) para la planta SLVJ, donde la línea Naranja es la serie filtrada

(GEN_KWF) y la azul la serie real (GEN_KW). Si bien las series en gran parte se superpone, existen secciones en las que se suaviza la serie y es donde hay cambios bruscos en la generación de energía que se ve reflejado en picos que están por fuera de su desviación media absoluta, calculada de forma interna por el filtro.

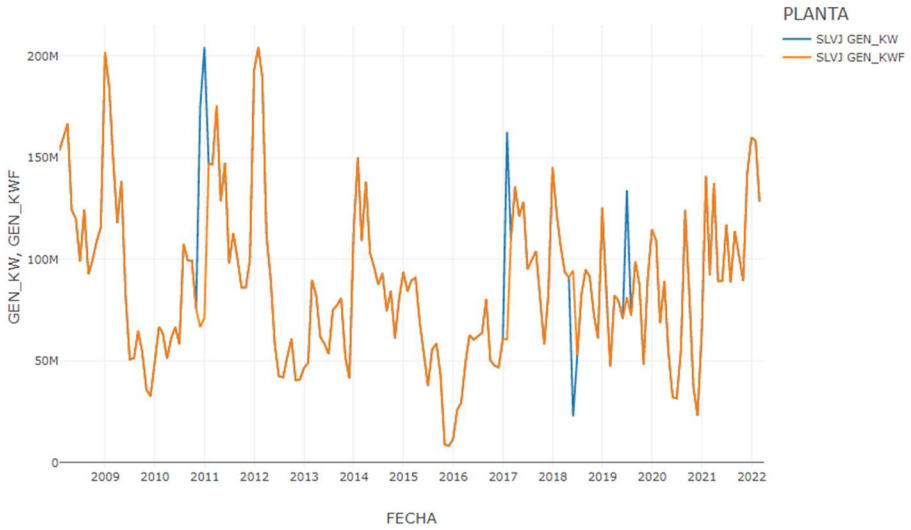


Figura 9. Imputación de valores atípicos a través del filtro de Hampel para SLVJ

Los periodos de sequía se obtuvieron a partir de diagramas de cajas y bigotes para cada una de las plantas, determinando el promedio de generación de energía en cada uno de los meses de todos los años con los que se cuenta información histórica, en la Figura 10, se muestra dos ejemplos para las plantas PRDO y HMO1.

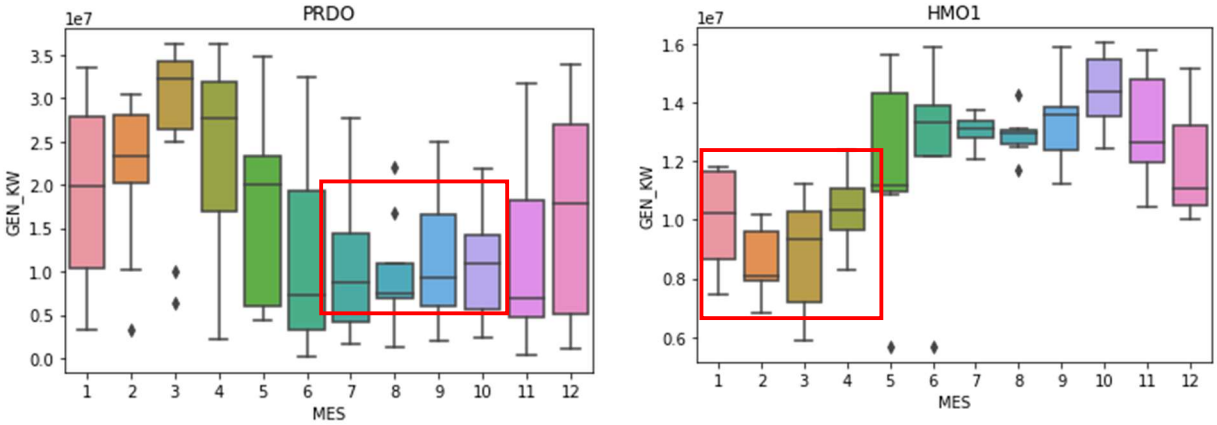


Figura 10. Promedio de generación de energía para las plantas PRDO y HMO1

De la Figura 10, se evidenció que existe un patrón que se repite anualmente, en cada una de las plantas, en este caso se muestra un ejemplo para las plantas PRDO y HMO1, donde los promedios de la generación mensual son menores que los demás promedios mensuales. En el caso de PRDO disminuyo en el rango del mes 7 al 10 y para HMO1 es menor en el rango del mes 1 al 4. Esto debido a que,

los niveles de hidrología para dichos periodos de tiempo son menor que la de los demás meses, acorde a la ubicación geográfica en la que se encuentran cada una de las centrales hidroeléctricas.

Luego de haber realizado la limpieza y preparación de los datos históricos, se obtiene como resultado el siguiente set de datos; en la Figura 11, se puede evidenciar una muestra de su estructura, la cual es la información que se les entrega a los diferentes modelos, para realizar el pronóstico de la generación de energía.

PLANTA	GEN_KW	FECHA	ANIO	MES	MANTENIMIENTO	TIPO	SEQUIA	FEN_NINIA	FEN_NINIO
2QV2	4822562	2015-01-31	2015	1	0	train	1	0	1
2QV2	531163.2	2016-01-31	2016	1	0	train	1	0	0
2QV2	9302488	2017-01-31	2017	1	0	train	1	0	0
2QV2	11916525	2018-01-31	2018	1	0	train	1	0	1
2QV2	4292909.5	2019-01-31	2019	1	0	test	1	0	0
2QV2	4845930.5	2020-01-31	2020	1	0	valid	1	1	0

Figura 11. Muestra del set de datos

5.2. Desempeño de modelos

Con el fin de validar el desempeño que tiene el modelo, se genera una serie de tablas comparativas de los resultados obtenidos que garanticen un alto nivel de confiabilidad en los pronósticos de la generación de energía para cada una de las plantas. Inicialmente se muestra el resultado de la generación de energía para el año de validación 2020 en dos escenarios: real y pronóstico ensamble.

Tabla 3. Generación real año 2020

PLANTA	GENERACIÓN MENSUAL [MW] AÑO 2020 (REAL)												Total
	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC	
ALBG	125,9	81,3	110,4	143,2	127,7	130,1	153,7	114,6	151,3	167,8	181,3	163,8	**1651,1**
SLVJ	109,3	68,6	89,3	54,7	32,2	31,6	55,2	124,2	83,8	36,9	23,1	64,2	**773,1**
CLMG	21,6	42,6	33,2	31,7	5,3	0,1	0	4,1	10,6	29,7	7,3	0,7	**186,9**
PRDO	20,3	22,5	27,3	18,4	5,4	0,9	2,0	16,0	10,8	17,2	3,4	8,8	**152,9**
CUC1	8,0	7,2	8,3	12,5	21,5	22,1	25,4	18,3	18,0	12,3	22,5	17,5	**193,5**
PRD4	3,7	3,2	3,7	1,7	3,7	3,5	3,8	3,7	3,2	2,9	3,5	3,9	**40,4**
NIM1	0,5	0,4	0,5	1,9	2,2	2,6	3,0	2,2	1,8	1,4	2,0	1,8	**20,1**
RCL1	1,2	1,0	1,1	1,1	1,0	1,1	1,0	0,6	0,8	0,9	0,9	1,0	**11,6**
AMA1	3,7	2,8	2,8	3,0	3,8	5,1	5,6	4,3	2,9	1,3	4,1	4,9	**44,4**
ALTG	3,7	2,1	1,9	2,8	4,5	5,6	5,8	4,2	3,0	0,4	5,5	8,1	**47,7**
2QV2	4,8	3,5	4,2	5,2	5,9	6,6	7,0	5,2	4,8	2,1	6,4	9,1	**64,9**
RFR1	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,6	0,7	0,5	**5,6**
RFR2	4,7	2,5	3,4	4,8	4,0	4,3	3,6	2,5	3,0	3,0	3,1	4,1	**43,0**
RMR1	1,4	0,7	0,0	0,9	1,4	1,4	1,4	1,0	1,1	0,7	1,0	1,3	**12,3**
RPD1	6,2	2,5	2,5	6,8	9,9	11,6	14,4	14,4	14,0	13,7	14,1	14,1	**124,3**
HMO1	8,4	6,3	5,8	9,4	9,6	12,1	13,9	8,4	11,5	14,4	14,1	13,1	**127,0**

Tabla 4. Generación pronostico ensamble año 2020

PLANTA	GENERACIÓN MENSUAL [MW] AÑO 2020 (PRONOSTICO)												Total
	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC	
ALBG	128,2	110,7	115,9	144,3	148,8	128,5	143,0	112,0	134,5	154,5	166,0	165,9	**1652,2**
SLVJ	93,2	71,0	85,2	73,2	54,5	39,8	62,5	89,9	79,7	29,1	51,1	74,9	**804,1**
CLMG	20,8	38,3	31,7	27,5	4,1	3,4	3,2	89,3	89,0	16,0	7,3	10,5	**341,1**
PRDO	19,2	19,6	24,0	20,6	9,9	6,9	3,7	15,0	8,7	9,5	5,3	11,4	**153,8**
CUC1	10,4	8,5	10,3	15,1	19,8	22,6	23,9	17,6	15,7	15,9	21,9	13,3	**195,0**
PRD4	3,5	3,0	3,6	2,6	3,6	3,4	3,8	3,5	2,9	3,2	3,3	3,5	**39,9**
NIM1	1,4	0,9	0,6	1,6	2,2	2,3	2,8	2,0	1,3	1,2	1,7	1,8	**19,8**
RCL1	1,1	0,9	1,1	1,1	1,0	1,0	0,9	0,6	0,7	0,9	1,0	1,0	**11,2**
AMA1	3,6	2,8	3,3	3,7	4,4	5,1	4,7	4,1	2,9	2,3	3,6	4,0	**44,5**
ALTG	3,4	2,9	3,1	3,8	4,6	6,3	5,7	4,0	2,5	2,0	4,9	6,2	**49,4**
2QV2	4,3	3,8	4,5	6,0	5,6	7,0	7,2	4,9	4,0	3,4	6,7	7,4	**64,8**
RFR1	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,6	0,6	0,5	**5,3**
RFR2	4,0	3,2	3,7	4,5	3,9	3,8	3,1	2,9	2,9	3,4	3,6	3,9	**42,9**
RMR1	1,3	0,6	0,3	1,0	1,0	1,3	1,3	1,0	1,0	0,9	1,0	1,1	**11,8**
RPD1	6,6	1,6	2,9	8,4	10,9	11,0	13,7	13,8	14,2	14,0	13,4	14,5	**125,0**
HMO1	7,9	6,6	6,1	9,3	11,1	11,6	13,3	9,6	11,8	14,7	12,6	11,9	**126,5**

Con los resultados de las tablas 3 y 4 se calcula el error relativo porcentual de la generación del modelo ensamble con respecto a la generación real, a partir de un mapa de color, que se muestra en la tabla 5, donde el color azul representa menor error, es decir el modelo ensamble para estos meses se acerca más al valor real y el color rojo implica que hay una mayor desviación en los resultados, siendo el caso más representativo el de la planta CLMG en los meses de JUN, AGO, SEP y DIC , debido a que hubo movimientos en la bolsa de valores que generaron cambios en las en la toma de decisiones y por ende en el funcionamiento de la planta.

Tabla 5. Error relativo porcentual entre generación real y generación pronosticada modelo ensamble año 2020.

PLANTA	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
ALBG	1,8%	36,1%	5,0%	0,8%	16,5%	1,2%	7,0%	2,3%	11,1%	7,9%	8,4%	1,3%
SLVJ	14,7%	3,5%	4,6%	33,8%	69,3%	25,9%	13,2%	27,6%	4,9%	21,1%	121,2%	16,7%
CLMG	3,7%	10,1%	4,5%	13,2%	22,6%	3300,0%	100,0%	2078,0%	739,6%	46,1%	0,0%	1400,0%
PRDO	5,4%	12,9%	12,1%	12,0%	83,3%	666,7%	85,0%	6,3%	19,4%	44,8%	55,9%	29,5%
CUC1	30,0%	18,1%	24,1%	20,8%	7,9%	2,3%	5,9%	3,8%	12,8%	29,3%	2,7%	24,0%
PRD4	5,4%	6,3%	2,7%	52,9%	2,7%	1,8%	0,0%	5,4%	9,4%	10,3%	5,7%	10,3%
NIM1	180,0%	125,0%	20,0%	15,8%	0,0%	11,5%	6,7%	9,1%	27,8%	14,3%	15,0%	0,0%
RCL1	8,3%	12,6%	0,0%	0,0%	0,0%	9,1%	10,0%	0,0%	12,5%	0,0%	7,0%	0,0%
AMA1	2,7%	0,0%	17,9%	23,3%	15,8%	0,0%	16,1%	4,7%	0,0%	76,9%	12,2%	18,4%
ALTG	8,1%	38,1%	63,2%	35,7%	2,2%	12,5%	1,7%	4,8%	16,7%	400,0%	10,9%	23,5%
2QV2	10,4%	8,1%	7,1%	15,4%	5,1%	6,1%	2,9%	5,8%	16,7%	61,9%	4,7%	18,7%
RFR1	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	14,3%	0,0%
RFR2	14,9%	28,0%	8,8%	6,3%	2,5%	11,6%	13,9%	16,0%	3,3%	13,3%	16,1%	4,9%
RMR1	7,1%	14,3%	30,0%	11,1%	28,6%	7,1%	7,1%	0,0%	9,1%	28,6%	0,0%	15,4%
RPD1	6,5%	36,0%	16,0%	23,5%	10,1%	5,2%	4,9%	4,2%	1,4%	2,2%	5,0%	2,8%
HMO1	6,0%	4,8%	5,2%	1,1%	15,3%	4,1%	4,3%	14,3%	2,6%	2,1%	10,6%	9,2%

Para determinar de forma más precisa el desempeño del modelo ensamble, con respecto a la generación real del año 2020, para cada una de las plantas se calcula la métrica de MAPE, que este se calcula a partir de los resultados obtenidos en la tabla 5. Los resultados se muestran en la tabla 6.

Como criterio de aceptación del modelo, se definió un umbral de precisión del 84%, es decir que el MAPE se encuentre por debajo del 16%, de las 16 plantas a las cuales se les pronóstico la generación de energía, 11 de ellas superan este umbral, mientras que 5 de ellas se encuentran por debajo, sin embargo, se evidenció que los errores, se deben acontecimientos específicos que sucedieron en las plantas, como se expuso anteriormente con CLMG, el cual afectaron el desempeño del modelo.

Tabla 6. Cálculo del MAPE validación año 2020 modelo ensamble

PLANTA	MAPE	Precisión
ALBG	8,3%	92%
SLVJ	29,7%	70%
CLMG	643,2%	-543%
PRDO	86,1%	14%
CUC1	15,1%	85%
PRD4	9,4%	91%
NIM1	35,4%	65%
RCL1	5,0%	95%
AMA1	15,7%	84%
ALTG	51,4%	49%
2QV2	13,6%	86%
RFR1	1,2%	99%
RFR2	11,6%	88%
RMR1	13,2%	87%
RPD1	9,8%	90%
HMO1	6,6%	93%

Por otra parte, otra medida que permitir definir el comportamiento y desempeño del modelo es el Error cuadrático medio que se calcula a partir de los datos de la tabla 4 y 5. El mapa de calor indica, que entre menor variabilidad exista entre ambos valores, mejor será el resultado obtenido por el modelo ensamble presentado en la tabla 7:

Tabla 7. Error cuadrático entre generación real y generación pronosticada modelo ensamble año 2020.

PLANTA	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
ALBG	5,25	861,5	30,2	1,21	445,21	2,56	114,49	6,76	282,2	176,9	234,1	4,41
SLVJ	259,21	5,76	16,8	342,25	497,29	67,24	53,29	1176,49	16,57	60,84	784,0	114,5
CLMG	0,64	18,49	2,25	17,64	1,44	10,89	10,24	7259,04	6146,5	187,69	0,00	96,04
PRDO	1,21	8,41	10,9	4,84	20,25	36,00	2,89	1,00	4,41	59,29	3,61	6,76
CUC1	5,76	1,69	4,00	6,76	2,89	0,25	2,25	0,49	5,29	12,96	0,36	17,64
PRD4	0,04	0,04	0,01	0,81	0,01	0,00	0,00	0,04	0,09	0,09	0,04	0,16
NIM1	0,81	0,25	0,01	0,09	0,00	0,09	0,04	0,04	0,25	0,04	0,09	0,00
RCL1	0,01	0,02	0,00	0,00	0,00	0,01	0,01	0,00	0,01	0,00	0,00	0,00
AMA1	0,01	0,00	0,25	0,49	0,36	0,00	0,81	0,04	0,00	1,00	0,25	0,81
ALTG	0,09	0,64	1,44	1,00	0,01	0,49	0,01	0,04	0,25	2,56	0,36	3,61
2QV2	0,25	0,08	0,09	0,64	0,09	0,16	0,04	0,09	0,64	1,69	0,09	2,89
RFR1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00
RFR2	0,49	0,49	0,09	0,09	0,01	0,25	0,25	0,16	0,01	0,16	0,25	0,04
RMR1	0,01	0,01	0,09	0,01	0,16	0,01	0,01	0,00	0,01	0,04	0,00	0,04
RPD1	0,16	0,81	0,16	2,56	1,00	0,36	0,49	0,36	0,04	0,09	0,49	0,16
HMO1	0,25	0,09	0,09	0,01	2,16	0,25	0,36	1,44	0,09	0,09	2,25	1,44

De manera más concreta, para tener un mejor panorama del desempeño del modelo ensamble, se utiliza la métrica de desempeño RMSE para medir la

variabilidad que obtiene cada una de las plantas, con respecto a la generación real, en el año 2020 de validación.

Tabla 8. Cálculo del RMSE validación año 2020 modelo ensamble

PLANTA	RMSE
ALBG	13,43
SLVJ	16,82
CLMG	33,85
PRDO	3,65
CUC1	2,24
PRD4	0,33
NIM1	0,38
RCL1	0,07
AMA1	0,58
ALTG	0,94
2QV2	0,75
RFR1	0,03
RFR2	0,44
RMR1	0,18
RPD1	0,75
HMO1	0,84

En la tabla 8 se observa, las plantas que posee una mayor variabilidad en la generación de energía en el periodo de validación son ALBG, SLVJ y CLMG. Un caso particular es el de ALBG, si bien presentó un bajo MAPE (8,3%) según la tabla 7 con respecto a las demás plantas, su medida de RMSE (13,43) es elevada, esto indica que aunque el modelo posee un error absoluto bajo, la variabilidad en la generación de energía es alta, ya que la planta posee una mayor capacidad de generación , por lo que el MAPE al desviaciones de alto volumen subestima el error global, mientras que el RMSE es sensible a los cambios que puedan presentarse entre la generación real y del modelo ensamble. Dando esto como resultado que, aunque son métricas dadas en diferentes unidades, permiten analizar el comportamiento del modelo desde otras perspectivas.

Otra manera de validar el modelo obtenido es comparar las medidas de desempeño MAPE y RMSE con respecto al modelo clásico de promedios históricos, el cual ha sido el modelo que "La Empresa" ha venido utilizando a través de los años. El resultado se muestra en la tabla 9.

De la tabla 9, se puede notar que, para cada una de las plantas, hubo una mejora tanto en la medida del RMSE, como del MAPE del modelo ensamble con respecto al modelo de promedios históricos, para la etapa de validación en el año 2020, con una mejora porcentual promedio del 11%.

Tabla 9. Comparación de métricas entre Modelo ensamble y modelo promedio

	MODELO ENSAMBLE		MODELO PROMEDIO	
PLANTA	RMSE	MAPE	RMSE	MAPE
ALBG	13,43	8%	26,59	10%
SLVJ	16,82	30%	36,15	17%
CLMG	33,85	643%	11,36	848%
PRDO	3,65	86%	3,8	90%
CUC1	2,24	15%	2,48	23%
PRD4	0,33	9%	0,58	17%
NIM1	0,38	35%	1,14	49%
RCL1	0,07	5%	0,11	10%
AMA1	0,58	16%	5,56	22%
ALTG	0,94	51%	2,41	62%
2QV2	0,75	14%	2,39	33%
RFR1	0,03	1%	0,06	11%
RFR2	0,44	12%	0,95	28%
RMR1	0,18	13%	0,39	20%
RPD1	0,75	10%	1,55	18%
HMO1	0,84	7%	1,2	12%

BackTesing año 2017

Dado que el 2017 fue un año similar al 2020 en cuanto al fenómeno de la niña, se parte de la distribución histórica de energía para cada una de las plantas a partir de un diagrama de cajas y bigotes y se superponen 4 series temporales identificadas de diferentes colores: la Azul representa uno de los modelos que mejor desempeño obtuvo que fue el XGBoost con todas las variables foráneas , la Roja el resultado del modelo ensamble , la Verde la generación real para el año 2017 y la amarilla la generación real para el año 2020. El resultado se muestra en la Figura 11

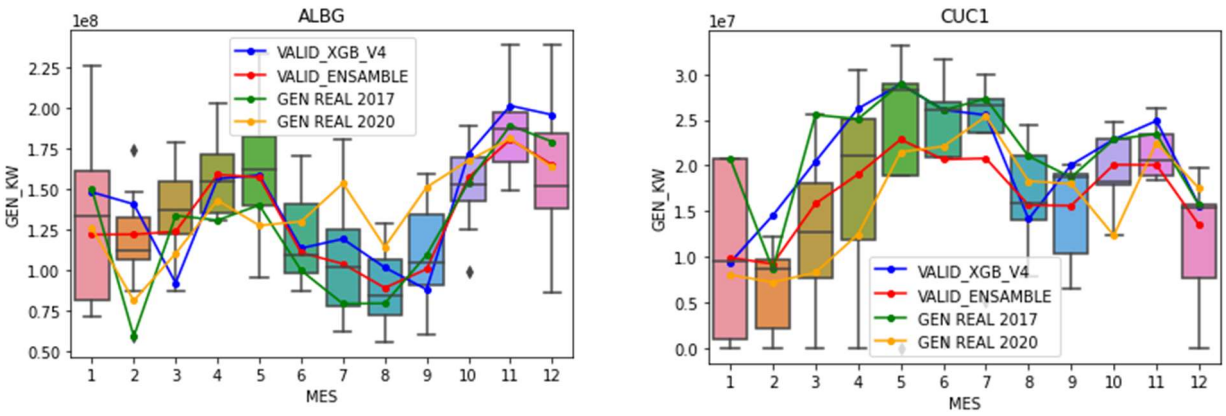


Figura 11. Gráfico de cajas y bigotes sobre el backtesting del año 2017 para ALBG y CUC1

En la Figura 11, se observa que al realizar el proceso del backtesting y comparar la generación real para el año 2017 (línea verde) para las plantas ALBG y CUC1

frente a la validación del modelo ensamble (línea roja), de manera general hay una muy buena captura de la tendencia que la serie. Sin embargo, se debe tener presente que según el diagrama de cajas y bigotes la generación real en varios meses está fuera de la distribución de los datos, que se pueden dar a acciones propias de "La Empresa", que no son capturadas por el modelo.

Pronóstico de generación de energía año 2022

Como el modelo fue desarrollado para pronosticar todo el año 2022, se puede realizar la comparación frente a la generación real, de los dos primeros meses que se tenía data confiable a la fecha de febrero 2022, los cuales arrojaron los siguientes resultados:

Tabla 10. Comparación de la generación real con el pronóstico modelo ensamble

	GENERACIÓN REAL [MW]		PRÓNOSTICO MODELO ENSAMBLE [MW]	
PLANTA	ene-22	feb-22	ene-22	feb-22
ALBG	191,9	150,3	128,2	110,7
SLVJ	158,2	128,3	117,2	99,3
CLMG	45,1	36,4	16	30,0
PRDO	23,0	22,4	21,4	22,1
CUC1	9,4	12,6	12,7	9,6
PRD4	3,8	3,4	3,1	3,2
NIM1	2,1	2,3	2,1	2,0
RCL1	0,9	0,8	1	0,9
AMA1	6,0	5,0	4,7	3,7
ALTG	7,1	4,9	4,8	3,8
2QV2	0,0	3,5	6,7	3,8
RFR1	0,7	0,7	0,6	0,5
RFR2	5,0	4,0	4,3	2,8
RMR1	1,4	1,0	1,2	1,0
RPD1	11,9	8,3	9,4	6,6
HMO1	9,8	8,3	9,5	8,0

Para comparar el desempeño que tuvo el pronóstico de la generación de energía se muestra el MAPE calculado para ambos meses, según se muestra en la tabla 13 con un MAPE promedio de 19%, es decir donde se tiene una precisión promedio del 81%, la cual es bastante significativa y permite dar una validez al desarrollo del modelo realizado. Sin embargo, para algunas plantas en específico como CLMG para el mes de enero se calcula un MAPE del 71%, el cual está asociado a que el pronóstico esperaba que hubiera un bajo nivel de hidrología y por ende menor generación, por condiciones climáticas ocurrió lo contrario, donde hubo gran cantidad de lluvias, que aumentaron el nivel de hidrología del embalse y por ende su generación. En otro caso particular, la planta 2QV2, no presentó generación de energía para el mes de enero, dado que ocurrieron fallas inesperadas en la planta obligándola a salir de producción.

Tabla 11. Cálculo del MAPE para los meses de enero y febrero año 2022

PLANTA	MAPE ENE - 22	MAPE FEB - 22
ALBG	33%	26%
SLVJ	26%	23%
CLMG	71%	18%
PRDO	7%	1%
CUC1	35%	24%
PRD4	18%	6%
NIM1	1%	12%
RCL1	8%	11%
AMA1	21%	26%
ALTG	32%	23%
2QV2		7%
RFR1	19%	26%
RFR2	13%	29%
RMR1	14%	2%
RPD1	21%	21%
HMO1	3%	4%

Para tener un abanico de posibilidades, que permita tener un mayor criterio en la toma de decisiones, se calcula los percentiles sobre el valor de generación de energía, una muestra del resultado obtenido se observa en la Figura 12, a partir de una de las gráficas realizadas en el tablero de Power BI.

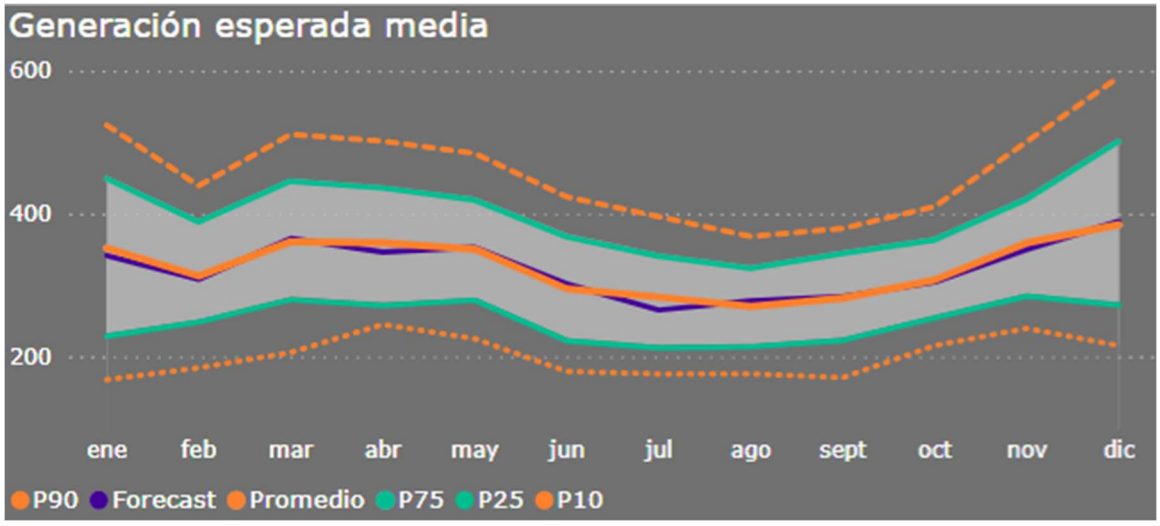


Figura 12. Generación por Percentiles visualizada en el tablero de Power BI

Por efectos de confidencialidad de los datos expuestos en el tablero de Power BI, solo se mostrará una pequeña parte de este. Es importante resaltar, que el tablero es interactivo con el usuario, el cual permite ir filtrando por periodos de fecha, por plantas y por percentiles para visualizar de forma detalla los resultados. El diseño y

la información expuesta se trabajó en conjunto con el equipo de "La Empresa", según las necesidades requeridas.

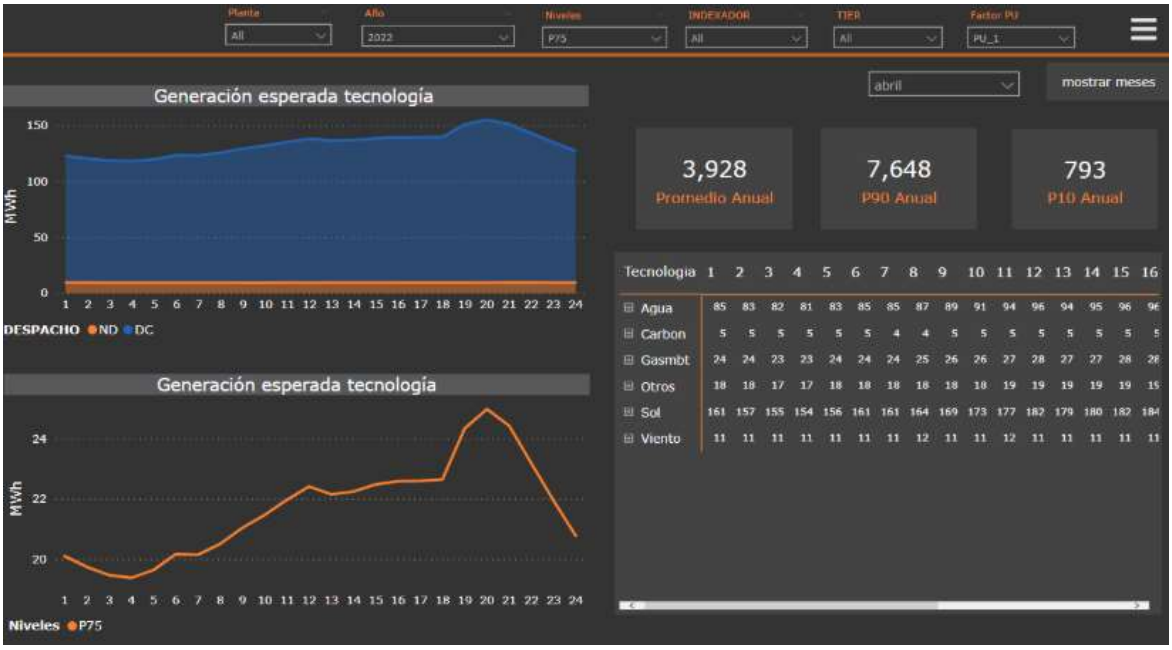


Figura 23. Visualización del tablero en Power BI

6. Conclusiones

- Los modelos de predicción de series temporales muestran una alternativa confiable para la toma de decisiones a futuro, en la generación de energía de plantas hidroeléctricas, minimizando la operatividad en el área funcional y escalando la productividad de su compañía, a través de una herramienta versátil automatizada que brinda un abanico de posibilidad para el análisis de sus procesos.

- Si bien los modelos ARIMA capturan muy bien los comportamientos de la serie, estos tienden a reducir la amplitud de la señal, debido a su operación de medias móviles ocasionando caídas en la generación de energía, por lo tanto, fue necesario darle poco peso a la hora de incluirlo en el modelo ensamble.

- El análisis exploratorio y la construcción de base analítica fueron el proceso más importante para tener un resultado positivo en la implementación del modelo, debido a que permiten garantizar una buena calidad de las variables que reciben los modelos y por ende un mayor grado de desempeño a la hora de capturar las tendencias de las series al incluir variables climatológicas que influían de forma directa en la generación de energía debido a los niveles de hidrología de las plantas hidroeléctricas.

- Un gran reto durante la realización del proyecto fue encontrar la manera apropiada para implementar el modelo ensamble, donde se lograra mejorar el rendimiento que se tenía en cada uno de los versionamientos de modelos, sin ocasionar que el modelo ensamble tuviera tendencias a la baja en la generación de energía. La mejor manera fue la implementación de un proceso que permitiera determinar el peso que debería tener cada uno de los modelos a partir de la optimización bayesiana.
- El ensamble de modelos permitió generar una mejora porcentual del 11% con respecto al modelo de promedio histórico que utilizaba "La Empresa" en la etapa de validación, gracias a que cada uno de los versionamientos de los modelos capturaban de diferente forma el comportamiento de la serie temporal, formando un acoplamiento que aprovechara las virtudes que entregaban cada uno de ellos.
- Gracias a la clusterización y automatización de los procesos que brindan los servicios de Microsoft Azure, los tiempos de respuesta en el procesamiento de la información permiten a "La Empresa" enfocar sus tareas en el análisis de los resultados, de manera que, se libere la carga de sus empleados, permitiéndoles trabajar en asuntos de mayor valor.
- La visualización a partir de tableros de Power BI, permiten una interacción rápida y concisa de los indicadores que tiene mayor incidencia y relevancia para "La Empresa", donde de forma automática según la necesidad requerida se actualizan los datos del último mes en curso.

7. Referencias Bibliográficas

[1] Hidroelectricidad, la mayor fuente de energía. La república. Febrero, 2020. Consultado Marzo, 2022. Disponible en: https://www.larepublica.co/especiales/colombia-potencia-energetica/hidroelectricidad-la-mayor-fuente-de-energia-renovable-2966269

[2] What is a forecasting model?. Indeed Editorial Team. Julio, 2021. Consultado Abril, 2022. Disponible en: https://www.indeed.com/career-advice/career-development/forecasting-models

[3] MEDINA, Fernando y GALVÁN Marco. Imputación de datos: teoría y práctica. *División Estadísticas y Proyecciones económicas*. Santiago de Chile, julio 2007. Disponible: https://repositorio.cepal.org/bitstream/handle/11362/4755/1/S0700590_es.pdf

[4] Detección de valores atípicos con filtro Hampel [anónimo]. Disponible en:

https://ichi.pro/es/deteccion-de-valores-atipicos-con-filtro-hampel-147188347029309

[5] Corporate Finance Institute. What is the Autoregressive Integrated Moving Average (ARIMA)? (2019). Disponible en: https://corporatefinanceinstitute.com/resources/knowledge/other/autoregressive-integrated-moving-average-arima/

[6] PRABHAKARAN, Selva. *ARIMA Model – Complete Guide to Time Series Forecasting in Python.* Machine learning + (22 de Agosto 2021). Disponible en: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

[7] MORDE,Vishal. *XGBoost Algorithm: Long May She Reign!* Towards Data Science. (7 de Abril 2019). Disponible en: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

[8] Microsoft Azure. *Productos de Azure* (2022). Disponible en: https://azure.microsoft.com/es-es/services/

[9] Atlas, Potencial Hidro energético de Colombia. *HIDROENERGÍA*. Cap. 1 (2015). Consultado Febrero, 2022. Disponible en: https://www1.upme.gov.co/Energia_electrica/Atlas/Atlas_p25-36.pdf

[10] Serna, P. Medidas de error en pronóstico de demanda. (agosto, 2020). Disponible en: https://www.datup.ai/logistica/medidas-de-error-en-pronosticos-de-demanda/#:~:text=El%20error%20porcentual%20absoluto%20medio,cada%20punto%20que%20se%20pronostica.

[11] Shrivastava, S. Cross Validation in Time Series. (Enero, 2020). Disponible en: https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4