CARRERA: ESPECIALIZACIÓN EN CIENCIA DE DATOS

TRABAJO FINAL INTEGRADOR

TITULO

Pronóstico del Salto Hidráulico de una planta de generación hidroeléctrica, utilizando algoritmos de Aprendizaje Automático

Nombre y Apellido del Alumno/a: José Luis Beltramone

Título de grado o posgrado (último): Especialista en Gestión de las Telecomunicaciones

Director; Mg. Ing. Gustavo Denicolay

Lugar y Fecha: La Plata, junio de 2022

Sumario Ejecutivo

El SALTO HIDRÁULICO de una central hidroeléctrica, definido como la diferencia entre los niveles del agua, a ambos lados del dique, es la variable de gestión más importante de su capacidad de generación. Por lo tanto, su pronóstico preciso, es indispensable para un aprovechamiento eficiente del caudal del hidráulico aguas arriba.

El presente trabajo aborda el pronóstico del salto hidráulico mediante métodos econométricos y de aprendizaje automático con el propósito de mejorar dicho pronóstico.

Se construyen modelos en código Python utilizando varios algoritmos regresores, contrastando y seleccionando los de mejor desempeño y optimizando su desempeño, para dos horizontes de pronóstico, a saber: siete (7) y cincuenta y seis (56) días, una (1) semana y ocho (8) semanas, respectivamente. Horizonte de una semana para la gestión operativa diaria de la planta y ocho semanas, para la gestión a mediano plazo de crecidas y bajantes.

Se trabaja sobre un conjunto de datos hidrológicos de un poco más de diez (10) años, pertenecientes a la Central Hidroeléctrica Yacyretá[13] y se obtienen resultados alentadores respecto de los obtenidos en la actualidad, que resultan de la aplicación de las reglas HEC-HMS[11].

Por lo tanto es, en el marco teórico de este trabajo, posible mejorar el pronóstico del salto hidráulico utilizando modelos de aprendizaje automático.

La lista de algoritmos ensayados no es exhaustiva, no incluye, por ejemplo, algoritmos de redes neuronales artificiales recurrentes (RNN) ni el algoritmo desarrollado por Facebook 'fb prophet'[22], por lo que se motiva y recomienda a eventuales lectores interesados completar este pendiente. A los Ingenieros Hidráulicos, responsables del Área de Hidrología de centrales hidroeléctricas, se los invita a ponerlo en práctica en simultáneo con sus procesos de cálculo actuales para contraste y eventuales mejora de sus propios resultados.

Dedicatoria

A mi queridísima esposa María Elha, quien me animó a tomar el desafío y me acompañó a lo largo de toda la especialización, dándome el ánimo y la contención necesaria para llegar hasta este ansiado momento.

Agradecimientos

Al Ingeniero Lucas Federico Chamorro Vega, Hidrólogo e Ingeniero Civil Ingeniero Sector Hidrológico - Departamento Técnico Entidad Binacional Yacyretá, mi “especialista de dominio” en este trabajo. Por su reconocida trayectoria en la industria y su avanzado conocimiento en la materia, que me motivaron a enfocar este trabajo en un asunto “no trivial”, donde el esfuerzo a invertir podría resultar en una oportunidad de mejora. Por su personalidad accesible, amable y abierta y su generosa capacidad de transmitir conocimiento, que me ayudaron a la comprensión de las reglas del negocio de la generación hidroeléctrica y su invaluable gestión, que me permitió el acceso a los datos que tanto necesitaba para la ejecución de este trabajo.

Al Mg. Ing. Gustavo Denicolay, mi profesor en la materia Minería de Datos y director de este trabajo: Por su vocación de maestro, su increíble pasión por el aprendizaje y el progreso de sus alumnos. Por su dedicación sin límites a las consultas y a las inquietudes de toda la clase (dentro y fuera del horario de cursada) y por la dirección precisa, las invaluables sugerencias y las necesarias correcciones al presente trabajo.

Finalmente, a las autoridades y docentes de la Especialización en Ciencia de Datos del Instituto Tecnológico de Buenos Aires, ITBA. Por su profesionalismo y vocación, que nos permitieron adaptarnos y continuar, sin sobresaltos, con la cursada de la carrera, pese a la inesperada irrupción de la terrible pandemia COVID-19.

Table of Contents

Lista de Tablas

Lista de ilustraciones

Capítulo I: Introducción

Introducción

El salto hidráulico, definido como la diferencia entre los niveles de agua a ambos lados del dique de contención (lado embalse y lado restitución) es la herramienta de gestión principal de una planta de generación hidroeléctrica.

En el caso de la 'Central Hidroeléctrica Yacyretá' (CHY) el salto hidráulico se predice a diario, con un horizonte de 5 días, para la estimación de su capacidad de generar energía en ese mismo horizonte temporal.

Los modelos hidráulicos actuales de pronóstico incorporan en el cálculo de estos niveles hidráulicos variables del tipo climatológico y meteorológico que, por su naturaleza, agregan una cuota de incertidumbre al resultado.

Esta incertidumbre obliga a la toma de márgenes de seguridad en las decisiones de operación de la planta (generar más o menos energía eléctrica implica liberar más o menos agua, respectivamente), afectando el rendimiento general de la planta.

En el presente T.F.I. se evalúa el desempeño de algoritmos de regresión econométricos y de Aprendizaje Automático ('Machine Learning') aplicados a el pronóstico de los niveles de embalse y restitución de una represa hidroeléctrica, con el objetivo de mejorar la exactitud de el pronóstico del salto hidráulico, permitiendo disminuir los márgenes de seguridad tomados en la gestión de la operación y, por ende, mejorando así la eficiencia general de la planta generadora.

Antecedentes: El Salto Hidráulico y su relación con la capacidad de generación

La energía eléctrica que consumimos en la industria, espacios públicos y en nuestros hogares proviene de la generación combinada de plantas que utilizan diversas fuentes de energía, entre ellas: gas, combustibles líquidos, carbón, luz solar, energía eólica, energía nuclear y energía hidráulica.

A su vez, la fuente de la energía hidráulica utilizada en la generación de energía eléctrica proviene de la energía potencial del agua acumulada en el embalse de un dique de contención.

Esta energía potencial es directamente proporcional a la masa de agua acumulada a un lado del dique (lado embalse) y a la diferencia de nivel con el agua del otro lado del dique (lado restitución).

El nivel del lado embalse varía con los caudales de agua que lo alimentan (afluentes) y con las precipitaciones pluviales, mientras que el de restitución, con el caudal de agua liberada para la generación de energía y/o voluntariamente (vertido), por excedente acumulado, por razones de navegación o por razones de preservación de los ecosistemas de aguas abajo (efluentes).

Las plantas generadoras hidroeléctricas deben administrar eficientemente el volumen de agua de sus embalses a fin de cumplir con sus compromisos de generación de energía, teniendo en cuenta, siempre, las condiciones y restricciones hidrológicas del ecosistema al cual pertenecen.

El caso particular de la CHY, siendo ésta una represa de "paso" (o de "pasada") (https://es.wikipedia.org/wiki/Central_hidroel%C3%A9ctrica_de_pasada) el aprovechamiento del caudal disponible es por demás importante ya que de no utilizarse se debe verter.

Por lo tanto, el pronóstico a corto y mediano plazo de los niveles de agua a ambos lados del dique de contención, embalse y restitución resulta la herramienta central para la gestión operativa de una planta de generación hidroeléctrica, tanto desde el punto de vista de la eficiencia como de la ecología su sistema. Su exactitud es clave en la toma de decisiones de generación, liberación de agua embalsada y en la programación de actividades de mantenimiento de infraestructura.

Definición del problema

"La incertidumbre en la exactitud de el pronóstico de los niveles de embalse y restitución de una central hidroeléctrica es tan alta (en términos estadísticos, tiene una alta desviación estándar) que obliga a la toma de márgenes de seguridad en la operación."

Objetivos del estudio

General

Desarrollar un prototipo para el pronóstico del Salto Hidráulico de una planta generadora hidroeléctrica, utilizando algoritmos de Aprendizaje Automático.

Específicos

En razón de alcanzar el objetivo general se establecen los siguientes objetivos específicos, que demarcarán las etapas del trabajo y permitirán evaluar el avance en cada una de ellas y del trabajo en general.

- Entender el proceso de la generación hidroeléctrica y el impacto del salto hidráulico en la gestión del negocio

Se trabajará con el Especialista de Dominio desde una perspectiva técnico-comercial, en el marco de la explotación industrial del mercado energético argentino. Administración de la oferta y demanda de energía y requerimientos hídricos y ambientales.

- Entender el rol de las variables que componen el conjunto de datos

Se trabajará con el Especialista de Dominio desde el punto de vista técnico y conceptual y con la profundidad que resulte del conjunto de datos que finalmente se utilice en el ejercicio.

- Contar con un conjunto de datos hidrológicos consolidado

En función de las conversaciones mantenidas con el Especialista de Dominio se presume que la información histórica de la CHY está registrada de forma asincrónica y disponible en varios sistemas de procesamiento de datos, de arquitectura diversa. Por lo tanto, será necesario un proceso de ETL para contar con un conjunto de datos (dataset) unificado, coherente en las muestras y listo para su explotación.

- Contar con un conjunto de datos hidrológicos reducido y explicativo

Para alcanzar este objetivo se realizará un Análisis Exploratorio de Datos (EDA) sobre el conjunto obtenido en la etapa anterior, con etapas de limpieza, detección y tratamiento de valores extremos y de valores inexistentes ('NA').

El dataset resultante facilitará las tareas de 'Feature Engineering' durante la etapa de ajuste del modelo y reducirá los requerimientos de procesamiento (CPU y memoria).

- Seleccionar un conjunto básico de modelos de Aprendizaje Automático

Este objetivo se alcanzará realizando pruebas preliminares en el conjunto de modelos de regresión aprendidos durante la cursada de la especialización, con particular atención en aquellos recomendados en las referencias bibliográficas aplicables. Se aspira a contar con una terna de modelos a desarrollar en mayor profundidad.

- Seleccionar el modelo de mejor desempeño

Para alcanzar este objetivo se evaluarán métricas de desempeño para algoritmos de regresión, tales como MAE, RSME, MAE, etc.

- Contar con un modelo ajustado

En esta etapa se optimizarán los hiperparámetros del modelo de regresión seleccionado, trabajando con 'Optimización Bayesiana'.

Justificación del estudio

La CHY tiene una capacidad de generación instalada de 3.200 MW y aporta en promedio aproximadamente la energía eléctrica equivalente al 22% del consumo total de Argentina (13).

La diferencia entre sus niveles de embalse y restitución debe mantenerse entre 22,50m y 24,10m para una operación segura y amigable con el medio ambiente (14). Valores inferiores pueden provocar problemas con la navegación y, superiores, problemas en los ecosistemas aguas abajo.

Actualmente, la exactitud promedio del pronóstico de la diferencia entre los niveles de embalse y restitución - denominado salto hidráulico- a cinco días visto, es de aproximadamente 12cm.

El equivalente de 1 centímetro de salto hidráulico en términos energéticos equivale a 100MWh/día de energía eléctrica (que podrían alimentar, aproximadamente, a una ciudad de 50.000 habitantes) y, desde lo hidrológico, a una masa de agua de 180.000 m3/seg/día.

De estos números puede apreciarse el impacto de la gestión del salto hidráulico en la economía energética y en la ecología del entorno, y de la necesidad de aprovechar al máximo este recurso.

Ilustración 1, Vista de la Central Hidroeléctrica Yacyretá (CHY) [7]

Definiciones

- Salto hidráulico de una central hidroeléctrica: Diferencia entre los niveles de embalse y restitución
- CHY: Central Hidroeléctrica Yacyretá

Alcance del trabajo y limitaciones

Se trabajará en la creación de un prototipo para el pronóstico del salto hidráulico a partir del contraste del desempeño de varios modelos de aprendizaje automático, no estando incluidas pruebas estadísticas de contraste con modelos actuales de pronóstico, pero se tendrán en cuenta para una futura y eventual implementación operativa en la CHY y/o continuación de este trabajo, fuera del ámbito académico.

Se trabajará con el apoyo técnico del jefe de Hidrología de la CHY como Especialista de Dominio (o Informante Clave) y se solicitó, al portal unificado de información pública y transparencia del gobierno paraguayo [16], el acceso a los registros hídricos de los últimos diez años.

A nivel de procesamiento se trabajará sobre el equipo de computación disponible, una PC con procesador i7 y 32GB de RAM, que se estima suficiente para el caso en cuestión y similar en capacidad los equipos presentes en el área de Hidrología de la CHY.

Capítulo II: Literatura de antecedentes de estudios similares

Introducción

Se hizo una búsqueda por internet, utilizando motores de uso general, Google (http://www.google.com) y otros específicos de la academia, tales como Google Scholar (https://scholar.google.es/) y Academia.edu - Share research (https://www.academia.edu/) y Home - Springer (https://link.springer.com/).

¿Cómo se predice el salto hidráulico en la actualidad?

Teniendo en cuenta la definición del salto hidráulico (como la diferencia de los niveles de agua de embalse y de restitución) su pronóstico resulta de manera indirecta.

Actualmente, se realiza siguiendo las reglas técnicas del buen arte, es decir, calculando los caudales afluentes mediante el modelo HEC-HMS de la 'Hydrologic Engineering Center'[11] y la del de restitución a partir del plan de generación y del plan de vertido (agua liberada sin pasar por las turbinas).

La aplicación de este modelo se trata en detalle en capítulo 9 del libro 'Diseño Hidrológico' de los Ing. Sergio Fattorelli y Pedro C. Fernandez[12].

Con estas estimaciones se afectan los niveles de embalse y restitución actuales, resolviendo la ecuación de balance hidráulico para obtener las predicciones deseadas.

Dado que estos cálculos de nivel incluyen variables del tipo meteorológico y climático, muy difíciles de predecir con exactitud, el pronóstico del salto hidráulico también conlleva inexactitudes y, en consecuencia, deben tomarse los márgenes de seguridad que impactan, tal cual se mencionó anteriormente, en la eficiencia de la planta generadora.

Aplicación de Técnicas de Aprendizaje Automático al pronóstico de niveles

Hay trabajos de investigación recientes que exploran la aplicación de modelos predictivos de regresión lineal múltiple[1, 2 y 5] y árboles de regresión[1] a el pronóstico de caudales fluviales y subterráneos, aplicados sobre el mismo conjunto de variables que se utilizan en la ecuación de equilibrio hidráulico. Otros trabajos utilizan, sobre el mismo conjunto de variables, modelos de regresión basados en redes neuronales artificiales [1, 3 y 4].

También existen trabajos que exploran el pronóstico de los niveles de agua en reservorios naturales (lagos y lagunas) utilizando los registros periódicos de estos como una serie temporal y prediciendo con algoritmos de auto regresión del tipo ARIMA[5 y 6].

En el trabajo 'Analysis and Prediction of Dammed Water Level in a Hydropower Reservoir Using Machine Learning and Persistence-Based Techniques.MDPI' [5] , pp. 21, se concluye: 'En el análisis a corto plazo, abordamos un problema de predicción del nivel del agua en el embalse con un horizonte de tiempo de predicción de una semana, utilizando diferentes técnicas de regresión ML y variables hidro-meteorológicas exógenas. En este caso, demostramos que los algoritmos de ML son capaces de obtener resultados extremadamente precisos.

En el trabajo 'Applications of different machine learning methods for water level predictions'[1], pp. 41, se concluye: 'Tras aplicar estos cuatro métodos se ha demostrado que los algoritmos de aprendizaje automático son capaces de crear buenos modelos de pronóstico para la predicción del nivel del agua y hacerlo mejor que los modelos de regresión lineal múltiple'

En el estudio 'A Time-Series Water Level Forecasting Model Based on Imputation and Variable Selection Method', pp. 9, se concluye 'Estos resultados experimentales indican que el modelo de pronóstico Random Forest cuando se aplica con variables completas tiene mejor desempeño de pronóstico que los modelos listados. '…Esto demuestra que el modelo de pronóstico de series de tiempo propuesto es factible para el pronóstico de los niveles de agua en el embalse de Shimen'.

Capítulo III: Metodología

Introducción

Para estructurar las fases de desarrollo del proyecto y la gestión de sus objetivos específicos, se utilizará la metodología CRISP-DM, cuyas etapas de proyecto se asemejan a las etapas necesarias para alcanzar los objetivos específicos planteados.

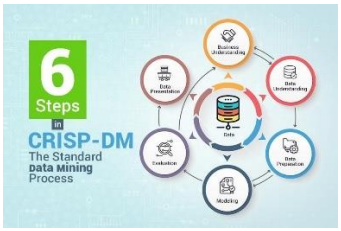

Ilustración 2: Modelo CrispDM para gestión de proyectos de datos

Ilustración 3: Librerías de Python para proyectos de datos

Etapa 1: Entendimiento del Negocio

Esta fase inicial se enfoca en la comprensión de las necesidades del cliente y de los objetivos de proyecto. Luego se define el problema de aprendizaje automático a resolver y finalmente, se prepara un plan de proyecto preliminar, diseñado para alcanzar el objetivo general y los objetivos específicos.

Objetivo específico: Entender en detalle el negocio de la generación hidroeléctrica y el impacto del salto hidráulico en la gestión del negocio. Ver adicionalmente el Capítulo I.

Etapa 2: Entendimiento de los datos

La fase comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Objetivo específico: Entender el rol de las variables que componen el conjunto de datos Objetivo específico:

Etapa 3: Preparación de los datos

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Objetivos específicos:

- Contar con un conjunto de datos hidrológicos consolidado
- Contar con un conjunto de datos hidrológicos reducido y explicativo

Etapa 4: Modelado

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de aprendizaje automático. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos.

Objetivo específico: Seleccionar un conjunto básico de modelos de Aprendizaje Automático

Etapa 5: Evaluación

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos. Antes de proceder al despliegue final del modelo, es importante

evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Objetivos específicos:

- Seleccionar el modelo de mejor desempeño
- Contar con un modelo ajustado

Etapa 6: Despliegue (puesta en producción)

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

Objetivo específico: Escribir la recomendación de aplicación del algoritmo optimizado

Nota esta fase no está incluida en el presente TFI, aunque se considera en la descripción dado que se pretende continuar luego de la aprobación de éste, fuera del ámbito académico.

Técnicas

Para el análisis exploratorio de los datos (EDA) y tratamiento de valores nulos se utilizarán gráficos del tipo línea en el tiempo, histogramas y box-plot.

Herramientas

Se trabajará en Python, utilizando los cuadernos de Jupyter (Jupyter Notebooks), con los paquetes: 'numpy' y 'pandas'[17] para el armado del conjunto de datos y su exploración, 'mathplotlib', 'Cufflinks[20] y plotly' para la parte gráfica y 'scikit-learn'[16] y 'scipy' para los algoritmos de aprendizaje automático.

Cuadernos con código de cada capítulo

El análisis de cada capítulo fue hecho código Python, ejecutado en cuadernos Jupyter ('Jupyter Notebooks') disponibilizados en GitHub, repositorio: jbeltramone/ITBA-TFI: TFI for Data Science Titulation at ITBA (github.com).A continuación se indica una tabla de referencia tema / cuaderno:

Cuaderno	Capítulo / Tema
TFI - 01	Lectura de los datos y armado del dataset
TFI - 02	Análisis Exploratorio de Datos
TFI - 03	Verificación de la condición de Random-Walk
TFI - 04	Descomposición de la serie temporal SALTO
TFI - 05	Naive Forecasting
TFI - 06	ARIMA
TFI - 07	Reducción y Pronóstico por Regresión
TFI - 08G	Modelo por desplazamiento diario del vector de predictoras
TFI - 09G	Modelo de pronóstico con offset variable
TFI - 09I	Modelo de pronóstico con offset variable y feature-engineering
TFI - 09J	Multivariable con Test de causalidad de Granger
TFI - 10	Análisis del desempeño de los modelos de pronóstico
TFI - 11	XGBRegressor, modelo por desplazamiento diario del vector de predictoras, Optimización
TFI -11B	KNRegressor, modelo por desplazamiento diario del vector de predictoras, Optimización

Tabla 1: Lista de cuadernos de código, por tema

Hipótesis de la Investigación

"Es posible pronosticar el salto hidráulico de una central hidroeléctrica (la diferencia entre los niveles de embalse y restitución), utilizando algoritmos de Aprendizaje Automático (Machine Learning) con una exactitud similar o mayor a la actual".

Nota: Fuera del alcance del presente trabajo, si la CHY autoriza el acceso a sus datos históricos y permite una implementación experimental del modelo desarrollado, se pretende extender la hipótesis de trabajo como sigue:

"Es posible predecir el salto hidráulico (la diferencia entre los niveles de embalse y restitución) de una central hidroeléctrica, utilizando algoritmos de Aprendizaje Automático (Machine Learning), con al menos 10% más de exactitud.

Recolección de los datos.

Crisp-DM - Etapa 3: Preparación de los datos

Los datos fueron provistos por la CHY, a través de una solicitud hecha al Gobierno de la República del Paraguay en el Portal de acceso a la información pública (Portal Paraguay - Acceso a la Información Pública) (16)

Del portal unificado de información pública y transparencia del gobierno paraguayo se recibieron varios archivos Excel con registros diarios de las variables mencionadas anteriormente, conteniendo, en general, una variable por archivo y con formatos tabulares diversos a los que se debió adaptar los algoritmos de lectura.

Mediante un programa realizado en Python se constató la frecuencia y las fechas de registros, de cada variable, resultando diarias y desde el 2011-01-01 al 2021-06-30 para todas ellas.

Se detectaron valores nulos en 7 posiciones de las variables COTA_EMBALSE, COTA_RESTITUCION, CAUDAL_ENTRANTE, CAUDAL_TURBINADO y CAUDAL_VERTIDO que, luego de analizarlos con el especialista de dominio, se acordó en completarlos con valores interpolados mediante polinomios de orden 2 (cuadráticos).

Para unificar todas las variables en un solo conjunto (dataset de 'pandas'[17]) y verificar que no quedaran valores faltantes en las variables, se utilizó una serie temporal con valores que van desde la primera hasta la última fecha (2011-01-01 al 2021-06-30) de los registros, en una operación de unión 'merge' de 'pandas'[17].

Posteriormente se guardó el dataset en un único archivo separado por comas ('.csv') 'CHY_dataset.csv en la carpeta 'output' del proyecto, para su posterior lectura y explotación.

Análisis Exploratorio de los datos (EDA)

Crisp-DM - Etapa 2: Entendimiento de los datos

Variables recibidas del Gobierno de la República del Paraguay.

Cota (nivel) de embalse, medido en metros

- Relación de paralelismo (+…+)
- Tipo: Independiente
- Definición nominal y operacional: Cuantitativa, medida (no-calculada) directamente en la represa, aguas arriba.

Cota (nivel) de restitución, medido en metros

- Relación de paralelismo (+…-)
- Tipo: Independiente (¿dependiente?)
- Definición nominal y operacional: Cuantitativa, medida (no-calculada) directamente en la represa, aguas abajo.

Caudal afluente, medido en m3/seg

- Relación de paralelismo (+…+)
- Tipo: Independiente
- Definición nominal y operacional: Cuantitativa, calculada.

Caudal turbinado, medido en m3/seg

- Relación de paralelismo (+…-)
- Tipo: Dependiente
- Definición nominal y operacional: Cuantitativa, medida.

Caudal vertido, medido en m3/seg

- Relación de paralelismo (+…-)
- Tipo: Independiente
- Definición nominal y operacional: Cuantitativa, medida.

Energía generada lado Argentina, medida en MWh

- Relación: Paralelismo (+…-)
- Tipo: Independiente
- Definición nominal y operacional: Cuantitativa, medida (no-calculada) directamente en la represa.

Energía generada lado Paraguay, medida en MWh

- Relación: Paralelismo (+…-)
- Tipo: Independiente
- Definición nominal y operacional: Cuantitativa, medida (no-calculada) directamente en la represa.

Nota: Para simplificar la comprensión del objeto central del TFI, que es el pronóstico de los valores futuros de la serie SALTO el análisis exploratorio de los datos se documenta en el Anexo-I del presente TFI, incluyendo verificación de las condiciones de 'Ruido Blanco' y de 'Paseo Aleatorio (random-walk) de la variable SALTO), dejando en este cuerpo solamente algunas conclusiones sobresalientes de dicho anexo para la variable, 'SALTO'. El Anexo-I se encuentra disponibilizado, junto con los cuadernos de código, en el repositorio GitHub (TBD).

EDA específico para la variable SALTO

Dado que la capacidad de generar energía de la CHY depende de la diferencia entre las cotas de embalse y de restitución y que ésta es la variable de pronóstico para la gestión eficiente de la planta, se define una nueva variable en el conjunto de datos igual a la diferencia entre ambas cotas, llamada 'salto hidráulico', con nomenclatura 'SALTO' dentro del dataset: SALTO = COTA_EMBALSE – COTA_RESTITUCION,

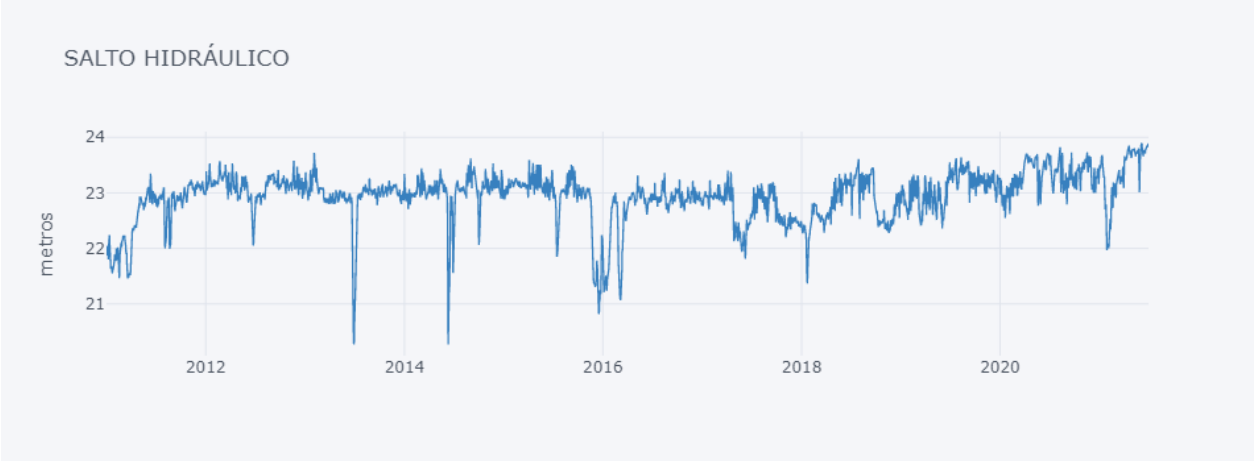

Ilustración 4: Variable 'SALTO', en el tiempo

Ilustración 5: Diagrama de caja (boxplot) de la variable SALTO

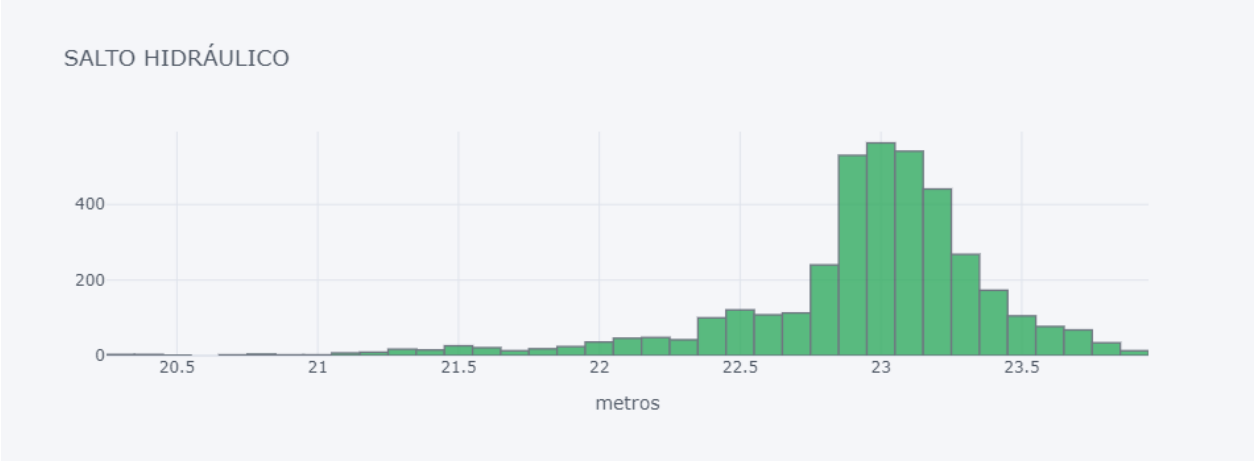

Ilustración 6: Histograma de valores de la variable SALTO

Puede verse de los diagramas anteriores, que el valor del salto hidráulico está alrededor de los 23 metros (correspondiente a la diferencia de los niveles de embalse y de restitución), con cuartiles, Q1: 22.89 metros, Q2: 23.01 metros y Q3: 23.19 metros y que no experimenta variaciones estacionales importantes a lo largo de los diez años de la muestra.

Los valores que aparecen más allá de los bigotes inferior y superior del diagrama de boxplot fueron validados con el especialista de dominio, justificados y tomados como verdaderos.

Adicionalmente puede verse en el histograma que los valores más allá de los bigotes (atípicos) son muy poco frecuentes comparados con el central (~15 vs. 663).

Notar que dadas las limitaciones (ya mencionadas anteriormente) en la capacidad máxima de generación y en la variación de los niveles de embalse y restitución (cotas máximas y mínimas), la relación entre el salto hidráulico y la energía generada es positiva para bajos niveles del salto, pero negativa para valores altos (gráfico inferior izquierdo).

Capítulo IV: Investigación

Modelado

Crisp-DM - Etapa 4: Modelado y ensayo

Teniendo ya un buen conocimiento de la serie SALTO y de su comportamiento, el próximo paso es ensayar diversas técnicas y algoritmos de pronóstico ('forecasting') y observar cómo se comporta cada uno en las diferentes condiciones de ensayo, para luego seleccionar el más apropiado para nuestro caso.

Definición de las métricas de evaluación de los métodos de pronóstico ('forecasting')

Para poder comparar objetivamente el resultado de cada ensayo es necesario definir herramientas de evaluación a aplicar a los resultados de cada ensayo.

Entre las métricas comúnmente utilizadas en procesos de forecasting y regresión (Forecast KPI: RMSE, MAE, MAPE & Bias | Towards Data Science) elegimos MAE ('Mean Absolute Error') y RMSE ('Root Mean Squared Error').

Definido el error 'e_t' de cada pronóstico en el instante de tiempo 't', como el valor real de la serie 'f_t' y la determinación 'd_t' para dicho instante, es decir:

$$e_t = f_t - d_t$$

El MAE es definido como:

$$MAE = \frac{1}{n}\sum |e_t|$$

Y que nos permite observar el error medio de nuestras predicciones en la misma unidad de medida que nuestra serie. Los valores de la serie SALTO están en metros, por lo tanto, el MAE se medirá en metros también.

Y el RMSE definido como:

$$RMSE = \sqrt{\frac{1}{n}\sum e_t^2}$$

Y que, además de expresar el error en metros, nos pondrá en evidencia la existencia de valores pico en los errores, por la elevación al cuadrado de cada error. Es decir, la existencia de un solo error de gran magnitud respecto de los demás se hará evidente en el RMSE, pues no se enmascarará en el promedio.

Horizontes de pronóstico

Dado que la CHY hace diariamente un pronóstico 'rolling' del valor del salto hidráulico a cinco días visto (al día siguiente, al siguiente más uno, más dos y hasta más cinco), se utilizará en este trabajo un horizonte de pronóstico ('forecasting horizon') de siete días (fh = 7, 1 semana), que cubre el período actualmente en uso y también otro de largo plazo, 56 días, para analizar los pronósticos a largo plazo (8 semanas).

Entonces, en el contexto de este trabajo, pronosticar a 7 días significa: con los valores de SALTO y demás variables predictoras del día 'n' (medidos ese mismo día), pronosticar el valor del nivel de SALTO al día siguiente, n+1, y posteriores, n+2, n+3 hasta n+7. Hay que recordar que los valores del conjunto de datos son diarios, tomados a la misma hora del día. Lo mismo aplica al horizonte de pronóstico de 56 días, parado en el días 'n' con los datos de dicho día, se trata de pronosticar los niveles de SALTO para los días n+1, n+2, n+3, …., hasta n+56.

Pronóstico ingenuo ('naive forecasting')

Se establecerá un piso, o línea base, de desempeño para las futuras evaluaciones, mediante una pronóstico ingenuo de valores futuros usando las diferentes estrategias de pronóstico naive (NaiveForecaster — sktime documentation) de la librería SKTime[18] (https://www.sktime.org/).

Estos mecanismos de pronóstico son sencillos y no requieren mayor esfuerzo, porque asumen que el comportamiento de la serie seguirá patrones evidentes de su visualización, tales como repetición de valores previos, absolutos o promedios, estacionalidad y tendencia.

Pronóstico a 7 días (1 semana)

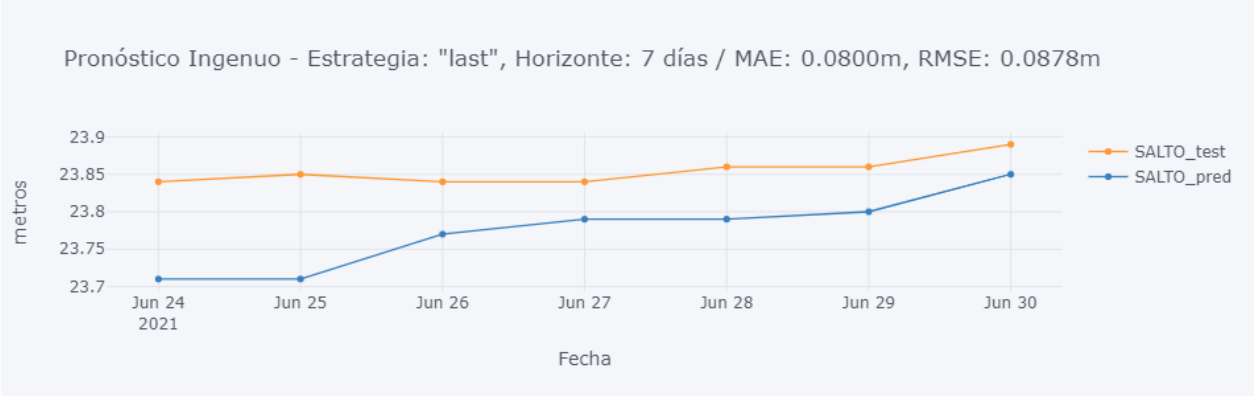

Ilustración 7: Pronóstico Ingenuo a 7 días, estrategia 'last'

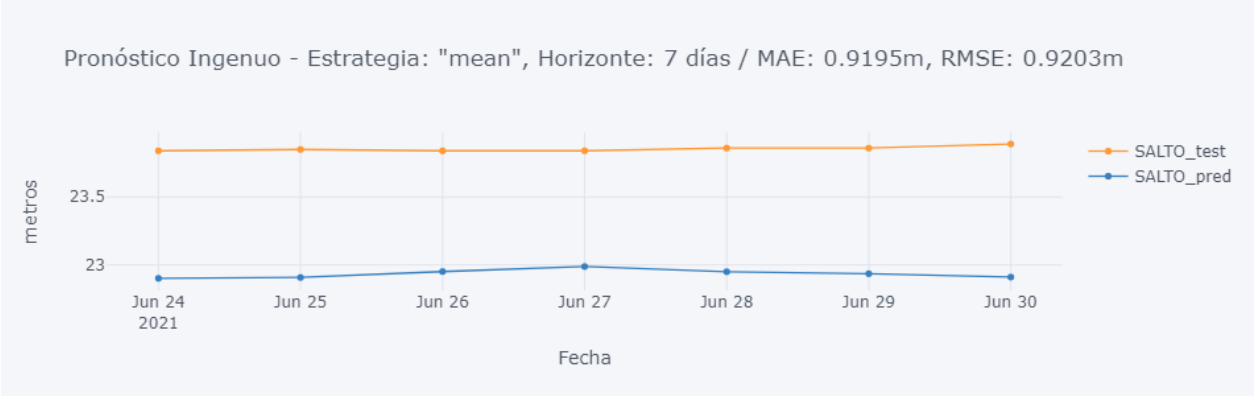

Ilustración 8: Pronóstico Ingenuo a 7 días, estrategia 'mean'

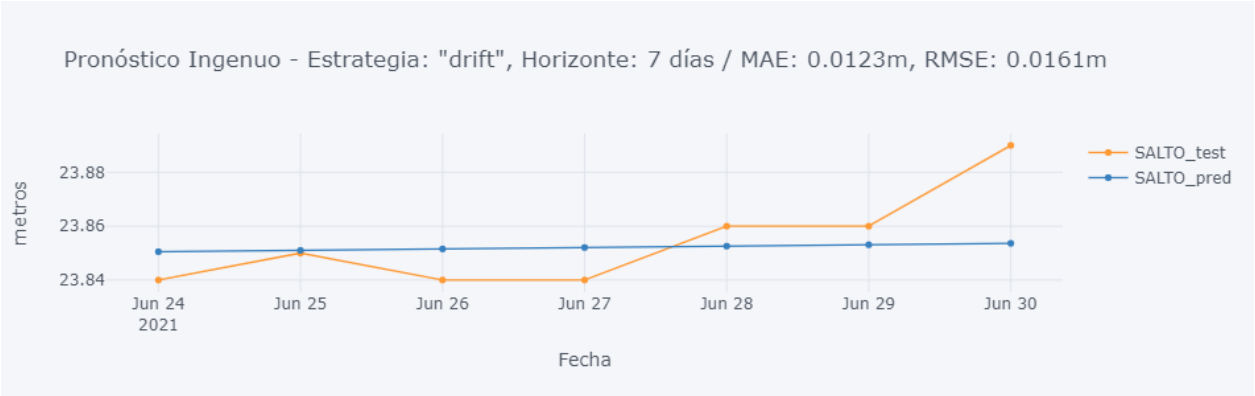

Ilustración 9: Pronóstico Ingenuo a 7 días, estrategia 'drift'

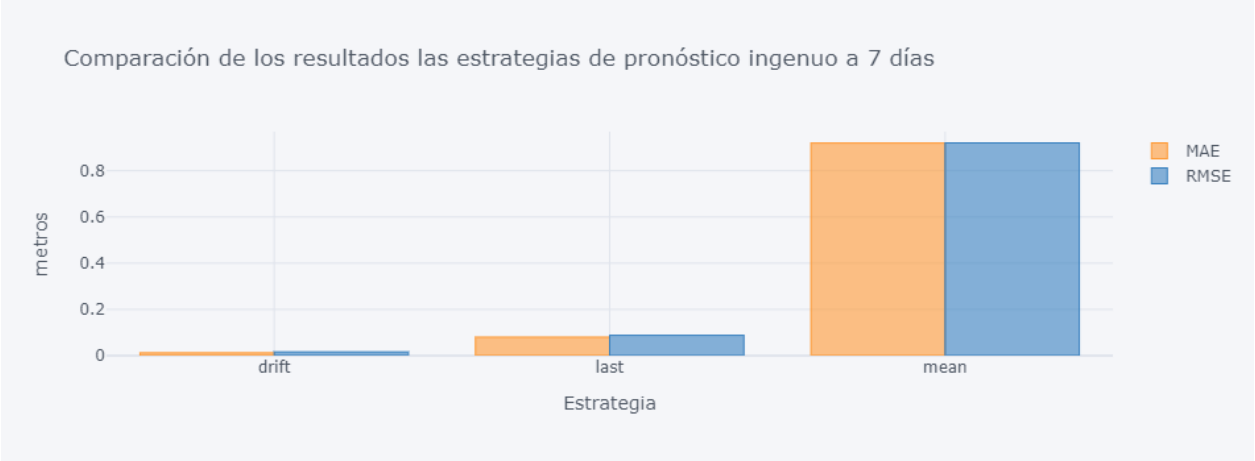

Ilustración 10: Comparación de las diferentes estrategias de Pronóstico Ingenuo a 7 días

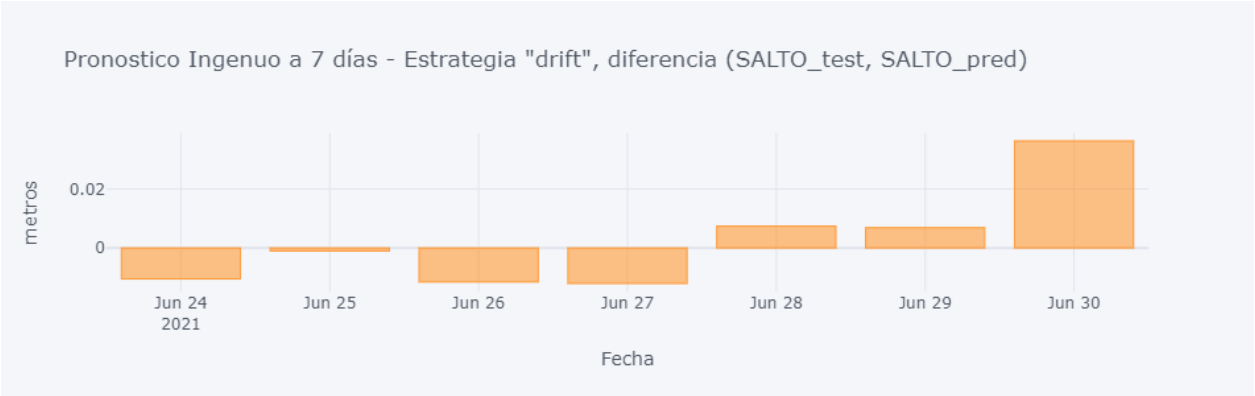

Ilustración 11: Diferencias de pronóstico ingenuo a 7 días, de la estrategia 'drift'

Pronóstico a 56 días (8 semanas)

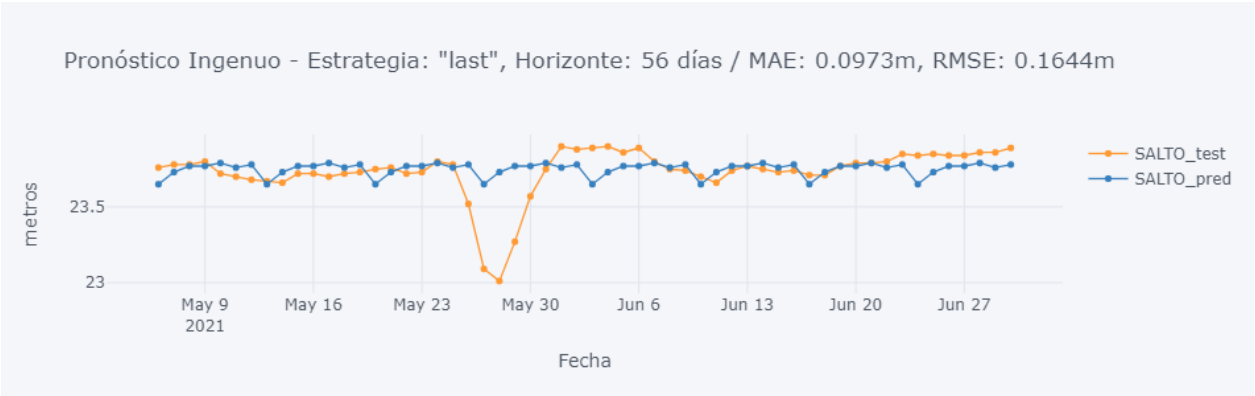

Ilustración 12: Pronóstico Ingenuo a 56 días, estrategia: 'last'

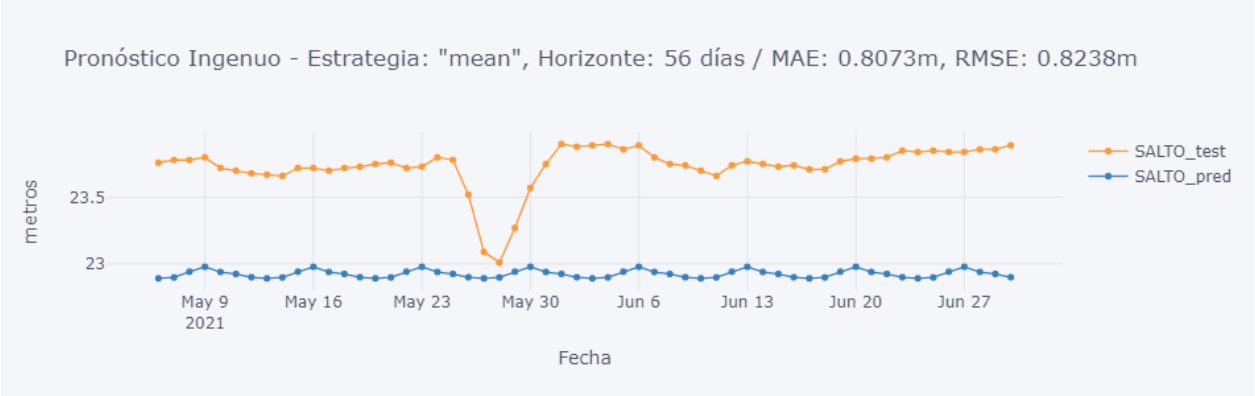

Ilustración 13: Pronóstico Ingenuo a 56 días, estrategia: 'mean'

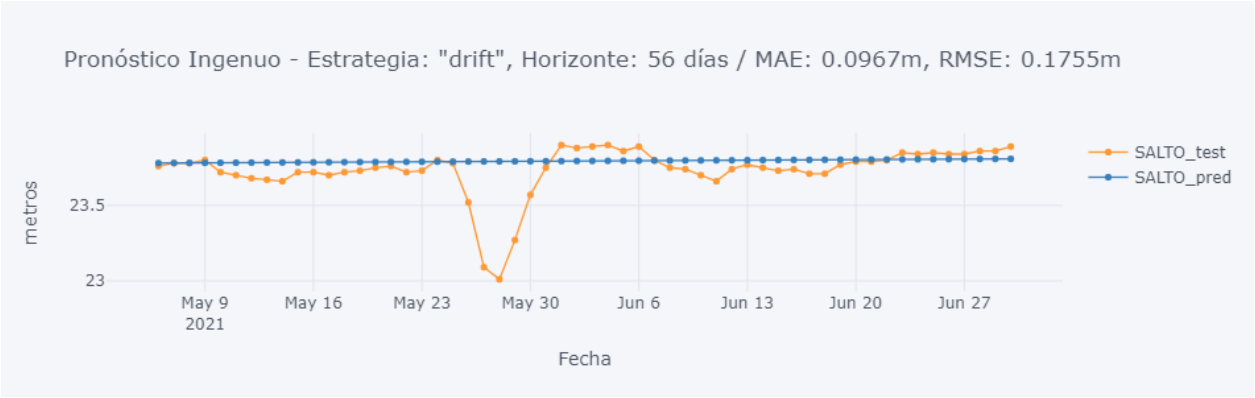

Ilustración 14: Pronóstico Ingenuo a 56 días, estrategia: 'drift'

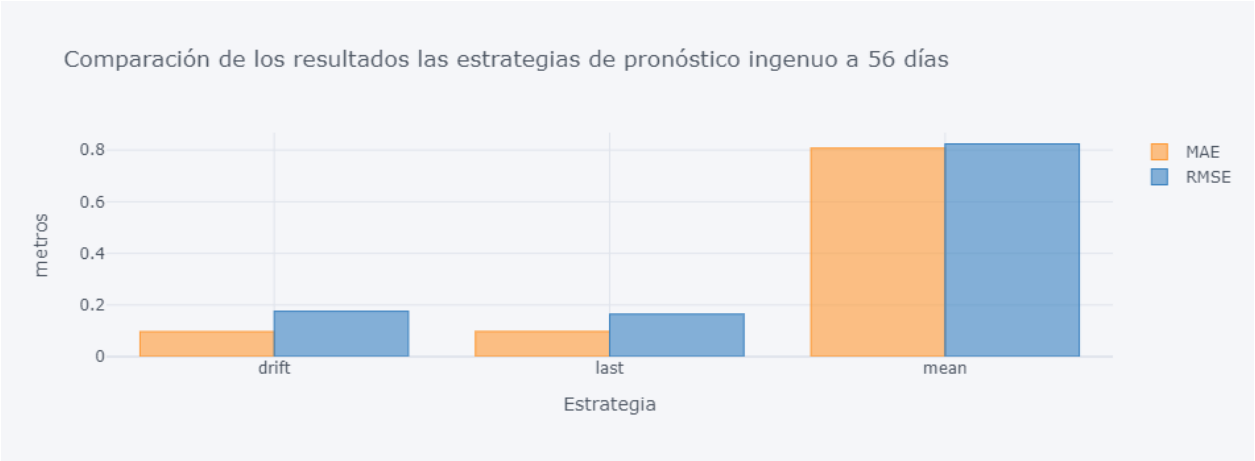

Ilustración 15: Comparación de las diferentes estrategias de Pronóstico Ingenuo a 56 días

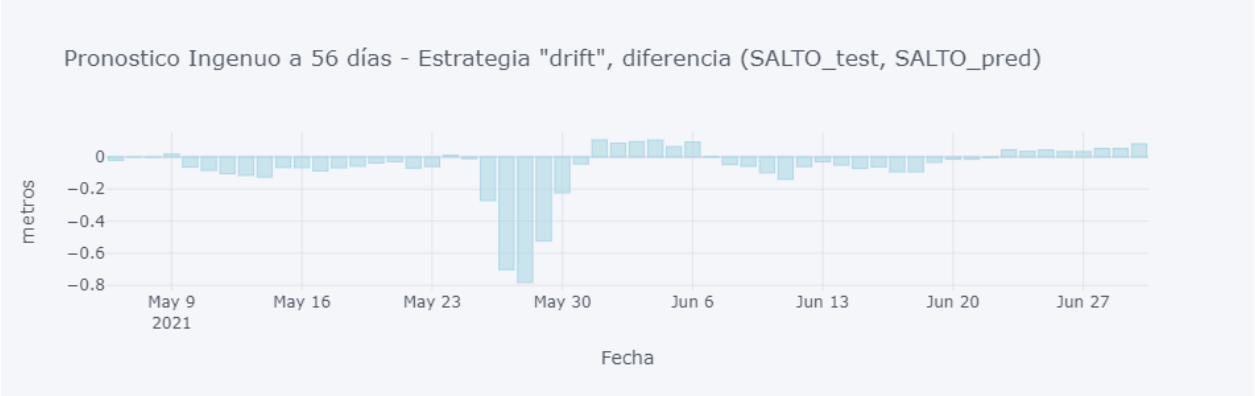

Ilustración 16: Ilustración 11: Diferencias de pronóstico ingenuo a 56 días, de la estrategia 'drift'

Vemos que para ambos horizontes de pronóstico las diferencias entre los valores reales (SALTO_test) y los pronosticados (SALTO_pred) para la estrategia 'drift', son las menores.

Estrategia	Horizonte(días)	MAE(m)	RSME(m)
'last'	7	0.0800	0.0878
	56	0.0973	0.1644
'mean'	7	0.9195	0.9203
	56	0.8073	0.8238
'drift'	7	0.0123	0.0161
	56	0.0967	0.1755

Tabla 2: Comparación de desempeño de las estrategias de Pronóstico Ingenuo

La estrategia de pronóstico ingenuo que mejor desempeño tuvo es la 'drift'. Esta estrategia ajusta una recta entre el primer y el último punto de la ventana de datos de entrenamiento y extrapola valores para el pronóstico (futuro)[23].

Métodos econométricos (ARIMA)

En esta sección se realizarán los pronósticos haremos pronóstico de valores futuros utilizando modelos autorregresivos integrados de media móvil. Primeramente, se analizarán las condiciones de estacionariedad, necesarias y luego se construirá el modelo correspondiente con la función AutoARIMA de la librería 'statsmodel'.

Verificación de condiciones de estacionariedad

Las condiciones de estacionariedad para una serie temporal son:

1. Media constante
2. Varianza constante
3. Covarianza constante entre períodos de idéntica distancia

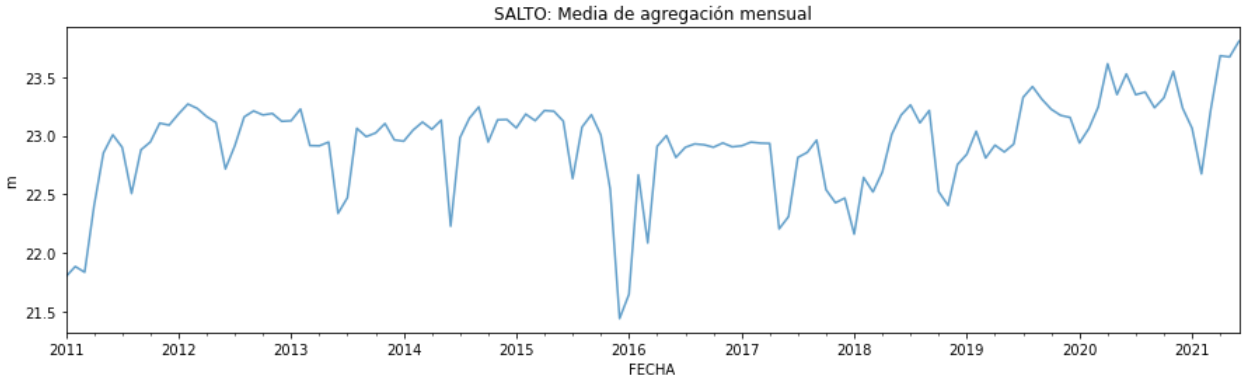

Ilustración 17: Media de agregación mensual de SALTO, en el tiempo.

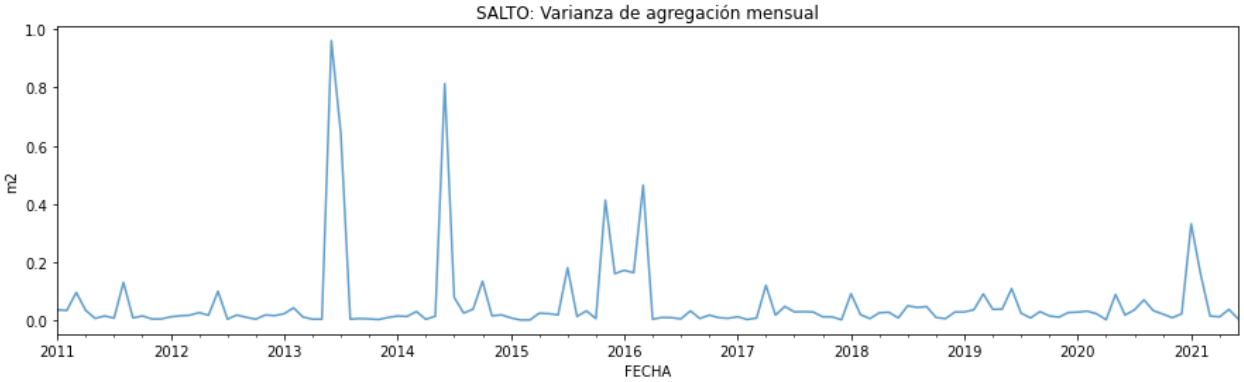

Ilustración 18: Varianza de agregación mensual de SALTO, en el tiempo.

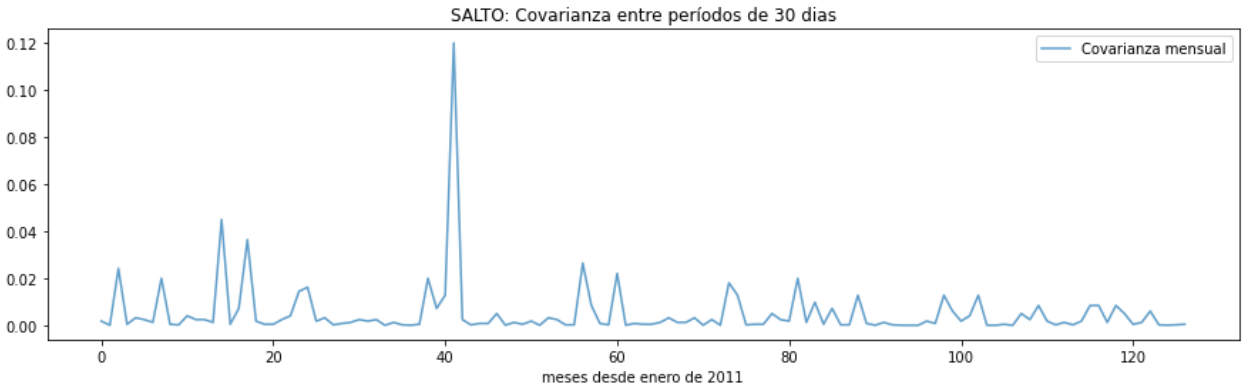

Ilustración 19: Covarianza entre períodos de 30 días de SALTO, en el tiempo.

Puede verse que las condiciones de estacionariedad no se cumplen. Por lo tanto habrá que diferenciar la serie SALTO antes de aplicar los modelos ARIMA.

Modelo ARIMA

Se dividirá SALTO en dos conjuntos, uno para entrenamiento del modelo ARIMA (SALTO_train) y otro para control de los valores pronosticados (SALTO_test).

Entonces, en el tiempo, la (secuencia de las muestras de SALTO serán: SALTO = SALTO_train + SALTO_test

Como el conjunto de datos de SALTO va desde el 01/enero/2011 al 30/junio/2021 y el horizonte de pronóstico es de siete días, SALTO_train contendrá los valores de SALTO desde el 01/enero/2011 al 23/junio/2021 y SALTO_test los valores de SALTO desde el 24/junio/2021 al 30/junio/2021.

Y, para el caso del pronóstico a 56 días, SALTO_train contendrá los valores de SALTO desde el 01/enero/2011 al 05/mayo/2021 y SALTO_test los valores de SALTO desde el 06/mayo/2021 al 30/junio/2021.

Utilizando la función AutoARIMA de la librería SKTime[18] (AutoARIMA — sktime documentation), creamos un modelo con estacionalidad de siete períodos (días en este caso, condición ya verificada anteriormente) que se ajustará al conjunto SALTO_train, para luego pronosticar los valores para el horizonte de pronóstico, que en este caso serán, primero a siete días y luego a cincuenta y seis.

El algoritmo de AutoARIMA selecciona como mejor modelo un ARIMA con (P=2, D=1, Q=1) (diferenciando una vez como se infirió anteriormente) y se obtienen los siguientes gráficos:

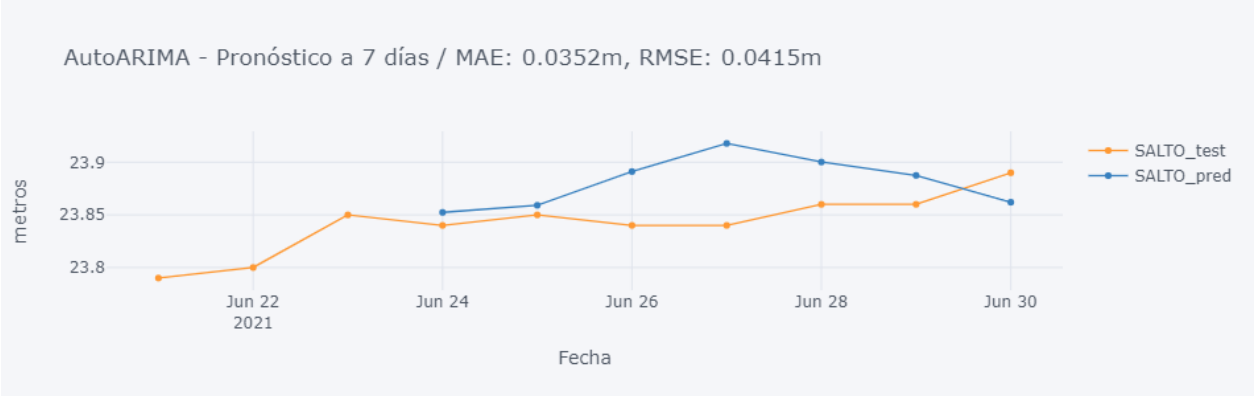

Ilustración 20: Pronóstico con AutoARIMA a 7 días

Se observa que los valores pronosticados están, salvo para el último día, por debajo de los reales, es decir, el modelo tiene un sesgo.

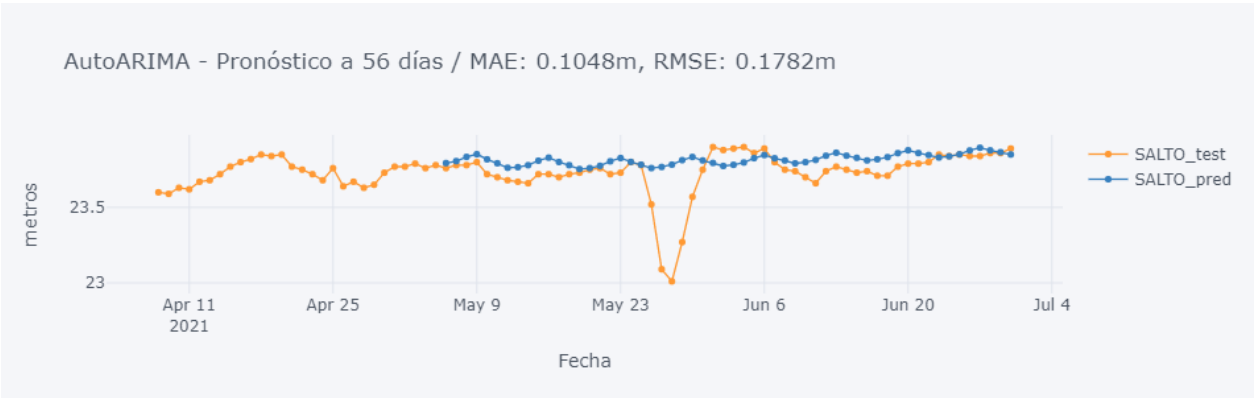

Ilustración 21: Pronóstico con AutoARIMA a 56 días

Para el pronóstico a 56 días, los valores pronosticados están tanto por debajo como por encima de los reales. Podría descartarse un sesgo para este caso.

Modelo	Horizonte(días)	MAE(m)	RSME(m)
AutoARIMA	7	0.0351	0.0415
	56	0.1048	0.1782

Tabla 3: Comparación de desempeño, AutoARIMA, pronósticos a 7 y a 56 días

Se observa que, en término de las métricas MAE y RSME, el desempeño del pronóstico a 7 días es mejor que el de a 56.

Comparación de resultados

En la tabla siguiente se listan los mejores valores obtenidos hasta el momento de cada modelo, para los horizontes de pronóstico de 7 y de 56 días.

Modelo	Horizonte(días)	MAE(m)	RSME(m)
AutoARIMA	7	0.0351	0.0415
	56	0.1048	0.1782
Pronóstico Ingenuo 'drift'	7	0.0123	0.0161
	56	0.0967	0.1755

Tabla 4: Comparación de desempeño, entre modelos de Pronóstico Ingenuo y AutoArima

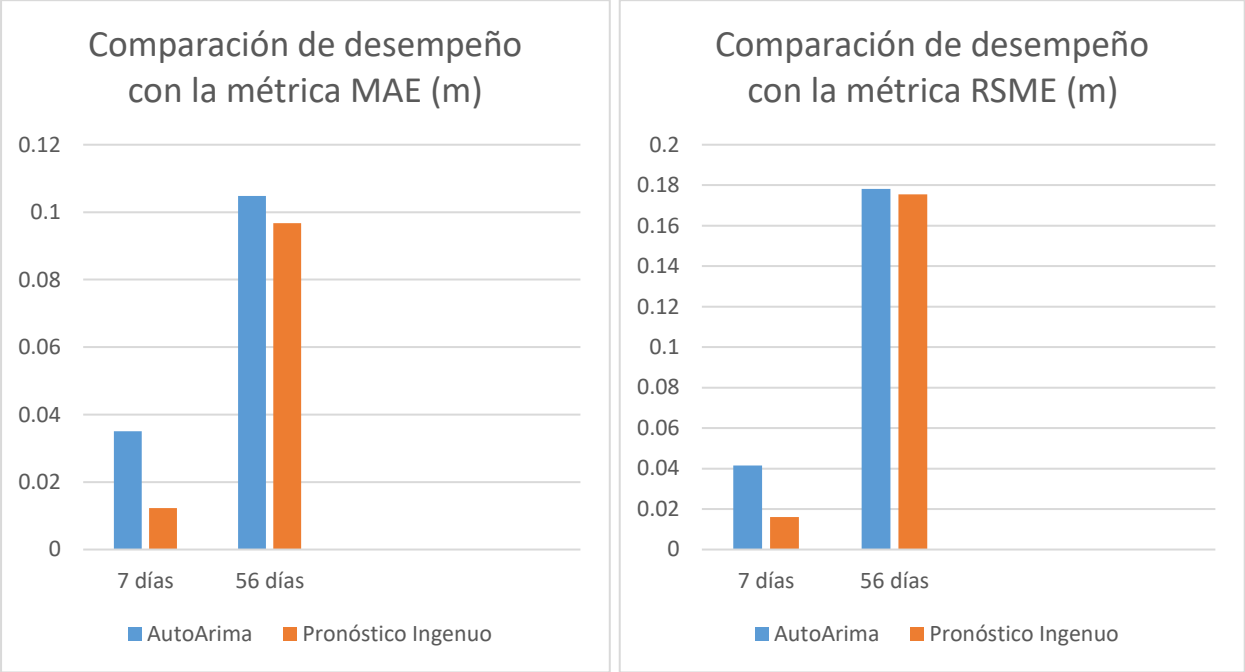

Ilustración 22: Comparación de desempeño, Pronosticador Ingenuo vs. AutoARIMA

Se aprecia que el desempeño del Pronosticador Ingenuo con estrategia 'drift' es mejor para ambos horizontes en términos de MAE y también de RSME.

Según la librería 'SKTime'[18], el modelo AutoARIMA tiene un paso interno de optimización. No obstante, convendría ensayar un paso de optimización de sus hiperparámetros, para ver si es posible mejorar su desempeño. Por el momento este paso se considera fuera del alcance del presente TFI, enfocado a modelos de Aprendizaje Automático.

Aprendizaje Automático (Machine Learning)

En esta fase se ensayarán diversos modelos y algoritmos de pronóstico de aprendizaje automático supervisado, desde una pasada de evaluación rápida, con parámetros default, continuando con el pronóstico más allá del dataset de entrenamiento, con ingeniería de variables, para luego seleccionar y optimizar los parámetros del modelo que mejor se desempeñe. Se continuará trabajando con los dos horizontes de pronóstico utilizados, es decir a siete (7) y a cincuenta y seis (56) días.

Plan de ensayo

Paso 1

Primeramente se hará una pasada sencilla en una configuración clásica de aprendizaje automático supervisado, dividiendo el conjunto de datos en conjuntos de entrenamiento y de testeo, generando las predictoras a partir de versiones retrasadas de la variable 'SALTO'. El pronóstico se hará sobre las predictoras del conjunto de testeo y la evaluación del desempeño sobre el vector de target del conjunto de testeo.

Dicho de otra manera, este paso puede imaginarse como una sucesión de pronósticos a 'un día', sobre los valores 'reales' de la variable, que se van incorporando a diario luego de cada medición, situación distinta a la de los próximos pasos, donde no se cuenta con esta información, porque simplemente es del futuro y no existe al momento de pronosticar.

Paso 2

Luego se continuará con pronósticos 'más allá del conjunto de datos existente' (modo real de trabajo), donde, del conjunto de testeo, se utilizará únicamente el vector de 'target' para evaluación de desempeño. Se ensayarán dos modelos:

1) Con un modelo ajustado a predecir el valor de SALTO al día siguiente, el vector de predictoras para el pronóstico del próximo valor de 'SALTO' se irá generando día a día, desplazando el último vector utilizado (en un día) y completando el valor de la casilla correspondiente al último valor de SALTO ('SALTO-01'), que a consecuencia del desplazamiento se transformará en 'NaN', con el último valor pronosticado. Se realizarán tantos pronósticos a un día, como días tenga el horizonte de pronóstico.

Paso 3

2) Con un ensamble de modelos ajustados cada uno a un offset de pronóstico distinto. Es decir, en un horizonte de pronóstico de siete días, habrá siete modelos, uno ajustado para el pronóstico al día siguiente (+1), otro al siguiente de éste (+2), otro al siguiente de éste también (+3) y así hasta el último modelo, que pronosticará para el último día del horizonte (+7 y +56 para este trabajo).

Paso 4

En la modalidad de ensamble de modelos se ensayarán también:

a) Ingeniería de variables ('feature-engineering') sobre la misma serie 'SALTO' generando nuevas columnas de predictoras. Es decir, se mantendrá el entorno univariable, dado que toda la información de las predictoras partirá de la misma serie de trabajo 'SALTO'.

Paso 5

b) Un entorno multivariable, agregando al conjunto de entrenamiento, las variables exógenas del conjunto, con un tratamiento previo de evaluación de su inclusión, con una prueba de causalidad de Granger.

Paso 6

A continuación, se evaluarán los resultados obtenidos para cada horizonte de pronóstico.

Paso 7

Finalmente, se concluirá con la optimización del modelo que mejor se haya desempeñado en cada horizonte de pronóstico.

Algoritmos (regresores) a utilizar

De la librería 'sci-kit learn' se utilizarán:

- KNeighborsRegressor()
- Lasso()
- LinearRegression()
- MLPRegressor()
- RandomForestRegressor()

- Ridge()
- SVR()

De la librería LightGBM:

- LGBMRegressor()

Y, de la librería dmlc XGBoost:

- XGBRegressor()

Paso 1: Pronóstico de valores futuros utilizando algoritmos de regresión, con conjuntos de datos de entrenamiento y de testeo, en modo univariable

En este capítulo se ensayarán diversos algoritmos de regresión, sobre un esquema de variables predictoras y un vector objetivo ('target'), generados a partir de la misma serie bajo análisis, SALTO, siguiendo un proceso de 'Reducción'. El conjunto de datos se subdividirá en uno de entrenamiento del modelo y en otro de testeo, sobre el que se hará el pronóstico y la evaluación de su desempeño. Nota: Recordar del párrafo anterior, que, en este paso, el pronóstico se realiza contando con los valores reales de las predictoras del período que se quiere pronosticar.

La matriz de variables predictoras ('features') se construye con versiones retrasadas ('lags') de la serie SALTO. y el vector objetivo ('target') con la versión corriente de esta misma serie. De aquí la denominación 'univariable': solo la información contenida en la serie SALTO y sus retrasos es utilizada para el pronóstico.

A partir de allí, el proceso de pronóstico continúa como un proceso normal de aprendizaje automático supervisado de regresión, repetido sucesivamente para los algoritmos de regresión seleccionados, para concluir con un análisis del desempeño de cada uno.

Conceptualmente, las variables predictoras serán los valores de SALTO en sucesivos días anteriores, comenzando con el del día inmediatamente anterior ('SALTO-01') y terminando con un número a determinar a partir del diagrama de autocorrelación.

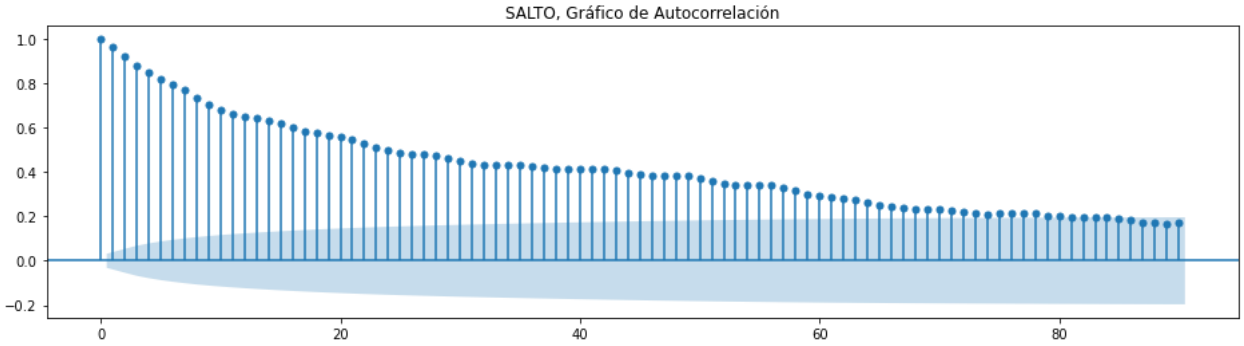

Ilustración 23: Diagrama de Autocorrelación de 'SALTO'

Se aprecia que hay correlación estadísticamente significativa del valor presente de SALTO con hasta los primeros sesenta (60) valores retrasados, aproximadamente.

Se experimentará con dicho conjunto de versiones retrasadas y con uno menor de cuarenta (40).

Nota: En este paso, los algoritmos de regresión se utilizan con sus hiperparámetros 'por defecto', sin ningún tipo de optimización.

Los resultados se presentan en las siguientes tablas, donde se observa que:

1) los mejores resultados se obtienen con un número menor de versiones retrasadas, cuarenta, en lugar de con sesenta, para ambos horizontes de pronóstico.
2) para ambos horizontes de pronóstico los algoritmos que mejor se desempeñan son la regresión lineal y la regresión Ridge.

Regresor	Lags	H(días)	MAE(m)	RMSE(m)
LinearRegression	40	7	0.035102	0.001556
LinearRegression	60	7	0.035668	0.001537
Ridge	40	7	0.035687	0.001611
Ridge	60	7	0.036341	0.001582
LGBMRegressor	60	7	0.057050	0.004568
LGBMRegressor	40	7	0.059598	0.004516
XGBRegressor	60	7	0.066337	0.005045
RandomForestRegressor	60	7	0.075600	0.006627
RandomForestRegressor	40	7	0.077100	0.006222
MLPRegressor	60	7	0.079329	0.007386
XGBRegressor	40	7	0.079906	0.007520
SVR	60	7	0.100191	0.010375
SVR	40	7	0.101530	0.010563
MLPRegressor	40	7	0.176510	0.046396
KNeighborsRegressor	60	7	0.190571	0.041938
KNeighborsRegressor	40	7	0.234571	0.060018
Lasso	60	7	0.902157	0.814169
Lasso	40	7	0.907603	0.824024

Tabla 5: Desempeño de los regresores en pronóstico a 7 días, con 40 y con 60 versiones retrasadas ('lags'). Sin optimización.

Regresor	Lags	H(días)	MAE(m)	RMSE(m)
LinearRegression	40	56	0.052908	0.007460
LinearRegression	60	56	0.053651	0.007674
Ridge	40	56	0.053785	0.007709
Ridge	60	56	0.054611	0.007932
LGBMRegressor	40	56	0.081537	0.011708
LGBMRegressor	60	56	0.098023	0.014325
RandomForestRegressor	40	56	0.110236	0.017030
MLPRegressor	40	56	0.117849	0.032541
SVR	40	56	0.139742	0.023856
RandomForestRegressor	60	56	0.153821	0.034220
SVR	60	56	0.154896	0.029365
XGBRegressor	40	56	0.166334	0.040505
KNeighborsRegressor	40	56	0.218143	0.054011
MLPRegressor	60	56	0.253137	0.076036
XGBRegressor	60	56	0.324172	0.156967
KNeighborsRegressor	60	56	0.404536	0.223904
Lasso	60	56	0.789883	0.651470
Lasso	40	56	0.795346	0.660131

Tabla 6: Desempeño de los regresores en pronóstico a 56 días, con 40 y con 60 versiones retrasadas ('lags'). Sin optimización.

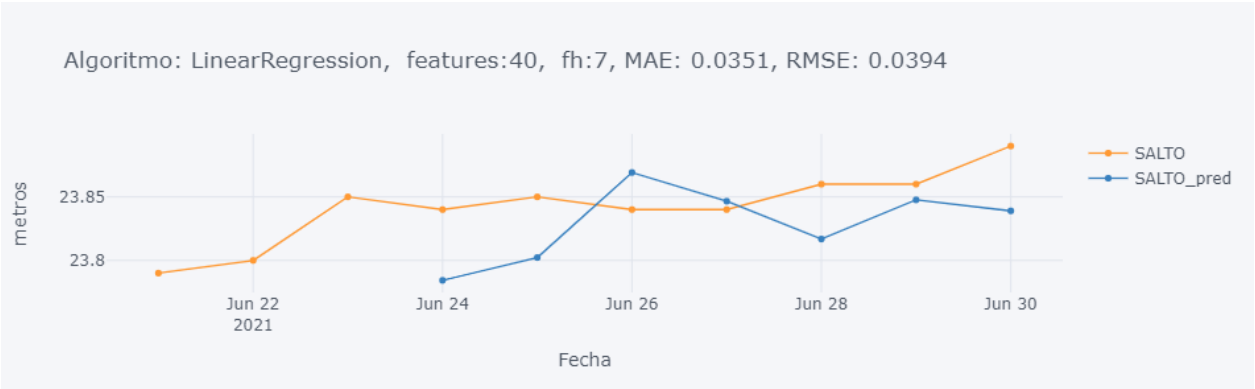

Ilustración 24: Pronóstico a siete días utilizando Regresión Lineal. Sin optimización.

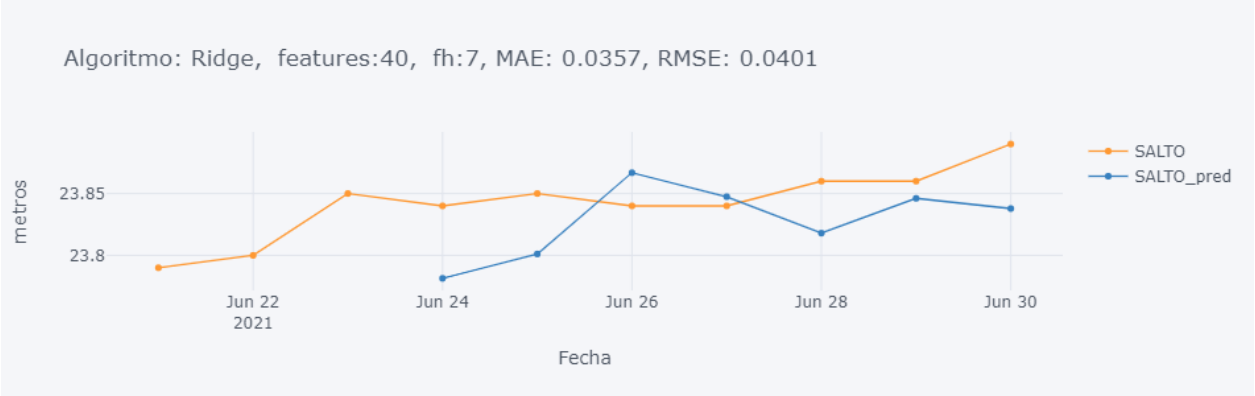

Ilustración 25: Pronóstico a sietedías, utilizando Regresión Ridge. Sin optimización.

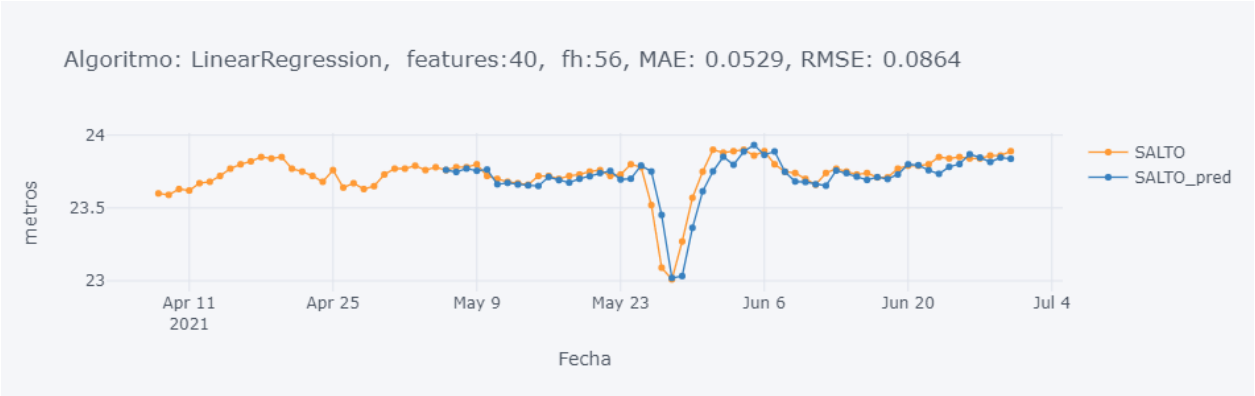

Ilustración 26: Pronóstico a cincuenta y seis días, utilizando Regresión Lineal. Sin optimización.

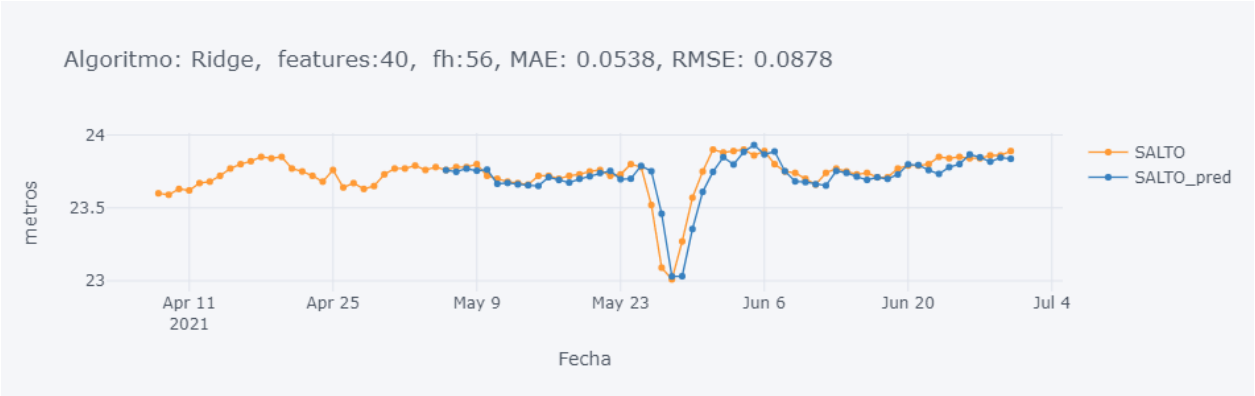

Ilustración 27: Pronóstico a cincuenta y seis días, utilizando Regresión Ridge. Sin optimización.

Se muestran las regresiones de los dos mejores desempeños, Regresión Lineal y Regresión Ridge. No obstante, los demás diagramas pueden verse en el cuaderno de código correspondiente.

Se observa que los valores del MAE son mayores para el horizonte de pronóstico de cincuenta y seis días.

Cross-Validation de los resultados

Para evaluar la varianza de los pronósticos según los datos de cálculo, se realizó un análisis de verificación cruzada ('Cross–Validation'), de 20 corridas. A continuación, se grafican los resultados en un diagrama de boxplot, mostrando por separado el correspondiente al regresor MLPRegressor, dado que, por el tamaño de sus bigotes, modifica la escala de manera de empequeñecer el gráfico de los demás.

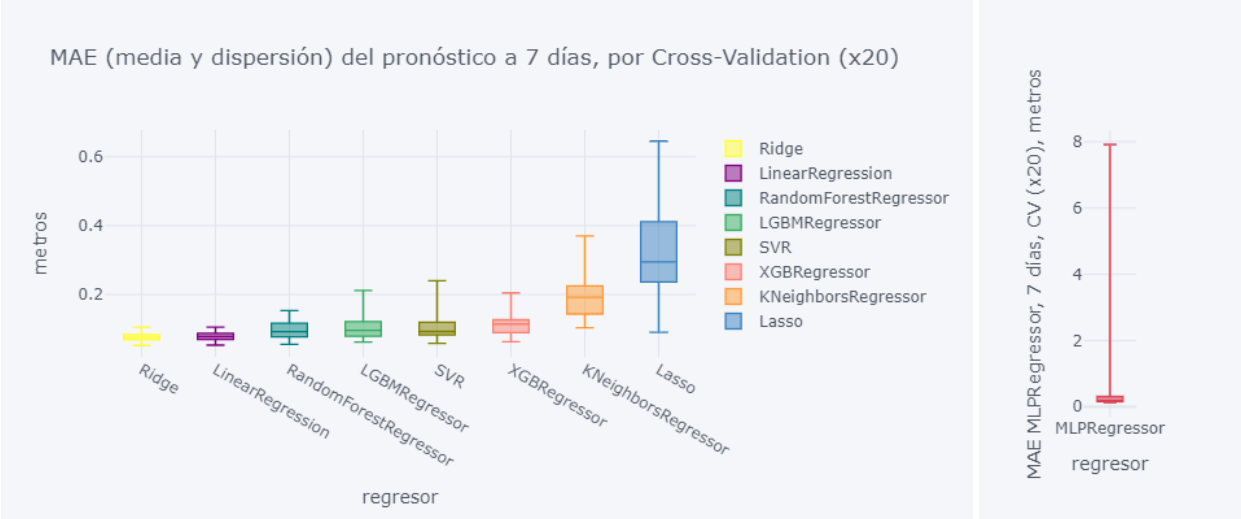

Ilustración 28: Pronóstico a siete días, Cross-Validation x20. Sin optimización.

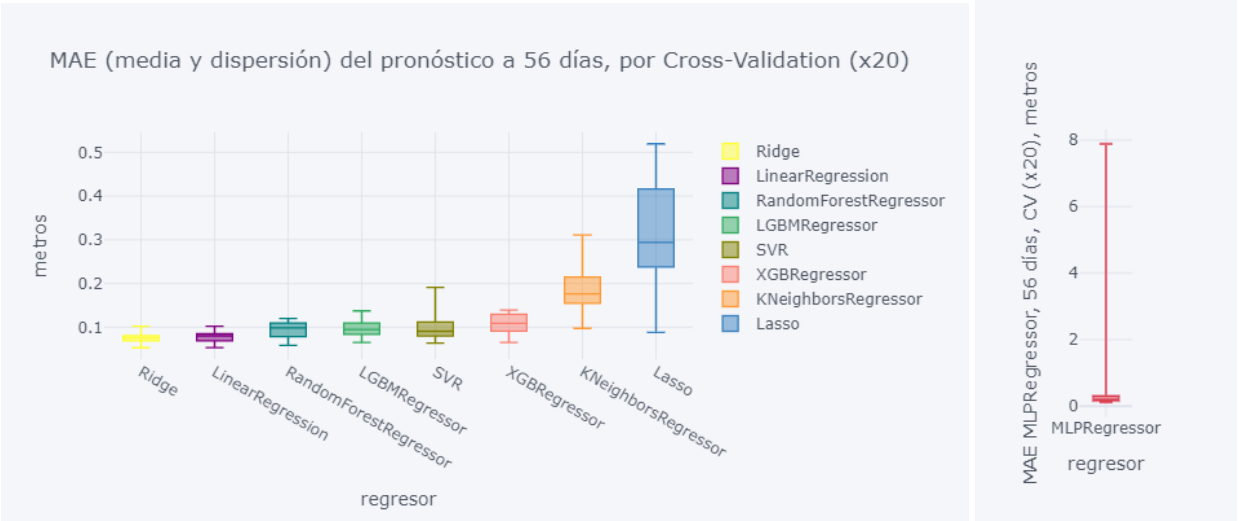

Ilustración 29: Pronóstico a cincuenta y seis días, Cross-Validation x20. Sin optimización.

Es importante tener presente dos observaciones importantes:

1) El modelo construido es capaz de pronosticar el valor de SALTO a un día visto (o al día siguiente). Por eso se utiliza la matriz de predictoras de X_test como input para pronosticar los 'n' días (7 o 56 en este caso) de cada horizonte. Es decir, se hacen 'n' pronósticos a un día, para cada vector (fila) de la matriz X_test.
2) Esto es posible porque se utiliza un conjunto de valores reales del pasado (conjunto de testeo), para pronosticar también valores del pasado. Pero en condiciones normales esto no es posible, porque se tratará de pronosticar valores del futuro y por lo tanto no se contará con el conjunto de predictoras y habrá que generarlo o utilizar otra estrategia de pronóstico.

Se podrían mejorar los resultados del desempeño incluyendo un paso de ingeniería de variables y otro de optimización de hiperparámetros de los algoritmos. Precisamente el anterior y estos dos serán tópicos de los capítulos siguientes.

Pronóstico de valores futuros utilizando algoritmos de regresión, más allá del conjunto completo de datos.

Continuando con el plan de ensayo, en el paso anterior la matriz de predictoras se construyó con versiones retrasadas de SALTO, que constituyen los valores de los últimos siete y/o cincuenta y seis días de SALTO, respectivamente, en cada horizonte de pronóstico.

Como se comentó anteriormente, en la operación real, estos valores no se tendrán disponibles y se deberá bien generarlos, bien pronosticar de otra manera, 'a futuro'.

A tal efecto, se dividirá el conjunto de datos en entrenamiento y verificación, tal cual el paso anterior, y se procederá según dos modalidades a saber:

- Modelo por desplazamiento del vector de X_pred
- Ensamble de modelos de pronósticos individuales con offset, para cada día del horizonte de pronóstico (offset según el día en el horizonte de pronóstico).

En ambas, la serie de tiempo 'y_pred' se construirá adjuntando ('appending') los valores pronosticados para cada día. La evaluación del desempeño se hará contrastando esta 'y_pred' con los valores de SALTO del conjunto de validación 'y_test' (últimos valores del conjunto de datos y que intentamos pronosticar).

Dado el mejor desempeño de los modelos con predictoras de 40 versiones retrasadas ('lags') verificado en el paso anterior, se ensayarán otros conjuntos, tales como 10, 20 y 30 también.

Paso 2: Modelo por desplazamiento diario del vector de predicciones, X_pred

Se inicializa X_pred (vector horizontal de predictoras) con la última fila del conjunto de entrenamiento, desplazada en el tiempo por un día ('shift=1').

Esto libera la casilla correspondiente al valor de SALTO, para completarlo luego, con el valor pronosticado, con los valores de X_pred (desde t=-n hasta t=-1) y continuar repitiendo el proceso para cada día del horizonte de pronóstico.

Nota: En este paso, los algoritmos de regresión se utilizan con sus hiperparámetros 'por defecto', sin ningún tipo de optimización.

La salida del cuaderno de este capítulo es un archivo de datos '.csv' con las métricas de desempeño de cada modelo. Luego, en el cuaderno número 10, se realiza el análisis y ploteo de los desempeños, resultando en este caso, los cuatro mejores para cada horizonte de pronóstico. Se observa que para ambos horizontes de pronóstico, el conjunto de versiones retrasadas con mejor desempeño es de 20. En el cuaderno correspondiente se muestran los resultados completos.

Regresor	lags	H(días)	MAE(m)	RMSE(m)
XGBRegressor	20	7	0.046314	0.052527
LGBMRegressor	20	7	0.056664	0.059468
XGBRegressor	40	7	0.072174	0.085822
MLPRegressor	10	7	0.080116	0.081512
KNeighborsRegressor	20	56	0.088393	0.156881
KNeighborsRegressor	30	56	0.091750	0.169967
KNeighborsRegressor	10	56	0.092250	0.162111
LGBMRegressor	10	56	0.092551	0.167566

Tabla 7: Desempeño de los regresores en pronóstico por desplazamiento a 7 y a 56 días, con 10 - 40 'lags'. Sin optimización.

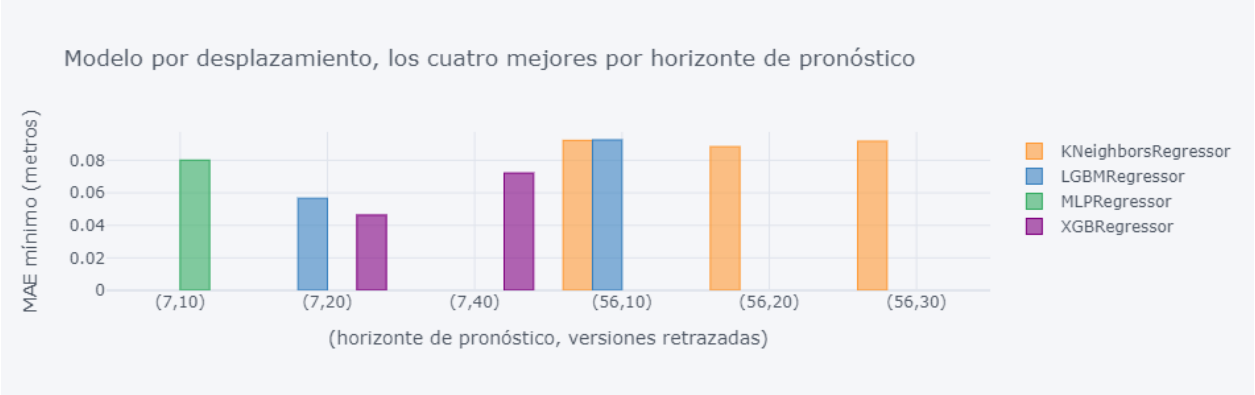

Ilustración 30: Desempeño de los regresores en pronóstico por desplazamiento a 7 y 56 días, con 10 - 40 'lags'. Sin optimización.

A continuación, se presentan las mejores predicciones para cada horizonte:

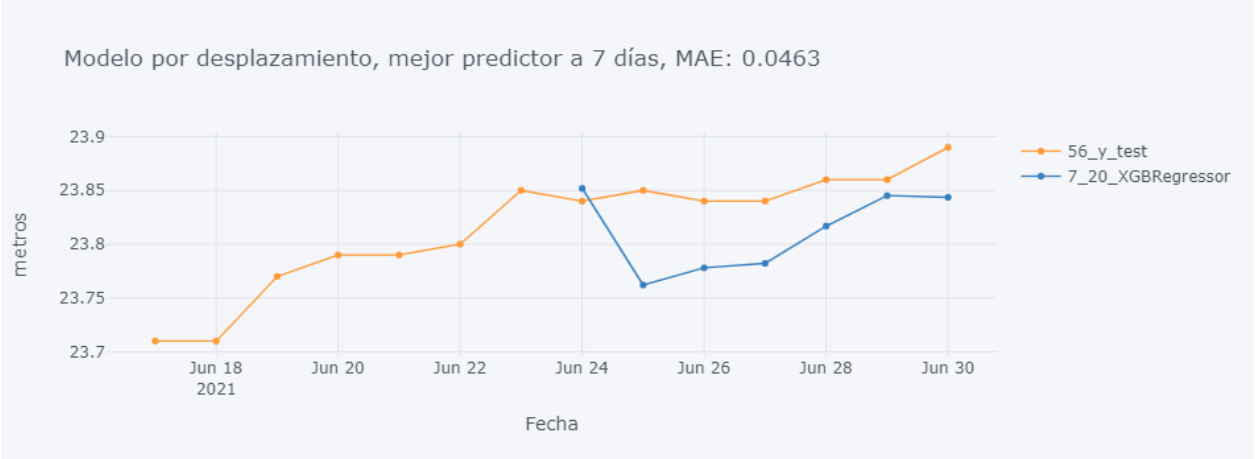

Ilustración 31: Mejor predicción por desplazamiento a siete días. Sin optimización.

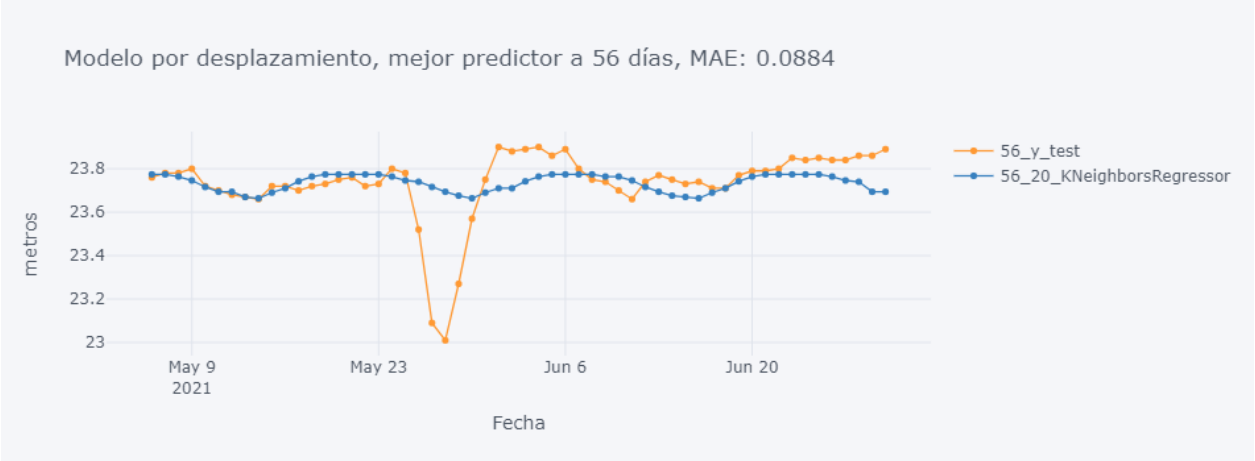

Ilustración 32: Mejor predicción por desplazamiento a cincuenta y seis días. Sin optimización.

Paso 3: Ensamble de modelos individuales para cada día dentro del horizonte de pronóstico: al día siguiente, a dos días, a tres, …, a siete, … y a cincuenta y seis.

Se crea un modelo de pronóstico hacia valores de SALTO desplazados en dirección futura, tomando como 'target' de la regresión, los valores de SALTO, SALTO+1, SALTO+2, …, SALTO+7, …, SALTO+56.

De manera similar al caso anterior, se implementa el modelo en un cuaderno y los datos generados se analizan en el cuaderno 10. Resultan de dicho experimento, los tres mejores desempeños, por horizonte de pronóstico:

Nota: En este paso, los algoritmos de regresión se utilizan con sus hiperparámetros 'por defecto', sin ningún tipo de optimización.

Regresor	lags	H(días)	MAE(m)	RMSE(m)
MLPRegressor	20	7	0.069504	0.075086
XGBRegressor	10	7	0.071549	0.097546
MLPRegressor	10	7	0.075249	0.080368
SVR	40	7	0.084388	0.094725
MLPRegressor	10	56	0.093691	0.164531
MLPRegressor	20	56	0.100005	0.172538
MLPRegressor	30	56	0.100554	0.166794
MLPRegressor	40	56	0.184925	0.221436

Ilustración 33: Desempeño de los regresores en pronóstico por ensamble de modelos diarios, a 7 y a 56 días, con 10 - 40 'lags'. Sin optimización.

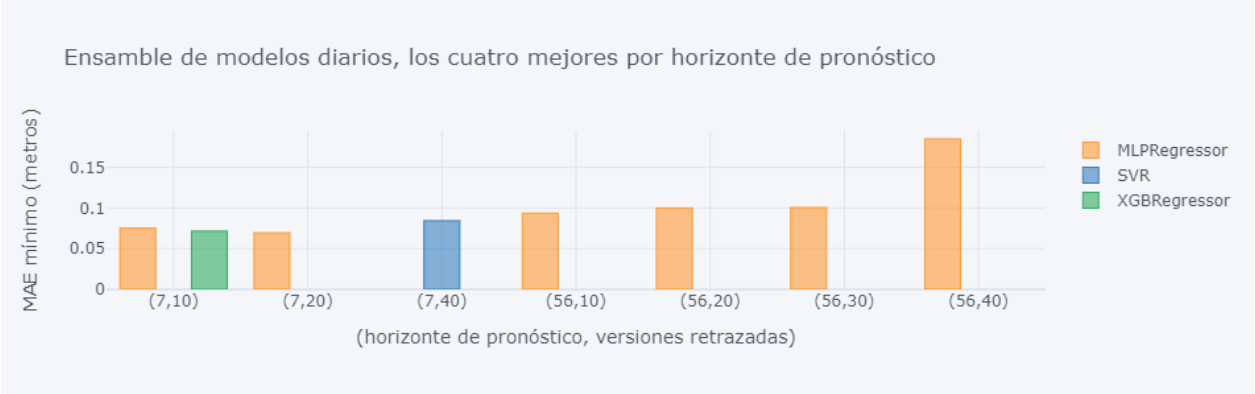

Ilustración 34: Desempeño de los regresores en el pronóstico por ensamble de modelos diarios, a 7 y 56 días, con 10 - 40 'lags'. Sin optimización.

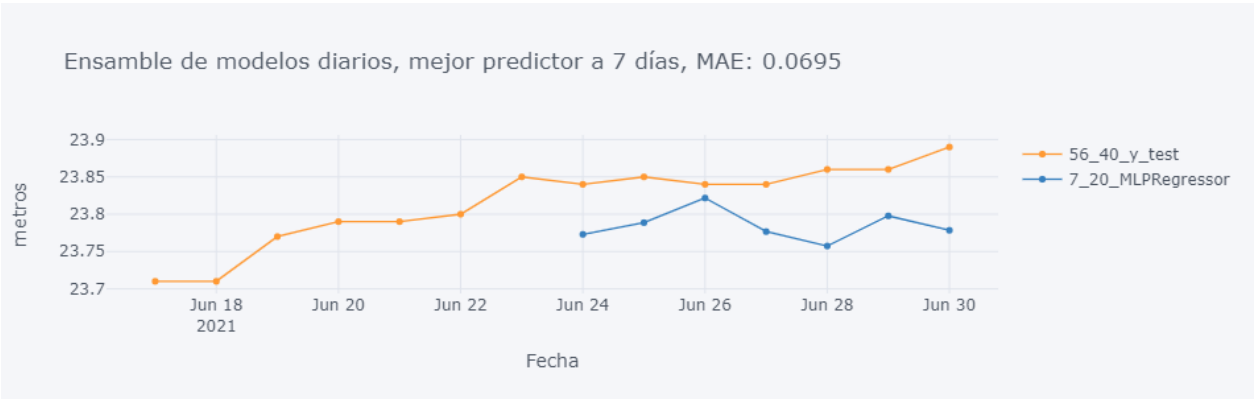

Ilustración 35: Mejor pronóstico por ensamble de modelos diarios, a siete días. Sin optimización.

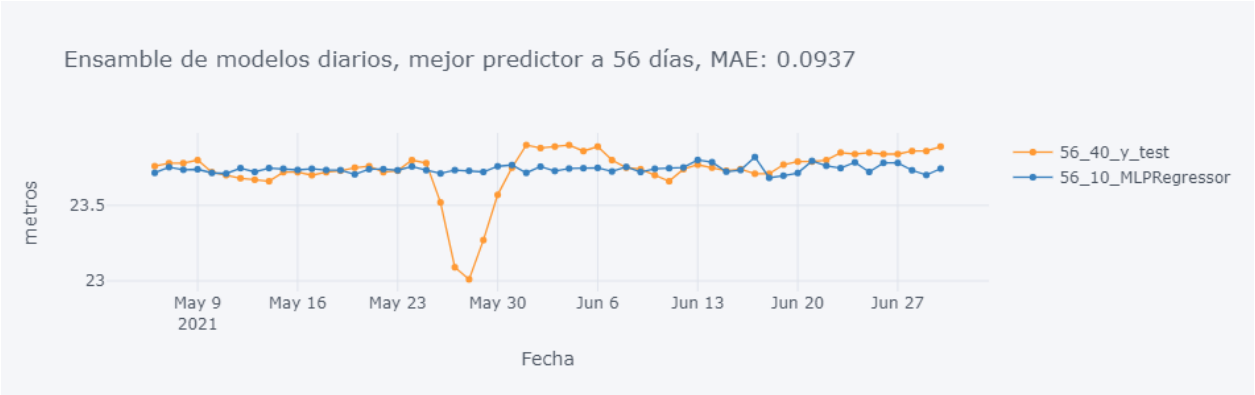

Ilustración 36: Mejor pronóstico por ensamble de modelos diarios, a cincuenta y seis días. Sin optimización.

Paso 4: Ensamble de modelos individuales para cada día dentro del horizonte de pronóstico: al día siguiente, a dos días, a tres, …, a siete, … y a cincuenta y seis, con ingeniería de variables ('feature engineering').

En este paso se agregan a las versiones retrasadas de SALTO ('lags') otras variables calculadas a partir de operaciones aritméticas sobre la columna de la primera versión retrasada, SALTO-1, y de fecha sobre la columna 'FECHA'.

Entre las operaciones aritméticas sobre 'SALTO-1' se encuentran medias, varianzas, máximos, mínimos, etc. rodantes, en grupos tomados de a 3, de a 5 y de a 7 elementos y entre las operaciones sobre la columna 'fecha', se aplican la extracción del número de día en el año, del número de semana, del número de mes y algunas otras más.

Nota: En este paso, los algoritmos de regresión se utilizan con sus hiperparámetros 'por defecto', sin ningún tipo de optimización.

El proceso de aprendizaje supervisado aplicado es similar al del paso anterior, resultando:

Regresor	lags	H(días)	MAE(m)	RMSE(m)
KNeighborsRegressor	10	7	0.082286	0.083980
KNeighborsRegressor	20	7	0.082286	0.083980
KNeighborsRegressor	30	7	0.082286	0.083980
KNeighborsRegressor	40	7	0.082286	0.083980
KNeighborsRegressor	10	56	0.093107	0.169060
KNeighborsRegressor	20	56	0.093107	0.169060
KNeighborsRegressor	40	56	0.093107	0.169060
KNeighborsRegressor	30	56	0.093107	0.169060

Ilustración 37:Desempeño de los regresores en pronóstico con emsambles diarios con 'feature-engineering'. Sin optimización.

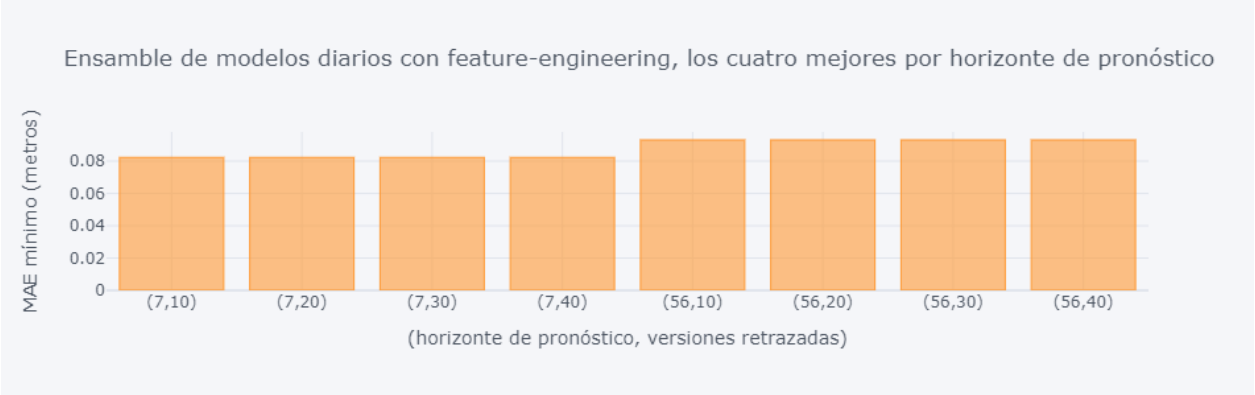

Ilustración 38: Desempeño de los regresores en el pronóstico por ensamble de modelos diarios, con feature-engineering, a 7 y 56 días, con 10 - 40 'lags, 'Algoritmo KNeighborsRegressor (en todos los casos) . Sin optimización.

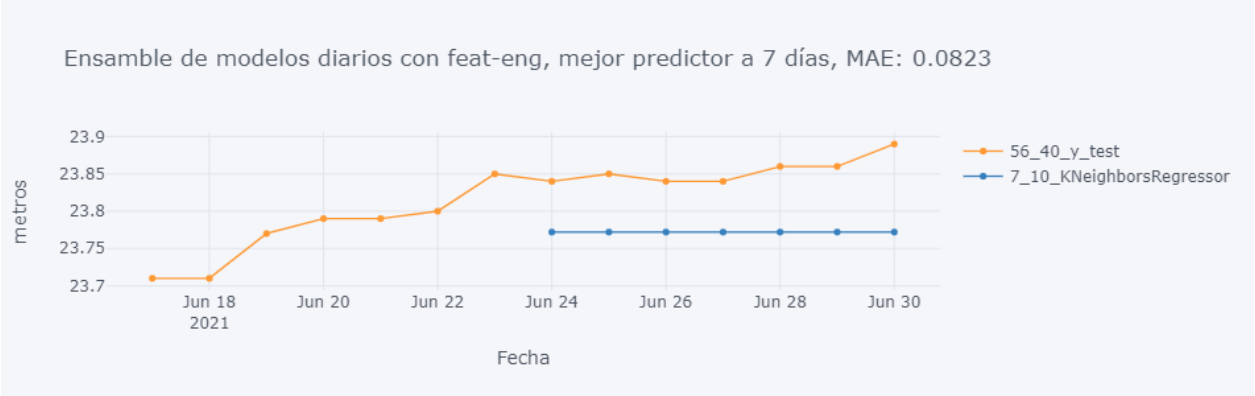

Ilustración 39: Ilustración 35: Mejor pronóstico por ensamble de modelos diarios con feat-eng,, a siete días. Sin optimización.

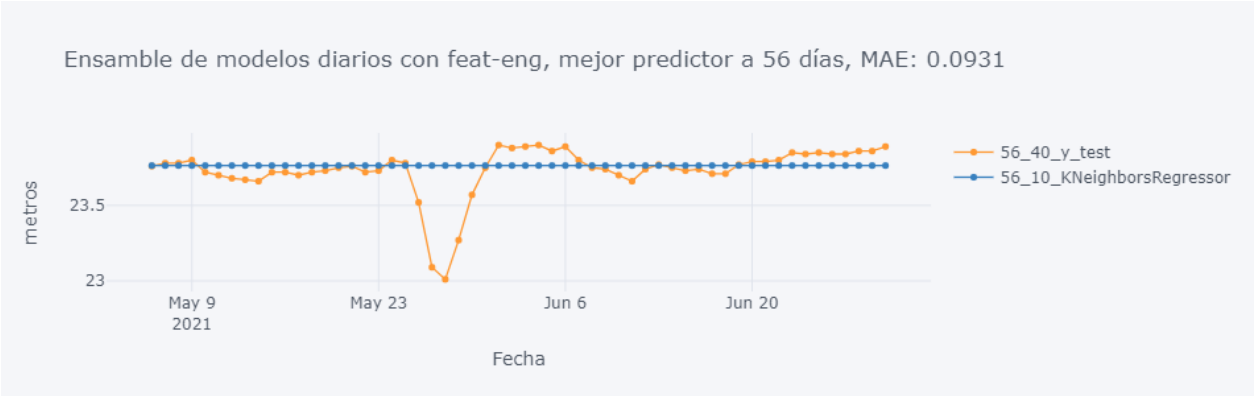

Ilustración 40: Ilustración 35: Mejor pronóstico por ensamble de modelos diarios con feat-eng., a cincuenta y seis días. Sin optimización.

Paso 5: Pronóstico de valores futuros utilizando algoritmos de regresión, en modo multivariable, más allá del conjunto completo de datos.

En este capítulo se agregarán a las versiones retrasadas ('lags') de 'SALTO', las variables restantes del conjunto de datos de la CHY, llamadas 'variables exógenas'.

Antes de hacerlo y con el fin de maximizar la contribución que cada variable puede aportar, se realizará un test de Causalidad Granger (Causalidad de Granger - Wikipedia, la enciclopedia libre y Using Granger Causality Test to Know If One Time Series Is Impacting in Predicting Another? | by Feroz Kazi | The Startup | Medium), que indicará que versión en el tiempo utilizar para cada una.

A continuación, el resultado del test de Granger sobre cada una de las variables exógenas, indicando el orden de retraso óptimo, para incorporar al conjunto de predictoras de versiones retrasadas:

Nota: En este paso, los algoritmos de regresión se utilizan con sus hiperparámetros 'por defecto', sin ningún tipo de optimización.

Variable	Versión retrasada
COTA_EMBALSE	1
COTA_RESTITUCION	1
QAFLR	5
QVBP	1
QVBAC	1
QTURB	2
SINP	1
SADI	2

Tabla 8: Versión retrasada a usar de cada variable exógena. Sin optimización.

A continuación, los cuatro mejores desempeños para cada horizonte:

Regresor	Lags	H(días)	MAE(m)	RMSE(m)
Ridge	40	7	0.125003	0.129043
LinearRegression	40	7	0.126541	0.130551
Ridge	30	7	0.137421	0.141339
LinearRegression	30	7	0.138863	0.142784
LGBMRegressor	40	56	0.220546	0.257833
LGBMRegressor	30	56	0.271625	0.300049
RandomForestRegressor	40	56	0.342159	0.373567
LGBMRegressor	20	56	0.347497	0.383712

Tabla 9: Desempeño de los regresores en pronóstico con variables exógenas, a 7 y a 56 días, con 10 - 40 'lags'. Sin optimización.

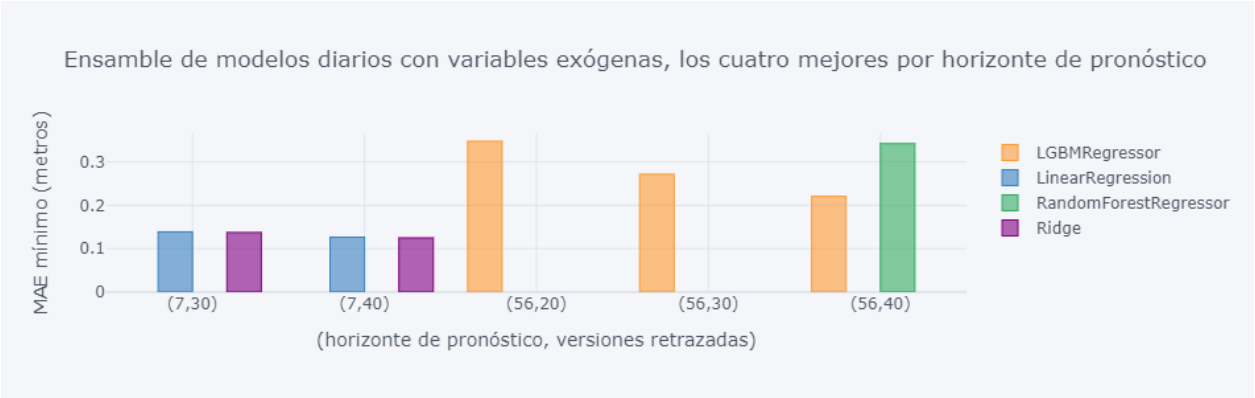

Ilustración 41: Desempeño de los regresores en el pronóstico por ensamble de modelos diarios, a 7 y a 56 días, con 10 - 40 'lags' y variables exógenas. Sin optimización.

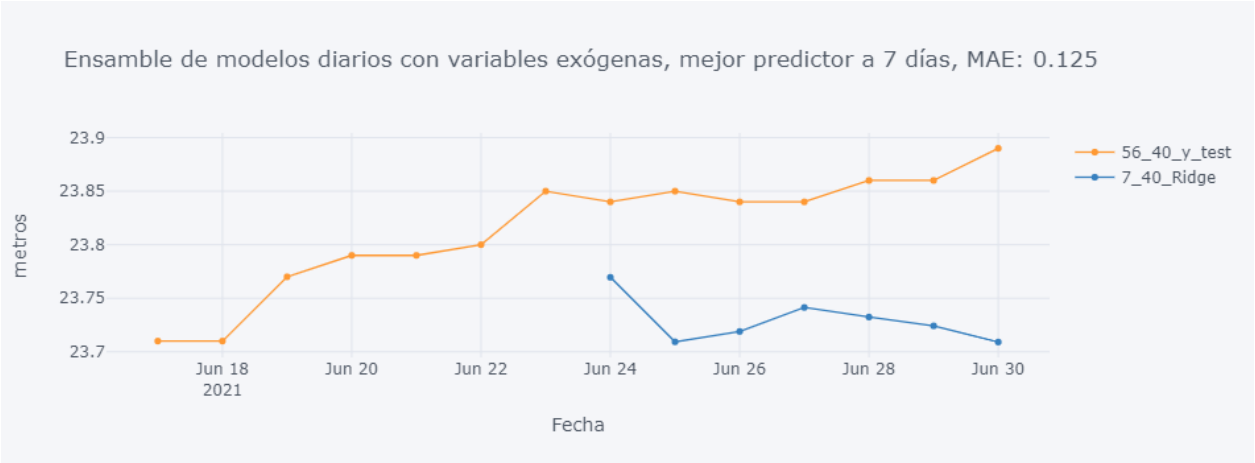

Ilustración 42: Mejor pronóstico por ensamble de modelos diarios con variables exógenas, a siete días. Sin optimización.

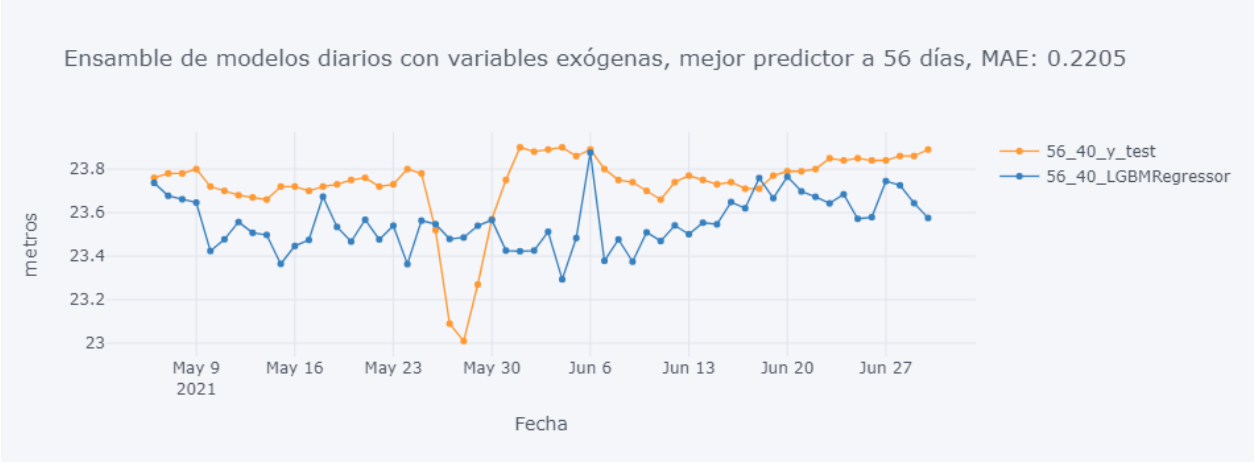

Ilustración 43: Mejor pronóstico por ensamble de modelos diarios con variables exógenas, a cincuenta y seis días. Sin optimización.

Paso 6: Evaluación y comentarios a los resultados

El mejor desempeño para el pronóstico a siete días se logra con un modelo de desplazamiento, con veinte (20) lags de predicción y un algoritmo XGBRegressor, y el mejor desempeño para el pronóstico a cincuenta y seis días, también se logra con un modelo de desplazamiento con veinte variables, pero con un regresor KNeighborsRegressor.

Nota: En todos los casos se utilizó los regresores sin optimización de sus hiperparámetros.

Regresor	lags	H(días)	MAE(m)	RMSE(m)	Modelo
XGBRegressor	20	7	0.046314	0.052527	Desplazamiento
MLPRegressor	20	7	0.069504	0.075086	Ensamble
KNeighborsRegressor	40	7	0.082286	0.083980	Ensamble+featEng
Ridge	40	7	0.125003	0.129043	Ensamble+exógenas
KNeighborsRegressor	20	56	0.088393	0.156881	Desplazamiento
KNeighborsRegressor	40	56	0.093107	0.169060	Ensamble+featEng
MLPRegressor	10	56	0.093691	0.164531	Ensamble
LGBMRegressor	40	56	0.220546	0.257833	Ensamble+exógenas

Tabla 10: Desempeño de los mejores modelos en los pronósticos a siete y a cincuenta y seis días. Sin optimización.

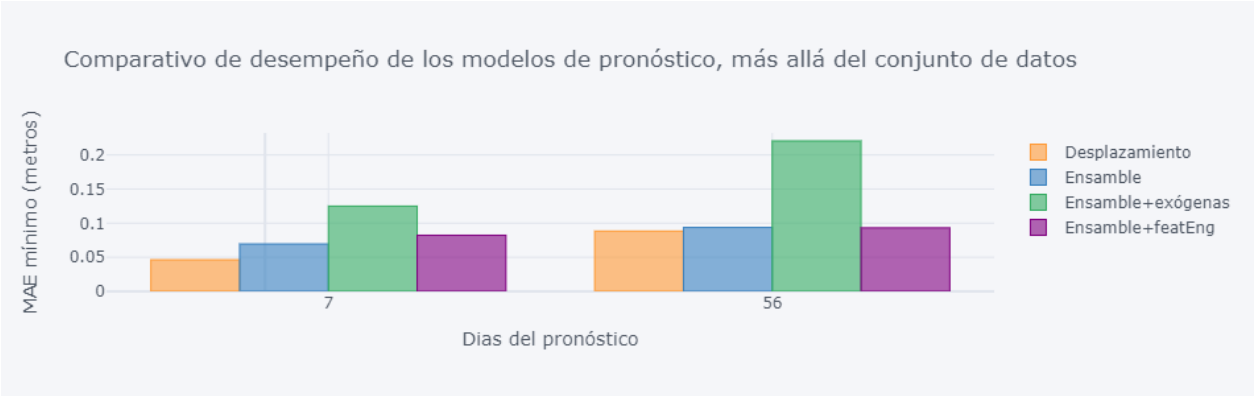

Ilustración 44: Desempeño de los mejores modelos en los pronósticos a siete y a cincuenta y seis días. Sin optimización.

En el próximo capítulo se trabajará en la optimización de los parámetros de los algoritmos utilizados en los modelos de pronóstico a siete y a cincuenta y seis días, que utilizan la modalidad de desplazamiento del vector de predicciones y el algoritmo de regresión XGBRegressor y KNeighborsRegressor, con un conjunto de veinte versiones retrasadas de SALTO ('lags').

Modelo	Horizonte(días)	MAE(m)	RSME(m)
AutoARIMA	7	0.0351	0.0415
	56	0.1048	0.1782
Pronóstico Ingenuo 'drift'	7	0.0123	0.0161
	56	0.0967	0.1755
Aprendizaje Automático	7	0.0463	0.0525
	56	0.0883	0.1568

Tabla 11: Comparación de desempeño, entre modelos de Pronóstico Ingenuo, AutoARIMA y Aprendizaje Automático. Sin optimización.

Paso 7: Optimización de los mejores modelos para cada horizonte de pronóstico

Según se vio en el apartado anterior, el modelo que mejor se desempeñó en el pronóstico a siete días fue el de desplazamiento diario de los valores de 'SALTO', utilizando un regresor Xtreme Gradient Boosting o XGBRegressor, con una matriz de predictoras compuesta por veinte (20) versiones retrasadas de SALTO.

Y, para el pronóstico a cincuenta y seis días, el modelo que mejor se desempeñó fue el de desplazamiento diario de los valores de 'SALTO', utilizando un regresor K Nearest Neighbors (KNeighborsRegressor), también con una matriz de predictoras compuesta por veinte (20) versiones retrasadas de SALTO.

Utilizando de la librería 'scikit-optimize'[21], la función 'gp_minimize' se realizó una optimización Bayesiana de ambos modelos, como sigue:

Pronóstico a siete días:

Par el caso del XGBRegressor, los hiperparámetros 'max_depth', 'learning_rate', 'n_estimators' y 'colsample_bytree', resultando una mejora apreciable en el MAE a siete días, con 20 versiones retrasadas ('lags') de SALTO.

Puede verse también la evolución del mínimo obtenido con las iteraciones de la optimización.

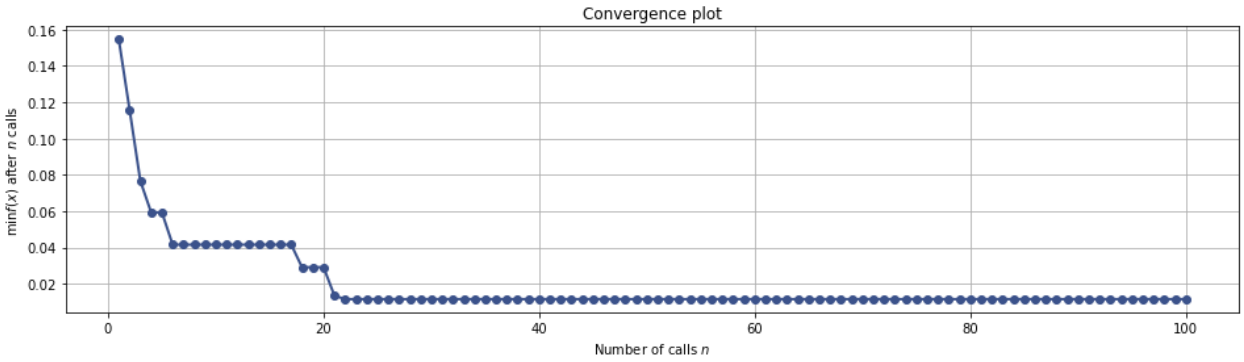

Ilustración 45: Evolución del mínimo encontrado por la función de gp_minimize() de optimización Bayesiana, para XGBoost

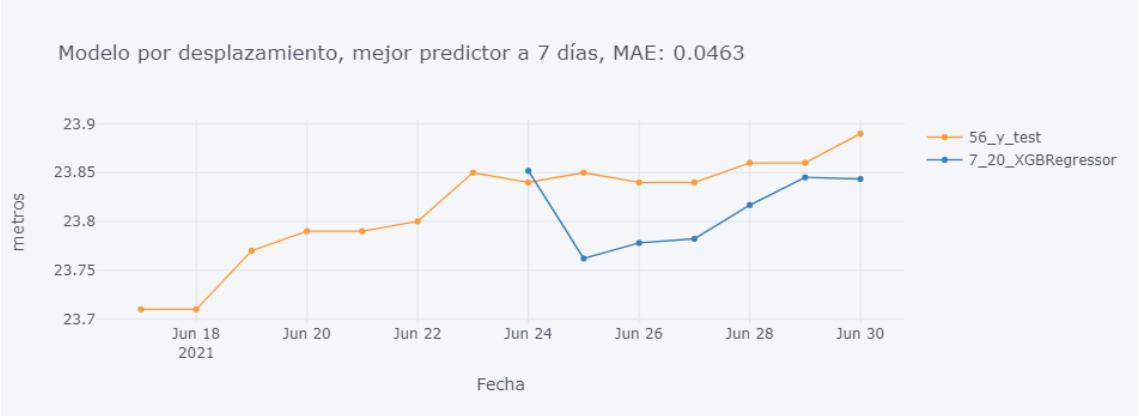

Ilustración 46: Mejor predicción por desplazamiento a siete días (repetición). Con optimización.

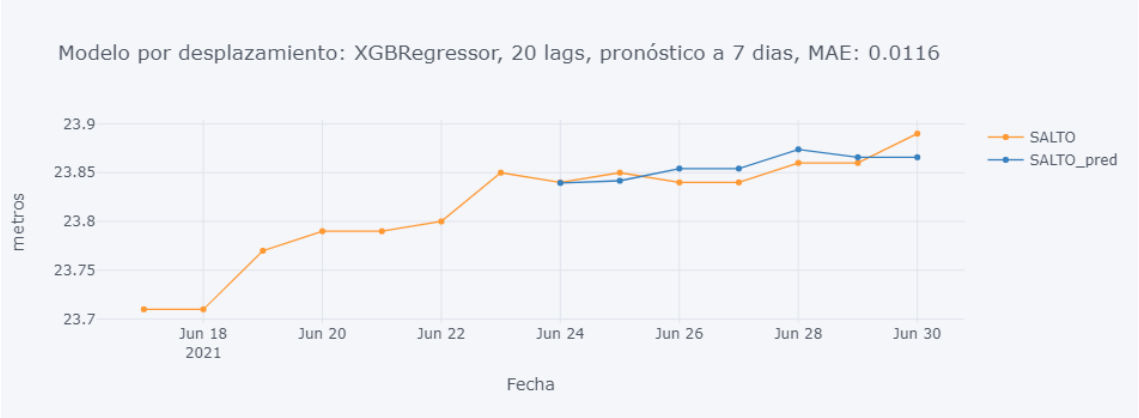

Ilustración 47: Mejor predicción por desplazamiento a siete días, Optimizado. Con optimización.

Hiperparámetros	Default	Optimizado
max_depth	6	1
learning_rate	0.300000012	0.33960893653732727
n_estimators	100	400
Colsample_bytree	1	1

Tabla 12: Tabla comparativa de parámetros XGBoost por defecto y optimizados

Pronóstico a cincuenta y seis días:

Par el caso del KNeighborsRegressor, los hiperparámetros 'n_neighbors', 'p', 'weights', 'algorithm' y 'metric', resultando en una pequeña mejora en el MAE a cincuenta y seis días, con 20 versiones retrasadas ('lags') de SALTO.

Puede verse también la evolución del mínimo obtenido con las iteraciones de la optimización.

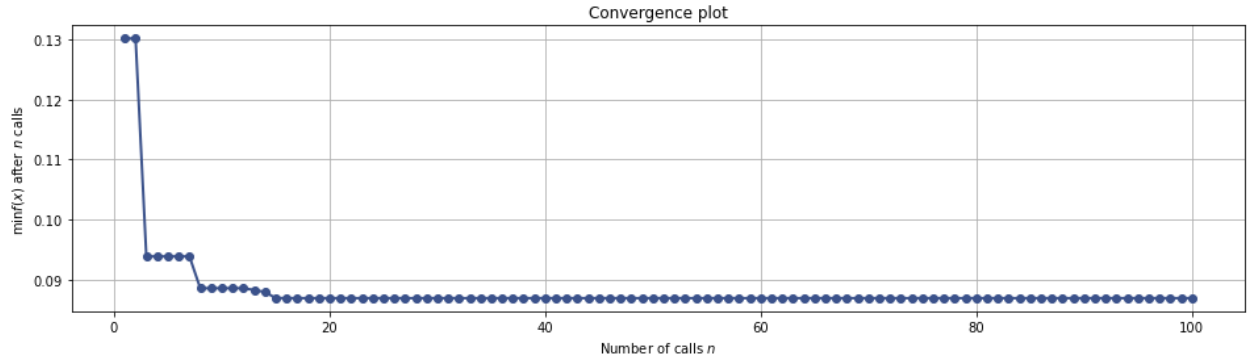

Ilustración 48: Evolución del mínimo encontrado por la función de gp_minimize() de optimización Bayesiana, para KNN

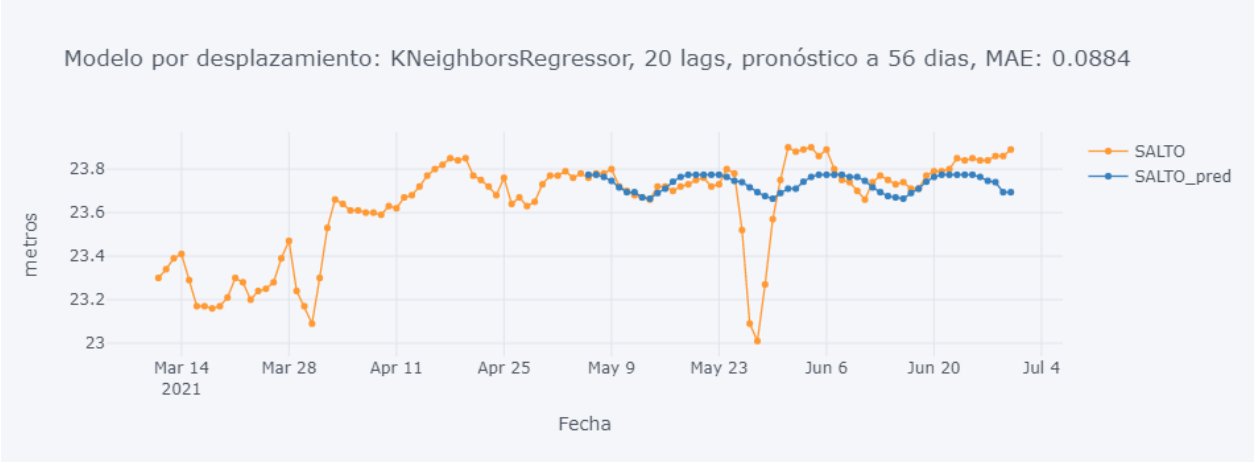

Ilustración 49: Mejor predicción por desplazamiento a cincuenta y seis días (repetición) . Sin optimización..

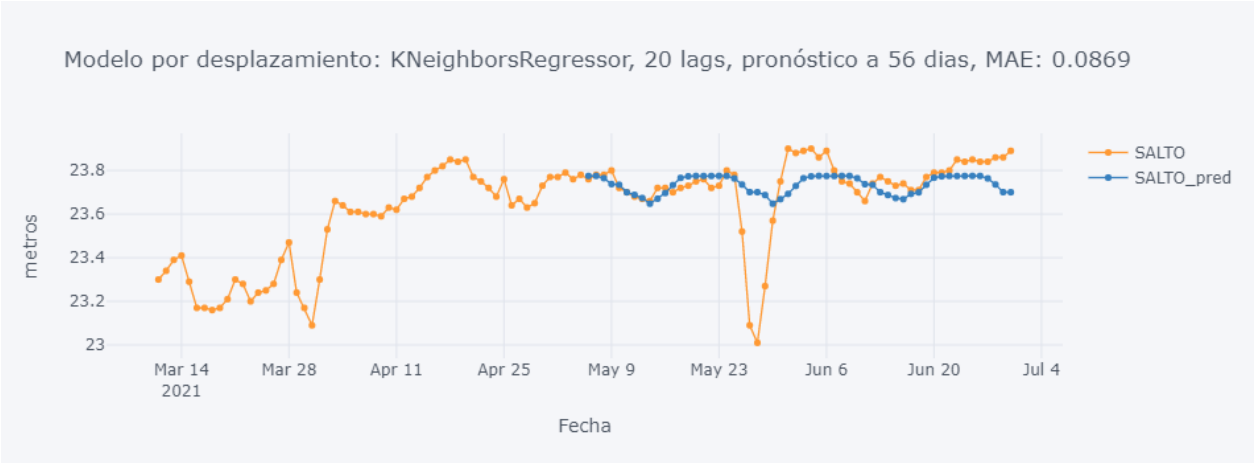

Ilustración 50: Mejor predicción por desplazamiento a cincuenta y seis días, Optimizado

Hiperparámetros	Default	Optimizado
n_neighbors	5	4
p	2	3
weights	'uniform'	'distance'
algorithm	'auto'	'brute'
metric	'minkowski'	'euclidean'

Tabla 13: Tabla comparativa de parámetros KNN por defecto y optimizados

Conclusión

Crisp-DM - Etapa 5: Evaluación Conclusiones del trabajo

Considerando que la exactitud promedio del sistema de pronóstico a siete días actualmente en uso es de 12 cm (0.12 m) y visto que los resultados obtenidos con el modelo por desplazamiento del vector de predictoras, de veinte versiones retrasadas de SALTO, que utiliza un algoritmo de regresión 'XGBoost' es de 0.016 m (1.6 cm) se demuestra la hipótesis del presente TFI, incluyendo su extensión práctica.

Capítulo V: Resumen, resultados e implicaciones

Introducción

Se hizo un análisis exploratorio de los datos de la central, para familiarizarse con sus roles en el proceso de la generación de la energía y para entender su impacto en el negocio.

Se verificó que la serie SALTO (el salto hidráulico de la Central Hidroeléctrica Yacyretá) no es un 'ruido blanco' ni se comporta como un 'paseo aleatorio'. Cualquiera de estas dos condiciones, de haberse comprobado, hubieran invalidado el estudio, dado que su implicancia es que no es posible pronosticar valores futuros por ninguno de los métodos utilizados.

Se estableció un piso de desempeño utilizando 'pronósticos ingenuos'. Cualquier método que se ensayare a continuación se consideraría de interés si su desempeño fuera mejor que el de los pronósticos ingenuos.

Se realizaron pronósticos mediante métodos econométricos, tales como ARIMA y para terminar se ensayaron varios métodos y algoritmos de aprendizaje automático supervisado.

Con todos y cada uno de los ensayos se trabajó con dos horizontes de pronóstico, uno a corto plazo, siete días (una semana), sindicado como operativo o del día a día' y otro de cincuenta y seis días (ocho semanas), de mediano plazo, con un eventual uso para la predicción de crecidas o bajantes significativas, que puedan afectar la generación de energía o la eficiencia en la utilización de la masa de agua embalsada.

Resultados

El resumen de los mejores desempeños puede verse en la siguiente tabla:

Modelo	Horizonte(días)	MAE(m)	RSME(m)
HEC-HMS	7	0.120	N/A
	56	N/A	N/A
Pronóstico Ingenuo 'drift'	7	0.0123	0.0161
	56	0.0967	0.1755
AutoARIMA	7	0.0351	0.0415
	56	0.1048	0.1782
Aprendizaje Automático	7	0.0463	0.0525
	56	0.0883	0.1568
Aprendizaje Automático con Optimización	7	0.0115	0.0135
	56	0.0869	0.1556

Tabla 14: Comparación de desempeño, entre el actual y los mejores modelos desarrollados

Implicancias

A la luz de los resultados y considerando que la exactitud promedio del cálculo actual, por métodos hidráulicos e hidrológicos es del orden de 0.120m, podría esperarse una mejora en la gestión de la CHY si se implementara el mejor modelo de pronóstico desarrollado en este trabajo.

Podría administrarse mejor los caudales de agua afluentes, en el día a día (horizonte de pronóstico a siete días) y hasta detectar eventuales crecidas y o bajantes con tiempo suficiente para tomar los recaudos correspondientes para mantener el máximo aprovechamiento del recurso (horizonte de pronóstico a cincuenta y seis días).

Sugerencias para futuras investigaciones

Ensayar otros algoritmos de regresión tales como las redes neuronales recurrentes, que no han sido contempladas en el presente trabajo.

Referencias-Bibliografía

(1) Marinósdóttir, H. (2019). Applications of different machine learning methods for water level predictions [Unpublished Master of Science (M.Sc.) Thesis in Management Engineering]. Reykjavík University.

(2) Irving, K., Kuemmerlen, M., Kiesel, J., Kakouei, K., Domisch, S. y Jähnig, S. (2018). A high-resolution streamflow and hydrological metrics dataset for ecological modeling using a regression model. Scientific Data, *Vol.*(5), § Background & Summary. doi: 10.1038/sdata.2018.224

(3) Valizadeh, N., El-Shafie, A., Mirzaei, M., Galavi, H., Mukhlisin, M., Jaafar, O. (2014). Accuracy Enhancement for Forecasting Water Levels of Reservoirs and River Streams Using a Multiple-Input-Pattern Fuzzification Approach. Hindawi, The Scientific World Journal, *Vol.*(2014), § Abstract. https://doi.org/10.1155/2014/432976

(4) BAE, D., JEONG, D. y KIM, G. (2007). Monthly dam inflow forecasts using weather forecasting information and neuro-fuzzy technique. Hydrological Sciences Journal, *Vol.*(52), § Abstract https://doi.org/10.1623/hysj.52.1.99

(5) Castillo-Botón, C., Casillas-Pérez, D., Casanova-Mateo L., Moreno-Saavedra, L., Morales-Díaz, B., Sanz-Justo, J., Gutiérrez, P. y Salcedo-Sanz, S. (2020). Analysis and Prediction of Dammed Water Level in a Hydropower Reservoir Using Machine Learning and Persistence-Based Techniques.MDPI *Water Journal, Vol.*(12), § Abstract. https://doi.org/10.3390/w12061528

(6) Yang, J., Cheng, C. y Chan, C. (2017). A Time-Series Water Level Forecasting Model Based on Imputation and Variable Selection Method. *Hindawi Computational Intelligence and Neuroscience, Vol.*(2017), § Abstrack. https://doi.org/10.1155/2017/8734214

(7) S/A (23 de septiembre de 2018). *Desde Yacyretá anunciaron aumento de caudal desde el sábado pasado*. Reporte Misiones. https://reportemisiones.com.ar/desde-yacyreta-anunciaron-aumento-de-caudal-desde-el-sabado-pasado/

(8) S/A (S/F). *6 Major Phases in CRISP-DM: The Standard Data Mining Process*. Pro Global Business Solutions. https://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process/

(9) Patidar, S. (31 de enero de 2019).*5 Steps to Learn Python for Data Science*. LaptrinhX. https://laptrinhx.com/5-steps-to-learn-python-for-data-science-2237653156/

(10) konivatsara y probe_international. Three Gorges Dam Water Data, Inflow Outflow Rates and Water Levels, 2011-2020. [threegorges-water-storage.csv], https://www.kaggle.com/konivat/three-gorges-dam-water-data.

(11) Hydrologic Engineering Center, 609 Second Street, Davis, CA 95616-4687, USA, Hydrologic Engineering Center (army.mil)

(12) Sergio Fattorelli, Pedro C. Fernandez. (2011). Diseño Hidrológico. WASA-GN (Water Assesstment & Advisory Global Network, ISBN: 978-987-05-2738-2. (Diseño Hidrológico - Sergio Fatorelli - pdf Docer.com.ar).

(13) Represa de Yacyretá. (2022, 18 de mayo). Wikipedia, La enciclopedia libre. Fecha de consulta: 11:55, junio 12, 2022 desde: Represa de Yacyretá - Wikipedia, la enciclopedia libre

(14) Consorcio Harza y Asociados. (diciembre 1973). Estudio de Factibilidad Técnico-Económico-Financiero del Aprovechamiento del Río Paraná a la Altura de las Islas Yacyretá y Apipé.

(15) Central hidroeléctrica de pasada. (2020, 10 de diciembre). Wikipedia, La enciclopedia libre. Fecha de consulta: 11:52, junio 12, 2022 desde: https://es.wikipedia.org/wiki/Central_hidroel%C3%A9ctrica_de_pasada

(16) Portal de Unificado de Información Pública. Gobierno de la República del Paraguay Fecha de consulta: 10 de agosto, 2011 desde https://informacionpublica.paraguay.gov.py/portal/

(16) Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830 desde scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation

(17) McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56) desde pandas documentation — pandas 1.4.2 documentation (pydata.org).

(18) Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, Franz Király (2019): "sktime: A Unified Interface for Machine Learning with Time Series", desde Welcome to sktime — sktime documentation

(19) Markus Löning, Tony Bagnall, Sajaysurya Ganesh, George Oastler, Jason Lines, ViktorKaz, …, Aadesh Deshmukh (2020). alan-turing-institute/sktime. Zenodo. http://doi.org/10.5281/zenodo.3749000

(20) Jorge Santos, Cufflinks, 'This library binds the power of 'plotly' with the flexibility of 'pandas' for easy plotting.'dede https://github.com/santosjorge/cufflinks

(21) Tim Head; MechCoder; Gilles Louppe; Iaroslav Shcherbatyi; fcharras; Zé Vinícius; cmmalone; Christopher Schröder; nel215; Nuno Campos; Todd Young; Stefano Cereda; Thomas Fan; rene-rex; Kejia (KJ) Shi; Justus Schwabedal; carlosdanielcsantos; Hvass-Labs; Mikhail Pak; SoManyUsernamesTaken; Fred Callaway; Loïc Estève; Lilian Besson; Mehdi Cherti; Karlson Pfannschmidt; Fabian Linzberger; Christophe Cauet; Anna Gut; Andreas Mueller; Alexander Fabisch. scikit-optimize, Sequential model-based optimization in Python , https://zenodo.org/record/1207017 desde

(22) Sean J. Taylor & Benjamin Lethan. (2018). Forecasting at scale. The American Statistitian. Forecasting at Scale: The American Statistician: Vol 72, No 1 (tandfonline.com)

(23) Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, Franz Király (2019): "sktime: A Unified Interface for Machine Learning with Time Series", desde NaiveForecaster — sktime documentation.

Anexos

Anexo – 1: Análisis exploratorio de los Datos (EDA)