UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

Predicción de químicos en reactores biológicos

TRABAJO FIN DE MÁSTER

Master en Inteligencia Artificial, Recnocimiento de Formas e Imagen Digital

Autor: Alejandro Bañuls Ordóñez
Tutor: Roberto Paredes Palacios
Tutor Externo: Jorge Mansanet Sandin

Curso 2020-2021

Resum

L'objectiu d'este TFM consistix a predir huit químics de quatre reactors biològics diferents i independents que es troben en una Estació Depuradora d'Aigües Residuals (EDAR) . Tot això resulta en un total de trenta-dos variables a predir, a més de nou variables addicionals que són comuns als quatre reactors. Pel que el total de variables a predir són quaranta-un. Així mateix, es tindran en compte cinquanta variables més que es podran usar com a input per a ajudar-nos a realitzar esta predicció. De cada variables hi ha una sèrie temporal, és a dir, una col·lecció d'observacions de la dita variable que han sigut preses de forma seqüencial i ordenades en el temps. S'avaluaran diferents models d'aprenentatge automàtic com és el cas de Prophet, xarxes neuronals recurrents (LSTM) i XGBoost. Per a cada un d'ells s'analitzaran els resultats obtinguts per cada variable utilitzant distints paràmetres i es realitzarà una comparació dels tres usant el millor en cada cas. Caldrà realitzar diverses operacions de normalitzat a les dades degut a la naturalesa dels mateixos (alt nivell de nuls, granularidades distintes, etc) . Amb tot això, es pretén construir un sistema que siga capaç de predir quin valor tindran eixes variables descrites anteriorment tenint en compte l'històric anterior.

Paraules clau: xarxes recurrents, sèries temporals, predicció, machine learning, deep learning

Resumen

El objetivo de este TFM consiste en predecir ocho químicos de cuatro reactores biológicos diferentes e independientes que se encuentran en una Estación Depuradora de Aguas Residuales (EDAR). Todo ello resulta en un total de treinta y dos variables a predecir, además de nueve variables adicionales que son comunes a los cuatro reactores. Por lo que el total de variables a predecir son cuarenta y uno. Asimismo, se tendrán en cuenta cincuenta variables más que se podrán usar como input para ayudarnos a realizar esta predicción. De cada variables existe una serie temporal, es decir, una colección de observaciones de dicha variable que han sido tomadas de forma secuencial y ordenadas en el tiempo. Se evaluarán diferentes modelos de aprendizaje automático como es el caso de Prophet, redes neuronales recurrentes (LSTM) y XGBoost. Para cada uno de ellos se analizarán los resultados obtenidos por cada variable utilizando distintos parámetros y se realizará una comparación de los tres usando el mejor en cada caso. Habrá que realizar diversas operaciones de normalizado a los datos debido a la naturaleza de los mismos (alto nivel de nulos, granularidades distintas, etc). Con todo ello, se pretende construir un sistema que sea capaz de predecir qué valor tendrán esas variables descritas anteriormente teniendo en cuenta el histórico anterior.

Palabras clave: redes recurrentes, series temporales, predicción, machine learning, deep learning

Abstract

The aim of this TFM is to predict eight chemicals from four different and independent biological reactors located in a Wastewater Treatment Plant (WWTP). This results in a total of thirty-two variables to be predicted, plus nine additional variables that are common to all four reactors. Therefore, the total number of variables to be predicted is forty-one. In addition, fifty more variables will be taken into account that can be used as input to help us make this prediction. For each variable there is a time series, i.e. a collection of observations of that variable that have been taken sequentially and ordered

in time. Different machine learning models will be evaluated, such as Prophet, recurrent neural networks (LSTM) and XGBoost. For each of them, the results obtained for each variable will be analysed using different parameters and a comparison of the three will be made using the best one in each case. Various normalisation operations will have to be performed on the data due to the nature of the data (high level of nulls, different granularities, etc.). The aim is to build a system that is capable of predicting the value of the variables described above, taking into account the previous history.

Key words: recurrent networks, time series, prediction, machine learning, deep learning

Índice general

Índice general v
Índice de figuras vii
Índice de tablas vii

1 Introducción 1
2 Objetivos 3
3 Marco teórico 5
4 Metodología 9
4.1 Prophet . . . 9
4.2 Keras . . . 9
4.3 XGBoost . . . 10
4.4 MLFlow . . . 10
5 Desarrollo 11
5.1 Datos . . . 11
5.2 Experimentación . . . 17
5.2.1 Prophet . . . 17
5.2.2 LSTM . . . 21
5.2.3 XGBoost . . . 27
5.3 Comparación de modelos . . . 34
6 Conclusiones 37
Bibliografía 39

Índice de figuras

5.1 Ejemplo de variable con cambio brusco en el histórico 15
5.2 Ejemplo de predicciones de Prophet 20
5.3 Predicciones con LSTM de dos de las variables de RTC ON 23
5.4 Ejemplo de predicciones de LSTM usando 30 días para predecir 25
5.5 Predicciones de SSLM Air 4 usando 30 (arriba) y 60 (abajo) días. 27
5.6 Ejemplo de predicciones con R^2 a 0 29
5.7 Predicciones de la mejor y la peor variable usando 30 días 29
5.8 Predicciones de la mejor y la peor variable usando 60 días 31
5.9 Comparación R^2 para SSVLM Air 3 usando 30 (arriba) y 60 (abajo) días . . 33
5.10 Comparación R^2 para SSVLM Air 2 usando 30 (arriba) y 60 (abajo) días . . 34
5.11 Comparación de predicciones entre Prophet y LSTM 36

Índice de tablas

5.1 Variables medidas mediante sensor 12
5.2 Variables medidas en el laboratorio 13
5.3 Porcentaje de nulos variables 14
5.4 Máximos y mínimos posibles de cada variable 16
5.5 Resultados obtenidos en la experimentación de Prophet 19
5.6 Resultados obtenidos en la experimentación de LSTM usando 30 días para predecir 22
5.7 Resultados obtenidos en la experimentación de LSTM usando 60 días para predecir 24
5.8 Comparación R^2 para predicciones usando 30 y 60 días 26
5.9 Resultados obtenidos en la experimentación de XGBoost usando 30 días para predecir 28
5.10 Resultados obtenidos en la experimentación de XGBoost usando 60 días para predecir 31
5.11 Comparación R^2 para predicciones usando 30 y 60 días 32
5.12 Comparación R^2 entre el modelo Prophet y LSTM 35

CAPÍTULO 1
Introducción

Una EDAR es una Estación Depuradora de Aguas Residuales en la que se pueden recoger y tratar aguas residuales e industriales. Es decir, una EDAR sirve para eliminar desperdicios, grasas y aceites flotantes, arenas y todos los elementos gruesos que pueda contener el agua; eliminar materiales decantables, tanto orgánicos como inorgánicos; y eliminar materia orgánica biodegradable disuelta en agua, etc.

En las estaciones depuradoras de aguas residuales se trabaja en cuatro fases:

- Pretratamiento: consiste en separar sólidos voluminosos (botellas, telas, plásticos) utilizando rejas y tamices.

- Tratamiento primario: son tratamientos físicos-químicos para sedimentar y precipitar sólidos en suspensión y para reducir la demanda bioquímica de oxígeno por parte de los sólidos orgánicos. Además, se utiliza para neutralizar aguas, eliminar contaminantes volátiles, para el desengrasado, el desaceitado, etc.

- Tratamiento secundario: son tratamientos biológicos que reducen la materia orgánica en las aguas residuales. Se utiliza tanto procesos aerobios, en presencia de oxígeno para degradar la materia orgánica, como anaerobios oxidando la materia orgánica sin oxígeno, seguido de una decantación secundaria.

- Tratamiento terciario: son procesos físicos, químicos y biológicos avanzados, donde se eliminan los metales pesados, nitrógeno, fósforo y patógenos. En algunas EDAR el agua se somete a un grado de tratamiento mayor para su utilización en el riego de parques, baldeo de calles, usos industriales o en zonas de escasez de agua.

En nuestro caso nos encontramos en los reactores del tratamiento secundario y los objetivos de este proyecto consisten en predecir mediante una serie histórica diferentes químicos o componentes que se encuentran en estos reactores como puede ser el amonio, el fósforo, etc. Tenemos un total de 4 reactores.

El porqué de esta predicción se basa en que dependiendo de los valores de estos componentes, el tratamiento del agua será diferente, de manera que al predecir las variables también podrán saber con anticipación que tratamiento deberán realizarle al agua. En los

siguientes apartados veremos cuáles son esas variables.

Lo primero que hay que decir, es que una serie temporal es una colección de observaciones de una variable tomadas de forma secuencial y ordenada en el tiempo. Estas, pueden tener una periodicidad anual, semestral, trimestral, mensual, etc., según los periodos de tiempo en los que están recogidos los datos que la componen, en nuestro caso será una periodicidad diaria como se verá más adelante. Las series temporales se pueden definir como un caso particular de los procesos estocásticos, ya que un proceso estocástico es una secuencia de variables aleatorias, ordenadas y equidistantes cronológicamente referidas a una característica observable en diferentes momentos.

El objetivo de una serie temporal reside en estudiar los cambios en esa variable con respeto al tiempo (descripción), y en predecir sus valores futuros (predicción). Por lo tanto, el análisis de series temporales presenta un conjunto de técnicas estadísticas que permiten extraer las regularidades que se observan en el comportamiento pasado de la variable, para tratar de predecir el comportamiento futuro.

Una serie temporal se representa mediante un gráfico temporal, con el valor de la serie en el eje de ordenadas y los tiempos en el eje de abscisas. Esta es la forma más sencilla de comenzar el análisis de una serie temporal y permite detectar las características y componentes más importantes de una serie. Estos gráficos se usarán para ver alguna de las predicciones que se obtengan de los modelos, por lo que a lo largo de este trabajo este tipo de gráfico aparecerá, de manera que se podrá comparar estas predicciones con el groundtruth y con otras predicciones.

Los componentes que forman una serie temporal son los siguientes:

- Tendencia: Se puede definir como un cambio a largo plazo que se produce en relación al nivel medio, o el cambio a largo plazo de la media. La tendencia se identifica con un movimiento suave de la serie a largo plazo. Esta componente será de vital importancia en la valoración de las predicciones ya que será un buen indicador de la calidad de las mismas.

- Estacionalidad: Se puede definir como cierta periodicidad de corto plazo, es decir, cuando se observa en la serie un patrón sistemático que se repite periódicamente (cada año, cada mes, etc., dependiendo de las unidades de tiempo en que vengan recogidos los datos). Por ejemplo, el paro laboral aumenta en general en invierno y disminuye en verano.

- Ciclo: Similar a la estacionalidad, ya que se puede definir como una fluctuación alrededor de la tendencia, pero de una duración irregular (no estrictamente periódica).

- Irregular: Son factores que aparecen de forma aleatoria y que no responden a un comportamiento sistemático o regular y por tanto no pueden ser predecidos. No se corresponden a la tendencia ni a la estacionalidad ni a los ciclos.

CAPÍTULO 2
Objetivos

Los objetivos de este proyecto son:

- Realizar un agrupamiento de los datos, así como su preprocesado.
- Efectuar una búsqueda exhaustiva de parámetros mediante un modelo de Prophet.
- Entrenar un modelo de LSTM
- Efectuar también, una búsqueda exhaustiva de parámetros para entrenar un modelo de XGBoost.
- Llevar a cabo una comparación entre los modelos obtenidos.
- Probar otros modelos.

CAPÍTULO 3
Marco teórico

Como se ha comentado, una serie temporales o serie cronológica, es una colección de datos numéricos (observaciones) al cual se asocia con un instante específico del tiempo. A través del análisis de estos, se puede construir un modelo que explique la estructura y el comportamiento de la serie de datos. Además, se puede proyectar la evolución de una variable a lo largo del tiempo.

Los modelos de series temporales suponen desconocimiento acerca de las relaciones causales que en la realidad afectan a la variable que se trata de pronosticar. Sin embargo, examinan el comportamiento de una serie temporal en el pasado para inferir cual será su comportamiento futuro.

El análisis de datos de series temporales ha sido un tema de interés popular en otros campos como la economía, la ingeniería y la medicina. Las técnicas tradicionales de manipulación de este tipo de datos se pueden encontrar en [4], y la aplicación de técnicas tradicionales de redes neuronales artificiales se describe en [5].

La mayoría de los trabajos que utilizan redes neuronales artificiales para manipular datos de series temporales se centran en la modelización y la previsión. Como primer intento de utilizar las redes neuronales para estas tareas, [6] modeló los precios de la harina a lo largo de 8 años. Todavía en los años 90, [7] delineó ocho pasos sobre el "diseño de un modelo de previsión de redes neuronales utilizando datos de series temporales económicas". Los enfoques más recientes incluyen el uso de redes recurrentes de *Elman* para predecir series temporales caóticas [8], el empleo de métodos de conjuntos de redes neuronales para predecir el tráfico de Internet [9], el uso de perceptrones multicapa simples para modelar la cantidad de basura en el Mar del Norte [10], e implementando en FPGA un algoritmo de predicción utilizando *Echo State Networks* para ”explotar el paralelismo inherente a estos sistemas” [11].

Los enfoques híbridos para el análisis de series temporales que utilizan redes neuronales no son infrecuentes. [12] presenta un modelo de previsión de series temporales utilizando redes neuronales y modelos ARIMA y [13] aplica los mismos tipos de modelos a la predicción de series temporales de la calidad del agua. En otros ejemplos de las mismas ideas, [14] compara el rendimiento de los modelos ARIMA y de las redes neuronales para realizar predicciones a corto plazo sobre generadores de energía fotovoltaica, mientras que [15] compara ambos modelos con el rendimiento de los Splines de Regresión Adaptativa Multivariante. [16] realiza predicciones de series temporales utilizando un modelo difuso híbrido: mientras que el método *Fuzzy C-means* se utiliza para la *fuzzifi-*

cación, la redes neuronales se emplea para la *defuzzificación*. Finalmente, [17] pronostica la velocidad del viento utilizando un híbrido de *Support Vector Machines*, *Ensemble Empirical Mode Decomposition* y *Partial Autocorrelation Function*.

El campo del *Deep Learning* ha atraído mucho interés en los últimos años. Una revisión muy completa de toda la historia de los desarrollos que han llevado al campo a su estado actual se puede encontrar en [18]. Vamos a ver las distintas aplicaciones del *Deep Learning* en el ámbito de las series temporales.

- **Clasificación**: La tarea de clasificación de cualquier tipo de datos se ha visto beneficiada por la llegada de las CNN. Anteriormente, los métodos de clasificación existentes se basaban en el uso de características específicas del dominio, normalmente elaboradas manualmente por expertos humanos. Encontrar las mejores características era objeto de mucha investigación y el rendimiento del clasificador dependía en gran medida de su calidad. La ventaja de las CNN es que pueden aprender esas características por sí mismas

  Un ejemplo de la aplicación de este tipo de aprendizaje de características no supervisado para la clasificación de señales de audio se presenta en [28]. En [29], las características aprendidas por la CNN se utilizan como entrada a un Modelo de Markov Oculto, logrando una caída en la tasa de error de más del 10 %. La aplicación de las CNN en estos trabajos presupone la restricción de que la serie temporal está compuesta por un solo canal. Una arquitectura que resuelve esta restricción se presenta en [30]. En [31] se compara el rendimiento de las CNNs con el de las LSTM para la clasificación de datos visuales y hápticos en un entorno robótico, y en [32] las señales producidas por sensores vestibles se transforman en imágenes para que las *Deep CNN* puedan ser utilizadas para la clasificación.

- **Forecasting**: En la literatura se pueden encontrar diferentes enfoques de *Deep Learning* para realizar tareas de *forecasting*. Por ejemplo, las redes de creencia profunda se utilizan en el trabajo de [19] junto con RBM (*Restricted Boltzmann Machines*). En [20] también se compara el rendimiento de las redes de creencia profunda con el de los *Denoising Autoencoders* apilados. Este último tipo de red también se emplea en [21] para predecir la temperatura de un ambiente interior. Otra aplicación de la predicción de series temporales se puede encontrar en [22], que utiliza Autoencoders apilados para predecir el flujo de tráfico a partir de un conjunto de datos.

  Una aplicación popular a la tarea de predicción de series temporales es en predicción meteorológica. En [23], algunas predicciones preliminares sobre datos meteorológicos proporcionados por El Observatorio de Hong Kong se hace a través del uso de Autoencoders apilados. En un trabajo posterior, los autores utilizan ideas similares para realizar predicciones sobre Big Data [24]. En lugar de Autoencoders, [25] utiliza redes de creencia profunda para construir un modelo híbrido en el que la redes neuronales modelan la distribución conjunta entre las variables de predicción meteorológica.

  Asimismo, existe [1] donde se describe un enfoque práctico de la previsión "a escala" que combina modelos configurables con el análisis de rendimiento.

- **Detección de anomalías**: Los trabajos que aplican técnicas de *Deep Learning* a la detección de anomalías en series temporales no son muy abundantes en la literatura. Todavía es difícil encontrar trabajos como [26], que utiliza *Denoising Autoencoders* apilados para realizar la detección de anomalías de algoritmos de seguimiento de

bajo nivel.

Sin embargo, hay muchas similitudes entre la detección de anomalías y las dos tareas anteriores. Por ejemplo, la identificación de una anomalía podría transformarse en una tarea de clasificación, como se hizo en [27]. Alternativamente, la detección de una anomalía podría considerarse lo mismo que encontrar regiones en las series temporales para las que los valores previstos son demasiado diferentes de los reales.

CAPÍTULO 4
Metodología

En este proyecto se han utilizado las siguientes herramientas:

4.1 Prophet

Prophet [1] es un toolkit de código abierto (lanzado por el equipo de Core Data Science de Facebook) de previsión de datos de series temporales basado en un modelo aditivo en el que las tendencias no lineales se ajustan a la estacionalidad anual, semanal y diaria, además de los efectos de las vacaciones. Funciona mejor con series temporales que tienen fuertes efectos estacionales y varias temporadas de datos históricos. Prophet es resistente a los datos que faltan y a los cambios de tendencia, y suele manejar bien los valores atípicos.

Prophet se utiliza en muchas aplicaciones de Facebook para producir previsiones fiables para la planificación y el establecimiento de objetivos. Han comprobado que su rendimiento es mejor que el de cualquier otro método en la mayoría de los casos. Ajustan los modelos en Stan para que obtengamos previsiones en tan solo unos segundos.También es resistente a los valores atípicos, los datos que faltan y los cambios drásticos en las series temporales, cosa que nos vendrá muy bien para este proyecto por las características de los datos.

Otra de las virtudes de Prophet es que incluye muchas posibilidades para que los usuarios retoquen y ajusten las previsiones. Puede utilizar parámetros interpretables por el ser humano para mejorar su previsión añadiendo su conocimiento del dominio.

4.2 Keras

Keras [2] es una API de redes neuronales de alto nivel escrita en python que puede funcionar sobre Tensorflow. Está diseñada para permitir una rápida experimentación con redes neuronales y entre sus principales virtudes se encuentra la modularidad, la extensibilidad y el ser fácil de usar para los usuarios.

En este proyecto se ha utilizado la parte de LSTM de Keras. Una LSTM es una arquitectura de red neuronal recurrente (RNN) utilizada en el campo del aprendizaje profundo. A diferencia de las redes neuronales estándar, la LSTM tiene conexiones de retroali-

mentación.

Las redes LSTM son muy adecuadas para clasificar, procesar y hacer predicciones basadas en datos de series temporales, ya que puede haber desfases de duración desconocida entre los eventos importantes de una serie temporal. Las LSTM se desarrollaron para hacer frente al problema del gradiente de evanescente que se puede encontrar al entrenar las RNN tradicionales. La relativa insensibilidad a la longitud de los desfases es una ventaja de los LSTM sobre las RNN, los modelos de Markov ocultos y otros métodos de aprendizaje de secuencias en numerosas aplicaciones.

4.3 XGBoost

XGBoost [3] es una librería distribuida optimizada de gradient boosting diseñada para ser altamente eficiente, flexible y portable. Implementa algoritmos de aprendizaje automático bajo el marco del Gradient Boosting. XGBoost proporciona un boosting de árbol paralelo (también conocido como GBDT, GBM) que resuelve muchos problemas de ciencia de datos de una manera rápida y precisa. El mismo código se ejecuta en los principales entornos distribuidos (Hadoop, SGE, MPI) y puede resolver problemas más allá de miles de millones de ejemplos.

XGBoost funciona como Newton-Raphson en el espacio de funciones a diferencia del gradient boosting que funciona como descenso de gradiente en el espacio de funciones, se utiliza una aproximación de Taylor de segundo orden en la función de pérdida para hacer la conexión con el método Newton-Raphson.

4.4 MLFlow

MLflow es una plataforma de código abierto para gestionar el ciclo de vida de ML, incluyendo la experimentación, la reproducibilidad, el despliegue y un registro central de modelos. MLflow ofrece actualmente cuatro componentes:

- Tracking: Registro y consulta de experimentos: código, datos, configuración y resultados.
- Proyectos: Empaquetar el código de la ciencia de los datos en un formato que reproduzca las ejecuciones en cualquier plataforma.
- Modelos: Implantar modelos de aprendizaje automático en diversos entornos de servicio.
- Registro: Almacenar, anotar, descubrir y gestionar modelos en un repositorio central

CAPÍTULO 5
Desarrollo

5.1 Datos

Lo que se recibió referido a los datos fue un conjunto de archivos los cuales se dividían en: un archivo para cada mes de las variables de nexus (medidas por un sensor) y cuatro archivos con las variables que provienen del laboratorio (medidas a mano). Estos cuatro archivos contienen todo el histórico, pero las variables están divididas en estos cuatro archivos.

Lo primero que se ha realizado ha sido juntar en un mismo dataframe todas las variables de nexus concatenando los archivos de todos los meses. Por otra parte, se ha juntado a mano el histórico de todas las variables de laboratorio en otro dataframe distinto, de manera que tenemos dos dataframe mucho más manejables con todos los datos.

A lo largo de este trabajo se denominará outputs a las variables a predecir e inputs a aquellas que ayudarán en la predicción de los outputs. Ambas se han recibido de dos fuentes distintas:

- Variables nexus: Estas variables eran medidas por sensores en el propio reactor, por lo que se tiene una granularidad alta (5 minutos). Los datos poseen buena calidad ya que tienen pocos o ningún nulo. Sin embargo solo se tiene un año de histórico (2020/03 - 2021/02).
- Variables laboratorio: Estas variables eran medidas a mano en un laboratorio, por lo que la granularidad es en el mejor de los casos diaria, pero tenemos una mala calidad del dato. En este caso, el histórico es mayor, comprendido entre 2018 y 2021. No obstante, como se verá más adelante, al juntar todos los datos se usará solamente el histórico comprendido en el mismo rango de las variables de nexus.

Para juntar todos los datos en un mismo dataframe, se han utilizado los valores históricos de las variables medidas en laboratorio y los valores sensorizados extrayendo la mediana por día, ya que su granularidad es de 5 minutos.

Entre las variables de nexus se encuentran con las que muestra la Tabla 5.1. Aquí se visualiza la descripción que tienen así como si son variables outputs o inputs. Las que aparecen como RX significa que esa misma variable está en los cuatro reactores de manera independiente.

Variable	Descripción	Output/Input
AB_QB_PLC3_RX_AMONIO	Valor de amonio (NH4) monitorizado en reactor biológico	Output
AB_QB_PLC3_RX_NITRATO	Valor de nitrato (NO3) monitorizado en reactor biológico	Output
AB_QB_PLC3_Q_AGUA_BRUTA	Caudal de entrada a la depuradora	Input
AB_QB_PLC3_Q_RECIRC_X	Caudal de recirculación de reactor.	Input
AB_QB_PLC3_Q_RECIRC_T	Suma de caudal de las cuatro recirculaciones	Input
AB_QB_PLC3_RX_CLORURO	Valor de cloruro monitorizado en reactor biológico	Input
AB_QB_PLC3_RX_CNSG_OD	Valor de consigna de oxígeno en el reactor biológico	Input
AB_QB_PLC3_RX_FOSFORO	Valor de ortofosfato (PO4) en el reactor biológico	Output
AB_QB_PLC3_RX_DOSIS_RTC_P	Dosis de cloruro férrico requerida y calculada por el equipo RTC en l/h	Output
AB_QB_PLC3_RX_ON_RTC_P	Orden de marcha de la bomba de dosificación on/off (0/1)	Input
AB_QB_PLC3_RX_QB_RTC_P	Caudal consigna para la bomba de dosificación de bomba en l/h.	Input
AB_QB_PLC3_RX_OD	Valor de oxígeno (O2) en el reactor biológico	Output
AB_QB_PLC3_RX_POTASIO	Valor de potasio en el reactor biológico	Input
AB_QB_PLC3_RX_TEMP	Valor de temperatura (ºC) en el reactor biológico	Input
AB_QB_PLC3_RX_RTC_ON_N	Orden de nitrificación del RTC (1= Nitrificar, 0= Desnitrificar)	Output

Tabla 5.1: Variables medidas mediante sensor

En este caso, se extraen seis variables a modelizar. Como se ha dicho, al ser variables independientes de cada reactor, se obtienen un total de veinticuatro variables a modelizar provenientes de los datos medidos mediante sensores. Cabe destacar que estos datos no contienen nulos como ya se ha indicado previamente.

Vistos los datos de nexus, se ha de centrar la atención en los de laboratorio, reflejados en la Tabla 5.2. Del mismo modo que la tabla anterior, estos aparecen con su descripción y si se trata de un input o de un output.

Variable	Descripción	Output/Input
SSLM Air X	Concentración de sólidos suspendidos en reactor	Output
SSVLM Air X	Concentración de sólidos suspendidos volátiles en reactor	Output
V30 Air X	Volumen de licor mezcla en 30 minutos	Input

V5 Air X	Volumen de licor mezcla en 5 minutos	Input
SSLM en recirculación (SSLMr)	Concentración de sólidos suspendidos en la recirculación	Input
Caudal purga	Caudal de purga de fangos de los decantadores secundarios	Output
%MS	Concentración de la purga de fangos secundarios	Output
SS E	Sólidos Suspendidos de entrada a la depuradora	Input
SS Dec	Sólidos Suspendidos tras decantación primaria y de entrada a reactores biológicos	Input
SS S	Sólidos Suspendidos de salida del reactor biológico	Output
DBO5 E	DBO5 de entrada a la EDAR	Input
DBO5 Dec	DBO5 tras decantación primaria y de entrada a reactores biológicos	Input
DBO5 S	DBO5 de salida del reactor biológico	Output
DQO E	DQO de entrada a la EDAR	Input
DQO Dec	DQO tras decantación primaria y de entrada a reactores biológicos	Input
DQO S	DQO de salida del reactor biológico	Output
Nt E	Nitrógeno Total de entrada a la EDAR	Input
Nt Dec	Nitrógeno Total tras decantación primaria y de entrada a reactores biológicos	Input
Nt S	Nitrógeno Total de salida de los reactores biológicos	Output
N-NH4 E	Amonio de entrada a la EDAR	Input
N-NH4 Dec	Amonio tras decantación primaria y de entrada a reactores biológicos	Input
N-NH4 S	Amonio de salida de los reactores biológicos	Output
N-NO3 S	Nitratos de salida de los reactores biológicos	Output
Pt E	Fósforo Total de entrada a la EDAR	Input
Pt Dec	Fósforo Total tras decantación primaria y de entrada a reactores biológicos	Input
Pt S	Fósforo Total de salida de los reactores biológicos	Output
pH E	pH de entrada a la EDAR	Input
pH Dec	pH tras decantación primaria y de entrada a reactores biológicos	Input

Tabla 5.2: Variables medidas en el laboratorio

En este tabla aparecen once variables a predecir, de las cuales dos aparecen en todos los reactores por lo que valdrán por 8. Ello lleva a un total de 17 variables medidas en el laboratorio. Para este tipo de variables si existirá el problema con los nulos. Más adelante se analizará y conocerá cuál es dicho porcentaje.

Juntando estas variables de laboratorio con las de nexus se obtienen un total de 41 variables a modelizar, lo que en el futuro de la experimentación significará un alto coste computacional por la cantidad de modelos que habrá que entrenar. Esto se debe a los diferentes comportamientos de las variables según el reactor, lo que imposibilitará utilizar el mismo modelo para todos los reactores.

De acuerdo con lo ya avanzado en el inicio del apartado, uno de los principales problemas que tienen estos datos es que existe un muy alto porcentaje de nulos. Como se puede advertir en la Tabla 5.3, 16 variables tienen más de un 50 % de nulos, de las cuales, 7 variables son de output y el resto de input. Viendo estos datos, se presupone desde un principio que será complicado obtener buenos resultados en estas variables debido a la falta de datos, ya que por encima de un 40 % de nulos ya es muy complicado obtener buenas predicciones.

Variable	Porcentaje nulos
Nt Dec	86.06
Pt Dec	84.57
DBO5 Dec	63.68
DQO Dec	60.69
SS Dec	60.19
Nt S	58.35
Pt S	57.47
SS E	57.36
DBO5 S	56.60
N-NO3 S	56.01
N-NH4 S	55.71
DQO S	55.42
SS S	55.42
DBO5 E	51.84
Nt E	50.74
Pt E	50.74
DQO E	47.98
V30 Air X	32.84
SSLM Air X	31.96
SSVLM Air X	31.96
V5 Air X	31.96
%MS	18.76

Tabla 5.3: Porcentaje de nulos variables

Otro de los problemas que tienen estos datos es que en algunas de las variables, la serie histórica puede no tener sentido debido a los múltiples picos que contienen por valores anómalos o cambios muy bruscos de la tendencia. Un ejemplo de esto se observa en la Figura 5.1. En el caso del Nitrato en el reactor 3 se puede observar picos por encima

de la media del resto del histórico, algo que también ocurre con el Cloruro del reactor 3. Este además,tiene un cambio de tendencia después del pico. En el Amonio, también se encuentran estos picos y en el caso del Potasio se ve que pega un cambio de tendencia extremo. Este problema, sumado al de los nulos, es algo que va a dificultar mucho la experimentación. No obstante, al tratarse de datos reales en un entorno del mundo real, es algo que siempre puede ocurrir.

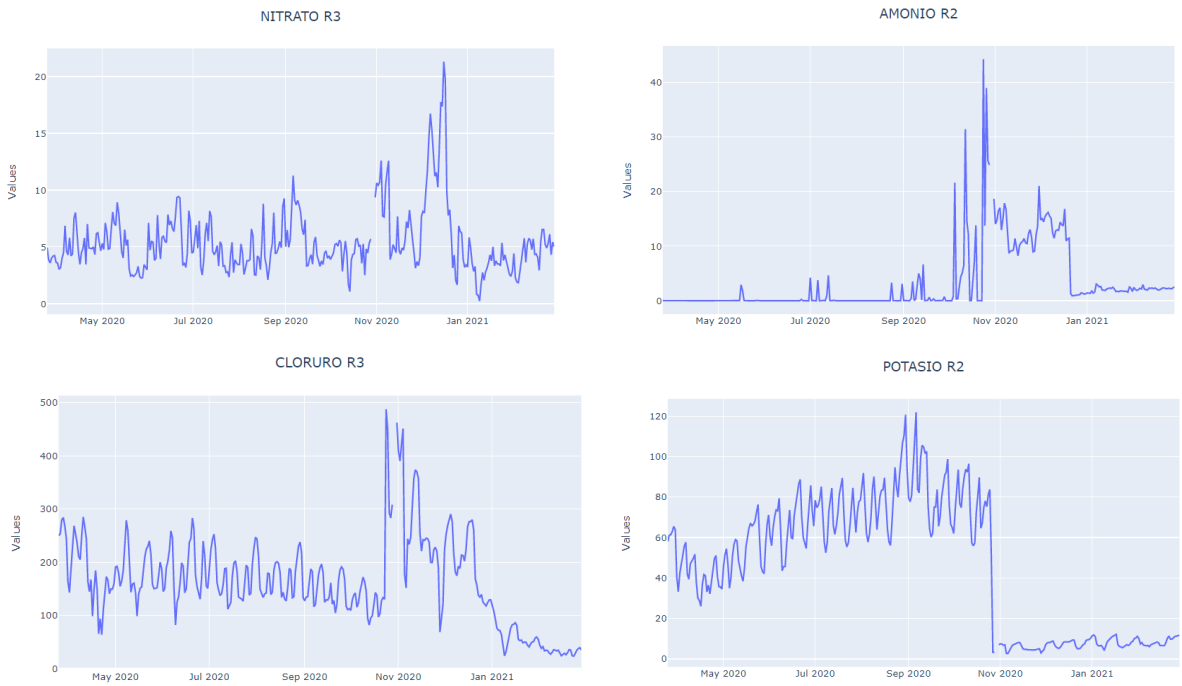


Figura 5.1: Ejemplo de variable con cambio brusco en el histórico

El problema de los nulos, de los picos y cambios de tendencia se van a intentar solventar de la siguiente manera: se interpolarán los datos nulos y se hará una limpieza de espurios, es decir, eliminación de aquellos valores anormales. Estos procesos se ejecutan con la intención de depurar la información aportada por la serie histórica de manera que no perjudique el análisis llevado a cabo en fases posteriores.

En el interpolado de datos, el objetivo del pre-proceso es el de completar los datos históricos en los huecos en los que no se dispone de dato, ya que en el análisis de series temporales, como es el enfoque de este estudio, la continuidad de datos es relevante. El interpolado aplicado es lineal, generando así datos distribuidos regularmente para el intervalo de valores faltante.

Para el caso de la limpieza de datos, el pre-proceso llevado a cabo aquí es el de eliminar datos espurios, es decir, datos considerados como improbables o incorrectos. Se han empleado criterios basados en los niveles máximos y mínimos que se conocen que determinadas variables en el agua pueden tener. Este criterio se ha obtenido mediante conocimiento experto. Estos valores de máximos y mínimos para las variables se pueden ver en la Tabla 5.4. Para las variables vistas que no aparecen en esta tabla no se conocen esos datos.

Variable	Mínimo	Máximo
AB_QB_PLC3_RX_AMONIO AB_QB_PLC3_RX_NITRATO	0	12
AB_QB_PLC3_Q_RECIRC_X	300	600
AB_QB_PLC3_Q_RECIRC_T	1200	1800
AB_QB_PLC3_RX_CLORURO	0	1000
AB_QB_PLC3_RX_CNSG_OD	0.2	2.2
AB_QB_PLC3_RX_FOSFORO	0	1
AB_QB_PLC3_RX_DOSIS_RTC_P	0	25
AB_QB_PLC3_RX_QB_RTC_P	0	55
AB_QB_PLC3_RX_OD	0.1	4
AB_QB_PLC3_RX_POTASIO	0	360
SSLM Air X	1500	4000
V30 Air X	30	1000
V5 Air X	180	800
SSLM en recirculación (SSLMr)	2000	8000
%MS	0.8	8
SS E	0	600
SS Dec	0	400
SS S	0	35
DBO5 E	0	500
DBO5 Dec	0	400
DBO5 S	0	25
DQO E	0	700
DQO Dec	0	500
DQO S	0	125
Nt E	0	90
Nt Dec	0	85
Nt S	0	10
N-NH4 E	0	50
N-NH4 Dec	0	45
N-NH4 S	0	3
N-NO3 S	0	7
Pt E	0	12
Pt Dec	0	10
Pt S	0	1
pH E y ph Dec	7	8

Tabla 5.4: Máximos y mínimos posibles de cada variable

El criterio empleado en estos casos es el de aplicar un valor NaN, el cual será rellenado de nuevo usando la interpolación antes descrita. Haciendo esto y basándonos en el ejemplo de la Figura 5.1, se observa que en el caso del Nitrato y del Amonio, esos picos desaparecen porque son datos erróneos. Sin embargo, para el caso del Cloruro y del Potasio, los picos y los cambios de tendencia no se van con este criterio, además de no poder eliminarlo debido a que esos resultados si son ”normales” porque son valores que esas variables pueden tener en una situación real.

5.2 Experimentación

5.2.1. Prophet

Los primeros experimentos que se realizaron fueron con Prophet. Se decidió hacer una búsqueda exhaustiva de hiperparámetros para obtener los mejores resultados posibles para cada variable. Los hiperparámetros que se estudiaron fueron los siguientes:

- **seasonality_mode** ["additive", "multiplicative"]: una estacionalidad aditiva significa que el efecto de la estacionalidad se añade a la tendencia para obtener la previsión. En ocasiones una serie temporal tiene un claro ciclo, pero la estacionalidad en la previsión es demasiado grande al principio de la serie temporal y demasiado pequeña al final. En esta serie temporal, la estacionalidad no es un factor aditivo constante como supone Prophet, sino que crece con la tendencia. Esto es una estacionalidad multiplicativa.

- **changepoint_prior_scale** [0.0001, 0.001, 0.2, 0.4]: este es probablemente el parámetro más impactante. Determina la flexibilidad de la tendencia, y en particular, cuánto cambia la tendencia en los puntos de cambio de esta. Si es demasiado pequeño, la tendencia estará infraajustada y la varianza que debería haber sido modelada con los cambios de tendencia, acabará siendo tratada con el término de ruido. Si es demasiado grande, la tendencia se sobreajustará y, en el caso más extremo, puede acabar con la tendencia capturando la estacionalidad anual.

- **n_changepoints** [1, 2, 4]: número de puntos donde se producen cambios abruptos en la trayectoria y que permitirá que la tendencia se adapte adecuadamente.

A estos hiperparámetros, también se añaden lo que en Prophet se conoce como *additional regressors*, que pueden utilizarse como uno de los componentes para predecir el resultado. Simplemente se añaden las columnas con el nombre del regresor a utilizar. Lo que se ha hecho en este proyecto para elegir que combinación de regresores usar, ha sido lo siguiente: primero se ha comprobado qué variables tenían una mayor correlación, es decir, por cada variable de output, se ha visto qué variables de input (del mismo reactor) tenían mayor correlación. Se han cogido las 5 variables que más correlacionaban y se han hecho todas las combinaciones posibles de esas cinco variables. Asimismo, se ha añadido la opción de ningún regresor para saber si verdaderamente la información extra de otras variables ayudaba a mejorar los resultados. Para las concentraciones de salida (variables que no pertenecen a ningún reactor y que acaban en S), se usan las concentraciones de entrada y decantación de cada variable, además del caudal de agua. Esta decisión fue tomada a partir de conocimiento experto.

Juntando todos los hiperparámetros descritos y haciendo las combinaciones posibles salen aproximadamente unas 744 para cada variable, exceptuando las concentraciones de salida por lo explicado en el anterior párrafo.

Una vez tuvimos preparados los hiperparámetros a probar, se hizo lo que se conoce como *GridSearch*, que es la búsqueda exhaustiva sobre los valores de los parámetros especificados para un modelo. Las métricas que se usaron para medir los modelos fueron las siguientes:

- R^2: el coeficiente de determinación. Determina la capacidad de un modelo para predecir futuros resultados. El mejor resultado posible es 1.0, y ocurre cuando la predicción coincide con los valores de la variable objetivo. R^2 puede tomar valores negativos, pues la predicción puede ser arbitrariamente mala. Cuando la predicción coincide con la esperanza de los valores de la variable objetivo, el resultado de R^2 es 0. Se define como:

$$R^2 = 1 - \frac{\sum (y_i - p_i)^2}{\sum (y_i - \bar{y})^2}$$

donde:

y_i son los valores objetivo,
p_i son los valores predichos,
$\bar{y}$ es la media de los valores objetivo.

- MSE: mide el error cuadrado promedio de las predicciones. Para cada punto, calcula la diferencia cuadrada entre las predicciones y el objetivo y luego promedia esos valores. Se define como:

$$MSE = \frac{1}{N} \sum (y_i - \bar{y})^2$$

- MAE: el error se calcula como el valor medio de la suma de los valores absolutos de la diferencia entre los valores predichos y los valores reales:

$$MAE = \frac{1}{N} \sum |y_i - \bar{y}|$$

Ahora que se ha visto las métricas a utilizar y que hiperparámetros se han usado para evaluar los modelos, se van a presentar los resultados obtenidos. Se ha dividido el *dataset* en un 80 % para *train* y un 20 % para *test*.Los resultados ordenados en función de la métrica de R^2 se pueden ver en la Tabla 5.5. Estos resultados pueden ser recuperados gracias al empleo de la herramienta MLFlow.

Variable	R^2	MAE	MSE
AB_QB_PLC3_R2_DOSIS_RTC_P	0,998	0,100	0,016
AB_QB_PLC3_R3_DOSIS_RTC_P	0,921	0,231	1,680
AB_QB_PLC3_R4_DOSIS_RTC_P	0,869	0,248	1,934
SSLM (mg/l) Air 3	0,846	121,980	23502,129
SSVLM (mg/l) Air 2	0,822	97,226	16922,009
SSVLM (mg/l) Air 3	0,783	115,619	21657,561
SSLM (mg/l) Air 1	0,749	165,765	38941,292
SSVLM (mg/l) Air 1	0,748	128,433	26322,148
SSVLM (mg/l) Air 4	0,678	141,158	31033,786
SSLM (mg/l) Air 2	0,671	175,838	45962,993
SSLM (mg/l) Air 4	0,657	180,523	48924,575
AB_QB_PLC3_R1_FOSFORO	0,656	0,070	0,010
AB_QB_PLC3_R4_FOSFORO	0,594	0,116	0,025
AB_QB_PLC3_R3_FOSFORO	0,587	0,138	0,030

AB_QB_PLC3_R2_FOSFORO	0,554	0,087	0,014
AB_QB_PLC3_R1_DOSIS_RTC_P	0,413	1,124	5,713
Caudal de purga (m3/día)	0,306	393,302	282154,624
AB_QB_PLC3_R2_OD	0,251	0,191	0,072
AB_QB_PLC3_R4_OD	0,166	0,246	0,109
AB_QB_PLC3_R1_NITRATO	0,149	1,585	4,872
AB_QB_PLC3_R1_RTC_ON_N	0,098	0,240	0,113
AB_QB_PLC3_R3_AMONIO	0,086	1,641	4,058
SSLM (mg/l) Recirculación	0,0838	974,452	1430348,123
AB_QB_PLC3_R2_NITRATO	0,073	1,724	4,535
N-NH4 (mg/l) S	0,0701	0,505	0,362
DBO5 (mg/l) S	0,069	1,111	2,232
N-NO3 (mg/l) S	-0,082	0,997	1,162
AB_QB_PLC3_R3_NITRATO	-0,085	1,2582	2,403
AB_QB_PLC3_R4_RTC_ON_N	-0,141	0,465	0,222
Nt (mg/l) S	-0,143	0,303	0,141
AB_QB_PLC3_R3_OD	-0,156	0,297	0,182
AB_QB_PLC3_R3_RTC_ON_N	-0,214	0,060	0,017
Pt (mg/l) S	-0,226	0,057	0,005
%MS (%)	-0,242	0,350	0,242
SS (mg/l) S	-0,271	1,351	3,480
AB_QB_PLC3_R4_AMONIO	-0,479	2,841	10,917
AB_QB_PLC3_R2_RTC_ON_N	-0,485	0,433	0,201
AB_QB_PLC3_R4_NITRATO	-0,517	2,792	10,692
AB_QB_PLC3_R1_OD	-1,525	0,417	0,227
DQO (mg/l) S	-1,560	5,437	41,701
AB_QB_PLC3_R1_AMONIO	-1,821	1,203	2,236
AB_QB_PLC3_R2_AMONIO	-37,359	2,973	9,072

Tabla 5.5: Resultados obtenidos en la experimentación de Prophet

Como se observa en la tabla, hay bastante disparidad en los resultados. Por un lado, están las primeras 12, con más de 0.60 de R^2 (el máximo es 1 y el mínimo es menos infinito), teniendo las primeras casi una resultado perfecto. En las últimas variables de la tabla tienen un resultado negativo. Por último tenemos otras variables que no llegan a ser negativas pero siguen siendo resultados no muy buenos. Muchas de ellas corresponden con las variables que tenían muchos nulos, como es el caso del N-NH4 S o el Pt S.

También se puede ver que algunos malos resultados pueden venir por comportamientos extraños, como era el caso del Amonio del reactor 2, que hemos visto su gráfica de datos en la Figura 5.1. No había ninguna tendencia o patrón claro, por lo que es posible que el modelo tampoco lo haya podido encontrar.

En la Figura 5.2 muestran las variables que se visualizan en las imágenes. Las dos primeras corresponden con variables con un R^2 alto. En la primera se observa que prácticamente la línea del *forecast* (línea roja) se solapa con la línea del test (línea azul), en este caso esta variable tenía un R^2 de **0.998**. En la siguiente no se solapan, pero no es mal resultado y aunque no consigue afinar del todo, es capaz de captar y seguir las tendencias. En este caso esta variable tenía un R^2 de **0.846**.

Figura 5.2: Ejemplo de predicciones de Prophet

Las dos siguiente variables son dos que no van bien. En ambas tenemos un R^2 negativo, siendo el amonio del reactor 2 la que peor resultado daba con mucha diferencia, con un R^2 de **-37.358**. Los resultados de la predicción están bastante por encima de los valores de test y tampoco capta la tendencia. Lo mismo ocurre con el siguiente ejemplo, con un R^2 de **-1.57**. Los resultados de la predicción están un poco más cerca y parece que en ocasiones capte un poco la tendencia, sobretodo al final de la predicción, donde a veces

si que capta un poco los picos hacia arriba o hacia abajo.

Básicamente, con este ejemplo se confirma lo que se ha visto en la tabla de resultados. Hay mucha disparidad en los resultados. Estos son los que se han obtenido durante la experimentación con Prophet. En los siguientes apartados se analizarán los resultados obtenidos con otros modelos.

5.2.2. LSTM

Una vez hecha la experimentación con el modelo de Prophet, se decidió hacer experimentos con un modelo de LSTM, para ver si los resultados mejoraban o empeoraban, qué variables en concreto, etc.

En este caso lo que se ha hecho con los datos ha sido normalizarlos. Ya que las LSTMs son sensibles a la escala de los datos de entrada, especialmente cuando se utilizan las funciones de activación sigmoide (por defecto) o tanh, lo primero que se ha hecho ha sido normalizar los datos entre 0 y 1.

En las LSTM, a diferencia de un modelo de Prophet, las fechas desaparecen, y lo que entra a la red es un vector con los datos en orden. Hay que predecir un valor en el tiempo T, basado en los datos de los días T_n, donde n puede ser cualquier número de pasos. Más adelante se ha probado a predecir los datos en función del los últimos 30 y 60 días. El conjunto de características (vector de entrada a la red) debe contener los valores iniciales de las variables a predecir de los últimos 30 o 60 días, mientras que la etiqueta o variable dependiente debe ser el valor de la variable el día 31 o 61.

Los primeros experimentos que se hicieron fueron en función de 30 días y usando solamente las variables a predecir como entrada a la red. Sin embargo, se observó que los resultados eran muy malos, empeorando todas las variables por mucho a los resultados obtenidos con el modelo de Prophet. Por ello, se descartó este modelo y lo que se hizo fue, aprovechando toda la experimentación realizada con el Prophet, usar lo que hemos llamado regresores adicionales. En este caso se han utilizado las variables de input del mejor modelo de Prophet de cada variable y se han añadido al vector de características. Por supuesto, estas nuevas variables añadidas también están normalizadas de 0 a 1.

La red que se ha usado tiene 2 capas de LSTM con 50 unidades cada una y los resultados se reflejan en la Tabla 5.6. Se han realizado 100 épocas para cada variable y en este caso se ha utilizado los 30 días anteriores para predecir el siguiente. Más adelante se verán los resultados con 60 días.

Variable	R^2	MAE	MSE
AB_QB_PLC3_R4_RTC_ON_N	0,998	0,001	0,011
SSVLM (mg/l) Air 3	0,975	0,001	0,019
SSLM (mg/l) Air 3	0,974	0,001	0,020
SSLM (mg/l) Air 1	0,967	0,001	0,0257
DBO5 (mg/l) S	0,958	0,001	0,0194
N-NO3 (mg/l) S	0,942	0,001	0,022
Pt (mg/l) S	0,934	0,001	0,023
AB_QB_PLC3_R2_NITRATO	0,929	0,002	0,034

SSLM (mg/l) Air 2	0,928	0,001	0,032
SSLM (mg/l) Air 4	0,918	0,002	0,033
SSVLM (mg/l) Air 1	0,887	0,003	0,043
AB_QB_PLC3_R4_NITRATO	0,869	0,003	0,043
DQO (mg/l) S	0,861	0,0021	0,034
N-NH4 (mg/l) S	0,835	0,006	0,06
SSVLM (mg/l) Air 2	0,829	0,002	0,035
AB_QB_PLC3_R3_AMONIO	0,821	0,002	0,037
SS (mg/l) S	0,816	0,004	0,051
AB_QB_PLC3_R3_NITRATO	0,765	0,002	0,036
AB_QB_PLC3_R1_DOSIS_RTC_P	0,758	0,003	0,029
Caudal de purga (m3/día)	0,757	0,010	0,064
AB_QB_PLC3_R4_AMONIO	0,755	0,009	0,0731
Nt (mg/l) S	0,743	0,001	0,021
AB_QB_PLC3_R1_AMONIO	0,729	0,001	0,022
AB_QB_PLC3_R1_NITRATO	0,725	0,006	0,0597
AB_QB_PLC3_R3_DOSIS_RTC_P	0,720	0,007	0,056
AB_QB_PLC3_R4_FOSFORO	0,683	0,018	0,105
AB_QB_PLC3_R3_OD	0,669	0,003	0,049
AB_QB_PLC3_R3_FOSFORO	0,661	0,022	0,122
AB_QB_PLC3_R1_OD	0,598	0,003	0,034
AB_QB_PLC3_R2_DOSIS_RTC_P	0,555	0,004	0,029
AB_QB_PLC3_R4_OD	0,514	0,005	0,059
AB_QB_PLC3_R2_OD	0,492	0,004	0,048
AB_QB_PLC3_R4_DOSIS_RTC_P	0,339	0,009	0,0624
AB_QB_PLC3_R1_FOSFORO	0,295	0,009	0,0701
SSVLM (mg/l) Air 4	0,205	0,005	0,056
AB_QB_PLC3_R2_FOSFORO	0,153	0,010	0,078
SSLM (mg/l) Recirculación	0,015	0,014	0,102
AB_QB_PLC3_R2_RTC_ON_N	-0,336	0,239	0,267
AB_QB_PLC3_R2_AMONIO	-1,188	0,004	0,054
%MS (%)	-5,112	0,005	0,0557
AB_QB_PLC3_R1_RTC_ON_N	-243,425	0,113	0,145
AB_QB_PLC3_R3_RTC_ON_N	-254976,028	0,016	0,017

Tabla 5.6: Resultados obtenidos en la experimentación de LSTM usando 30 días para predecir

Lo primero que hay que destacar de estos nuevos resultados es que a simple vista se puede observar una mejora general. Si bien es cierto que algunas variables empeoran un poco, hay muchas menos con un R^2 negativo.

Otra cosa que llama la atención de los resultados son las dos últimas variables, que tienen un resultado muy malo. Sin embargo, en el caso de estas dos puede llegar a tener sentido. Para recordar, estas dos variables eran binarias, es decir, podían tomar valores de 0 y 1, por lo que puede pasar que la tendencia si que la capte bien, pero se quede corto al llegar a ese 1 o a ese 0. Para visualizarlo mejor, el conjunto de test y las predicciones se observan en la Figura 5.3 para poder ver que ha ocurrido exactamente. En el caso del RTC ON del reactor 3 lo que ocurre es que todo el conjunto de test es 1 menos un día que vale 0, por lo que el valor del R^2 es muy engañoso en ese caso porque solo se ha equivocado en un día. Por otro lado, en el caso del reactor 1, la tendencia si que la capta, pero como ya se

ha comentado, no baja lo suficiente a 0. En estos casos particulares, una posible solución sería hacer que si la predicción es cercana a 1 vale 1, y sino vale 0. De esa manera, como las tendencias si que las capta bien, el resultado sería bueno. Esto puede ocurrir con el RTC ON del reactor 2 que también tiene un resultado negativo pero en menor medida.

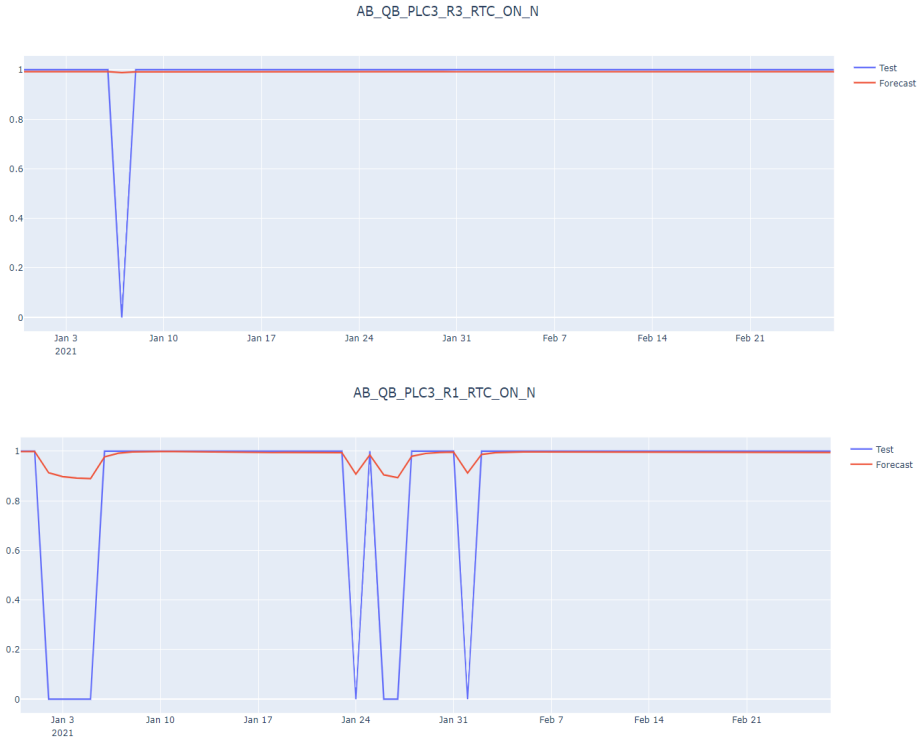


Figura 5.3: Predicciones con LSTM de dos de las variables de RTC ON

Es curioso, sin embargo, que la que mejor resultado de, sea el RTC ON del reactor 4, pero esto puede deberse a que oscile mucho más que las otras y sea capaz más fácilmente de saber que es 0 o 1. En los otros casos, es posible que la mayoría de datos sean 1 o 0 y no existan esas oscilaciones que se han comentado. Otro ejemplo de buenos resultados se puede encontrar para las variables de SSVLM y SSLM del reactor 3, localizado en la Figura 5.4. Captan perfectamente la tendencia y hay en algunos casos que llega clavar los resultados. La mayoría de los resultados son positivos, exceptuando 5 casos y más de la mitad superan el 0.6 de R^2.

Vistos los resultados de LSTM usando 30 días, se probará que las predicciones dependan de 60 días en vez de los 30 que se han visto y comentado. En este caso se hace lo mismo, se usan los regresores adicionales que se han obtenido en los mejores modelos de Prophet de cada variable, normalizamos los datos para tenerlos de 0 a 1. Lo que cambia es que cuando construimos el dataset. Se usarán los datos de 60 días para predecir. Dicho todo esto, los resultados con esos 60 días se pueden ver en la Tabla 5.7.

Variable	R^2	MAE	MSE
AB_QB_PLC3_R4_RTC_ON_N	0,998	0,001	0,011
N-NO3 (mg/l) S	0,954	0,001	0,031
AB_QB_PLC3_R2_NITRATO	0,935	0,001	0,033
SSLM (mg/l) Air 1	0,923	0,002	0,038
SSLM (mg/l) Air 3	0,922	0,002	0,034

DBO5 (mg/l) S	0,915	0,001	0,025
SSVLM (mg/l) Air 1	0,895	0,002	0,042
Pt (mg/l) S	0,895	0,001	0,030
N-NH4 (mg/l) S	0,866	0,007	0,065
SS (mg/l) S	0,865	0,003	0,048
SSVLM (mg/l) Air 2	0,832	0,001	0,032
DQO (mg/l) S	0,795	0,003	0,045
AB_QB_PLC3_R4_AMONIO	0,774	0,007	0,067
Caudal de purga (m3/día)	0,746	0,009	0,075
AB_QB_PLC3_R1_NITRATO	0,704	0,007	0,072
Nt (mg/l) S	0,701	0,001	0,022
AB_QB_PLC3_R3_DOSIS_RTC_P	0,700	0,006	0,050
AB_QB_PLC3_R1_AMONIO	0,654	0,001	0,021
SSVLM (mg/l) Air 3	0,646	0,004	0,055
AB_QB_PLC3_R3_AMONIO	0,642	0,004	0,054
AB_QB_PLC3_R3_FOSFORO	0,611	0,023	0,126
AB_QB_PLC3_R4_DOSIS_RTC_P	0,607	0,006	0,052
AB_QB_PLC3_R3_NITRATO	0,605	0,003	0,042
AB_QB_PLC3_R2_DOSIS_RTC_P	0,544	0,004	0,029
AB_QB_PLC3_R4_NITRATO	0,529	0,009	0,080
AB_QB_PLC3_R1_DOSIS_RTC_P	0,487	0,005	0,038
SSLM (mg/l) Air 2	0,437	0,008	0,074
AB_QB_PLC3_R2_OD	0,362	0,004	0,041
SSVLM (mg/l) Air 4	0,312	0,003	0,046
AB_QB_PLC3_R1_FOSFORO	0,226	0,009	0,072
AB_QB_PLC3_R4_FOSFORO	0,182	0,040	0,159
AB_QB_PLC3_R4_OD	0,134	0,006	0,054
AB_QB_PLC3_R2_FOSFORO	0,039	0,012	0,082
AB_QB_PLC3_R1_OD	0,035	0,007	0,065
AB_QB_PLC3_R3_OD	-0,190	0,005	0,040
SSLM (mg/l) Recirculación	-0,279	0,016	0,111
SSLM (mg/l) Air 4	-2,268	0,026	0,132
%MS (%)	-3,053	0,006	0,063
AB_QB_PLC3_R2_RTC_ON_N	-3,311	0,347	0,505
AB_QB_PLC3_R2_AMONIO	-8,465	0,029	0,162
AB_QB_PLC3_R1_RTC_ON_N	-91,553	0,110	0,126
AB_QB_PLC3_R3_RTC_ON_N	-102776,552	0,017	0,029

Tabla 5.7: Resultados obtenidos en la experimentación de LSTM usando 60 días para predecir

También se obtienen buenos resultados usando para predecir 60 días en lugar de 30. Hay muchas variables con más de 0.60 en R^2. En este caso, ocurre lo mismo que se ha visto antes con las variables del RTC ON de los reactores 1 y 3, así como también ocurre con el del reactor 2. Como se ha dicho antes, obtiene buenos resultados, pero lo que interesa saber es si realmente ha mejorado a lo que teníamos con 30 días. Para ello está la Tabla 5.8. En ella se encuentran las variables, el valor que han tenido de R^2 dependiendo de los días utilizados y la última columna es la diferencia, es decir, si el resultado es negativo, es que los 60 días han empeorado a los 30 días; si es positivo al revés, los 60 días han mejorado a los 30 días.

Figura 5.4: Ejemplo de predicciones de LSTM usando 30 días para predecir

Variable	R^2 30 días	R^2 60 días	Incremento/Decremento
%MS (%)	-5,112	-3,053	2,058
AB_QB_PLC3_R1_AMONIO	0,729	0,654	-0,075
AB_QB_PLC3_R1_DOSIS_RTC_P	0,757	0,487	-0,270
AB_QB_PLC3_R1_FOSFORO	0,295	0,226	-0,068
AB_QB_PLC3_R1_NITRATO	0,724	0,704	-0,020
AB_QB_PLC3_R1_OD	0,598	0,035	-0,563
AB_QB_PLC3_R1_RTC_ON_N	-243,425	-91,553	151,872
AB_QB_PLC3_R2_AMONIO	-1,188	-8,465	-7,277
AB_QB_PLC3_R2_DOSIS_RTC_P	0,554	0,544	-0,010
AB_QB_PLC3_R2_FOSFORO	0,153	0,039	-0,117
AB_QB_PLC3_R2_NITRATO	0,928	0,935	0,007
AB_QB_PLC3_R2_OD	0,492	0,362	-0,130
AB_QB_PLC3_R2_RTC_ON_N	-0,336	-3,311	-2,975
AB_QB_PLC3_R3_AMONIO	0,820	0,642	-0,178
AB_QB_PLC3_R3_DOSIS_RTC_P	0,720	0,70	-0,019
AB_QB_PLC3_R3_FOSFORO	0,661	0,611	-0,050
AB_QB_PLC3_R3_NITRATO	0,765	0,605	-0,160
AB_QB_PLC3_R3_OD	0,669	-0,190	-0,859
AB_QB_PLC3_R3_RTC_ON_N	-254976,028	-102776,552	152199,475
AB_QB_PLC3_R4_AMONIO	0,755	0,774	0,018
AB_QB_PLC3_R4_DOSIS_RTC_P	0,339	0,607	0,267
AB_QB_PLC3_R4_FOSFORO	0,683	0,182	-0,501
AB_QB_PLC3_R4_NITRATO	0,869	0,529	-0,340
AB_QB_PLC3_R4_OD	0,513	0,134	-0,378
AB_QB_PLC3_R4_RTC_ON_N	0,9981	0,9988	0,0007

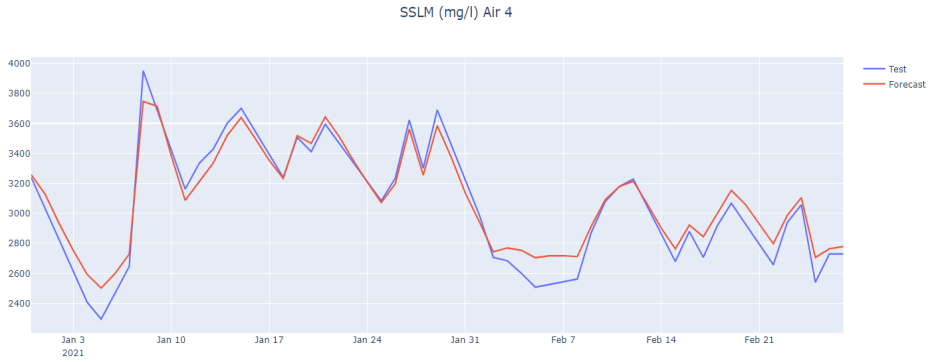
Caudal de purga (m3/día)	0,757	0,746	-0,010
DBO5 (mg/l) S	0,958	0,915	-0,042
DQO (mg/l) S	0,860	0,795	-0,065
N-NH4 (mg/l) S	0,835	0,866	0,031
N-NO3 (mg/l) S	0,942	0,954	0,011
Nt (mg/l) S	0,743	0,701	-0,042
Pt (mg/l) S	0,934	0,895	-0,038
SS (mg/l) S	0,816	0,865	0,048
SSLM (mg/l) Air 1	0,967	0,923	-0,044
SSLM (mg/l) Air 2	0,927	0,437	-0,490
SSLM (mg/l) Air 3	0,974	0,922	-0,051
SSLM (mg/l) Air 4	0,918	-2,268	-3,186
SSLM (mg/l) Recirculación	0,015	-0,279	-0,294
SSVLM (mg/l) Air 1	0,887	0,895	0,008
SSVLM (mg/l) Air 2	0,829	0,832	0,003
SSVLM (mg/l) Air 3	0,974	0,646	-0,327
SSVLM (mg/l) Air 4	0,205	0,312	0,106

Tabla 5.8: Comparación R^2 para predicciones usando 30 y 60 días

En algunos casos mejora usando 60 días y en otros casos empeora. Sin embargo, la variación es muy pequeña en prácticamente todos los casos, excepto en algunos, como ocurre con las variables de RTC ON. Estas ya se sabe lo que ocurría con ellas, por lo que es normal esta diferencia.

Otros casos que llaman mucho la atención son los de SSLM Air 4, que empeora de 30 días a 60 días en 3.18, pasando de un muy buen resultado con 0.918 a -2.26; o el caso del Amonio en el reactor 2, que empeora pasando de -1.18, que ya era un muy mal resultado, a -8.46 el cual es aun peor. En la Figura 5.5 se ven las predicciones de la variable SSLM Air 4, básicamente la tendencia es capturada por ambas más o menos, pero usando 30 días la predicción es mucho más precisa, llegando incluso a acertar el resultado, cosa que no ocurre cuando usamos 60 días. Con 60 días logra capturar la tendencia en ocasiones como se ha dicho, pero los resultados están muy alejados del *groundtruth*.

Para el resto de variables prácticamente no afecta. viendo las únicas que afecta, la opción de usar 30 días es la mejor. Se usarán los resultados de esta prueba para realizar la comparación entre los modelos que se hará más adelante.



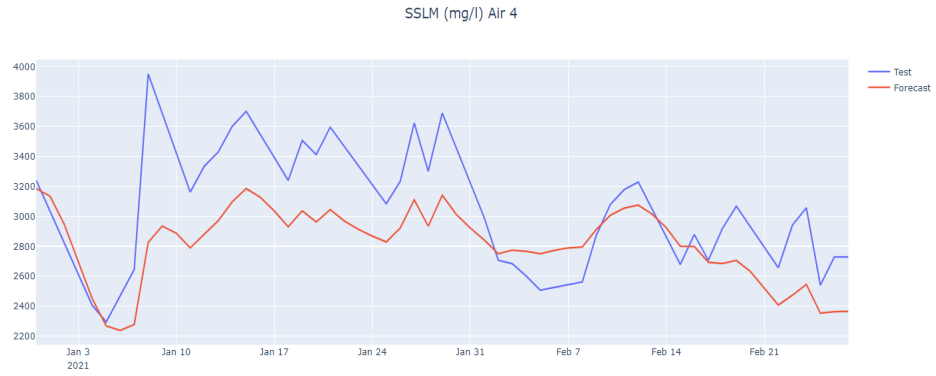


Figura 5.5: Predicciones de SSLM Air 4 usando 30 (arriba) y 60 (abajo) días.

5.2.3. XGBoost

Hecha la experimentación con las LSTM, se pasó a realizar una prueba con un modelo basado en árboles, para ello se usó un modelo de XGBoost. La experimentación que se ha realizado con él es muy parecida a lo que hemos hecho con las LSTM. Se han utilizado los regresores adicionales obtenidos en el Prophet para aportarle al modelo más información. También se normalizan los datos y se generan los datasets utilizando 30 o 60 días para predecir.

En la experimentación también se ha hecho un *GridSearch* entre los siguientes parámetros:

- n_estimators [50, 100, 150, 200]: Número de árboles reforzados por gradiente. Equivale al número de rondas de refuerzo.
- max_depth [2, 4, 6, 8]: Profundidad máxima de los árboles.
- min_child_weight [1, 5, 10]: Suma mínima de peso de instancia (hessian) necesaria en un niño.
- gamma [0.5, 1, 1.5, 2, 5]: Reducción mínima de loss necesaria para realizar una nueva partición en un nodo hoja del árbol.

Vistos los parámetros que se van a usar, los resultados obtenidos se van a visualizar al mismo nivel que lo visto con las LSTM. Primero se verán los resultados usando 30 días, después usando 60 días y por último la diferencia entre ambos modelos. Los resultados que se muestren de cada variable corresponderán con el mejor modelo resultante de la búsqueda exhaustiva de hiperparámetros.

En la Tabla 5.9 se pueden ver los resultados del mejor modelo de cada variable obtenido en el *GridSearch*.

Variable	R^2	MAE	MSE
SSLM (mg/l) Air 1	0,866	103,576	17508,171
SSLM (mg/l) Air 2	0,830	115,141	20691,304
AB_QB_PLC3_R2_NITRATO	0,811	0,613	0,607

AB_QB_PLC3_R4_AMONIO	0,798	0,767	1,165
N-NO3 (mg/l) S	0,775	0,309	0,190
AB_QB_PLC3_R1_AMONIO	0,774	0,247	0,095
SSVLM (mg/l) Air 1	0,749	106,425	17363,545
SSLM (mg/l) Air 3	0,737	123,023	28458,678
AB_QB_PLC3_R1_DOSIS_RTC_P	0,6705	0,830	2,092
AB_QB_PLC3_R4_OD	0,626	0,105	0,020
AB_QB_PLC3_R2_DOSIS_RTC_P	0,604	0,873	2,287
SSVLM (mg/l) Air 2	0,547	122,574	26248,617
N-NH4 (mg/l) S	0,518	0,212	0,084
SSVLM (mg/l) Air 4	0,509	124,170	28828,218
AB_QB_PLC3_R4_NITRATO	0,496	0,947	1,425
SSLM (mg/l) Air 4	0,467	148,909	41240,032
SSVLM (mg/l) Air 3	0,465	116,242	30245,056
AB_QB_PLC3_R4_RTC_ON_N	0,443	0,146	0,027
AB_QB_PLC3_R2_OD	0,400	0,088	0,028
Caudal de purga (m3/día)	0,328	292,667	135337,505
AB_QB_PLC3_R1_NITRATO	0,237	1,233	2,354
Nt (mg/l) S	0,225	0,178	0,045
AB_QB_PLC3_R3_AMONIO	0,196	0,783	1,076
DBO5 (mg/l) S	0,186	0,428	0,868
AB_QB_PLC3_R1_FOSFORO	0	0,154	0,032
AB_QB_PLC3_R2_FOSFORO	0	0,158	0,034
AB_QB_PLC3_R3_FOSFORO	0	0,229	0,080
AB_QB_PLC3_R4_FOSFORO	0	0,208	0,066
AB_QB_PLC3_R3_RTC_ON_N	0	0,175	0,038
Pt (mg/l) S	0	0,092	0,010
DQO (mg/l) S	-0,036	1,826	7,244
SS (mg/l) S	-0,075	0,769	0,920
AB_QB_PLC3_R2_AMONIO	-0,241	1,053	1,420
AB_QB_PLC3_R3_NITRATO	-0,278	0,897	1,412
SSLM (mg/l) Recirculación	-0,414	749,691	923013,002
AB_QB_PLC3_R4_DOSIS_RTC_P	-0,452	1,638	6,339
AB_QB_PLC3_R1_OD	-0,511	0,181	0,0467
AB_QB_PLC3_R3_OD	-0,551	0,401	0,209
AB_QB_PLC3_R3_DOSIS_RTC_P	-1,210	2,019	10,552
%MS (%)	-1,608	0,279	0,125
AB_QB_PLC3_R2_RTC_ON_N	-1,855	0,243	0,080
AB_QB_PLC3_R1_RTC_ON_N	-1,943	0,193	0,062

Tabla 5.9: Resultados obtenidos en la experimentación de XGBoost usando 30 días para predecir

Hay 12 R^2 negativos, algo que significa que ni se obtienen buenas predicciones ni el modelo logra captar la tendencia. Además, algo que llama la atención es que hay 6 ceros justos. En la Figura 5.6, se ve que lo que hace es una línea recta, lo que significa que es muy mal resultado ya que ni da buenas predicciones ni capta la tendencia, por lo que prácticamente la mitad de las variables con malos resultados.

Para ver otro ejemplo de estas predicciones se centra la atención en la diferencia entre la variable que obtiene el mejor resultado y la que peor lo obtiene. En este caso, no se

Figura 5.6: Ejemplo de predicciones con R^2 a 0

van a usar las dos últimas ya que se ha visto en anteriores apartados que estas variables tienen un tratamiento especial pues son binarias. Las predicciones se pueden ver en la Figura 5.7.

Figura 5.7: Predicciones de la mejor y la peor variable usando 30 días

En el primer caso, pese a no dar con resultado exacto, que en ocasiones lo hace, si que capta bien la tendencia. Por el otro lado, ni capta la tendencia ni da buenos resultados.

Visto esto, significa que este modelo no está dando muy buenos resultados usando 30 días para predecir, ya que prácticamente la mitad de las variables no servirían. Lo siguiente que se hizo fue tratar de mejorarlo usando 60 días. El procedimiento ha sido el mismo que con 30 días y que con el caso de las LSTM: se ha creado el dataset para entrenar usando 60 días de datos para predecir uno y se ha realizado un *GridSearch* con parámetros anteriores. En la Tabla 5.10 se ven los resultados que se han obtenido.

Variable	R^2	MAE	MSE
AB_QB_PLC3_R4_AMONIO	0,751	0,688	1,218
AB_QB_PLC3_R2_NITRATO	0,694	0,725	0,925
N-NO3 (mg/l) S	0,686	0,352	0,180
N-NH4 (mg/l) S	0,680	0,183	0,068
AB_QB_PLC3_R1_AMONIO	0,604	0,377	0,209
SSLM (mg/l) Air 4	0,587	130,557	38547,997
AB_QB_PLC3_R4_OD	0,520	0,116	0,023
AB_QB_PLC3_R1_DOSIS_RTC_P	0,515	0,768	2,405
AB_QB_PLC3_R2_DOSIS_RTC_P	0,495	0,907	2,506
SSLM (mg/l) Air 1	0,410	181,462	55374,491
AB_QB_PLC3_R4_NITRATO	0,317	1,177	2,088
AB_QB_PLC3_R2_AMONIO	0,195	0,677	0,658
AB_QB_PLC3_R2_OD	0,098	0,089	0,032
AB_QB_PLC3_R3_AMONIO	0,068	0,957	1,549
AB_QB_PLC3_R1_FOSFORO	0	0,163	0,034
AB_QB_PLC3_R2_FOSFORO	0	0,166	0,036
AB_QB_PLC3_R3_FOSFORO	0	0,229	0,077
AB_QB_PLC3_R4_FOSFORO	0	0,207	0,063
AB_QB_PLC3_R3_RTC_ON_N	0	0,170	0,036
Pt (mg/l) S	0	0,0929	0,010
SSVLM (mg/l) Air 1	-0,008	171,443	47704,481
AB_QB_PLC3_R1_NITRATO	-0,020	1,131	2,341
SSLM (mg/l) Air 3	-0,062	174,686	64929,142
AB_QB_PLC3_R4_RTC_ON_N	-0,119	0,189	0,039
DBO5 (mg/l) S	-0,167	0,481	1,085
SS (mg/l) S	-0,184	0,760	0,951
SSVLM (mg/l) Air 4	-0,215	159,281	45257,620
Caudal de purga (m3/día)	-0,270	330,915	164717,787
Nt (mg/l) S	-0,324	0,186	0,061
AB_QB_PLC3_R3_NITRATO	-0,475	0,972	1,683
AB_QB_PLC3_R1_OD	-0,485	0,177	0,049
SSLM (mg/l) Recirculación	-0,775	806,763	1031164,866
AB_QB_PLC3_R4_DOSIS_RTC_P	-0,821	1,978	7,406
AB_QB_PLC3_R3_OD	-1,012	0,308	0,137
DQO (mg/l) S	-1,041	2,418	13,098
SSLM (mg/l) Air 2	-1,053	200,352	80882,430
SSVLM (mg/l) Air 2	-1,498	191,298	70934,900
AB_QB_PLC3_R1_RTC_ON_N	-1,592	0,199	0,058
AB_QB_PLC3_R3_DOSIS_RTC_P	-1,634	2,631	13,711

%MS (%)	-1,910	0,317	0,163
AB_QB_PLC3_R2_RTC_ON_N	-2,122	0,283	0,116
SSVLM (mg/l) Air 3	-2,313	189,719	77329,700

Tabla 5.10: Resultados obtenidos en la experimentación de XGBoost usando 60 días para predecir

Los resultados son malos. Llama la atención que las mismas variables que antes tenían un R^2 de 0 en este caso también ocurra, por lo que puede ser que por la naturaleza de la serie temporal, al modelo le sea imposible captar las tendencias independientemente de los parámetros o valores usados para predecir. Además vuelven a verse muchos valores negativos que, sumados a los valores iguales a 0, hacen un total de 29 variables con malos resultados. Los valores positivos tampoco son los mejores ya que en ningún caso han conseguido llegar a un R^2 de 0.80.

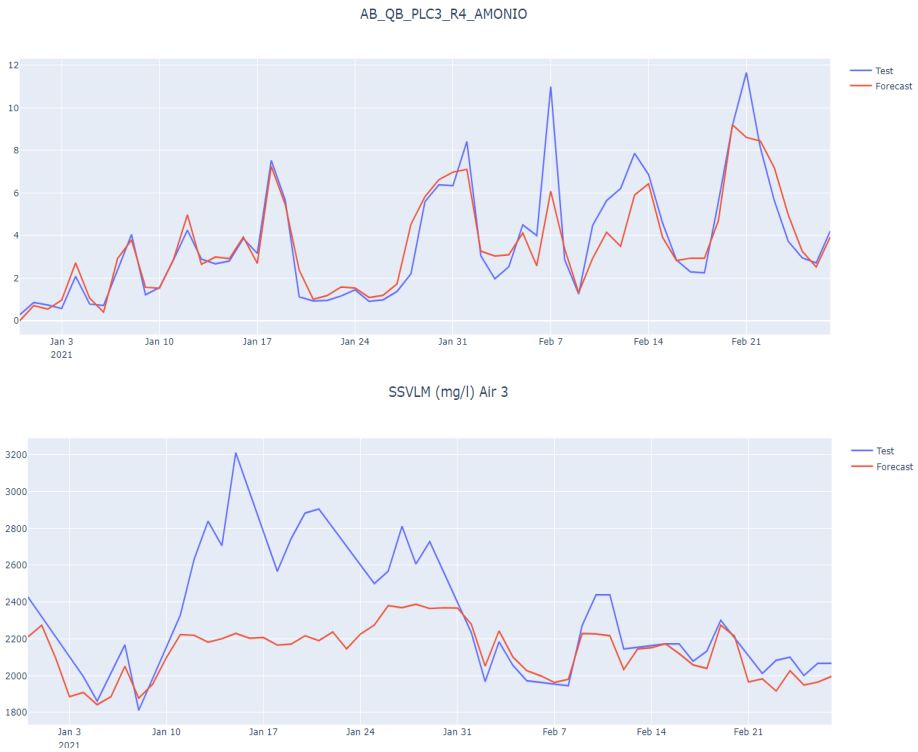


Figura 5.8: Predicciones de la mejor y la peor variable usando 60 días

En la Figura 5.8 se muestran los resultados de la mejor variable y de la peor usando 60 días. El Amonio capta muy bien la tendencia y en ocasiones consigue dar un resultado muy preciso. En el caso del SSVLM, no ha ocurrido como en otras gráficas de predicciones que se han visto. Logra captar la tendencia en alguna parte de la predicción pero esto deja de ocurrir y se aleja mucho del resultado real, haciendo que empeore.

Al igual que se ha hecho en el caso de las LSTM, se ha realizado una comparación entre los R^2 de cada variable usando 30 y 60 días respectivamente.En la Tabla 5.11 aparecen los valores con cada opción y en la última columna la diferencia. Si el valor es negativo significa que los 30 días dará mejor resultado; si es positivo, significa que 60 días dará un mejor modelo.

Variable	R^2 30 días	R^2 60 días	Incremento/Decremento
%MS (%)	-1,608	-1,910	-0,302
AB_QB_PLC3_R1_AMONIO	0,774	0,604	-0,169
AB_QB_PLC3_R1_DOSIS_RTC_P	0,670	0,515	-0,154
AB_QB_PLC3_R1_FOSFORO	0	0	0
AB_QB_PLC3_R1_NITRATO	0,237	-0,020	-0,258
AB_QB_PLC3_R1_OD	-0,511	-0,485	0,025
AB_QB_PLC3_R1_RTC_ON_N	-1,943	-1,592	0,351
AB_QB_PLC3_R2_AMONIO	-0,241	0,195	0,436
AB_QB_PLC3_R2_DOSIS_RTC_P	0,604	0,495	-0,109
AB_QB_PLC3_R2_FOSFORO	0	0	0
AB_QB_PLC3_R2_NITRATO	0,811 0,694	-0,116	
AB_QB_PLC3_R2_OD	0,400	0,098	-0,301
AB_QB_PLC3_R2_RTC_ON_N	-1,855	-2,122	-0,266
AB_QB_PLC3_R3_AMONIO	0,196	0,068	-0,127
AB_QB_PLC3_R3_DOSIS_RTC_P	-1,210	-1,634	-0,423
AB_QB_PLC3_R3_FOSFORO	0	0	0
AB_QB_PLC3_R3_NITRATO	-0,278	-0,475	-0,197
AB_QB_PLC3_R3_OD	-0,551	-1,012	-0,461
AB_QB_PLC3_R3_RTC_ON_N	0	0	0
AB_QB_PLC3_R4_AMONIO	0,798	0,751	-0,047
AB_QB_PLC3_R4_DOSIS_RTC_P	-0,452	-0,821	-0,368
AB_QB_PLC3_R4_FOSFORO	0	0	0
AB_QB_PLC3_R4_NITRATO	0,496	0,317	-0,178
AB_QB_PLC3_R4_OD	0,626	0,520	-0,105
AB_QB_PLC3_R4_RTC_ON_N	0,443	-0,119	-0,562
Caudal de purga (m3/día)	0,328	-0,270	-0,599
DBO5 (mg/l) S	0,186	-0,167	-0,354
DQO (mg/l) S	-0,036	-1,041	-1,005
N-NH4 (mg/l) S	0,518	0,680	0,162
N-NO3 (mg/l) S	0,775	0,686	-0,088
Nt (mg/l) S	0,225	-0,324	-0,549
Pt (mg/l) S	0	0	0
SS (mg/l) S	-0,075	-0,184	-0,108
SSLM (mg/l) Air 1	0,866	0,410	-0,456
SSLM (mg/l) Air 2	0,830	-1,053	-1,883
SSLM (mg/l) Air 3	0,737	-0,062	-0,799
SSLM (mg/l) Air 4	0,467	0,587	0,119
SSLM (mg/l) Recirculación	-0,414	-0,775	-0,360
SSVLM (mg/l) Air 1	0,749	-0,008	-0,758
SSVLM (mg/l) Air 2	0,547	-1,498	-2,045
SSVLM (mg/l) Air 3	0,465	-2,313	-2,779
SSVLM (mg/l) Air 4	0,509	-0,215	-0,725

Tabla 5.11: Comparación R^2 para predicciones usando 30 y 60 días

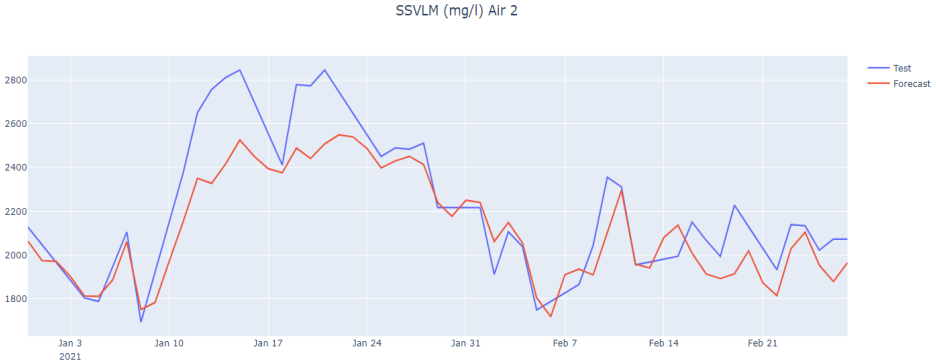
En la tabla se ve que usando 60 días da peores resultados, y en las variables que mejora, en ambos casos no son buenos resultados. Asimismo, llama la atención lo que se ha comentado antes, las mismas variables que obtienen un 0 para 30 días también lo obtienen para 60. Además, coincide con las variables de todos los reactores del fósforo, por lo

que puede tener sentido ya que si se trata de la misma variable, aunque estén en distintos reactores, como se trata de los mismos rangos de valores, suelen ser parecidas las series temporales.

En la Figura 5.9 se puede ver la variable en la que más diferencia ha habido (-2.77). En este caso, la diferencia ha sido pasar de un R^2 de 0.465 usando 30 días a uno de -2.318 usando 60 días. Ambos tienen partes en los que captan bien la tendencia, pero el de 60 días tiene una parte que se aleja mucho del valor real y el de 30 días consigue acercarse más, aún manteniéndose bastante lejano también.

En la Figura 5.10 se puede ver otro caso, donde la diferencia entre 30 y 60 días es grande, de -2.045. Esta variable es la SSVLM Air 2. Ocurre lo mismo que en la anterior descrita, solo que en este caso no capta tan bien la tendencia la de 60 días.

Figura 5.9: Comparación R^2 para SSVLM Air 3 usando 30 (arriba) y 60 (abajo) días



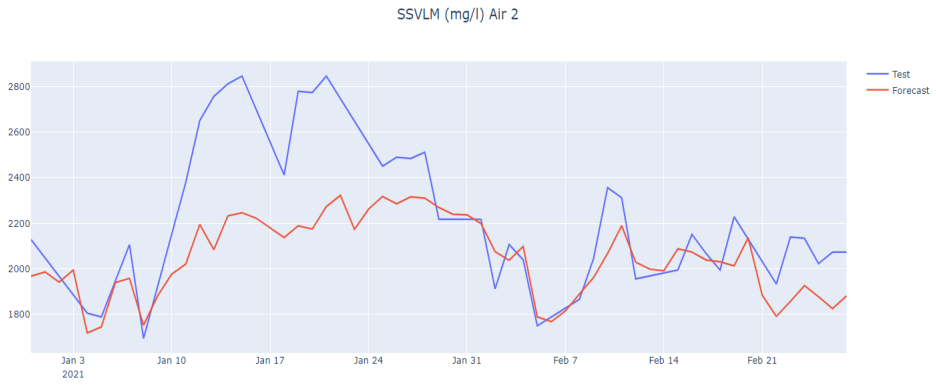


Figura 5.10: Comparación R^2 para SSVLM Air 2 usando 30 (arriba) y 60 (abajo) días

5.3 Comparación de modelos

Vista la experimentación realizada con los 3 modelos; vistos los distintos parámetros con los que se ha buscado el mejor modelo; y los resultados que se han obtenido con cada uno de ellos, se puede hacer una comparativa para saber cuál, por las características de este problema y por los resultados obtenidos, se debería utilizar.

Lo primero que habría que comentar es que el modelo XGBoost es el que peores resultados da. Prácticamente más de la mitad de las variables daban un R^2 de 0 o negativo, lo que significa que eran resultados muy malos, además esto ocurría tanto para la versión utilizando 30 días como para la versión de 60 días. Si bien es cierto, que este modelo ha sido el más rápido a la hora de entrenar, no merece la pena debido a que los resultados son mucho peores en comparación con los otros dos modelos.

Dicho esto, queda el modelo de Prophet y el modelo de LSTM. En este último se ha optado por la versión que utiliza 30 días para predecir un valor, ya que daba mejores resultados. Para poder ver las diferencias a nivel visual de los resultados se va a mostrar lo mismo que se ha mostrado con LSTM y XGBoost para comparar las versiones de 30 y 60 días. No obstante, aquí se hará entre los modelos. Estarán los R^2 de Prophet y LSTM y la diferencia, siendo esta negativa si el modelo de Prophet es mejor y si es positivo significará que el modelo de LSTM es mejor. Los resultados se reflejan en la Tabla 5.12.

Variable	R^2 Prophet	R^2 LSTM	Incremento/Decremento
%MS (%)	-0,241	-5,112	-4,871
AB_QB_PLC3_R1_AMONIO	-1,821	0,729	2,550
AB_QB_PLC3_R1_DOSIS_RTC_P	0,412	0,757	0,345
AB_QB_PLC3_R1_FOSFORO	0,656	0,295	-0,360
AB_QB_PLC3_R1_NITRATO	0,148	0,724	0,5756
AB_QB_PLC3_R1_OD	-1,525	0,5983	2,124
AB_QB_PLC3_R1_RTC_ON_N	0,097	-243,425	-243,523
AB_QB_PLC3_R2_AMONIO	-37,358	-1,188	36,170
AB_QB_PLC3_R2_DOSIS_RTC_P	0,998	0,554	-0,443
AB_QB_PLC3_R2_FOSFORO	0,553	0,153	-0,400
AB_QB_PLC3_R2_NITRATO	0,073	0,928	0,855
AB_QB_PLC3_R2_OD	0,251	0,492	0,240
AB_QB_PLC3_R2_RTC_ON_N	-0,485	-0,336	0,149

AB_QB_PLC3_R3_AMONIO	0,086	0,820	0,734
AB_QB_PLC3_R3_DOSIS_RTC_P	0,921	0,720	-0,201
AB_QB_PLC3_R3_FOSFORO	0,587	0,6611	0,073
AB_QB_PLC3_R3_NITRATO	-0,085	0,765	0,850
AB_QB_PLC3_R3_OD	-0,156	0,669	0,826
AB_QB_PLC3_R3_RTC_ON_N	-0,214	-254976,028	-254975,814
AB_QB_PLC3_R4_AMONIO	-0,478	0,755	1,234
AB_QB_PLC3_R4_DOSIS_RTC_P	0,869	0,339	-0,529
AB_QB_PLC3_R4_FOSFORO	0,594	0,683	0,089
AB_QB_PLC3_R4_NITRATO	-0,516	0,869	1,386
AB_QB_PLC3_R4_OD	0,166	0,513	0,346
AB_QB_PLC3_R4_RTC_ON_N	-0,141	0,998	1,139
Caudal de purga (m3/día)	0,306	0,757	0,450
DBO5 (mg/l) S	0,068	0,958	0,889
DQO (mg/l) S	-1,559	0,860	2,420
N-NH4 (mg/l) S	0,070	0,835	0,764
N-NO3 (mg/l) S	-0,082	0,942	1,024
Nt (mg/l) S	-0,143	0,743	0,887
Pt (mg/l) S	-0,226	0,934	1,160
SS (mg/l) S	-0,271	0,816	1,088
SSLM (mg/l) Air 1	0,749	0,967	0,217
SSLM (mg/l) Air 2	0,670	0,927	0,256
SSLM (mg/l) Air 3	0,846	0,974	0,127
SSLM (mg/l) Air 4	0,657	0,918	0,260
SSLM (mg/l) Recirculación	0,083	0,015	-0,068
SSVLM (mg/l) Air 1	0,748	0,887	0,138
SSVLM (mg/l) Air 2	0,822	0,829	0,006
SSVLM (mg/l) Air 3	0,783	0,974	0,191
SSVLM (mg/l) Air 4	0,677	0,205	-0,472

Tabla 5.12: Comparación R^2 entre el modelo Prophet y LSTM

Prácticamente en todos los casos usar LSTM mejora, esto se ve en que la mayoría de las variables tienen una diferencia positiva. Para ser exactos, 10 variables empeoran y 31 mejoran. Dentro de las con LSTM empeoran nos encontramos las ya comentadas RTC_ON, las cuales son binarias y ya se ha visto lo que ocurría con sus predicciones.

Hay casos como el Amonio 1 o el OD 1 donde el uso de las LSTM mejora en más de 2 el R^2, pasando incluso de un valor negativo a uno positivo alto como es el caso del Amonio 1. En la Figura 5.11 se ve como en el caso del Amonio 1 el LSTM hace una buena predicción. Es capaz de captar la tendencia incluso llegando a dar resultados muy próximos al valor real, mientras que el modelo Prophet no capta ni la tendencia ni es capaz ni tan siquiera de acercarse a los valores reales.

En el caso del OD 1 pasa algo parecido. El modelo Prophet no logra captar la tendencia y básicamente es una línea recta con algunas ondulaciones. Por otra parte, las LSTM si logran captar la parte inicial y luego se acercan bastante al valor real.

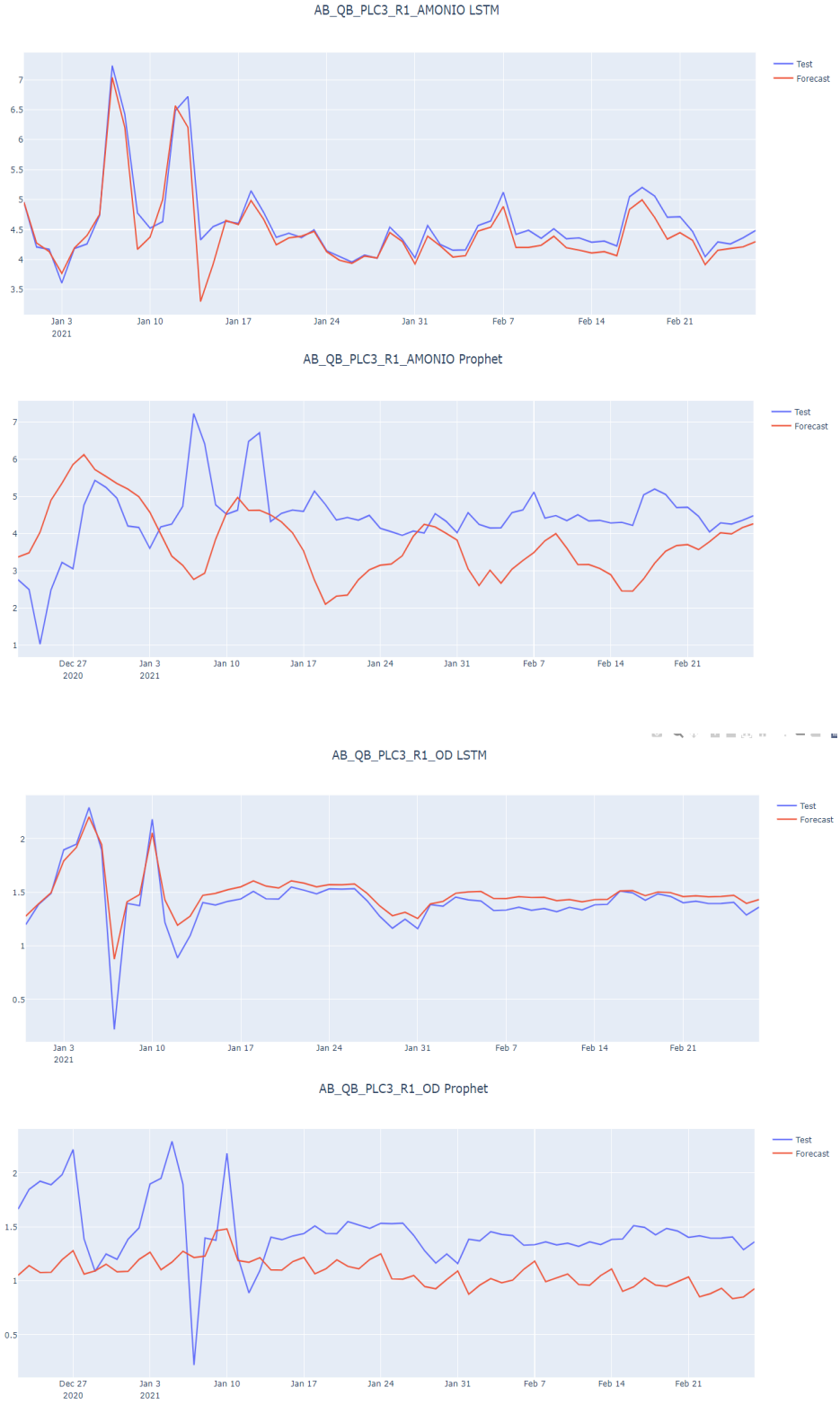


Figura 5.11: Comparación de predicciones entre Prophet y LSTM

Después de ver estos dos ejemplos de predicciones entre ambos modelos y viendo la tabla de resultados mostrada, la mejor solución sería utilizar un modelo LSTM.En casi todas las variables ofrece mejores resultados y en las que no, ocurren dos cosas: o en ambos modelos la variable tiene un R^2 negativo; o tenemos el caso de las RTC_ON, las cuales no necesariamente dan malos resultados, por lo que se podría encontrar alguna manera de normalizar la predicción para que fuese 0 o 1.

CAPÍTULO 6
Conclusiones

A lo largo de todo este trabajo se ha visto como, a partir de un conjunto de datos, se han evaluado tres tipos de modelos distintos. Se han analizado por separado y por último se han comparado para saber con cuál se obtienen mejores resultados. En primer lugar, cabe mencionar que la calidad del dato ha lastrado bastante toda la experimentación, ya que como se ha comentado, las variables a predecir contaban con hasta un 56 % de nulos y para las de input que se han usado como regresores adicionales, de hasta un 86 %. Estos datos tienen esta calidad ya que se tratan de datos reales para un problema real, por lo que haber conseguido unos resultados aceptables con esta calidad ya se podría considerar un logro.

Como se ha dicho, este es un problema real para una posible aplicación práctica, donde lo que verdaderamente interesaba no era tanto conseguir el resultado exacto, sino conseguir captar la tendencia de los químicos. De esta manera que se podrá prever que tratamientos deberá recibir el agua según dicha tendencia. Esto se ha conseguido tanto con el modelo de Prophet como con el modelo LSTM, aunque como se ha comprobado, al dar mejores resultados este último, se podrá no solo captar bien la tendencia, sino también dar un valor bastante aproximado al real, lo que vendrá muy bien en el futuro.

Hay que añadir que conforme pase el tiempo, si se siguen capturando y sacando los datos reales cada mes aproximadamente, ya que se utilizan 30 días para predecir, se podrán reentrenar los modelos de manera que progresivamente la baja calidad del dato quede paliada mediante reentrenamientos con nuevos y mejores datos hasta conseguir que el modelo vaya muy bien para todas las variables.

A parte de todo esto, merece la pena añadir que se ha realizado la experimentación con tres modelos muy distintos haciendo un *GridSearch* en cada uno de ellos, por lo que también se puede decir que se ha hecho una búsqueda exhaustiva. Lo que hace posible que el margen de mejora en este momento sea muy pequeño debido a la calidad del dato.

Respecto a posibles trabajos futuros estaría el ya comentado análisis sobre lo que supondría para los modelos aumentar la calidad del dato. Un posible trabajo futuro sería realizar un reentrenamiento en este caso del modelo LSTM, ya que es el que mejores resultados ha proporcionado. Como se ha dicho, en las variables que peores resultados se ha obtenido, también se han obtenido con el resto de modelos. Una causa de ello puede ser que las series temporales de esas variables en concreto no se puedan predecir bien debido a la interpolación que se realiza para rellenar los datos nulos. Otro posible trabajo futuro sería iniciar una línea de investigación en el caso de que fuese imposible obtener

una mejora de la calidad del dato. En ese caso habría que investigar las posibilidades de generarlos o aumentarlos de manera sintética, teniendo en cuenta que los resultados siempre serán peores que con datos reales.

Por último, se debe añadir que se entregó este proyecto en un docker con una API REST para el correcto funcionamiento del módulo. Las características de esta herramienta no han sido incluidas en el trabajo ya que se centra en la parte de investigación del mejor modelo.

Bibliografía

[1] Taylor SJ, Letham B. 2017. Forecasting at scale. *PeerJ Preprints* 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2

[2] Chollet, F., & others. (2015). Keras. https://keras.io

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

[4] Hamilton, J.D.: Time series analysis, vol. 2. Princeton university press Princeton (1994)

[5] Azoff, E.M.: Neural network time series forecasting of financial markets. John Wiley & Sons, Inc. (1994)

[6] Chakraborty, K., Mehrotra, K., Mohan, C.K., Ranka, S.: Forecasting the behavior of multivariate time series using neural networks. Neural networks 5(6), 961–970 (1992)

[7] Kaastra, I., Boyd, M.: Designing a neural network for forecasting financial and economic time series. Neurocomputing 10(3), 215–236 (1996)

[8] Chandra, R., Zhang, M.: Cooperative coevolution of elman recurrent neural networks for chaotic time series prediction. Neurocomputing 86, 116–123 (2012)

[9] Cortez, P., Rio, M., Rocha, M., Sousa, P.: Multi-scale internet traffic forecasting using neural networks and time series methods. Expert Systems 29(2), 143–155 (2012)

[10] Schulz, M., Matthies, M.: Artificial neural networks for modeling time series of beach litter in the southern north sea. Marine environmental research 98, 14–20 (2014)

[11] Alomar, M.L., Canals, V., Perez-Mora, N., Mart´ınez-Moll, V., Rossell´o, J.L.: Fpgabased stochastic echo state networks for time-series forecasting. Computational Intelligence and Neuroscience 501, 537267 (2015)

[12] Khashei, M., Bijari, M.: A novel hybridization of artificial neural networks and arima models for time series forecasting. Applied Soft Computing 11(2), 2664– 2675 (2011)

[13] Faruk, D.O.: A hybrid neural network and arima model for water quality time ¨ series prediction. Engineering Applications of Artificial Intelligence 23(4), 586–594 (2010)

[14] Kardakos, E.G., Alexiadis, M.C., Vagropoulos, S., Simoglou, C.K., Biskas, P.N., Bakirtzis, A.G., et al.: Application of time series and artificial neural network models in short-term forecasting of pv power generation. In: Power Engineering Conference (UPEC), 2013 48th International Universities'. pp. 1–6. IEEE (2013)

[15] Lin, C.J., Chen, H.F., Lee, T.S.: Forecasting tourism demand using time series, artificial neural networks and multivariate adaptive regression splines: evidence from taiwan. International Journal of Business Administration 2(2), p14 (2011)

[16] Egrioglu, E., Aladag, C.H., Yolcu, U.: Fuzzy time series forecasting with a novel hybrid approach combining fuzzy c-means and neural networks. Expert Systems with Applications 40(3), 854–857 (2013)

[17] Hu, J., Wang, J., Zeng, G.: A hybrid forecasting approach applied to wind speed time series. Renewable Energy 60, 185–194 (2013)

[18] Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks 61, 85–117 (2015)

[19] Kuremoto, T., Kimura, S., Kobayashi, K., Obayashi, M.: Time series forecasting using a deep belief network with restricted boltzmann machines. Neurocomputing 137, 47–56 (2014)

[20] Turner, J.T.: Time series analysis using deep feed forward neural networks. Ph.D. thesis, University of Maryland, Baltimore County (2014)

[21] Romeu, P., Zamora-Mart´ınez, F., Botella-Rocamora, P., Pardo, J.: Time-series forecasting of indoor temperature using pre-trained deep neural networks. In: Artificial Neural Networks and Machine Learning–ICANN 2013, pp. 451–458. Springer (2013)

[22] Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y.: Traffic flow prediction with big data: a deep learning approach. Intelligent Transportation Systems, IEEE Transactions on 16(2), 865–873 (2015)

[23] Liu, J.N., Hu, Y., You, J.J., Chan, P.W.: Deep neural network based feature representation for weather forecasting. In: Proceedings on the International Conference on Artificial Intelligence (ICAI). p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2014)

[24] Liu, J.N., Hu, Y., He, Y., Chan, P.W., Lai, L.: Deep neural network modeling for big data weather forecasting. In: Information Granularity, Big Data, and Computational Intelligence, pp. 389–408. Springer (2015)

[25] Grover, A., Kapoor, A., Horvitz, E.: A deep hybrid model for weather forecasting. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 379–386. ACM (2015)

[26] Feng, W., Han, C.: A novel approach for trajectory feature representation and anomalous trajectory detection. In: Information Fusion (Fusion), 2015 18th International Conference on. pp. 1093–1099. IEEE (2015)

[27] Längkvist, M., Karlsson, L., Loutfi, A.: Sleep stage classification using unsupervised feature learning. Advances in Artificial Neural Systems 2012, 5 (2012)

[28] Lee, H., Pham, P., Largman, Y., Ng, A.Y.: Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in neural information processing systems. pp. 1096–1104 (2009)

[29] Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Penn, G.: Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. pp. 4277–4280. IEEE (2012)

[30] Zheng, Y., Liu, Q., Chen, E., Ge, Y., Zhao, J.L.: Time series classification using multi-channels deep convolutional neural networks. In: Web-Age Information Management, pp. 298–310. Springer (2014)

[31] Gao, Y., Hendricks, L.A., Kuchenbecker, K.J., Darrell, T.: Deep learning for tactile understanding from visual and haptic data. arXiv preprint arXiv:1511.06065 (2015)

[32] Jiang, W., Yin, Z.: Human activity recognition using wearable sensors by deep convolutional neural networks. In: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference. pp. 1307–1310. ACM (2015)