



## Universidad de Granada

Escuela Técnica Superior de Ingenierías Informática y de  
Telecomunicación

### Trabajo de Fin de Máster

# **Uso de machine learning en la determinación de niveles de suciedad en sistemas fotovoltaicos a partir de parámetros ambientales**

**Alumna:** Luíza Araujo Costa Silva

**Tutora:** Prof<sup>a</sup>. Dr<sup>a</sup>. María del Carmen Pegalajar Jiménez

Departamento de Ciencias de la Computación e Inteligencia Artificial

**Julio / 2021**



UNIVERSIDAD  
DE GRANADA

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Departamento de Ciencias de la Computación e Inteligencia Artificial

### Declaración de Supervisión

La Prof. Dr<sup>a</sup>. D<sup>a</sup>. María del Carmen Pegalajar Jiménez del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada,

CERTIFICA:

Que la memoria titulada:

***"Uso de machine learning en la determinación de niveles de suciedad en sistemas fotovoltaicos a partir de parámetros ambientales"***

ha sido realizada por D<sup>a</sup>. Luíza Araujo Costa Silva bajo mi dirección en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada para optar al Máster en Ciencia de Datos e Ingeniería de Computadores por la Universidad de Granada.

Granada, 05 de Julio de 2021.

La tutora:

PEGALAJAR  
JIMENEZ MARIA  
CARMEN -  
26011572K

Firmado digitalmente por PEGALAJAR  
JIMENEZ MARIA CARMEN -26011572K  
Nombre de reconocimiento (DN): c=ES,  
serialNumber=IDCES-26011572K,  
givenName=MARIA CARMEN,  
sn=PEGALAJAR JIMENEZ cn=PEGALAJAR  
JIMENEZ MARIA CARMEN -26011572K  
Fecha: 2021.07.05 07:09:45 +0200

María del Carmen Pegalajar Jiménez

## Resumen

El constante crecimiento y desarrollo urbano e industrial de la sociedad, ha provocado un aumento de la demanda de más fuentes energéticas renovables. Actualmente una de las grandes apuestas de energía limpia, es la energía fotovoltaica (FV) que es generada a través de luz solar en los paneles solares fotovoltaicos.

Sin embargo, varios factores influyen en la capacidad de generación de energía FV. Uno de esos factores, conocido como *soiling*, es la suciedad que se deposita naturalmente sobre los paneles debido a factores meteorológicos y atmosféricos. El nivel de suciedad de un módulo es medido por una tasa conocida como *Soiling Ratio*, o simplemente “SR”. Diversos modelos matemáticos son utilizados para estimar las pérdidas por soiling.

Utilizándose datos de parámetros ambientales de satélites y datos de mediciones de SR realizadas en una estación de medición en Jaén (España), este estudio se propone a construir modelos que utilizan técnicas de *machine learning* para predecir los niveles de suciedad con base en la serie temporal de dichas medidas. En seguida, se verifica la calidad alcanzada por los modelos en la predicción del SR y se comparan los resultados obtenidos con resultados de modelos matemáticos actualmente utilizados.

Los datos obtenidos fueron divididos en tres conjuntos de entradas distintas, juntamente con ventanas deslizantes (las medidas de SR de los días anteriores): 1) parámetros seleccionados a través del algoritmo *Boruta*, 2) parámetros utilizados en los modelos matemáticos y 3) sólo las ventanas deslizantes.

A cada una de esas entradas fueron aplicados cinco métodos de regresión: Regresión Lineal, Árbol de Decisión, *Random Forest*, Perceptrón Multicapa (*MLP*) y *LSTM*. En los experimentos realizados se variaron los hiperparámetros de los algoritmos y el número de días en las ventanas deslizantes.

Tras los experimentos se presenta el análisis de los resultados obtenidos para cada uno de los grupos de entradas y modelos aplicados. Además, teniendo en cuenta todos los experimentos realizados, se concluye que los modelos con Regresión Lineal y *MLP* con 1 capa oculta y 30 neuronas, ambos utilizando el conjunto de datos 3 (sólo con las ventanas deslizantes) y ventanas de 1 día, han aportado los mejores resultados al problema planteado. El tercero mejor modelo es el *MLP* con 1 capa oculta y 30 neuronas, utilizando el conjunto de datos con las variables utilizadas por los modelos matemáticos y también ventana de 1 día.

Finalmente, estos tres mejores modelos son utilizados para hacer predicciones secuenciales. Es decir, se utilizan las salidas de las predicciones de los días anteriores para realizar las predicciones futuras. Los dos predictores que utilizan el conjunto de datos sólo con las ventanas deslizantes tienen resultados similares con buenos resultados, con una capacidad de predicción de hasta 5 días futuros con bajo RMSE (media de un 0.002). De igual forma en las predicciones a más largo horizonte, son satisfactorias con un RMSE de un 0.014 para predicciones de 90 días, frente a un error medio de un 0.026 de los modelos matemáticos utilizados. El predictor que utiliza el conjunto de datos con las variables utilizadas por los modelos matemáticos, tiene resultados menos preciso (una media de un 0.014 entre todas sus predicciones) sin embargo aún bastante mejores que el error de los modelos matemáticos.

Además, no se ha podido ratificar si los mejores resultados son obtenidos sin el uso de los parámetros ambientales por el hecho de que datos de satélite son más susceptibles a errores, o si realmente, se tornan ruido para el predictor al ser datos externos a la serie temporal de mediciones. Más investigación se hace necesaria para certificar su influencia en los resultados de los modelos de este trabajo.

**Palabras clave:** *machine learning*, redes neuronales artificiales, regresión lineal, MLP, *soiling*, suciedad, parámetros ambientales, fotovoltaica.

## Abstract

The constant and actual urban and industrial growth of our society have raised the demand for more renewable sources of energy. One of the biggest bets on clean energy sources is photovoltaic energy (PV) which is generated from solar light on photovoltaic solar panels.

However, many factors impact the PV power capacity. One of these, known as soiling, is the dirt that naturally settles over FV panels due to atmospheric and meteorological events. The measure of soiling level is known as Soiling Ratio, or simply “SR”. Many mathematical models are used to estimate power losses by soiling.

Using data of environmental parameters from satellites and SR measured in a soiling station in Jaén (Spain), this study proposes to build machine learning models to predict soiling based on the time series data of these measurements. Thus, the achieved metrics are evaluated and compared to the ones achieved by the mathematical models currently used.

The acquired data is split into 3 different inputs, together with sliding windows (the measured SR of previous days): 1) features selected by Boruta algorithm, 2) features used on the mathematical models and 3) only sliding windows.

To each of these inputs five regression methods were applied: Linear Regression, Decision Tree, Random Forest, Multi-layer Perceptron and LSTM. Through the experimentations many hyperparameters were tested.

After the experiments, an analysis of the results is presented to each one of the inputs and algorithms. Furthermore, when analyzing all experiments, it is stated that the Linear Regression and the MLP with 1 hidden layer and 30 neurons models have the best results for the proposed problem. In addition, the third best model is the MLP model, also with 1 hidden layer and 30 neurons, but using the Input with features used on the mathematical models.

Finally, these three best models are used to make sequential predictions. Meaning, the output of previous days predictions is used as input for future predictions. Both predictors that use the Input with only sliding windows present similar and great results, with an ability to predict until 5 future days with low RMSE (0.002 on average). Even the larger term predictions have satisfactory outcomes: RMSE of 0.0014 on the prediction of 90 days ahead, comparing to the 0.026 average error obtained with the mathematical models. The predictor that uses Input with features used on the mathematical models is less precise (an average of 0.014 among all predictions), however better than the error gotten with the mathematical models.

In addition, it could not be confirmed if the best results are obtained without the use of environmental parameters because of the fact that satellite data are more susceptible to errors, or if they actually became noise to the predictor, for being external data to the time series. Further research is needed to certify their influence on the results.

**Keywords:** machine learning, artificial neural network, linear regression, MLP, soiling, environmental parameters, photovoltaic.

## Agradecimientos

Doy gracias a la Universidad de Granada por haberme brindado la oportunidad de abrir un nuevo camino en mi jornada profesional y a los profesores de este máster por proporcionarme tanto conocimiento.

A mi tutora, María del Carmen, siempre tan receptiva en mis momentos de dudas. Gracias por su orientación y confianza en mi trabajo. Por su generosidad al compartir su experiencia, y por ser un ejemplo de profesional y persona: extremadamente talentosa y amable.

Doy gracias a los investigadores del CECTEMA de la Universidad de Jaén, que gentilmente facilitaron los datos para la realización de este estudio, Eduardo Fernández y João Gabriel Bessa. En especial, gracias a João, por aportar tanto conocimiento técnico sobre el tema del *soiling*, por ayudarme siempre con tan buena voluntad y por su acompañamiento constante en este trabajo.

Además, también doy gracias a mis padres, a mi hermana y a mi familia que, aunque a distancia, siempre estuvieron apoyándome de todas las maneras posibles. También a mis amigos, que han estado a mi lado para celebrar los logros y confortarme en momentos difíciles.

## Índice

Estructura del trabajo .....	13
Objetivos .....	14
1. INTRODUCCIÓN .....	15
2. LA TECNOLOGÍA SOLAR FOTOVOLTAICA .....	17
2.1. Energía Fotovoltaica .....	18
2.2. <i>Soiling</i> : La suciedad en los módulos fotovoltaicos .....	21
2.3. El uso de Machine Learning en estudios del <i>soiling</i> .....	23
3. ANÁLISIS EXPLORATORIO DE LOS DATOS .....	26
3.1. Informaciones sobre la base de datos .....	27
3.1.1. Bases de datos originales .....	27
3.1.1.1. Monitorización de <i>Soiling</i> .....	27
3.1.1.2. Datos Meteorológicos .....	28
3.1.1.3. Diagnósticos de Aerosoles .....	29
3.1.1.4. Índice de Mezcla de Aerosoles .....	29
3.2. Preparación del conjunto de datos .....	30
3.3. Análisis de los datos.....	31
4. PREPROCESAMIENTO DE LOS DATOS .....	41
4.1. Imputación de los valores no medidos.....	41
4.2. Normalización de los datos .....	41
4.3. Ventana deslizante.....	42
5. EXPERIMENTOS REALIZADOS .....	43
5.1. Diseño de los experimentos .....	44
5.1.1. Modelos Matemáticos .....	44
5.1.2. Selección de Características .....	46
5.1.3. Métricas utilizadas.....	48
5.2. Resultados obtenidos.....	49
5.2.1. Regresión Lineal .....	50
5.2.2. Árbol de Regresión .....	51
5.2.3. Random Forest .....	53
5.2.4. MLP – Perceptrón Multicapa .....	55
5.2.5. LSTM .....	57
6. ANÁLISIS DE LOS RESULTADOS .....	62
6.1. Análisis por grupos de entrada.....	62

6.2. Análisis por algoritmo .....	64
6.3. Análisis entre todos los experimentos .....	70
7. PREDICCIONES SECUENCIALES.....	73
7.1. Modelo con Regresión Lineal – Input_3 .....	74
7.2. Modelo con Perceptrón Multicapa – Input_3.....	77
7.3. Modelo con Perceptrón Multicapa – Input_2.....	80
7.4. Comparación entre las predicciones secuenciales .....	83
8. CONCLUSIONES.....	87
9. LÍNEAS DE FUTURO.....	88
10. GLOSARIO .....	89
11. BIBLIOGRAFÍA .....	90
ANEXO I: Diccionario de datos de la base utilizada .....	95

## Índice de Figuras

Figura 1.	Módulos en una planta fotovoltaica .....	17
Figura 2.	Evolución de la generación de energía fotovoltaica.....	19
Figura 3.	Fuentes de la energía eléctrica generada en marzo 2021 en España .....	19
Figura 4.	Evolución del precio medio de módulos FV solares.....	20
Figura 5.	Equipo de medición del nivel de suciedad en Jaén. ....	22
Figura 6.	Datos estadísticos de la variable “SR” .....	31
Figura 7.	Lista de <i>outliers</i> de la variable “SR”.....	32
Figura 8.	Datos estadísticos de la variable “SR” sin <i>outliers</i> .....	33
Figura 9.	Cuartiles de la variable “SR” .....	33
Figura 10.	Distribución general de los valores de “SR” .....	34
Figura 11.	Cuartiles por mes del “SR” durante el año de 2019 .....	35
Figura 12.	Cuartiles por mes del “SR” durante el año de 2020 .....	35
Figura 13.	Cuartiles por mes del “SR” durante el año de 2021 .....	36
Figura 14.	Evolución diaria del “SR” durante el año.....	37
Figura 15.	Evolución mensual del “SR” durante el año .....	37
Figura 16.	Matriz de correlación entre todas las variables .....	39
Figura 17.	Gráfico de correlación entre SR y otras variables .....	40
Figura 18.	Gráfico de evolución de las métricas por número de ventanas .....	42
Figura 19.	Gráfico de evolución de las métricas para hiperparámetros MLP .....	55
Figura 20.	Gráfico de evolución de las métricas por capas ocultas y neuronas en MLP .....	56
Figura 21.	Gráfico del ajuste del mejor modelo de Regresión Lineal.....	65
Figura 22.	Gráfico del ajuste del mejor modelo con Árbol de Decisión.....	66
Figura 23.	Gráfico del ajuste del mejor modelo con Random Forest .....	67
Figura 24.	Gráfico del ajuste del mejor modelo con MLP .....	69
Figura 25.	Gráfico del ajuste del mejor modelo con LSTM .....	70
Figura 26.	Ejemplo de entradas y salidas en una predicción secuencial de 90 días .....	73
Figura 27.	Gráfico de evolución: RMSE en la predicción secuencial de hasta 90 días con Regresión Lineal.....	76
Figura 28.	Gráfico de evolución: RMSE en la predicción secuencial de hasta 5 días con Regresión Lineal .....	77
Figura 29.	Gráfico de evolución: RMSE en la predicción secuencial de hasta 90 días con Perceptrón Multicapa .....	79
Figura 30.	Gráfico de evolución: RMSE en la predicción secuencial de hasta 5 días con Perceptrón Multicapa .....	80

Figura 31.	Gráfico de evolución: RMSE en la predicción secuencial de hasta 90 días con Perceptrón Multicapa para Input_2 .....	83
Figura 32.	Gráfico comparativo entre modelos: predicciones secuenciales hasta 90 días .....	84
Figura 33.	Gráfico comparativo entre modelos con Input_3: predicciones secuenciales hasta 7 días .....	84

## Índice de Tablas

Tabla 1. Resultados obtenidos en estudios que aplican Redes Neuronales Artificiales en la predicción del nivel de suciedad. Tomada de [8].....	23
Tabla 2. Resultados de métricas aplicadas a los modelos matemáticos utilizados.....	45
Tabla 3. Resultados obtenidos con la Regresión Lineal .....	50
Tabla 4. Resultados obtenidos con el Árbol de Decisión .....	51
Tabla 5. Resultados obtenidos con el Random Forest .....	53
Tabla 6. Resultados obtenidos con Perceptrón Multicapa .....	57
Tabla 7. Resultados obtenidos con LSTM .....	58
Tabla 8. Mejores resultados para Input_1 .....	62
Tabla 9. Mejores resultados para Input_2 .....	63
Tabla 10. Mejores resultados para Input_3 .....	63
Tabla 11. Mejores resultados para el algoritmo de Regresión Lineal.....	64
Tabla 12. Mejores resultados para el algoritmo de Árbol de Decisión.....	65
Tabla 13. Mejores resultados para el algoritmo Random Forest .....	67
Tabla 14. Mejores resultados para el algoritmo MLP - Perceptrón Multicapa .....	68
Tabla 15. Mejores resultados para el algoritmo LSTM .....	69
Tabla 16. Mejores resultados entre todos los experimentos realizados y modelos matemáticos .....	71
Tabla 17. Resultados de la predicción secuencial para el modelo con Regresión Lineal .....	75
Tabla 18. Resultados de la predicción secuencial para el modelo con MLP .....	78
Tabla 19. Resultados de la predicción secuencial para el modelo con MLP con Input_2.....	81
Tabla 20. Comparativo de resultados de la predicción secuencial .....	85

## Estructura del trabajo

Este trabajo está dividido en 9 capítulos. En el capítulo 1 se presenta una introducción a la importancia del uso de las energías renovables, su crecimiento y algunos de los factores que interfieren en su eficiencia. Se han nombrado las técnicas de análisis utilizadas y a continuación los objetivos del trabajo.

En el capítulo 2 se presenta el estado del arte de la tecnología fotovoltaica, su historia y evolución, y luego se describen de forma un poco más dedicada el concepto de *soiling* y sus particularidades e impactos en la generación de energía eléctrica, además de una visión sobre el uso de machine learning en este campo de estudio.

En el capítulo 3 se expone el análisis exploratorio de los datos, donde se explican el contexto de obtención de los datos, sus características cualitativas y cuantitativas.

En el capítulo 4 se describen los procedimientos de preprocesamiento de los datos.

El capítulo 5 expone los experimentos realizados, desde su diseño, los métodos elegidos y sus particularidades, las métricas utilizadas y los resultados obtenidos para cada algoritmo utilizado.

En el capítulo 6 se analizan los resultados obtenidos con tres sesgos distintos: por tipo de entrada utilizada, por algoritmo y entre todos los experimentos realizados.

En el capítulo 7 se describen los experimentos de predicciones secuenciales realizadas después del análisis de resultados, realizándolas con los mejores modelos presentados en el capítulo anterior.

En el capítulo 8 se exponen las conclusiones generales inferidas en este estudio.

Finalmente, en el capítulo 9 se describen las posibilidades de trabajos futuros, que se pueden desarrollar a partir de este trabajo realizado.

## Objetivos

En el actual Trabajo de Fin de Máster se estipularon los siguientes objetivos de proyecto:

- Estudiar y presentar un panorama general del estado del arte de la energía fotovoltaica, el *soiling* (niveles de suciedad) y el uso de modelos de *machine learning*.
- Realizar un análisis exploratorio de los datos facilitados para el estudio.
- Aplicar técnicas de *machine learning* para predecir el nivel de suciedad (SR) en placas fotovoltaicas utilizando los datos de satélites de parámetros ambientales.

## 1. INTRODUCCIÓN

El actual crecimiento y desarrollo urbano e industrial de la sociedad, demanda cada vez más fuentes energéticas, llevando también al aumento de la emisión de sustancias que contaminan el medio ambiente y que son grandes responsables por los temidos cambios climáticos.

Actualmente un 63% de las fuentes de energía primaria para generación de energía eléctrica en el mundo provienen de los combustibles fósiles: el carbón (36%), el gas natural (23%) y el petróleo (3%). [1]

La energía generada por fuentes de energía renovables, es conocida como energía limpia. Este término se refiere a una fuente de energía que no libere gases u otros residuos perjudiciales o que contribuyan para el llamado efecto invernadero durante su proceso de producción y su consumo. [2]

El uso de energías renovables es una solución que contribuye ampliamente no sólo para la reducción de los efectos nocivos al medio ambiente y a la salud, como también para el desarrollo económico y social sostenible.

Desde hace algunos años, muchos países hacen parte de acuerdos globales para la reducción de emisión de sustancias tóxicas. España es uno de ellos y tiene datos bastante expresivos. Según un informe de la Red Eléctrica de España [3], en marzo de 2021, el 76,9% la producción eléctrica española ha sido generada por las consideradas fuentes limpias de energía.

La energía solar es, como su propio nombre sugiere, obtenida del Sol, y sus principales tecnologías son la solar térmica, que se utiliza del calor del Sol, y la solar fotovoltaica, que se utiliza de la luz solar. A la energía generada por los paneles fotovoltaicos se denomina energía fotovoltaica (FV).

La radiación solar en la superficie terrestre es más que suficiente para suprir el consumo mundial de energía. Sin embargo, varios factores influyen en la capacidad de generación de energía, como la latitud de una ubicación, las estaciones del año y las condiciones climáticas y atmosféricas.

Uno de esos factores, es conocido por los expertos del área como *soiling*, que puede ser traducido, de manera muy simplificada como la “suciedad” de las placas FV. El nivel de suciedad de un módulo es medido por una tasa conocida como *Soiling Ratio*, o simplemente “SR”.

Tras ese panorama, y teniendo en cuenta la necesidad de que las energías renovables sean cada vez más eficientes y tengan un coste cada vez más bajo, mientras el coste de los combustibles fósiles tiende a seguir en la dirección opuesta, se propone en este trabajo un estudio de la relación entre el nivel de suciedad – el *Soiling Ratio* (SR) – y las condiciones de los parámetros atmosféricos provenientes de satélites en la capacidad de generación de energía en placas fotovoltaicas.

Un gran número de estudios ha sido realizado en todo el mundo para desarrollar maneras de predecir las pérdidas energéticas debido a esa suciedad [4]. Tanto modelos matemáticos [5]–[7] como modelos que utilizan técnicas de *machine learning* son construidos con la finalidad de minimizar las pérdidas energéticas y financieras y servir como soporte para acciones que las puedan mitigar.

Estudios indican que el uso del *machine learning* en la predicción del soiling son altamente benéficos por su gran capacidad de tratar de sistemas complejos y sin soluciones analíticas lineales [8]. Siendo este un campo de creciente atención en la comunidad académica, se resalta la importancia de la investigación para el desarrollo de técnicas de mitigación nuevas y cada vez más eficientes. [9]

Para este estudio, utilizaremos los datos facilitados por el Centro de Estudios Avanzados en Ciencias de la Tierra, Energía y Medio Ambiente (CEACTEMA) de la Universidad de Jaén y datos de satélites de la NASA.

A dichos datos, se aplican técnicas de machine learning con la finalidad de verificar la calidad alcanzada por los modelos en la predicción del *Soiling Ratio* y comparar los resultados obtenidos con modelos ya publicados.

Además, se propone realizar predicciones de días futuros con los mejores modelos obtenidos.

## 2. LA TECNOLOGÍA SOLAR FOTOVOLTAICA

Se denomina energía fotovoltaica a la transformación de la radiación solar en electricidad, que es generada en dispositivos conocidos como paneles o módulos fotovoltaicos. En esos paneles, se produce la energía eléctrica a través de la excitación de los electrones de un dispositivo semiconductor. [10]

En la **Figura 1** se ve un ejemplo de una planta o granja fotovoltaica, que es un conjunto de módulos fotovoltaicos para el suministro de energía eléctrica.



**Figura 1. Módulos en una planta fotovoltaica**  
Tomada de [11]

El término “fotovoltaica” tiene su origen del griego “*Phos*”, que significa “luz”, y “*Volt*”, la unidad de medida de fuerza eléctrica que tiene su nombre originado del físico italiano Alessandro Volta, el creador de las pilas. También es comúnmente presentado en la literatura con la sigla “FV”.

La producción de energía solar (llamado de efecto fotovoltaico) utiliza paneles solares producidos con material semiconductor para, cuando los fotones (o las partículas de luz solar) inciden, los electrones del material semiconductor entraren en movimiento, generando así electricidad.

La energía solar es generada por estos paneles solares y llevada a un equipamiento llamado inversor solar, que transforma la corriente eléctrica continua en una corriente alterna, pudiendo ser distribuida y utilizada por los equipamientos. [12]

A continuación, pasaremos a describir el estado del arte de la energía fotovoltaica, un poco de su historia, su evolución y *status quo* en España. En seguida, se explica el concepto de soiling, su impacto y su relación con la generación de energía fotovoltaica. Y finalmente, se hará una breve revisión de los trabajos existentes en la literatura relacionados con el uso de técnicas de machine learning en los estudios del *soiling*.

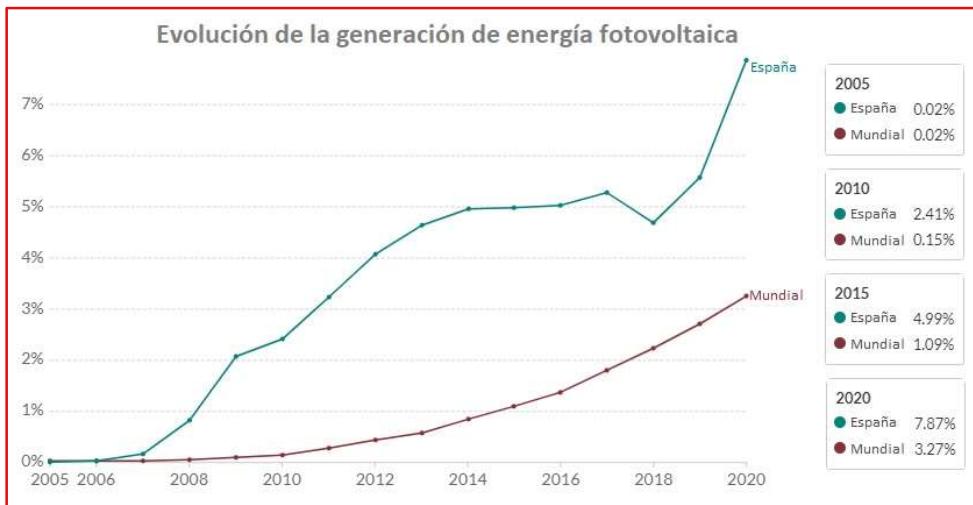
## 2.1. Energía Fotovoltaica

La historia de cómo se inventó de la energía fotovoltaica comienza en 1839, cuando el físico francés Alexandre Edmond Becquerel descubrió el efecto fotovoltaico. Más de 40 años después, en 1883, el inventor estadounidense Charles Fritts creó la primera celda fotovoltaica, utilizándose de selenio revestido de oro. Fritts fue la primera persona en generar una corriente constante, con una conversión eléctrica de 1% de eficiencia.

También en los Estados Unidos, pasado más de medio siglo, en 1954, el ingeniero Russel Ohl registra la primera patente de un sistema fotovoltaico que se asemeja a los sistemas que se conocen en la actualidad. Un año más tarde, en un proyecto de los científicos del laboratorio Bell Labs, Calvin Fuller, Gerald Pearson y Daryl Chapin, las celdas fotovoltaicas fueron utilizadas por primera vez como fuente de alimentación de una red telefónica en el estado de Georgia.

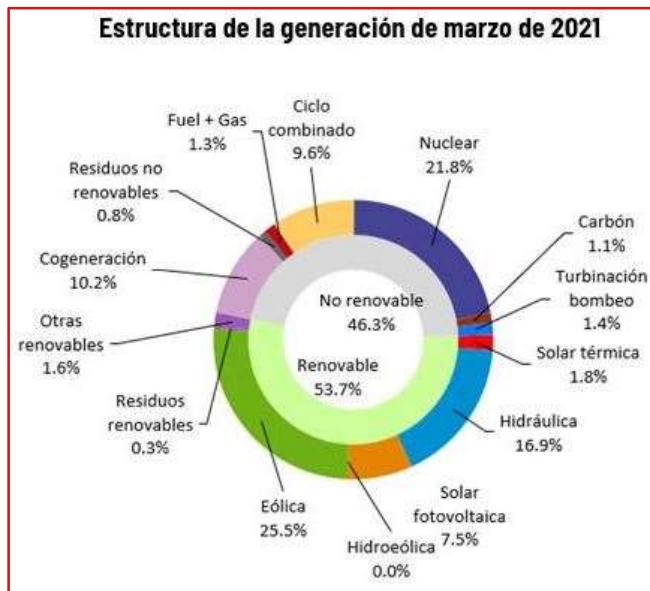
En 1958, se inició la utilización de los paneles solares para suministrar electricidad a satélites geoestacionarios de comunicaciones. El satélite Vanguard I fue lanzado con un panel solar para la alimentación energética de su radio durante el viaje. Con el éxito del proyecto, en la década de los años sesenta, se empezaron a implementar los primeros sistemas fotovoltaicos terrestres, para uso de la población en domicilios, comercios y hasta autobuses, barcos y aviones. [13]

Como podemos observar en la **Figura 2**, la producción de energía fotovoltaica ha crecido considerablemente en los últimos años a nivel mundial, y de forma notable en España.



**Figura 2. Evolución de la generación de energía fotovoltaica**  
Adaptada de [1]

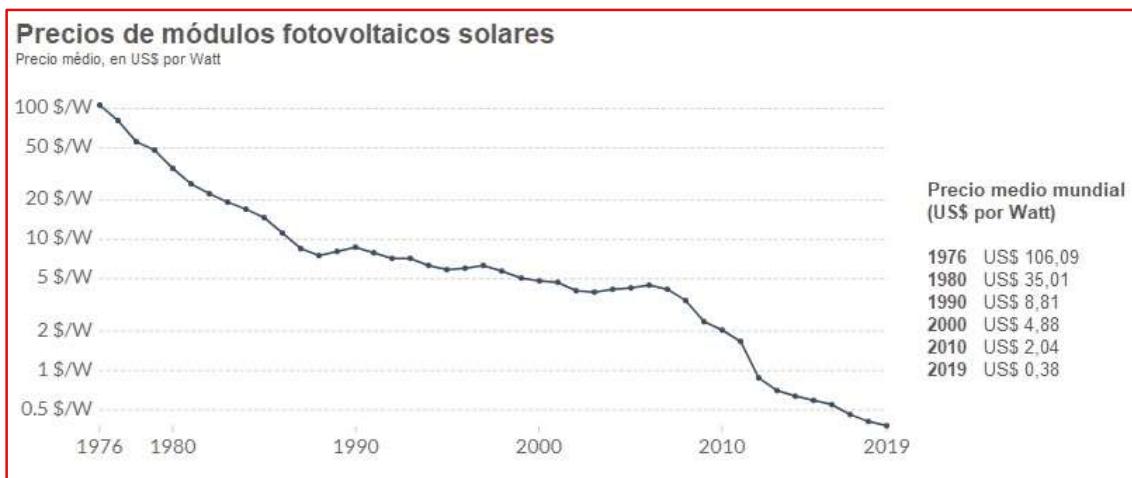
En la **Figura 3**, observamos los datos de la Red Eléctrica de España [3] que apuntan que, en marzo de 2021, la energía solar FV fue responsable por un 7,5% del total de energía generada en el país, siendo un 56% más que en el mismo período de 2020.



**Figura 3. Fuentes de la energía eléctrica generada en marzo 2021 en España**  
Tomada de [3]

En abril de 2020, el grupo Iberdrola ha puesto en marcha la granja fotovoltaica de Núñez de Balboa, en Badajoz. El proyecto, que se ha convertido en la mayor planta de Europa, ocupa una superficie cercana a las 1.000 hectáreas y tiene capacidad de suministrar energía limpia a 250.000 personas. [11]

Un importante factor para la ascensión y viabilidad del amplio uso de la energía FV ha sido la drástica reducción de su coste durante los años siguientes. [14] Podemos observar en la **Figura 4**, la caída del precio medio mundial de los módulos solares.



**Figura 4. Evolución del precio medio de módulos FV solares**  
Adaptada de [15]

En 1976, el precio aproximado de un módulo solar FV era de US\$106/W (dólares estadounidenses por Watt). En los siguientes años de 1980, 1990, 2000 y 2010 baja a aproximadamente US\$35/W, US\$8/W, US\$5/W y US\$2/W, respectivamente. Hasta llegar al coste de US\$0.38/W en 2019, representando un descenso del 99,6% del precio. [15]

La luz solar es considerada una fuente inagotable de energía. Sin embargo, esa radiación no alcanza de manera uniforme toda la superficie del planeta y muchos son los factores que influyen en la capacidad de generación de energía. Hay que tenerse en cuenta estas cuestiones, pues pueden suponer una pérdida energética y además financiera importante.

La ubicación de una granja solar, por ejemplo, puede afectar la cantidad de radiación solar recibida. También la posición en que está un panel, afecta el ángulo de incidencia

de los rayos solares. Factores como las estaciones del año y las condiciones climáticas y atmosféricas también pueden beneficiar o perjudicar la eficiencia de generación.

## 2.2. ***Soiling: La suciedad en los módulos fotovoltaicos***

De acuerdo con Bessa, *soiling* es “*el proceso en el cual la suciedad, el polvo y contaminantes orgánicos e inorgánicos se depositan en una superficie que, en el caso de los paneles fotovoltaicos, impide la transmisión de la luz en el módulo*”. [16]

Los primeros estudios sobre pérdidas de energía debido al *soiling* datan de los años setenta [17], como el estudio publicado en 1974 por Garg, nombrado “*Effect of dirt on transparent covers in flat-plate solar energy collectors*” [18]. Sin embargo, sus efectos siguen siendo tan relevantes para la comunidad fotovoltaica que, de 2012 a 2017, hubo un crecimiento superior a un 200% en las publicaciones científicas relacionadas al tema. [19]

Según Comerio [14], estudios demuestran que el *soiling* puede reducir hasta un 25% del rendimiento energético de las instalaciones fotovoltaicas. Además, apunta que hay un riesgo de daños a los módulos parcialmente sombreados, ya que la potencia eléctrica producida que no es entregada para consumo es disipada en las celdas afectadas. Ese efecto puede calentar en demasiá determinados puntos del panel y causar una ruptura en su cristal. De esa manera, el acumulo de esas substancias en los módulos FV tiene un impacto relevante en la generación de energía fotovoltaica.

Estudios recientes apuntan que en 2018 el *soiling* ha causado pérdidas financieras de 3 a 5 mil millones de euros, referentes a la pérdida de cerca del 3% a 4% de la energía fotovoltaica generada globalmente. Se estima que, en 2023, con el creciente uso de la energía fotovoltaica, estas pérdidas pueden llegar a hasta 7 mil millones de euros por año. [4]

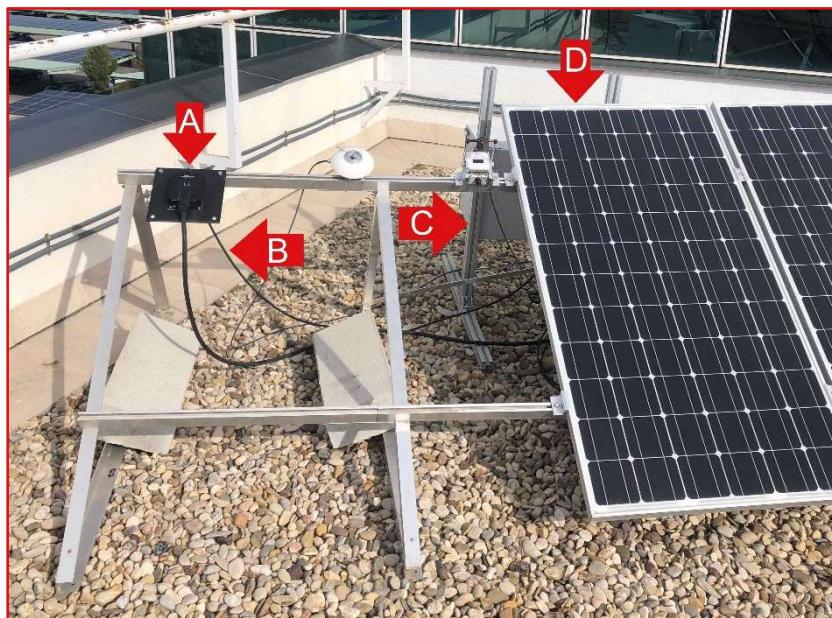
Hay muchas publicaciones científicas que estudian diversos métodos de mitigación del impacto del *soiling* en la generación energética. En ese trabajo se enfocarán los

impactos de los parámetros ambientales, como la lluvia, la temperatura y la contaminación atmosférica.

El nivel de suciedad en un módulo fotovoltaico es medido por una tasa conocida entre los expertos del asunto como *Soiling Ratio*, o por su sigla “SR”. Esa tasa se refiere a la cantidad de energía que no se ha generado en un módulo debido al acumulo de suciedad en su superficie.

En la teoría, el SR varía de 0 a 1, siendo 1 la representación de un módulo que se encuentre perfectamente limpio y generando el 100% de la energía posible. De esa manera, un SR de 0.95, indica que debido al acumulo de suciedad en su superficie, el módulo registra una pérdida de un 5% de la energía que podría ser generada.

Observamos en la **Figura 5** el equipo de medición del SR instalado en la azotea del edificio del CECTEMA, en Jaén. Se trata exactamente del equipo que realiza las mediciones utilizadas en este estudio.



**Figura 5. Equipo de medición del nivel de suciedad en Jaén.**  
Cortesía de CECTEMA/UJA

La estructura marcada como (A) en dicha ilustración es una célula fotovoltaica de referencia que se limpia diariamente por las tuberías etiquetadas como (B). La estructura (C) es la caja donde se encuentran el tanque de agua para la limpieza y el sistema de adquisición de datos. Finalmente, la estructura marcada como (D) es el

módulo fotovoltaico de silicio que sufre el acumulo de suciedad. Tanto la célula (A) como el módulo (D) son fabricado con el mismo material.

Por lo tanto, el SR es la razón de la salida eléctrica (*electrical output*) producida por (D) – sucio – y (A) – limpio –, como expuesto en la fórmula (1).

$$SR = \frac{Z_D}{Z_A} \quad (1)$$

### 2.3. El uso de Machine Learning en estudios del soiling

Younis y Alhorrr [8] destacan el uso de redes neuronales artificiales como una solución eficiente para problemas con múltiples variables. En su recién publicado artículo, realizaron una revisión bibliográfica de las publicaciones que abordan el uso de redes neuronales artificiales en la predicción de los niveles de suciedad en la performance FV.

En la **Tabla 1**, tomada de dicho estudio, se presenta un resumen de las configuraciones de la arquitectura de las redes neuronales utilizadas en el tema por diversos autores.

**Tabla 1.** Resultados obtenidos en estudios que aplican Redes Neuronales Artificiales en la predicción del nivel de suciedad. Tomada de [8].

Author	Modeled parameter	Training algorithm	Number of neurons		Performance Metrics		
			Input	Hidden	R <sup>2</sup> (%)	MSE (%)	MAPE (%)
(Javed et al., 2017)	daily change in the Cleanliness Index	Levenberg–Marquardt back-propagation	10	20 (Single layer)	53.7	0.0038	242
(Hammad et al., 2018)	daily conversion efficiency	Levenberg–Marquardt backpropagation and scaled conjugate gradient backpropagation	2	10 and 4 (two layers)	90	5.7	1.4
(Al-Kouz et al., 2019)	daily conversion efficiency	Scaled conjugate gradient backpropagation	2	70 (Single layer)	90.7	5.02	1.29
(Laarabi et al., 2019)	soiling rate	Levenberg–Marquardt back-propagation	6	35 (Single layer)	92.8	6.81	9.04
(Chiteka et al., 2019a)	daily conversion efficiency	Levenberg–Marquardt back-propagation	5	4 (Single layer)	97.91% (adjusted)	1.35	6.3
(Shapsough et al., 2019)	maximum power	–	4	8 (Single layer)	96.4	28.09	–
(Zitouni et al., 2020)	Soiling rate	Keras	6	10 (Single layer)	81.3	6.76	–
(Mohanty and Kale, 2020)	dust intensity, dust density, and area covered by dust	gradient descent	3	8 (Single layer)	–	18.4 14.6 4.41	–
(Arshad et al., 2020)	Maximum power	stochastic gradient descent with backpropagation	2	4 and 2 (two layers)	–	47.61	–
	Maximum power	stochastic gradient descent with backpropagation	2	Less than or equal to 10 (three layers)	–	47.61	–
(Pérez et al., 2021)	Energy losses due to soiling	Levenberg–Marquardt	6 (no short circuit current)	10 (single layer)	–	49	–
	Energy losses due to soiling	Levenberg–Marquardt	7 (with short circuit current)	10 (single layer)	–	49	–

Se observa que el artículo de Javed et al. [20], considerado el pionero en este tema, utiliza el algoritmo de entrenamiento Levenberg-Marquardt con retroalimentación. Se ve que este algoritmo es de los más utilizados y de manera general tiene buenos resultados.

En el artículo de Chiteka et al. [21], que presenta el mejor  $R^2$  ajustado segundo la tabla (un 97.91), se utiliza una combinación de técnicas bastante interesante. Con los datos recogidos de una estación meteorológica ubicada en la ciudad de Harare, en Zimbabue, inicialmente tenían catorce parámetros ambientales y meteorológicos disponibles para el análisis: precipitación, velocidad del viento, dirección del viento, temperatura máxima, humedad relativa, presión de la superficie, variación de la temperatura, temperatura media, velocidad máxima del viento, variación de la velocidad del viento, velocidad mínima del viento, temperatura mínima, tasa de limpieza y las partículas de PM10 (partículas de diámetro inferior a 10 micrómetros). A dichos parámetros es implementado el algoritmo *Boruta* al regresor Random Forest para la selección de parámetros relevantes, con lo que se quedan con las cinco variables más relevantes, en su caso, respectivamente: PM10, humedad relativa, velocidad del viento, precipitación y dirección del viento.

Luego, se hace un modelo de red neuronal artificial para establecer las relaciones entre los parámetros ambientales y meteorológicos para la predicción del nivel de suciedad, y se desarrolla una red neuronal con retroalimentación utilizándose el algoritmo de entrenamiento Levenberg-Marquardt. La arquitectura de su red cuenta con los cinco parámetros más relevantes mencionados anteriormente como entrada, una única capa oculta con variación de 10 a 25 neuronas y una capa de salida, con el nivel de suciedad alcanzado.

Es interesante recalcar que los artículos publicados hasta el momento utilizan mayoritariamente datos meteorológicos recogidos en suelo, mientras en este presente estudio se propone la utilización de datos recogidos por satélites, aportando así un nuevo abordaje al tema.

Algunos estudios apuntan que de manera general los datos de parámetros ambientales procedentes de satélites son menos sensibles que los datos procedentes

de suelo, lo que puede llevar a más errores [22]. Sin embargo, la posibilidad de utilizar informaciones de cualquier lugar del planeta que ya están disponibles y que no necesitan de instalaciones específicas, justifica que se plantee su utilización.

Las características ambientales de la ubicación geográfica también influyen en la calidad de la predicción. Es decir, cuantos menos factores de suciedad haya en la atmósfera, menos acierto se observa en la predicción. [23] Por ejemplo, en Jaén, donde la suciedad no es tan acentuada, los modelos suelen obtener resultados menos precisos que en regiones con mucho polvo, como en Qatar.

### 3. ANÁLISIS EXPLORATORIO DE LOS DATOS

En este apartado vamos a realizar un análisis de las características de la base de datos utilizada en ese trabajo, que proviene de dos fuentes distintas, aunque sean complementarias.

La primera parte fue gentilmente facilitada por el Centro de Estudios Avanzados en Ciencias de la Tierra, Energía y Medio Ambiente (CEACTEMA), de la Universidad de Jaén, representado por los investigadores MSc João Gabriel Bessa y Dr. Eduardo Fernández. La base nombrada Atonometrics contiene datos en serie temporal de la estación de monitorización de niveles de suciedad en un edificio de la UJA.

La segunda parte de datos proviene de la página online del Proyecto MERRA-2 de la NASA. La sigla MERRA significa “*Modern-Era Retrospective analysis for Research and Applications*”, que puede ser traducida libremente como “Análisis retrospectivo de la Era Moderna para investigación y aplicaciones”. El proyecto ofrece varias bases de datos que combinan informaciones recogidas por sus satélites, como datos ambientales, atmosféricos y otras observaciones espaciales relativas a la contaminación atmosférica, como la interacción entre los aerosoles y otros procesos físicos en el sistema climático. [24]

Para este trabajo se utilizan tres bases de datos de MERRA-2:

- M2T1NXFLX: datos meteorológicos. [25]
- M2T1NXAER: datos de diagnósticos de aerosoles. [26]
- M2I3NVAER: datos de los índices de aerosoles mezclados. [27]

A continuación, se presentarán más detalles e informaciones sobre los parámetros y variables disponibles en las bases de datos, un análisis cuantitativo y cualitativo de éstos y finalizaremos con el análisis exploratorio de la base final utilizada.

### 3.1. Informaciones sobre la base de datos

Como hemos comentado anteriormente, la base de datos utilizada en este trabajo es el resultado de la unión de otros conjuntos de datos para atender a las necesidades del estudio.

Contiene 742 filas y 24 columnas, indexadas por la fecha de medición. Las informaciones contenidas en la base, se refieren al índice de nivel de suciedad medido en una determinada fecha, vinculado a los factores ambientales y meteorológicos medidos por satélites en esta misma fecha. Es decir, fue realizado un *left join* de la base de medición (Atonometrics) con las bases de datos de satélites de la NASA.

Los datos contemplan el período del 01/03/2019 al 31/03/2021. En el Anexo I, se facilita una tabla informativa detallada de todas las columnas de estos datos.

#### 3.1.1. Bases de datos originales

En este apartado se presentará información relacionada con las cuatro bases de datos que ha sido utilizadas para construir la base de datos final:

- Monitorización de Soiling
- Datos Meteorológicos
- Diagnósticos de Aerosoles
- Índice de Mezcla de Aerosoles.

##### 3.1.1.1. Monitorización de Soiling

La base de datos nombrada Atonometrics, que será la referencia principal de este trabajo, contiene los datos de la estación de monitorización de *soiling* del CECTEMA de la Universidad de Jaén, en formato de serie temporal, que contempla el período del 30/01/2019 al 05/04/2021.

Estos datos han sido facilitados en formato de planilla Excel (extensión .xlsx) y posteriormente importada al ambiente Python. Consta de 797 filas y 7 columnas.

De acuerdo con los investigadores responsables, durante el primer mes de mediciones el equipo presentó diversas alteraciones y errores mientras se realizaba la calibración de su configuración. Por esta razón, se utilizarán los datos a partir del 01/03/2019.

Para la composición de la base de datos del estudio, se utilizaron 2 columnas: Fecha de medición (columna “Date”) y la medición del nivel de suciedad (columna “SR\_Isc”). Las columnas restantes no fueron consideradas ya que se trataban de datos meteorológicos que constan en la base de Datos Meteorológicos de la NASA. Además, son datos meteorológicos de medición en el suelo, y el objetivo en este trabajo es la utilización de datos de satélites.

### 3.1.1.2. Datos Meteorológicos

La base de datos nombrada M2T1NXFLX es parte del Proyecto MERRA-2 de la Agencia Espacial Estadounidense y contiene datos de los llamados “flujos de superficie”, que son los eventos atmosféricos que actúan en la superficie terrestre, como el viento, la lluvia y la temperatura.

Esta base está disponible a través de la página del proyecto Soda [28], para descarga en formatos más sencillos y ya agrupados por la periodicidad diaria. Para este estudio fueron descargados los datos del período de 01/03/2019 a 31/03/2021, referentes a las coordenadas geográficas de la ciudad de Jaén (37.78, -3.77).

Esta base de datos está compuesta por 1 fichero en formato csv, que después de importados al ambiente Python, se presenta en un total de 762 filas y 11 columnas.

De esta base de datos, serán utilizadas 7 columnas: “Temperature”, “Relative Humidity”, “Pressure”, “Wind speed”, “Wind direction”, “Rainfall”, “Short-wave irradiation”.

En el Anexo I, se listan y describen todas las columnas utilizadas de esta y de las demás bases de datos.

### **3.1.1.3. Diagnósticos de Aerosoles**

La base de datos de Diagnósticos de Aerosoles, originalmente nombrada M2T1NXAER, es parte del Proyecto MERRA-2 de la Agencia Espacial Estadounidense y contiene datos de partículas aerosoles encontradas en la atmósfera terrestre.

Para este estudio fueron descargados los datos del período de 01/01/2019 a 31/03/2021 referentes a las coordenadas geográficas de la ciudad de Jaén (37.78,-3.8,37.8,-3.77).

Los datos están distribuidos en 821 ficheros en formato netCDF (extensión .nc4), que después de importados al ambiente Python y unificados en un único Dataframe, se presenta en un total de 19704 filas y 51 columnas.

De estas columnas, será utilizada la columna “DUSMASS25”, que se refiere a la concentración de polvo en la superficie terrestre, y es especialmente utilizada para cálculo de la contaminación aérea.

En el Anexo I, se listan las columnas utilizadas de esta y de las demás bases de datos.

### **3.1.1.4. Índice de Mezcla de Aerosoles**

Originalmente nombrada M2I3NVAER, la base de datos de Índice de Mezcla de Aerosoles es parte del Proyecto MERRA-2 de la Agencia Espacial Estadounidense y contiene datos de los índices de mezcla de partículas aerosoles en la atmósfera terrestre.

Para este estudio fueron descargados los datos del período de 01/01/2019 a 31/03/2021 referentes a las coordenadas geográficas de la ciudad de Jaén (37.78,-3.8,37.8,-3.77).

Esta base de datos está compuesta por 821 ficheros en formato netCDF (extensión *.nc4*), que después de importarlos al ambiente Python y unificados en un único Dataframe, se presenta en un total de 6568 filas y 22 columnas.

De esta base de datos, serán utilizadas 14 columnas: “SO4”, “BCPHOBIC”, “BCPHILIC”, “OCPHOBIC”, “OCPHILIC”, “DU001”, “DU002”, “DU003”, “DU004”, “SS001”, “SS002”, “SS003”, “SS004”, “AIRDENS”. Estas variables también son utilizadas para cálculo de la contaminación aérea.

En el Anexo I, se listan y describen todas las columnas utilizadas de esta y de las demás bases de datos.

### **3.2. Preparación del conjunto de datos**

Tras conocer el origen de los datos, pasamos a la preparación del conjunto de datos que será utilizado.

Para unificar los datos, se ha utilizado como llave la fecha referente a la medición realizada en cada base original.

La base “Diagnóstico de Aerosoles” presenta mediciones diarias a cada 1 hora, mientras la base “Índice de Mezcla de Aerosoles” tiene sus mediciones diarias cada 3 horas. Por lo tanto, se ha realizado una agregación de los datos, calculando la media diaria de los valores obtenidos, dejando así cada una de las bases con 821 filas.

Tras la unificación de las cuatro bases utilizadas, se ha realizado la limpieza de los datos. Se ha fijado la fecha límite de los datos entre 01/03/2019 y 31/03/2021, periodo común entre todas las bases.

Para fines didácticos, se cambia el nombre del campo “SR\_Isc” a “SR”.

En seguida se incluye la columna “PM10”, que ha sido incluida y calculada basada en la fórmula (2), determinada por la NASA [29]. Las PM10 son las partículas aerosoles de diámetro entre 2,5 y 10 micrómetros, directamente relacionadas a la contaminación atmosférica. [30] Las variables utilizadas en este cálculo están presentes en nuestra base de datos.

$$PM10 = (1.375 * SO4 + BCphobic + BCphilic + OCphobic + OCphilic + DU001 + DU002 + DU003 + 0.74 * DU004 + SS01 + SS02 + SS03 + SS04) * AIRDENS \quad (2)$$

El valor de PM10 es también utilizado en los modelos matemáticos que vamos a utilizar en nuestro análisis comparativo final.

Con esos procedimientos, se obtiene una base de datos con 762 filas y 24 columnas.

### 3.3. Análisis de los datos

Se inicia el análisis de los datos con una breve visualización de los datos estadísticos. Como son muchas columnas, con informaciones muy específicas de datos ambientales, aquí se presenta solo el campo principal, SR.

Variable "SR"	
count	762
mean	0.979561
std	0.048577
min	0.595000
25%	0.977000
50%	0.988000
75%	0.999000
max	1.083000

**Figura 6. Datos estadísticos de la variable “SR”**  
Elaboración propia.

Como se observa, es un total de 762 registros, la media del SR es de aproximadamente 0.979 y presenta una desviación estándar de 0.048, es decir que de manera general el SR varía entre 0.931 y 1.027.

Se observa que el valor máximo del SR es de 1.083, contrariando la teoría de que el valor de SR varía de 0 a 1. Segundo aclarado por los investigadores del CECTEMA/UJA, esta situación sugiere que, por alguna razón el módulo sucio estaría generando más energía que un módulo limpio. Algunas de las razones pueden ser diferencias en la fabricación de los módulos (aunque sean del mismo material) o

pequeñas variaciones en el recurso solar entre el equipo de medición limpio y el sucio ([Figura 5](#), en el [Capítulo 2.2 Soiling: La suciedad en los módulos fotovoltaicos](#)).

Para el estudio realizado en este trabajo, se mantendrán como válidos los valores superiores a 1.

## Datos nulos

Solamente en la columna “SR” se han encontrado valores nulos. Estas 38 filas fueron tratadas por imputación de datos. La metodología utilizada está descrita en el apartado “Preprocesamiento de los datos”.

## *Outliers*

Para identificación de *outliers* en la variable SR, se ha utilizado el método conocido como Z-Score, que a través del cálculo de la desviación estándar y de la media, enseña los puntos que se alejan de la media de distribución normal. En este caso, el umbral de distancia utilizado es de 3 puntos, y nos ha señalado 20 registros.

Index / Fecha	SR	Index / Fecha	SR
2019-11-04	0.802	2019-12-29	0.682
2019-11-07	0.796	2019-12-30	0.678
2019-12-21	0.721	2019-12-31	0.679
2019-12-22	0.667	2020-01-01	0.687
2019-12-23	0.676	2020-01-02	0.697
2019-12-24	0.669	2020-01-03	0.812
2019-12-25	0.727	2020-01-04	0.710
2019-12-26	0.833	2020-01-05	0.682
2019-12-27	0.708	2020-01-06	0.704
2019-12-28	0.595	2020-12-07	0.737

**Figura 7. Lista de outliers de la variable “SR”**  
Elaboración propia.

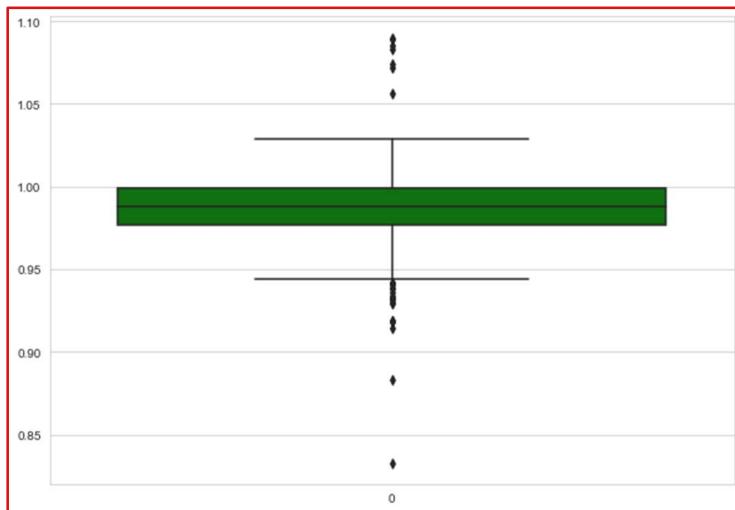
Los registros identificados como *outliers* coinciden con los apuntes de la documentación de la base de datos, en la que, en las fechas arriba mencionadas, hubo errores de medición por parte del equipo y por lo tanto no son datos fiables. Por esa razón, estas 20 filas fueron suprimidas de la base del estudio.

Tras la supresión de los *outliers*, se puede observar un ligero cambio en los datos estadísticos, principalmente con la alteración del valor mínimo para 0.867.

Variable "SR"	
count	742
mean	0.986743
std	0.019192
min	0.867000
25%	0.978000
50%	0.988000
75%	0.999000
max	1.083000

**Figura 8.** Datos estadísticos de la variable “SR” sin outliers  
Elaboración propia.

En la **Figura 9**, se demuestra en un *boxplot* los cuartiles de la variable “SR”.

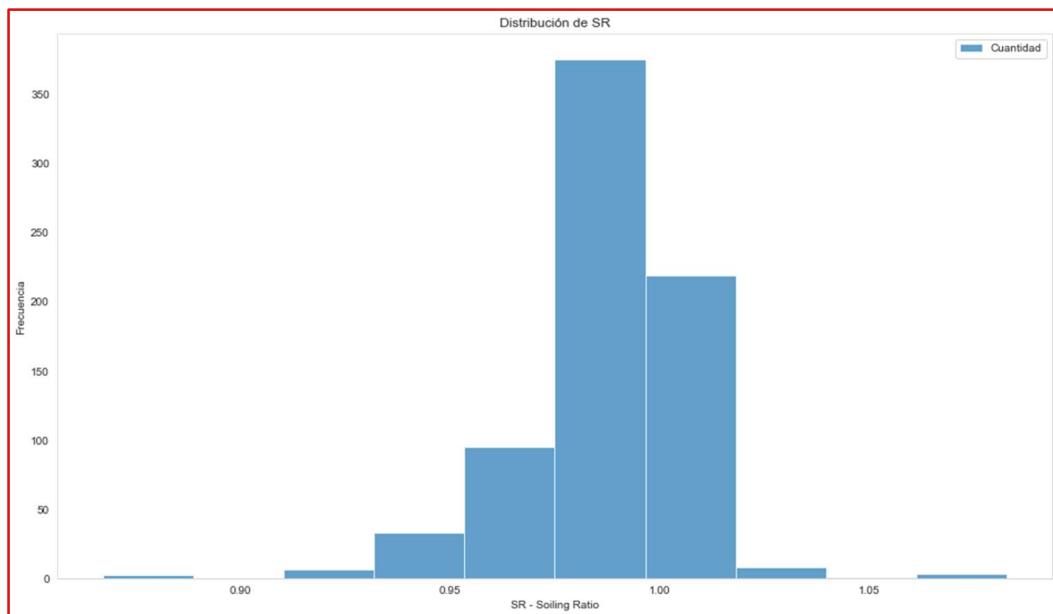


**Figura 9.** Cuartiles de la variable “SR”  
Elaboración propia.

De acuerdo con las orientaciones de los investigadores del CECTEMA/UJA, los datos fuera de los cuartiles no son *outliers* y son mediciones, aunque raras, correctas. Eso puede pasar a razón de que el nivel de suciedad es una medición acumulativa, o sea, con el pasar de los días, más sustancias se depositan sobre el módulo, y que también pueden ser removidas de un día al otro, con una limpieza programada o la lluvia. Así, un módulo puede haber quedado limpio tras algún evento que lo hubiera dejado bastante sucio.

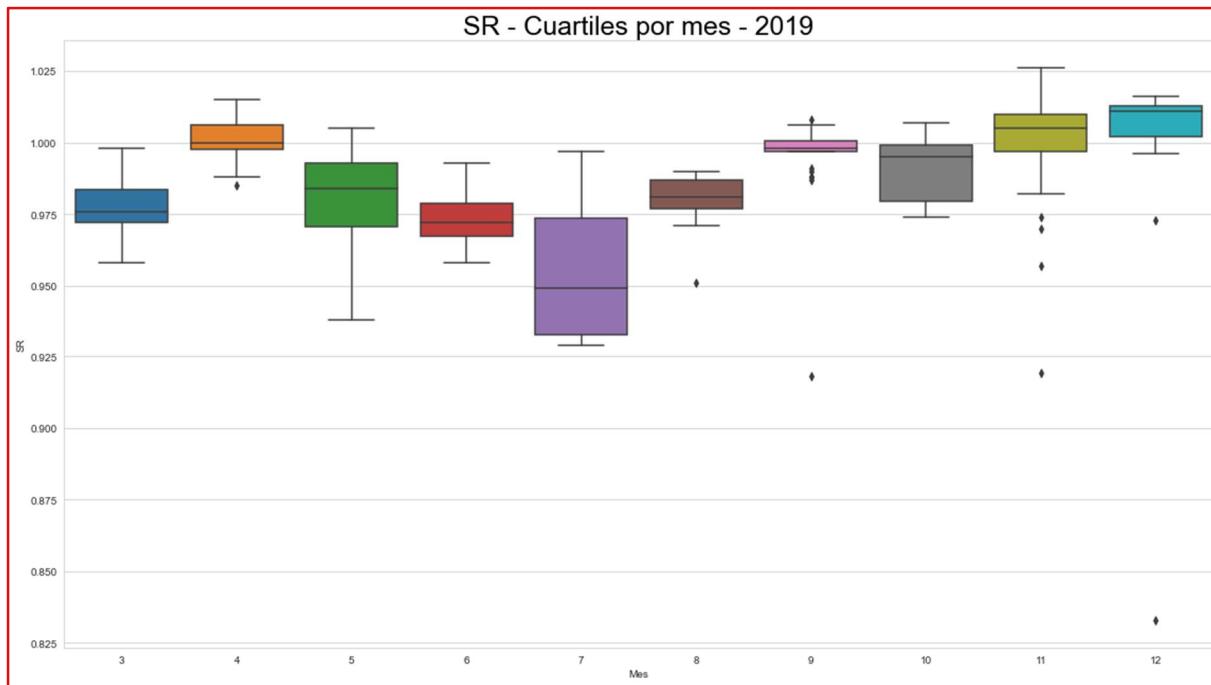
## Distribución de los datos

En la **Figura 10**, se observa que la gran mayoría de los SR tienen su valor situado en los rangos esperados.

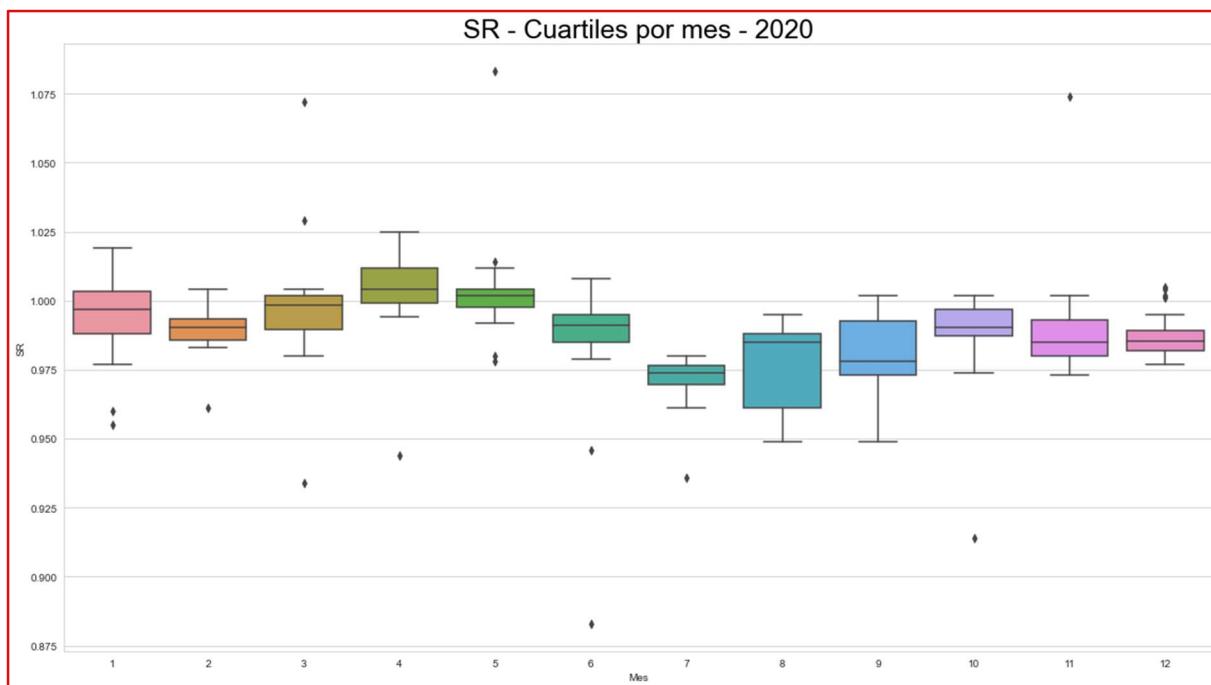


**Figura 10.** Distribución general de los valores de “SR”  
Elaboración propia.

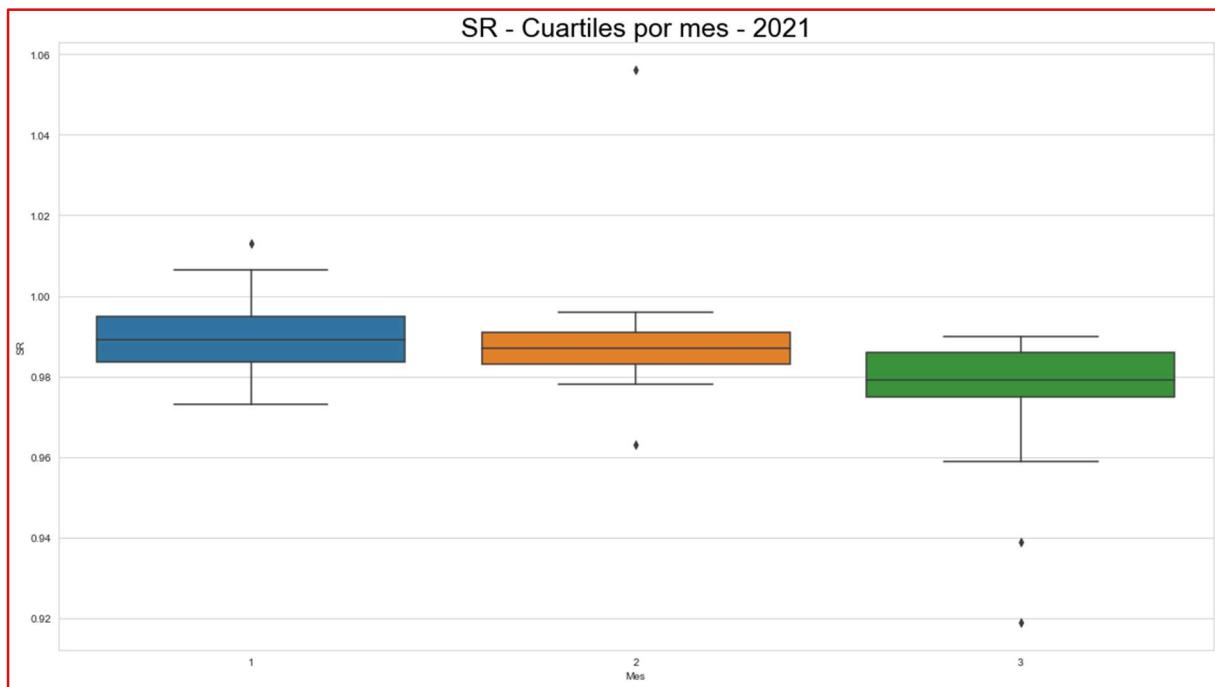
Observamos la distribución de los datos en los cuartiles, separados para cada mes de los años medidos. En las **Figuras 11, 12 y 13** se presentan los datos de los años 2019, 2020 y 2021, respectivamente.



**Figura 11. Cuartiles por mes del “SR” durante el año de 2019**  
Elaboración propia.



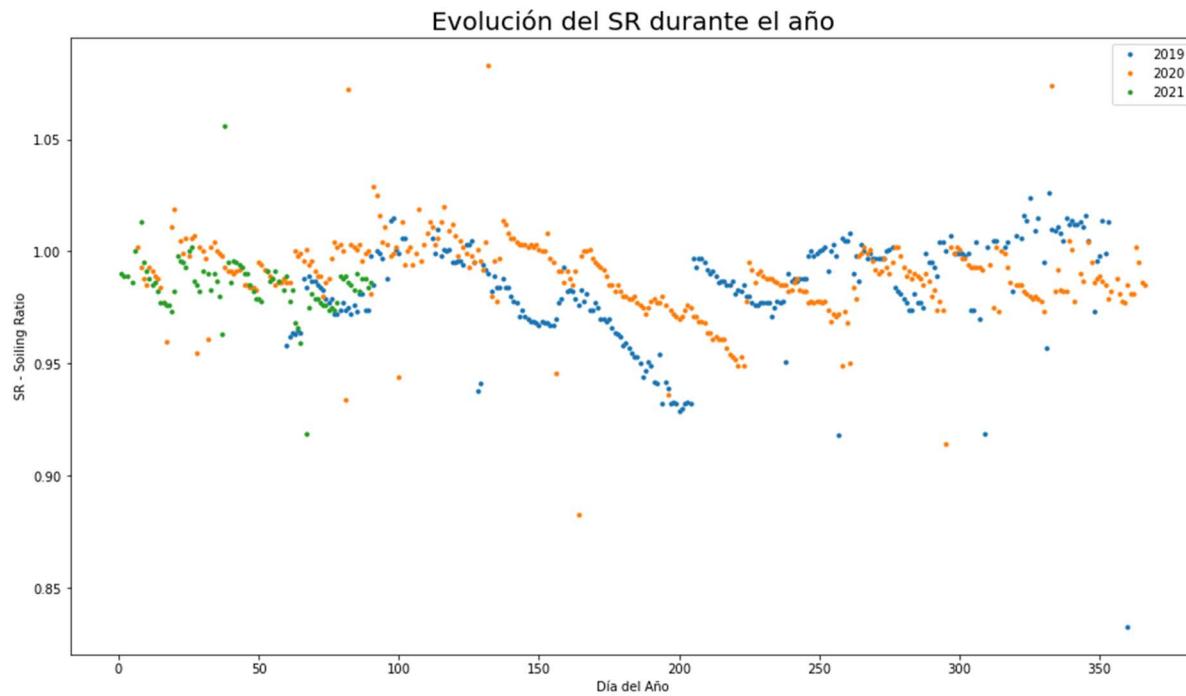
**Figura 12. Cuartiles por mes del “SR” durante el año de 2020**  
Elaboración propia.



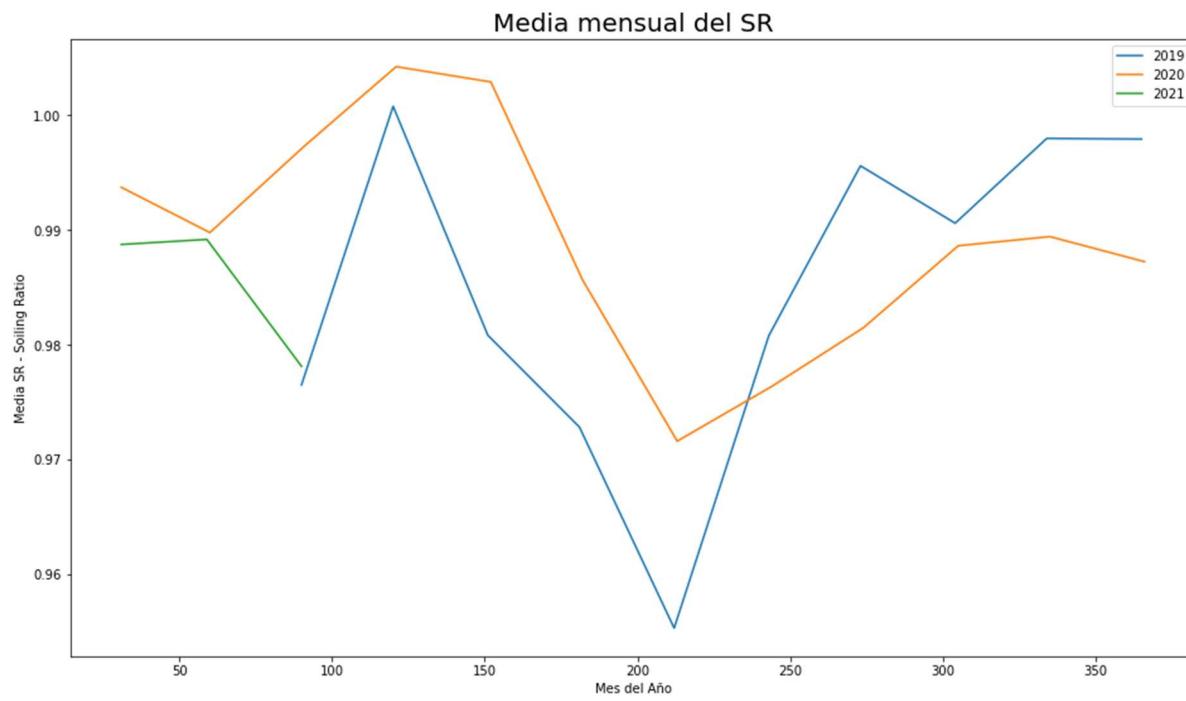
**Figura 13. Cuartiles por mes del “SR” durante el año de 2021**  
Elaboración propia.

Con la visualización de los cuartiles, nos queda claro que hay algunas pocas, sin embargo, constantes ocurrencias en los extremos. Eso nos indica que los datos no son *outliers*, como ya nos habían informado los investigadores responsables por las mediciones de los niveles de suciedad.

En las **Figuras 14 y 15**, se observa la estacionalidad del nivel de suciedad, siendo los meses del verano (más secos y, por lo tanto, sin lluvia) los que más influyen en el SR.



**Figura 14. Evolución diaria del “SR” durante el año**  
Elaboración propia.



**Figura 15. Evolución mensual del “SR” durante el año**  
Elaboración propia.

Para comprobar estadísticamente la estacionariedad observada, utilizamos la función `adfuller()` de la librería `statsmodels` y realizamos una prueba de Dickey-Fuller aumentada.

Se puede interpretar el resultado de la prueba, basándose en el *p-value* obtenido. Un *p-value* menor que el umbral de un 5% sugiere que se rechaza la hipótesis nula. Es decir, si se obtiene un *p-value* inferior a 0.05, los datos poseen estacionariedad. [31]

Se obtiene los resultados:

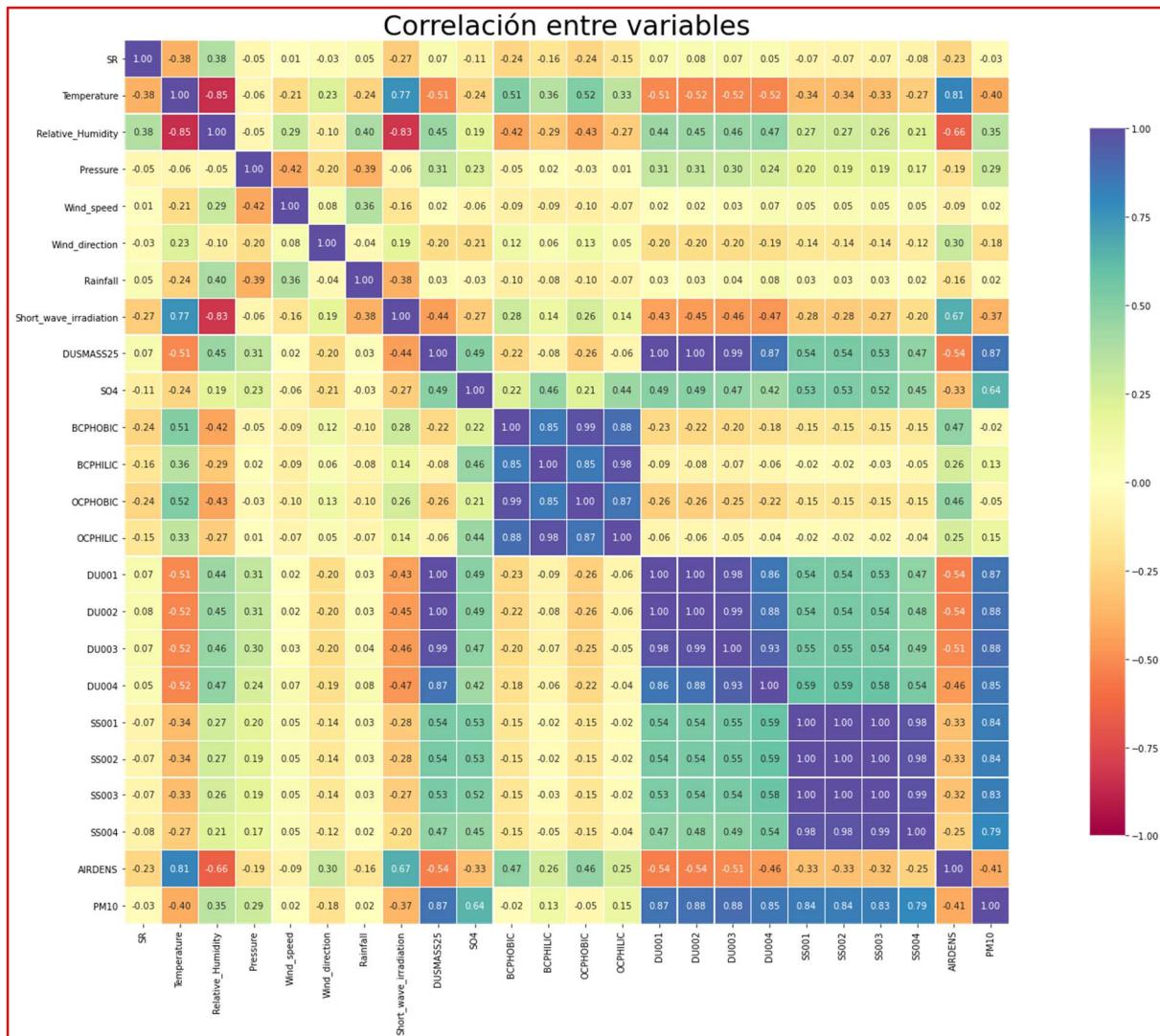
```
ADF Statistic: -5.415519
p-value: 0.000003
Critical Values:
    1%: -3.439
    5%: -2.865
   10%: -2.569
```

El resultado de la prueba nos confirma que se trata de una serie temporal estacionaria, con un *p-value* de 0.000003. Además, la función nos informa que nuestro valor estadístico obtenido es de -5, inferior al valor de -3.439 al 1%. O sea, la probabilidad de que los datos no sean estacionales es de menos de un 1%.

## Correlación entre las variables

Para analizar la correlación entre las variables, se produce una matriz de correlación (**Figura 16**).

Teniéndose en cuenta el contexto de la base de datos utilizada, es esperado que muchas variables presenten correlaciones entre ellas, ya que son datos ambientales de medición diaria. Por ejemplo, es esperado que la Temperatura y la Humedad Relativa presenten una fuerte correlación negativa. Sin embargo, no necesariamente esas correlaciones aportan alguna información relevante a nuestro análisis.

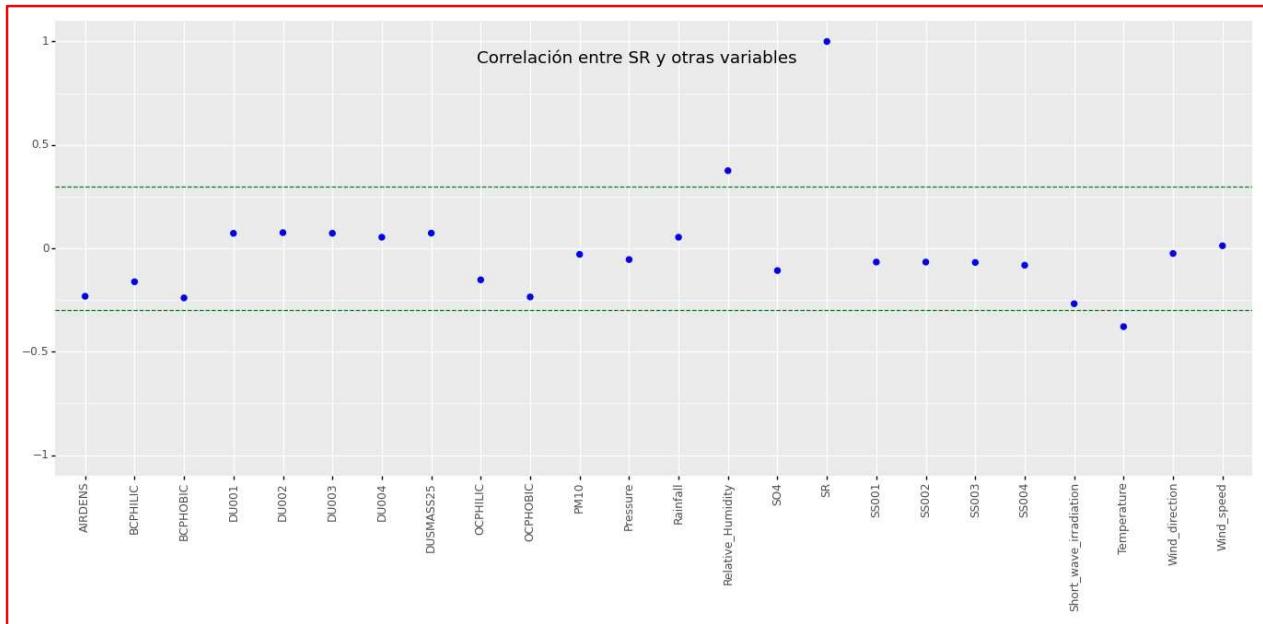


**Figura 16. Matriz de correlación entre todas las variables**  
Elaboración propia.

En este caso, nos fijamos en la correlación entre el nivel de suciedad medida (SR) y las demás variables. La observamos en la **Figura 17**.

Se puede observar que apenas 2 variables presentan un índice de correlación al menos mediano (utilizándose de un valor medio de 0.3):

- Humedad relativa: 0.375343
- Temperatura: -0.378858



**Figura 17.** Gráfico de correlación entre SR y otras variables  
Elaboración propia.

Tras los análisis realizados, ya se obtiene conocimiento de qué se puede esperar de los datos. De esa manera, se pasa a la fase siguiente, en la que se inicia el preprocesamiento de los datos y la predicción.

## 4. PREPROCESAMIENTO DE LOS DATOS

En la etapa de Preprocesamiento de los datos, se realizan procedimientos que tienen como objetivo dejar el conjunto de datos listo para ser utilizado, intentándose mitigar posibles ruidos en la predicción. En ese capítulo se explicarán las modificaciones aplicadas al conjunto de datos.

### 4.1. Imputación de los valores no medidos

Durante el análisis exploratorio de los datos, se han encontrado 38 filas que contenían valores nulos para la medición del nivel de suciedad (SR). Por tratarse de una pequeña cantidad de datos y por el hecho de que el SR es un valor que se acumula con el pasar del tiempo, se realiza una imputación de valores.

La regla utilizada ha sido la media de los valores del SR del día anterior y del día posterior, imputando así un SR medio para el día faltante.

### 4.2. Normalización de los datos

La normalización de los datos tiene como objetivo cambiar los valores de las columnas numéricas en el conjunto de datos utilizados a una escala común, manteniendo las diferencias en los rangos de valores para cada columna. La normalización es altamente recomendada para análisis de regresión y principalmente para el uso de las redes neuronales.

En este trabajo, se ha utilizado el RobustScaler, del paquete Scikit-learn [32], que escala las características utilizando estadísticas que son robustas contra los *outliers*. RobustScaler utiliza la mediana para escalar los datos en el rango entre el primer y el tercer cuartil.

Este método ha sido elegido en razón de que observamos valores fuera de los cuartiles, aunque no sean considerados *outliers*.

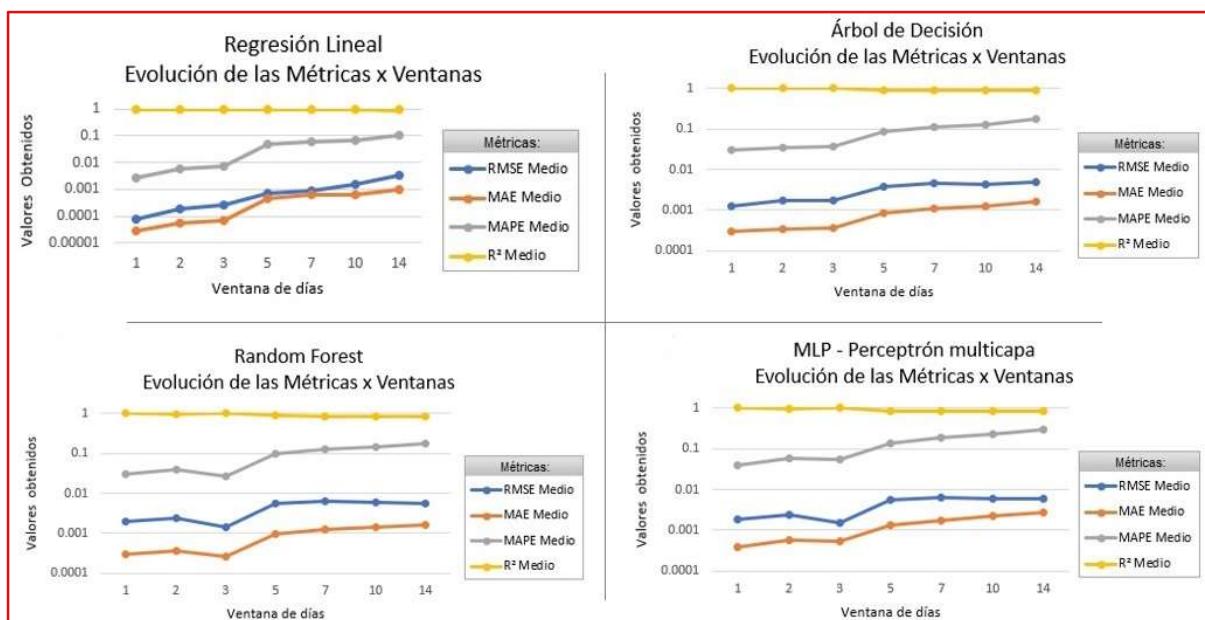
### 4.3. Ventana deslizante

La técnica de ventana deslizante tiene como objetivo facilitar como entrada al modelo los valores anteriores de medición. Es decir, en nuestro caso, se agrega al conjunto de datos los  $x$  valores anteriores de SR medidos.

Esta es una técnica bastante utilizada en conjuntos de series temporales debido a su naturaleza de tratar de datos continuos.

Fueron realizadas pruebas con diversas ventanas: 1, 2, 3, 5, 7, 10, 14 días anteriores, para de esa manera verificar con cuál se obtienen los mejores resultados.

Tras algunos experimentos, se hizo un análisis de la evolución de las métricas de acuerdo con el número de ventanas utilizadas. Se observa que los experimentos con el número de ventanas superior a 3, presenta un aumento notable en los errores, como se muestra en la **Figura 18**.



**Figura 18. Gráfico de evolución de las métricas por número de ventanas**  
Elaboración propia.

## 5. EXPERIMENTOS REALIZADOS

Para este estudio, se realizaron diversos experimentos que consisten en la ejecución de diferentes algoritmos de regresión para la predicción del Soiling Ratio, teniendo como partida los datos de medición en serie temporal facilitados por la Universidad de Jaén.

Como lenguaje de programación se ha elegido Python 3 [33], por su versatilidad de paquetes y funciones. Para el desarrollo de los scripts, se han utilizado la IDE Spyder [34] y Jupyter Notebook [35].

Los algoritmos utilizados pertenecen a los ampliamente conocidos paquetes Scikit-learn [32] y Keras [36].

- Scikit-learn
  - Regresión Lineal
  - Árbol de Regresión
  - Random Forest
  - MLP – Perceptrón Multi-Capa
- Keras
  - LSTM

En este capítulo se van a exponer los modelos matemáticos elegidos, detalles del diseño de los experimentos realizados, las métricas utilizadas y los resultados obtenidos de los experimentos de cada uno de los algoritmos, con cada uno de los *inputs* propuestos.

## 5.1. Diseño de los experimentos

Para el entrenamiento de los modelos de regresión propuestos, se van a dividir los conjuntos de datos en particiones de *train* y *test*, siendo utilizada la proporción 70/30, respectivamente.

Como ya mencionado en el capítulo “4. Preprocesamiento de datos”, ha sido aplicada la normalización de los datos y creada las ventanas de SR con pruebas de distintos tamaños.

Para cada algoritmo del experimento serán probados diferentes hiperparámetros y tamaños de ventana. De esa manera, será posible evaluar las mejores combinaciones para cada uno de los conjuntos de entrada propuestos.

### 5.1.1. Modelos Matemáticos

Los modelos matemáticos son herramientas importantes para el desarrollo de las investigaciones sobre el *soiling* y muchos estudios importantes se han publicado sobre este tema, como se observa en la revisión bibliográfica publicada por Bessa et al. [4]

Para este trabajo se eligen dos modelos matemáticos recientes y actualmente utilizados por los investigadores de la Universidad de Jaén para realizar sus predicciones: Coello y Boyle [5] y You [7].

El modelo Coello y Boyle [5] propone el cálculo del SR basado en la fórmula (3):

$$SR = 1 - 0.3437 \operatorname{erf}(0.17\omega^{0.8473}) \quad (3)$$

Siendo que:  $\omega$  el total de masa acumulada, en g/m<sup>2</sup>, en un dado momento.

El modelo de You [7] también utiliza el total de masa acumulada  $\omega$ , y además considera una pérdida de eficiencia energética de un 0.0139% por gramo de polvo depositado en cada metro cuadrado de un módulo FV. Para calcular el SR utiliza la fórmula (4):

$$SR = 0.0139\omega \quad (4)$$

El cálculo de  $\omega$  es realizado utilizando la fórmula (5):

$$\omega = V_d \cdot C \cdot N_D \cdot 10^{-6} \quad (5)$$

Siendo que:  $V_d$  es la velocidad de deposición,  $C$  es el total de partículas suspendidas y  $N_D$  es el número de días sin lluvia.

Para los dos modelos, cuando hay un registro de lluvia mayor que el umbral de 0,3mm de lluvia, automáticamente asumen un SR =1. O sea, consideran que el módulo está perfectamente limpio.

Haciendo relación con la base de datos que se utiliza en este presente trabajo, las variables que serán utilizadas para el cálculo del SR por los modelos matemáticos son: 'Temperature', 'Relative Humidity', 'Pressure', 'Wind speed', 'Short wave irradiation', 'AIRDENS' y 'Rainfall'.

Tras calcular los SR con los modelos matemáticos, se calculan sus métricas para posterior comparación (Tabla 2). Es importante aclarar que estos modelos fueron creados para ser utilizados con los datos meteorológicos de suelo. Para un breve análisis, se hace el cálculo utilizando ambos datos: los de satélite de la NASA que usaremos en nuestros experimentos y los datos de la estación meteorológica de la UJA.

**Tabla 2.** Resultados de métricas aplicadas a los modelos matemáticos utilizados

Modelo	RMSE	MAE	MAPE	R <sup>2</sup>	Origen
Coello y Boyle	0.016876	0.011128	1.136918	0.341808	Datos de Suelo
	0.038538	0.023643	2.420922	0.205066	Datos de Satélites
You	0.019935	0.013792	1.418861	0.347129	Datos de Suelo
	0.030156	0.01736	1.777113	0.185304	Datos de Satélites
Media:	<b>0.02637625</b>	<b>0.01648075</b>	<b>1.6825</b>	<b>0.2697185</b>	

Es importante recordar que es esperado que datos de parámetros ambientales procedentes de satélites puede llevar a más errores. De esa manera, y para seguir las orientaciones de los modelos desarrollados, se utilizarán para las comparaciones finales las métricas de los datos de suelo, aunque en los modelos de *machine learning* serán utilizados solamente los datos de satélites.

### 5.1.2. Selección de Características

Para la Selección de Características, se proponen 3 conjuntos de datos:

- Input 1: Selección de características a través del algoritmo *Boruta* [37].

*Boruta* es un extractor de características que utiliza la medida de importancia de cada característica generada por el algoritmo Random Forest y la compara con la importancia de copias aleatorias de sus otros atributos.

*Boruta* utiliza una relevante búsqueda *top-down* comparando la importancia de la variable original con la importancia que puede ser alcanzada de manera aleatoria, estimada pela media de sus permutaciones duplicadas, y progresivamente suprime características no relevantes [21]. Una extensa descripción del funcionamiento de este algoritmo puede ser contemplada en el artículo de Kursa [37].

Al final de su ejecución, el algoritmo retorna dos listas: el área verde y el área azul. El área verde contiene las características con clara importancia para la predicción, y el área azul contiene las características con un grado de importancia que el algoritmo sólo no puede definir si es o no relevante. Las características del área azul necesitan un análisis contextual de su importancia.

Para este trabajo, *Boruta* ha sido ejecutado utilizando como hiperparámetros:

- max\_depth = 5
- n\_estimators = 'auto'
- max\_iter = 100

Su ejecución ha resultado en la siguiente selección:

- Área verde: 'Temperature', 'Relative\_Humidity', 'Pressure', 'Wind\_speed', 'Short\_wave\_irradiation', 'AIRDENS'
- Área azul: 'Rainfall'

Como se sabe que la cantidad de lluvia influye en el valor del SR, ya que es el principal agente de limpieza del módulo fotovoltaico, se opta por mantener esta característica en nuestro conjunto de datos. Además, vamos a agregar las ventanas deslizantes del SR.

De esta manera, tenemos:

`Input_1 = ['Temperature', 'Relative_Humidity', 'Pressure',  
'Wind_speed', 'Short_wave_irradiation', 'AIRDENS', 'Rainfall',  
'Ventana_d-1', 'Ventana_d-2', ... 'Ventana_d-n']`, siendo **n** el número de días de la ventana deslizante del SR utilizada.

- Input 2: Selección de las mismas características utilizadas en los modelos matemáticos de Coello y Boyle[5] y de You [7].

Ambos modelos utilizan entrada de datos meteorológicos similares, con diferencias en sus cálculos. Más detalles de la fórmula y cálculo se encuentran en el apartado “5.1 Modelos Matemáticos”. Además, en nuestro Input\_2, vamos a agregar las ventanas deslizantes del SR.

De esa manera, tenemos:

`Input_2 = ['Temperature', 'Wind_speed', 'PM10', 'DUSMASS25',  
'Rainfall', 'Ventana_d-1', 'Ventana_d-2', ... 'Ventana_d-n']`, siendo **n** el número de días de la ventana deslizante del SR utilizada.

- Input 3: Selección únicamente de las ventanas deslizantes.

El último conjunto de datos tiene la intención de comprobar si los parámetros ambientales interfieren en la predicción en el caso de uso de ventanas deslizantes. Por lo tanto, se utiliza como entrada solamente las ventanas deslizantes de SR.

De esa manera, tenemos:

**Input\_3** = ['Ventana\_d-1', 'Ventana\_d-2', ... 'Ventana\_d-n'], siendo **n** el número de días de la ventana deslizante del SR utilizada.

### 5.1.3. Métricas utilizadas

Las métricas seleccionadas para medir la bondad de los modelos creados en este estudio son, en este orden de relevancia:

1. **RMSE**: *Root Mean Squared Error*, o Raíz del Error Cuadrático Medio.

Es una medida muy utilizada en la comparación de estimadores pues cuantifica cuán diferente son dos conjuntos de valores. Es especialmente recomendada cuando los grandes errores necesitan ser evitados, como es el caso del SR, debido a su pequeña amplitud de valores. Los valores de RMSE más pequeños son los más deseados, pues más se acercan del valor predicho.

2. **MAE**: *Mean Absolute Error*, o Error Medio Absoluto.

Es una medida que nos enseña el promedio de la diferencia absoluta entre el valor real y el valor predicho. Así como en el RMSE, los valores más pequeños son los mejores.

3. **MAPE**: *Mean Absolute Percentage Error*, o Error Porcentual Absoluto Medio.

Es una medida de precisión que nos informa el tamaño del error absoluto en términos porcentuales. En este trabajo el MAPE será informado ya en formato porcentual, o sea, sin la necesidad de su multiplicación por 100.

4. **R<sup>2</sup>**: R-cuadrado, o Coeficiente de Determinación

Es una medida que expresa el porcentaje de variación de la variable de predicha con su relación con una o más variables predictoras y se encuentra entre 0 y 100. Diferentemente de las medidas anteriores, en este caso, un mayor valor obtenido de R<sup>2</sup>, infiere un mejor ajuste del modelo a sus datos.

Para todas las métricas, los valores predichos tendrán su normalización deseada para reflejar mejor el ajuste realizado.

## 5.2. Resultados obtenidos

En este apartado serán expuestos los resultados obtenidos para cada uno de los algoritmos utilizados y para cada uno de los conjuntos de datos propuestos. En cada una de las tablas de resultados, se presentan las siguientes columnas en común:

- #: número secuencial de la ejecución, para simple identificación.
- Vent: número de ventanas utilizadas em la ejecución.
- Train: respectivas métricas para el conjunto de entrenamiento.
  - RMSE
  - MAE
  - MAPE
  - R<sup>2</sup>
- Test: respectivas métricas para el conjunto de test.
  - RMSE
  - MAE
  - MAPE
  - R<sup>2</sup>

También se presenta la información sobre los hiperparámetros utilizados para cada uno de los modelos.

### 5.2.1. Regresión Lineal

Para la regresión lineal, fueron utilizadas ventanas de 1, 2, 3, 5, 7, 10 y 14 días.

En la Tabla 3 se encuentran los resultados obtenidos de estos modelos, donde se señalan en verde y en rojo, respectivamente los 5 mejores y los 5 peores valores de RMSE obtenidos. Además, se incluyen las medias de las métricas al final de la tabla.

Se observa que los mejores resultados son obtenidos con la predicción utilizando ventanas de hasta 2 días, aunque la ventana de 3 días también presenta resultados satisfactorios. Los peores resultados son obtenidos con las ventanas de 10 y 14 días.

**Tabla 3.** Resultados obtenidos con la Regresión Lineal

#	Input	Vent	Train				Test			
			RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>
1	Input_1	1	0.000173	0.000029	0.002985	0.999931	0.000078	0.000037	0.003788	0.999971
2	Input_1	2	0.000263	0.00005	0.00506	0.999842	0.000201	0.000073	0.007447	0.999804
3	Input_1	3	0.00025	0.000055	0.005589	0.999856	0.00026	0.000086	0.008768	0.999672
4	Input_1	5	0.002032	0.000551	0.056301	0.990542	0.000747	0.000532	0.053962	0.997299
5	Input_1	7	0.002444	0.000759	0.077576	0.986323	0.000997	0.000771	0.078187	0.995188
6	Input_1	10	0.002446	0.000658	0.067153	0.986303	0.001541	0.000686	0.069519	0.988501
7	Input_1	14	0.001947	0.000562	0.057268	0.991318	0.003456	0.001081	0.109789	0.942148
8	Input_2	1	0.000174	0.000024	0.002415	0.99993	0.000073	0.000031	0.003113	0.999974
9	Input_2	2	0.000264	0.000042	0.004304	0.99984	0.000195	0.000064	0.006527	0.999816
10	Input_2	3	0.000252	0.000048	0.004968	0.999854	0.000255	0.000079	0.007992	0.999685
11	Input_2	5	0.001952	0.000625	0.06376	0.991275	0.000908	0.000548	0.055646	0.99601
12	Input_2	7	0.002391	0.000773	0.078824	0.986904	0.001062	0.000688	0.06977	0.994533
13	Input_2	10	0.002424	0.000683	0.069581	0.986548	0.001602	0.000698	0.070858	0.987573
14	Input_2	14	0.001958	0.000515	0.052523	0.991219	0.003397	0.000987	0.100253	0.944096
15	Input_3	1	0.000175	0.000016	0.001702	0.99993	0.000068	0.000013	0.001361	0.999977
16	Input_3	2	0.000266	0.00003	0.003069	0.999838	0.000192	0.000034	0.003427	0.999822
17	Input_3	3	0.000255	0.000035	0.003571	0.999852	0.000253	0.000045	0.004561	0.999689
18	Input_3	5	0.002068	0.000423	0.04348	0.990207	0.000577	0.000289	0.029185	0.998389
19	Input_3	7	0.002499	0.000586	0.060071	0.985694	0.000737	0.000397	0.04018	0.997367
20	Input_3	10	0.002461	0.000572	0.058414	0.986127	0.0015	0.000534	0.054144	0.989098
21	Input_3	14	0.001962	0.0005	0.050983	0.991187	0.003378	0.000946	0.096068	0.944734
<b>Media de las métricas:</b>			<b>0.0013646</b>	<b>0.0003589</b>	<b>0.0366475</b>	<b>0.9934533</b>	<b>0.0010227</b>	<b>0.0004104</b>	<b>0.041645</b>	<b>0.989207</b>

## 5.2.2. Árbol de Regresión

Para el método de Árbol de Regresión, se utilizaron ventanas de 1, 2, 3 y 5 días, debido al hecho de que ventanas mayores que 5 presentaron errores demasiadamente grandes, conforme explicado en el capítulo [4.3 Ventana deslizante](#).

Los hiperparámetros utilizados para este método son:

- maxDepth: Máxima profundidad del árbol, con valores: 7, 10, 20, 25 y 30.

En la Tabla 4 se encuentran los resultados obtenidos de estos modelos, donde se señalan en verde y en rojo, respectivamente los 5 mejores y los 5 peores valores de RMSE obtenidos. Además, se incluyen las medias de las métricas al final de la tabla.

Se observa que los mejores resultados son obtenidos con la predicción utilizando ventanas de hasta 3 días, mientras los peores resultados son obtenidos con las ventanas de 5 días.

También se observa que algunos modelos (por ejemplo, # 17, 18, 19 y 40) tienen un RMSE igual a cero en el entrenamiento, pero con sus métricas de teste muy bajas, lo que podría indicar algún *overfitting*.

**Tabla 4.** Resultados obtenidos con el Árbol de Decisión

#	Input	Vent	max Depth	Train				Test			
				RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>
1	Input_1	1	7	0.000255	0.000113	0.011436	0.99985	0.001945	0.000239	0.023309	0.98168
2	Input_1	1	10	0	0	0	1	0.001999	0.00028	0.027517	0.98064
3	Input_1	1	20	0	0	0	1	0.001936	0.00022	0.021345	0.98185
4	Input_1	1	25	0	0	0	1	0.001968	0.000256	0.025124	0.98125
5	Input_1	1	30	0	0	0	1	0.001255	0.000209	0.02085	0.99237
6	Input_1	2	7	0.000267	0.000121	0.012202	0.99984	0.001983	0.000291	0.028672	0.98096
7	Input_1	2	10	0	0	0	1	0.001236	0.000195	0.019442	0.9926
8	Input_1	2	20	0	0	0	1	0.001325	0.000211	0.020866	0.9915
9	Input_1	2	25	0	0	0	1	0.001955	0.000253	0.024811	0.98149
10	Input_1	2	30	0	0	0	1	0.001154	0.000146	0.014419	0.99356
11	Input_1	3	7	0.000267	0.00012	0.012169	0.99984	0.001977	0.000281	0.027601	0.98108
12	Input_1	3	10	0	0	0	1	0.001996	0.000278	0.027408	0.9807
13	Input_1	3	20	0	0	0	1	0.002002	0.000289	0.028521	0.98059
14	Input_1	3	25	0	0	0	1	0.001232	0.0002	0.020004	0.99265

15	Input_1	3	30	0	0	0	1	0.001291	0.000226	0.022679	0.99193
16	Input_1	5	7	0.000303	0.000171	0.017245	0.99979	0.00276	0.000612	0.061478	0.96311
17	Input_1	5	10	0	0	0	1	0.003075	0.000563	0.056198	0.9542
18	Input_1	5	20	0	0	0	1	0.003039	0.000547	0.054588	0.95527
19	Input_1	5	25	0	0	0	1	0.003031	0.000572	0.057045	0.9555
20	Input_1	5	30	0	0	0	1	0.002678	0.000513	0.051783	0.96526
21	Input_2	1	7	0.000266	0.000117	0.011857	0.99984	0.001176	0.000181	0.018079	0.9933
22	Input_2	1	10	0	0	0	1	0.001165	0.000159	0.015846	0.99343
23	Input_2	1	20	0	0	0	1	0.001156	0.000164	0.016282	0.99353
24	Input_2	1	25	0	0	0	1	0.001194	0.000191	0.019103	0.9931
25	Input_2	1	30	0	0	0	1	0.001865	0.000213	0.020978	0.98316
26	Input_2	2	7	0.000292	0.000125	0.012637	0.99981	0.001183	0.000189	0.018862	0.99322
27	Input_2	2	10	0	0	0	1	0.00187	0.000233	0.023053	0.98307
28	Input_2	2	20	0	0	0	1	0.00119	0.000191	0.019091	0.99314
29	Input_2	2	25	0	0	0	1	0.001325	0.000217	0.021534	0.99149
30	Input_2	2	30	0	0	0	1	0.001205	0.0002	0.020002	0.99297
31	Input_2	3	7	0.000267	0.00012	0.012169	0.99984	0.001195	0.000196	0.019531	0.99309
32	Input_2	3	10	0	0	0	1	0.001873	0.00024	0.023709	0.98301
33	Input_2	3	20	0	0	0	1	0.001176	0.000175	0.017403	0.9933
34	Input_2	3	25	0	0	0	1	0.00196	0.000258	0.025225	0.98139
35	Input_2	3	30	0	0	0	1	0.00119	0.000182	0.018122	0.99314
36	Input_2	5	7	0.000349	0.00022	0.022283	0.99972	0.003177	0.000759	0.076011	0.95111
37	Input_2	5	10	0	0	0	1	0.002769	0.00054	0.05427	0.96288
38	Input_2	5	20	0	0	0	1	0.002457	0.00046	0.045948	0.97076
39	Input_2	5	25	0	0	0	1	0.002396	0.000451	0.045371	0.97221
40	Input_2	5	30	0	0	0	1	0.003125	0.000567	0.056607	0.95269
41	Input_3	1	7	0.000285	0.000116	0.011705	0.99981	0.001126	0.000138	0.013589	0.99386
42	Input_3	1	10	0.000152	0.000012	0.001204	0.99995	0.001117	0.000114	0.011171	0.99396
43	Input_3	1	20	0.000152	0.000012	0.001204	0.99995	0.001117	0.000114	0.011171	0.99396
44	Input_3	1	25	0.000152	0.000012	0.001204	0.99995	0.001117	0.000114	0.011171	0.99396
45	Input_3	1	30	0.000152	0.000012	0.001204	0.99995	0.001117	0.000114	0.011171	0.99396
46	Input_3	2	7	0.000319	0.000126	0.012819	0.99977	0.001162	0.000167	0.016602	0.99346
47	Input_3	2	10	0.00019	0.000014	0.00147	0.99992	0.001294	0.000174	0.017038	0.99189
48	Input_3	2	20	0.00019	0.000014	0.00147	0.99992	0.001291	0.000167	0.01633	0.99193
49	Input_3	2	25	0.00019	0.000014	0.00147	0.99992	0.001131	0.000129	0.012689	0.9938
50	Input_3	2	30	0.00019	0.000014	0.00147	0.99992	0.001275	0.000163	0.015839	0.99213
51	Input_3	3	7	0.000311	0.000127	0.012865	0.99978	0.00195	0.000251	0.024452	0.98158
52	Input_3	3	10	0.000181	0.000015	0.001603	0.99993	0.001947	0.000233	0.022654	0.98164
53	Input_3	3	20	0.000181	0.000015	0.001603	0.99993	0.001151	0.000152	0.01506	0.99358
54	Input_3	3	25	0.000181	0.000015	0.001603	0.99993	0.001306	0.000182	0.017758	0.99174
55	Input_3	3	30	0.000181	0.000015	0.001603	0.99993	0.001946	0.000229	0.02219	0.98166
56	Input_3	5	7	0.001509	0.000352	0.035793	0.99478	0.002064	0.000399	0.039387	0.97937
57	Input_3	5	10	0.001479	0.000193	0.019776	0.99499	0.001445	0.000278	0.027463	0.98989

58	Input_3	5	20	0.001479	0.000193	0.019776	0.99499	0.002018	0.0003	0.029434	0.98028
59	Input_3	5	25	0.001479	0.000193	0.019776	0.99499	0.002011	0.000298	0.029276	0.9804
60	Input_3	5	30	0.001479	0.000193	0.019776	0.99499	0.002018	0.0003	0.02945	0.98027
<b>Media de las métricas:</b>		<b>0.000212</b>	<b>0.000046</b>	<b>0.00469</b>	<b>0.99953</b>	<b>0.001735</b>	<b>0.000269</b>	<b>0.02671</b>	<b>0.98359</b>		

### 5.2.3. Random Forest

Para el método Random Forest, fueron utilizadas para los experimentos finales las ventanas de 1, 2, 3 y 5 días, debido al hecho de que ventanas mayores que 5 presentaron errores demasiadamente grandes, conforme explicado en el capítulo [4.3 Ventana deslizante](#).

Los hiperparámetros utilizados para este método son:

- maxDepth: Máxima profundidad del árbol, con valores: 7, 10, 20, 25 y 30.
- Número máximo de estimadores: 100

En la Tabla 5 se encuentran los resultados obtenidos de estos modelos, donde se señalan en verde y en rojo, respectivamente los 5 mejores y los 5 peores valores de RMSE obtenidos. Además, se incluyen las medias de las métricas al final de la tabla.

Se observa que se mantienen los mejores resultados con las predicciones utilizando ventanas de hasta 3 días, y los peores con las ventanas de 5 días.

En este método se nota una concentración de los mejores resultados para el conjunto de datos Input\_3 y para las profundidades medianas de nuestros experimentos. Los peores resultados se concentran en el Input\_2.

Random Forest presenta su mejor RMSE de un 0.000576 y RMSE medio de 0.001065.

**Tabla 5.** Resultados obtenidos con el Random Forest

#	Input	Vent	max Depth	Train				Test			
				RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>
1	Input_1	1	7	0.001211	0.000201	0.020869	0.996642	0.001016	0.000183	0.018408	0.994997
2	Input_1	1	10	0.001051	0.00016	0.016993	0.997468	0.000866	0.000165	0.016882	0.996367
3	Input_1	1	20	0.000992	0.000159	0.016812	0.997744	0.000888	0.000166	0.01698	0.996177
4	Input_1	1	25	0.001095	0.000164	0.017405	0.997254	0.000959	0.000177	0.017917	0.995544

5	Input_1	1	30	0.000998	0.000167	0.017436	0.997717	0.000898	0.000178	0.018072	0.996093
6	Input_1	2	7	0.001224	0.000206	0.021178	0.996571	0.001151	0.000215	0.021775	0.99358
7	Input_1	2	10	0.001149	0.000191	0.019958	0.996978	0.000985	0.000184	0.018728	0.995296
8	Input_1	2	20	0.001295	0.000186	0.019519	0.996157	0.000839	0.000161	0.016313	0.996589
9	Input_1	2	25	0.001062	0.000176	0.018278	0.997418	0.000902	0.000175	0.017807	0.99606
10	Input_1	2	30	0.001292	0.000196	0.020654	0.996176	0.000889	0.000175	0.017737	0.996176
11	Input_1	3	7	0.001272	0.000216	0.022489	0.996297	0.000998	0.000193	0.019697	0.995178
12	Input_1	3	10	0.001251	0.000188	0.019728	0.996415	0.000991	0.000199	0.020203	0.995247
13	Input_1	3	20	0.001374	0.000207	0.021608	0.995674	0.001121	0.000211	0.021461	0.993916
14	Input_1	3	25	0.001054	0.000179	0.01879	0.997456	0.000894	0.000205	0.020889	0.996132
15	Input_1	3	30	0.001209	0.000193	0.020205	0.996653	0.000951	0.000193	0.019699	0.995622
16	Input_1	5	7	0.001552	0.000371	0.038337	0.994481	0.001215	0.000436	0.044348	0.992845
17	Input_1	5	10	0.001532	0.000353	0.036687	0.994627	0.001387	0.000505	0.051197	0.990678
18	Input_1	5	20	0.001402	0.00033	0.033997	0.995496	0.001347	0.000536	0.054336	0.991208
19	Input_1	5	25	0.001458	0.000323	0.033487	0.99513	0.001253	0.000455	0.046197	0.992396
20	Input_1	5	30	0.001397	0.00033	0.034362	0.995529	0.001235	0.000506	0.051409	0.992607
21	Input_2	1	7	0.000877	0.000169	0.017681	0.998237	0.001187	0.000217	0.022334	0.993179
22	Input_2	1	10	0.000983	0.000159	0.016465	0.997787	0.001427	0.000237	0.024457	0.99014
23	Input_2	1	20	0.000689	0.000119	0.012551	0.998914	0.000951	0.000171	0.017699	0.995619
24	Input_2	1	25	0.000905	0.000147	0.015424	0.998126	0.001021	0.000184	0.018924	0.994953
25	Input_2	1	30	0.000855	0.000139	0.014536	0.998326	0.001175	0.000218	0.022454	0.993317
26	Input_2	2	7	0.000825	0.000166	0.017173	0.998443	0.001199	0.00023	0.023724	0.993039
27	Input_2	2	10	0.001048	0.000161	0.016799	0.997484	0.001168	0.000213	0.021851	0.993393
28	Input_2	2	20	0.000948	0.000156	0.016314	0.997942	0.001142	0.000226	0.023381	0.993686
29	Input_2	2	25	0.001028	0.000172	0.018008	0.997578	0.001452	0.000259	0.026809	0.989795
30	Input_2	2	30	0.000957	0.000147	0.015211	0.997901	0.001038	0.000206	0.021091	0.994779
31	Input_2	3	7	0.00108	0.000179	0.018924	0.997327	0.001072	0.000219	0.022593	0.994431
32	Input_2	3	10	0.000981	0.000156	0.016374	0.997798	0.001378	0.000252	0.026201	0.990799
33	Input_2	3	20	0.001033	0.000161	0.017008	0.997556	0.001121	0.000219	0.022568	0.993913
34	Input_2	3	25	0.001033	0.000171	0.01756	0.997556	0.0011	0.000212	0.021683	0.994141
35	Input_2	3	30	0.000863	0.000149	0.015753	0.998294	0.001082	0.000196	0.020268	0.994328
36	Input_2	5	7	0.001351	0.000332	0.034236	0.99582	0.002139	0.000649	0.066221	0.977839
37	Input_2	5	10	0.001315	0.000274	0.028535	0.996039	0.001989	0.000524	0.05331	0.980842
38	Input_2	5	20	0.001441	0.000301	0.031329	0.995243	0.002062	0.000561	0.057221	0.979417
39	Input_2	5	25	0.001344	0.000287	0.02978	0.995863	0.002137	0.00059	0.060134	0.977888
40	Input_2	5	30	0.001314	0.000294	0.030366	0.996048	0.002065	0.000587	0.059873	0.979355
41	Input_3	1	7	0.000772	0.000116	0.011993	0.998634	0.00064	0.000115	0.011523	0.998014
42	Input_3	1	10	0.000637	0.000084	0.008673	0.999072	0.00062	0.000101	0.010146	0.998141
43	Input_3	1	20	0.000691	0.000087	0.008991	0.998907	0.000618	0.000102	0.010179	0.998149
44	Input_3	1	25	0.000619	0.000077	0.007983	0.999123	0.000642	0.000103	0.010256	0.998007
45	Input_3	1	30	0.000624	0.000081	0.008318	0.999109	0.000576	0.000096	0.009762	0.998393
46	Input_3	2	7	0.000818	0.000136	0.014182	0.998466	0.000734	0.000143	0.014392	0.997393
47	Input_3	2	10	0.000899	0.000125	0.01292	0.998149	0.000722	0.000126	0.012724	0.997473
48	Input_3	2	20	0.000912	0.000124	0.012961	0.998093	0.000613	0.000117	0.011862	0.99818
49	Input_3	2	25	0.000735	0.000109	0.011314	0.998762	0.000726	0.000134	0.013528	0.997449
50	Input_3	2	30	0.00076	0.000115	0.01203	0.998678	0.000639	0.000122	0.012464	0.998022
51	Input_3	3	7	0.000935	0.000155	0.015882	0.997996	0.000775	0.000162	0.016411	0.997089
52	Input_3	3	10	0.000977	0.000143	0.014962	0.997813	0.000706	0.000146	0.014858	0.997587
53	Input_3	3	20	0.000649	0.000112	0.011706	0.999036	0.000604	0.000123	0.012579	0.998233
54	Input_3	3	25	0.00086	0.000128	0.013256	0.998304	0.000685	0.00014	0.014215	0.997729
55	Input_3	3	30	0.000882	0.000134	0.013952	0.998218	0.000673	0.000139	0.014205	0.997804
56	Input_3	5	7	0.001758	0.000387	0.039753	0.99292	0.001006	0.00036	0.036503	0.995095
57	Input_3	5	10	0.001806	0.000369	0.037788	0.992531	0.00115	0.000351	0.035567	0.993599

58	Input_3	5	20	0.001845	0.000377	0.038735	0.992203	0.001052	0.000326	0.032967	0.994643
59	Input_3	5	25	0.001745	0.000356	0.036841	0.993025	0.001006	0.00032	0.032659	0.995094
60	Input_3	5	30	0.001794	0.000355	0.036572	0.992631	0.00104	0.000319	0.032467	0.994765
<b>Media de las métricas:</b>		<b>0.001108</b>	<b>0.000198</b>	<b>0.020627</b>	<b>0.996964</b>	<b>0.001065</b>	<b>0.000249</b>	<b>0.025369</b>	<b>0.993807</b>		

### 5.2.4. MLP – Perceptrón Multicapa

Para los experimentos con el Perceptrón Multicapa, se ha utilizado el *early\_stopping* con el 10% del conjunto de entrenamiento para validación, paciencia máxima de 3 épocas y tolerancia de 0.000001, utilizando el MSE como métrica. El número de épocas que se han utilizado en el entrenamiento de cada modelo se encuentra en la columna “epoch”.

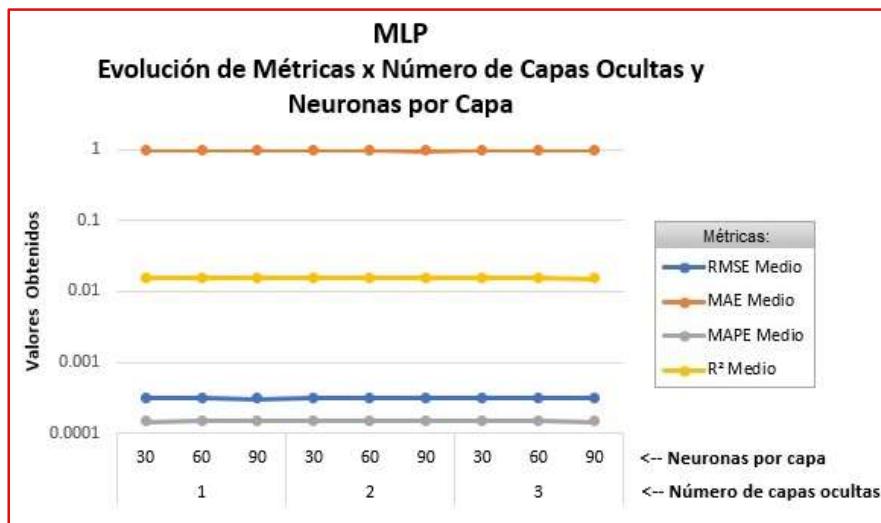
En los experimentos iniciales fueron probados diversos valores para *solver* (“lbfgs”, “sgd” y “adam”) y *activation* (“identity”, “logistic”, “tanh” y “relu”). Sin embargo, como se puede observar en la **Figura 19**, estos parámetros presentaron diferencias muy expresivas en sus métricas. De hecho, en la documentación oficial de *Sklearn* [32], es informado que el *solver* “lbfgs” es más indicado para conjuntos de datos pequeños, con menos de miles de registros.



**Figura 19.** Gráfico de evolución de las métricas para hiperparámetros MLP.  
Elaboración propia.

Por esta razón, para los experimentos finales fueron seleccionados los dos parámetros con mejor performance: “lbfgs” para *solver* y “identity” como función de activación.

Otra percepción obtenida de los experimentos iniciales es en relación a la cantidad de capas ocultas y neuronas por capa. Para nuestros conjuntos de datos, la variación de las métricas ha mostrado ser extremadamente baja. Es decir, para nuestros datos si insertamos más capas ocultas o neuronas no tenemos diferencias significativas en los resultados. En la **Figura 20** se muestra dicho análisis.



**Figura 20.** Gráfico de evolución de las métricas por capas ocultas y neuronas en MLP.  
Elaboración propia.

Tras este análisis se ha elegido construir los modelos con 1 capa oculta y 30 neuronas.

Así como para los anteriores métodos, fueron utilizadas para los experimentos finales las ventanas de 1, 2, 3 y 5 días, debido al hecho de que ventanas mayores que 5 presentaron errores demasiadamente grandes, conforme explicado en el capítulo [4.3 Ventana deslizante](#).

En la Tabla 6 se encuentran los resultados obtenidos de estos modelos, donde se señalan en verde y en rojo, respectivamente los 5 mejores y los 5 peores valores de RMSE obtenidos. Además, se incluyen las medias de las métricas al final de la tabla.

Se observa que se obtienen los mejores resultados con las predicciones utilizando ventanas de hasta 2 días, y los peores con las ventanas de 5 días.

En este método se nota por el número de épocas para entrenamientos de los modelos, que el conjunto de datos Input\_3 ha sido bastante más eficiente. Sin embargo, los mejores resultados se encuentran con el menor número de ventanas.

El RMSE medio obtenido ha sido de un 0.000317 y su mejor resultado ha sido un 0.000078.

**Tabla 6.** Resultados obtenidos con Perceptrón Multicapa

#	Input	Vent	Train				Test				epoch
			RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>	
1	Input_1	1	0.000173	0.000029	0.002995	0.99993	0.000078	0.000036	0.003712	0.99997	43
2	Input_1	2	0.000263	0.00005	0.005114	0.99984	0.000202	0.000075	0.007656	0.9998	50
3	Input_1	3	0.000251	0.000054	0.005498	0.99986	0.000259	0.000082	0.008307	0.99968	44
4	Input_1	5	0.002032	0.00055	0.056288	0.99054	0.000747	0.000532	0.053966	0.9973	59
5	Input_2	1	0.000174	0.000024	0.002418	0.99993	0.000073	0.000031	0.003119	0.99997	13
6	Input_2	2	0.000264	0.000042	0.004291	0.99984	0.000195	0.000064	0.006507	0.99982	18
7	Input_2	3	0.000252	0.000048	0.004971	0.99985	0.000255	0.000079	0.007996	0.99969	13
8	Input_2	5	0.001952	0.000625	0.063762	0.99128	0.000908	0.000548	0.055643	0.99601	19
9	Input_3	1	0.000175	0.000016	0.001706	0.99993	0.000068	0.000013	0.001365	0.99998	5
10	Input_3	2	0.000266	0.00003	0.003068	0.99984	0.000192	0.000034	0.003427	0.99982	6
11	Input_3	3	0.000255	0.000035	0.003572	0.99985	0.000253	0.000045	0.004561	0.99969	9
12	Input_3	5	0.002068	0.000423	0.043483	0.99021	0.000577	0.000289	0.029185	0.99839	9
<b>Media de las métricas:</b>			<b>0.000677</b>	<b>0.000161</b>	<b>0.016431</b>	<b>0.99757</b>	<b>0.000317</b>	<b>0.000152</b>	<b>0.015454</b>	<b>0.99918</b>	

### 5.2.5. LSTM

Para los experimentos con el LSTM, se ha utilizado el *early\_stopping* con el 20% del conjunto de entrenamiento para validación, paciencia máxima de 5 épocas y tolerancia de 0.000001, utilizando el MSE como métrica. El número de épocas que se han utilizado en el entrenamiento de cada modelo se encuentra en la columna “epoch”.

En los experimentos iniciales fueron probadas diversas configuraciones para los modelos, con hasta 3 capas y variaciones de sus neuronas. Sin embargo, los resultados no han variado significativamente (posiblemente porque estos modelos necesiten una mayor cantidad de datos para aprender correctamente).

Por esa razón, para los experimentos finales se ha utilizado el optimizador “adam” y una única capa oculta con las unidades LSTM para todos los experimentos.

Así como para los anteriores métodos, fueron utilizadas para los experimentos finales las ventanas de 1, 2, 3 y 5 días, debido al hecho de que ventanas mayores que 5 presentaron errores demasiadamente grandes, conforme explicado en el capítulo [4.3 Ventana deslizante](#).

Los hiperparámetros utilizados para este método son:

- Neur = Número de neuronas en la capa oculta, con valores: 3, 5, 10, 32.
- Activ = Función de activación en la capa oculta, con valores: “tanh”, “sigmoid”, “relu”.

En la Tabla 7 se encuentran los resultados obtenidos de estos modelos, donde se señalan en verde y en rojo, respectivamente los 5 mejores y los 5 peores valores de RMSE obtenidos. Además, se incluyen las medias de las métricas al final de la tabla.

Se observa que los mejores resultados no siguen un patrón como en los otros métodos. El mejor RMSE obtenido es de un 0.012109, siendo su media de un 0.013101.

**Tabla 7.** Resultados obtenidos con LSTM

#	Input	Vent	Neur	Activ	Train					Test					epoch
					RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	
1	Input_1	1	3	tanh	0.01487	0.007687	0.786195	0.492681	0.014003	0.00801	0.814949	0.050325	0.012109	0.00961	134
2	Input_1	1	3	sigmoid	0.016128	0.008198	0.839794	0.403182	0.012273	0.00608	0.616436	0.270483	0.013101	0.00961	111
3	Input_1	1	3	relu	0.015552	0.008072	0.824343	0.445081	0.013151	0.00667	0.676378	0.162385	0.012109	0.00961	32
4	Input_1	1	5	tanh	0.01595	0.008614	0.8809	0.416315	0.012554	0.0072	0.729616	0.236663	0.013101	0.00961	36
5	Input_1	1	5	sigmoid	0.015878	0.008052	0.824189	0.421576	0.012307	0.0063	0.638388	0.266431	0.012109	0.00961	168
6	Input_1	1	5	relu	0.014983	0.007969	0.814029	0.48495	0.013814	0.00743	0.753531	0.075777	0.013101	0.00961	36
7	Input_1	1	10	tanh	0.014309	0.007355	0.751233	0.530213	0.015374	0.00915	0.93025	-0.14469	0.012109	0.00961	66
8	Input_1	1	10	sigmoid	0.015817	0.008431	0.862001	0.426019	0.012539	0.00672	0.681561	0.238563	0.012109	0.00961	105
9	Input_1	1	10	relu	0.016185	0.009679	0.988081	0.398984	0.016054	0.01013	1.02	-0.24824	0.012109	0.00961	20
10	Input_1	1	32	tanh	0.014701	0.007889	0.804104	0.504179	0.0152	0.00956	0.972006	-0.11902	0.012109	0.00961	24
11	Input_1	1	32	sigmoid	0.016768	0.009196	0.942043	0.354904	0.012862	0.0069	0.700799	0.198778	0.012109	0.00961	21
12	Input_1	1	32	relu	0.016123	0.009913	1.01	0.403595	0.014477	0.00851	0.865242	-0.01511	0.012109	0.00961	11
13	Input_1	2	3	tanh	0.015209	0.007757	0.793968	0.468809	0.012459	0.00708	0.71859	0.248179	0.012109	0.00961	28
14	Input_1	2	3	sigmoid	0.016181	0.008354	0.856354	0.398772	0.012268	0.00604	0.611258	0.271061	0.012109	0.00961	78
15	Input_1	2	3	relu	0.014212	0.007814	0.796229	0.536165	0.014877	0.00799	0.811698	-0.07198	0.012109	0.00961	74
16	Input_1	2	5	tanh	0.01504	0.007639	0.78128	0.480572	0.012462	0.00719	0.730407	0.247844	0.012109	0.00961	32
17	Input_1	2	5	sigmoid	0.016107	0.008352	0.85564	0.404256	0.012263	0.00608	0.615474	0.271668	0.012109	0.00961	56
18	Input_1	2	5	relu	0.015048	0.007897	0.806854	0.480017	0.013697	0.00766	0.777571	0.091353	0.012109	0.00961	45
19	Input_1	2	10	tanh	0.014542	0.007452	0.761192	0.514374	0.014123	0.00826	0.840645	0.033925	0.012109	0.00961	57
20	Input_1	2	10	sigmoid	0.015401	0.007944	0.812022	0.455343	0.012919	0.00706	0.714559	0.191617	0.012109	0.00961	120
21	Input_1	2	10	relu	0.015142	0.008645	0.884169	0.473457	0.014471	0.00912	0.927252	-0.01428	0.012109	0.00961	20
22	Input_1	2	32	tanh	0.013992	0.007342	0.748911	0.55042	0.015381	0.00961	0.977858	-0.14574	0.012109	0.00961	27

23	Input_1	2	32	sigmoid	0.015368	0.008324	0.850583	0.457673	0.013612	0.00804	0.815199	0.102641	81
24	Input_1	2	32	relu	0.013517	0.007268	0.741855	0.580419	0.016131	0.00978	0.995724	-0.26025	34
25	Input_1	3	3	tanh	0.014895	0.00785	0.803349	0.49028	0.013211	0.00797	0.809116	0.154726	48
26	Input_1	3	3	sigmoid	0.015905	0.008107	0.830708	0.418833	0.012228	0.00607	0.613869	0.275792	105
27	Input_1	3	3	relu	0.014783	0.007577	0.77546	0.49789	0.012286	0.00656	0.665818	0.268904	71
28	Input_1	3	5	tanh	0.014858	0.007666	0.783374	0.492809	0.012738	0.00729	0.740162	0.214093	31
29	Input_1	3	5	sigmoid	0.015873	0.008143	0.833887	0.421157	0.012271	0.00616	0.623091	0.270759	83
30	Input_1	3	5	relu	0.014236	0.007136	0.728555	0.534401	0.015347	0.00781	0.792236	-0.14075	99
31	Input_1	3	10	tanh	0.014468	0.007545	0.770713	0.519078	0.013512	0.00841	0.853777	0.11573	57
32	Input_1	3	10	sigmoid	0.015767	0.008265	0.846306	0.428883	0.012632	0.00679	0.687476	0.227121	68
33	Input_1	3	10	relu	0.014169	0.007558	0.771383	0.538745	0.015323	0.00778	0.787101	-0.13715	43
34	Input_1	3	32	tanh	0.012687	0.006787	0.690996	0.630216	0.014976	0.00901	0.916756	-0.08632	50
35	Input_1	3	32	sigmoid	0.015849	0.008462	0.866103	0.422919	0.013035	0.00727	0.737059	0.17706	36
36	Input_1	3	32	relu	0.013351	0.00725	0.739193	0.590492	0.015922	0.00927	0.942022	-0.22784	38
37	Input_1	5	3	tanh	0.015364	0.008141	0.833032	0.457173	0.012988	0.00768	0.777897	0.183007	28
38	Input_1	5	3	sigmoid	0.015631	0.00798	0.816687	0.438146	0.012397	0.00648	0.656006	0.255612	84
39	Input_1	5	3	relu	0.014645	0.007549	0.771294	0.506771	0.013283	0.00705	0.714342	0.145468	149
40	Input_1	5	5	tanh	0.015447	0.008853	0.903826	0.451278	0.013824	0.00896	0.908321	0.074448	21
41	Input_1	5	5	sigmoid	0.015759	0.008287	0.848367	0.428843	0.012465	0.00671	0.679406	0.247502	52
42	Input_1	5	5	relu	0.016863	0.00933	0.960741	0.346046	0.013343	0.00698	0.708653	0.137738	48
43	Input_1	5	10	tanh	0.014552	0.00759	0.775227	0.51302	0.013566	0.00844	0.856556	0.108633	30
44	Input_1	5	10	sigmoid	0.015227	0.00791	0.808919	0.4668	0.012864	0.00733	0.742514	0.198478	107
45	Input_1	5	10	relu	0.014325	0.007683	0.785643	0.528053	0.014717	0.00758	0.768045	-0.04907	29
46	Input_1	5	32	tanh	0.013966	0.007514	0.767275	0.551431	0.01523	0.0096	0.975801	-0.12339	32
47	Input_1	5	32	sigmoid	0.016003	0.008681	0.888985	0.411066	0.012993	0.00738	0.748668	0.182365	20
48	Input_1	5	32	relu	0.014269	0.007639	0.779741	0.531766	0.013987	0.00833	0.845515	0.052472	17
49	Input_2	1	3	tanh	0.014584	0.007666	0.784481	0.511982	0.012887	0.00804	0.815797	0.19562	58
50	Input_2	1	3	sigmoid	0.016046	0.008202	0.839549	0.409234	0.012245	0.00634	0.641695	0.273824	237
51	Input_2	1	3	relu	0.013407	0.007717	0.788058	0.587599	0.014101	0.00909	0.923565	0.036905	153
52	Input_2	1	5	tanh	0.014829	0.007853	0.803013	0.495463	0.012695	0.00796	0.807716	0.219484	47
53	Input_2	1	5	sigmoid	0.015871	0.00813	0.832358	0.422099	0.012232	0.00656	0.663857	0.275359	313
54	Input_2	1	5	relu	0.016143	0.010389	1.05	0.402089	0.015579	0.01127	1.14	-0.17555	33
55	Input_2	1	10	tanh	0.014772	0.00767	0.783839	0.499324	0.013056	0.00843	0.855561	0.174365	28
56	Input_2	1	10	sigmoid	0.015731	0.00823	0.841934	0.432259	0.012248	0.00677	0.685231	0.273423	248
57	Input_2	1	10	relu	0.013097	0.00746	0.758926	0.606465	0.01421	0.00922	0.936024	0.021976	69
58	Input_2	1	32	tanh	0.014095	0.008461	0.860963	0.544216	0.014614	0.00983	0.996308	-0.03431	51
59	Input_2	1	32	sigmoid	0.015623	0.008517	0.870009	0.44003	0.01263	0.00753	0.763162	0.227411	172
60	Input_2	1	32	relu	0.013737	0.008537	0.868429	0.567068	0.015871	0.011	1.11	-0.21995	23
61	Input_2	2	3	tanh	0.014742	0.007823	0.800187	0.500956	0.012667	0.00753	0.763305	0.222828	94
62	Input_2	2	3	sigmoid	0.015083	0.00744	0.761433	0.477587	0.012339	0.00625	0.632065	0.262654	469
63	Input_2	2	3	relu	0.014695	0.007471	0.764444	0.504138	0.01268	0.00705	0.713785	0.221243	98
64	Input_2	2	5	tanh	0.014719	0.007676	0.785392	0.502473	0.012423	0.00748	0.758798	0.25258	76
65	Input_2	2	5	sigmoid	0.014845	0.00738	0.754794	0.493916	0.012496	0.00669	0.676637	0.243699	500
66	Input_2	2	5	relu	0.014713	0.007467	0.765332	0.502918	0.012652	0.0068	0.688889	0.224672	188
67	Input_2	2	10	tanh	0.014746	0.007851	0.802054	0.500661	0.013191	0.00871	0.883532	0.157205	37
68	Input_2	2	10	sigmoid	0.014783	0.007606	0.778034	0.498149	0.01268	0.00732	0.741486	0.221268	326
69	Input_2	2	10	relu	0.013594	0.007619	0.777224	0.57561	0.013925	0.00921	0.932635	0.060842	54
70	Input_2	2	32	tanh	0.013602	0.007259	0.741438	0.575162	0.013972	0.009	0.912272	0.054514	51
71	Input_2	2	32	sigmoid	0.016119	0.008923	0.908113	0.403377	0.013317	0.00729	0.73534	0.141089	59
72	Input_2	2	32	relu	0.013545	0.007802	0.793766	0.578681	0.014158	0.00933	0.944836	0.029221	30
73	Input_2	3	3	tanh	0.014852	0.007582	0.775296	0.493243	0.012385	0.00713	0.722218	0.257147	48
74	Input_2	3	3	sigmoid	0.015336	0.007753	0.793784	0.459652	0.012341	0.00651	0.658844	0.262353	284
75	Input_2	3	3	relu	0.014306	0.007411	0.757156	0.52978	0.013284	0.00812	0.82266	0.145341	58

76	Input_2	3	5	tanh	0.014063	0.007135	0.728873	0.54565	0.012506	0.00721	0.730371	0.242493	88
77	Input_2	3	5	sigmoid	0.015047	0.007526	0.770185	0.479793	0.012534	0.007	0.708811	0.239174	263
78	Input_2	3	5	relu	0.014315	0.007769	0.792494	0.529186	0.013171	0.00803	0.815514	0.159834	68
79	Input_2	3	10	tanh	0.014362	0.007389	0.755514	0.526139	0.012667	0.00779	0.790026	0.222895	62
80	Input_2	3	10	sigmoid	0.014363	0.006872	0.704332	0.526036	0.012316	0.00691	0.700916	0.265375	333
81	Input_2	3	10	relu	0.013723	0.007489	0.763934	0.567345	0.013712	0.00849	0.860424	0.089386	41
82	Input_2	3	32	tanh	0.014414	0.007635	0.779032	0.522684	0.013451	0.00879	0.890492	0.123683	39
83	Input_2	3	32	sigmoid	0.015607	0.008773	0.899512	0.440374	0.013268	0.00844	0.856675	0.147388	83
84	Input_2	3	32	relu	0.012968	0.006915	0.704738	0.613644	0.013139	0.00801	0.812982	0.163832	45
85	Input_2	5	3	tanh	0.014485	0.006967	0.713112	0.517458	0.01225	0.00672	0.681011	0.273245	106
86	Input_2	5	3	sigmoid	0.016292	0.008565	0.878917	0.389558	0.012326	0.0063	0.637064	0.264176	113
87	Input_2	5	3	relu	0.015107	0.00811	0.830946	0.475125	0.013292	0.00759	0.767736	0.144281	178
88	Input_2	5	5	tanh	0.014312	0.006987	0.71513	0.528931	0.012297	0.00706	0.71445	0.267613	108
89	Input_2	5	5	sigmoid	0.015074	0.007647	0.781437	0.477443	0.012571	0.00705	0.713599	0.234649	233
90	Input_2	5	5	relu	0.013814	0.007051	0.720554	0.561127	0.012738	0.00712	0.720437	0.214111	55
91	Input_2	5	10	tanh	0.014439	0.007144	0.730517	0.52052	0.012732	0.00757	0.766982	0.214923	71
92	Input_2	5	10	sigmoid	0.014853	0.007477	0.764017	0.49268	0.012502	0.00694	0.70268	0.242983	195
93	Input_2	5	10	relu	0.014441	0.007377	0.754142	0.520404	0.013348	0.00753	0.762494	0.137119	32
94	Input_2	5	32	tanh	0.014516	0.007631	0.778419	0.515447	0.013472	0.00848	0.859398	0.120907	32
95	Input_2	5	32	sigmoid	0.016084	0.008991	0.921689	0.405063	0.013387	0.00834	0.846191	0.132088	17
96	Input_2	5	32	relu	0.013812	0.007291	0.746996	0.561276	0.013098	0.00779	0.790921	0.169113	33
97	Input_3	1	3	tanh	0.016185	0.008275	0.848483	0.398949	0.01231	0.0061	0.6168	0.266059	37
98	Input_3	1	5	tanh	0.01605	0.00814	0.834166	0.409005	0.012245	0.00608	0.61503	0.273851	124
99	Input_3	1	10	tanh	0.016066	0.008134	0.833586	0.407755	0.012273	0.00612	0.618886	0.270444	137
100	Input_3	1	32	tanh	0.015961	0.008091	0.828709	0.415487	0.012268	0.0061	0.616688	0.271068	199
101	Input_3	2	3	tanh	0.015027	0.007411	0.758656	0.481431	0.01226	0.00638	0.645297	0.272027	91
102	Input_3	2	5	tanh	0.014775	0.00719	0.735781	0.498711	0.012428	0.00641	0.648853	0.251973	83
103	Input_3	2	10	tanh	0.0149	0.007435	0.760574	0.490145	0.01243	0.00653	0.660931	0.251721	70
104	Input_3	2	32	tanh	0.014981	0.007614	0.778822	0.484633	0.012745	0.0068	0.688496	0.213281	52
105	Input_3	3	3	tanh	0.014589	0.007151	0.731847	0.510997	0.012289	0.00675	0.68346	0.268582	107
106	Input_3	3	5	tanh	0.01463	0.007065	0.722853	0.508282	0.012265	0.00666	0.673799	0.271466	74
107	Input_3	3	10	tanh	0.014641	0.007178	0.734579	0.507524	0.01238	0.0068	0.688843	0.257721	60
108	Input_3	3	32	tanh	0.014742	0.007495	0.767157	0.500726	0.012584	0.00714	0.722669	0.233085	41
109	Input_3	5	3	tanh	0.014595	0.00706	0.722919	0.510093	0.012211	0.00649	0.656496	0.277871	111
110	Input_3	5	5	tanh	0.014627	0.007232	0.739541	0.507952	0.012385	0.00678	0.685822	0.257051	61
111	Input_3	5	10	tanh	0.014553	0.007017	0.717481	0.512909	0.0125	0.00682	0.690814	0.243232	55
112	Input_3	5	32	tanh	0.014799	0.00776	0.794007	0.496321	0.012952	0.00743	0.752444	0.18748	34
113	Input_3	1	3	sigmoid	0.016276	0.008413	0.861392	0.392241	0.01229	0.00623	0.630803	0.268454	92
114	Input_3	1	5	sigmoid	0.020384	0.015026	1.53	0.046711	0.014129	0.00867	0.881529	0.033196	7
115	Input_3	1	10	sigmoid	0.01648	0.008689	0.889279	0.376841	0.012434	0.00642	0.649842	0.251161	58
116	Input_3	1	32	sigmoid	0.01691	0.00925	0.945895	0.343966	0.012723	0.00674	0.683386	0.21604	36
117	Input_3	2	3	sigmoid	0.015879	0.008165	0.835379	0.420952	0.012311	0.00637	0.643966	0.266	128
118	Input_3	2	5	sigmoid	0.015432	0.007888	0.807372	0.453123	0.012526	0.00657	0.664314	0.240033	265
119	Input_3	2	10	sigmoid	0.015042	0.007517	0.76983	0.480432	0.012462	0.0064	0.647009	0.247875	224
120	Input_3	2	32	sigmoid	0.016312	0.008785	0.899272	0.388972	0.0129	0.00703	0.712108	0.194041	20
121	Input_3	3	3	sigmoid	0.01578	0.008156	0.835075	0.427933	0.012344	0.0065	0.657923	0.262001	56
122	Input_3	3	5	sigmoid	0.016003	0.008366	0.85677	0.411634	0.012537	0.00667	0.674639	0.238689	49
123	Input_3	3	10	sigmoid	0.014813	0.007348	0.751815	0.49585	0.012457	0.00681	0.689555	0.248409	231
124	Input_3	3	32	sigmoid	0.016016	0.008593	0.879392	0.410705	0.013003	0.00731	0.740697	0.181067	27
125	Input_3	5	3	sigmoid	0.016001	0.008432	0.864358	0.411175	0.012477	0.00656	0.663396	0.246016	68
126	Input_3	5	5	sigmoid	0.015182	0.007782	0.796002	0.469906	0.01262	0.00725	0.73453	0.22866	144
127	Input_3	5	10	sigmoid	0.014837	0.007466	0.763898	0.493769	0.01262	0.00734	0.743555	0.22866	203
128	Input_3	5	32	sigmoid	0.015564	0.008603	0.881833	0.442917	0.013471	0.00874	0.886534	0.121152	82

129	Input_3	1	3	relu	0.020903	0.015344	1.52	-0.002462	0.014567	0.00911	0.92685	-0.02777	44
130	Input_3	1	5	relu	0.015966	0.008039	0.823657	0.415133	0.012313	0.00606	0.612664	0.26566	160
131	Input_3	1	10	relu	0.015985	0.008053	0.825053	0.413719	0.012332	0.0061	0.616384	0.263411	142
132	Input_3	1	32	relu	0.015937	0.008033	0.82298	0.417226	0.012258	0.00608	0.614743	0.272201	108
133	Input_3	2	3	relu	0.015014	0.007426	0.760042	0.482371	0.012575	0.00642	0.649603	0.234184	93
134	Input_3	2	5	relu	0.014859	0.007405	0.757903	0.49301	0.012926	0.00658	0.665553	0.190784	61
135	Input_3	2	10	relu	0.014986	0.007562	0.773473	0.484299	0.012749	0.00651	0.6582	0.212836	102
136	Input_3	2	32	relu	0.01473	0.007156	0.731534	0.50177	0.012655	0.00634	0.641183	0.224375	62
137	Input_3	3	3	relu	0.014686	0.007068	0.723304	0.504489	0.012453	0.00661	0.668142	0.248937	79
138	Input_3	3	5	relu	0.014437	0.006798	0.695736	0.521137	0.012109	0.00639	0.646637	0.289877	91
139	Input_3	3	10	relu	0.01435	0.006765	0.692594	0.526915	0.012182	0.00632	0.639833	0.281291	84
140	Input_3	3	32	relu	0.014417	0.007056	0.723332	0.52246	0.012322	0.0065	0.658873	0.264646	45
141	Input_3	5	3	relu	0.014816	0.007605	0.777231	0.495152	0.012573	0.00734	0.7438	0.234394	77
142	Input_3	5	5	relu	0.014585	0.007209	0.737908	0.510794	0.012778	0.007	0.709449	0.209244	73
143	Input_3	5	10	relu	0.014744	0.007539	0.771175	0.500092	0.01291	0.00732	0.74138	0.192785	34
144	Input_3	5	32	relu	0.014728	0.007248	0.741866	0.501156	0.012975	0.00716	0.724618	0.184617	19
<b>Media de las métricas:</b>					<b>0.015089</b>	<b>0.007938</b>	<b>0.811283</b>	<b>0.474468</b>	<b>0.013101</b>	<b>0.007430</b>	<b>0.75297</b>	<b>0.164153</b>	

## 6. ANÁLISIS DE LOS RESULTADOS

Siguiendo los objetivos propuestos para ese trabajo, en este capítulo se analizarán los resultados obtenidos en los experimentos del capítulo anterior. Utilizaremos como métrica principal el RMSE y las demás métricas ya mencionadas en caso de empate, con la siguiente orden de prioridad:

- 1) RMSE
- 2) MAE
- 3) MAPE
- 4)  $R^2$

Se recuerda que: RMSE y MAE son medidas en las mismas unidades que los datos y MAPE es una medida de porcentaje. Las tres son métricas a minimizar. Sin embargo, el  $R^2$  es medido de 0 a 1, y se busca su medida máxima.

Iniciamos el capítulo con el análisis de los mejores modelos para cada uno de los conjuntos de entradas de los experimentos. Luego se analizan los mejores resultados obtenidos para cada modelo utilizado y finalmente se hará un análisis global de todos los modelos experimentados.

### 6.1. Análisis por grupos de entrada

Para este análisis, seleccionamos los modelos con los 5 mejores resultados para cada uno de los *Inputs*. Los modelos son ordenados del mejor al peor resultado.

#### Input\_1

En la Tabla 8 observamos que para el conjunto de entradas Input\_1, los mejores modelos son los con MLP y Regresión Lineal.

**Tabla 8.** Mejores resultados para Input\_1

Input_1		Vent	Train				Test			
#	Modelo		RMSE	MAE	MAPE	$R^2$	RMSE	MAE	MAPE	$R^2$
1	MLP (43 épocas)	1	0.000173	0.000029	0.002995	0.999931	0.000078	0.000036	0.003712	0.999971

2	Regresión Lineal	1	0.000173	0.000029	0.002985	0.999931	0.000078	0.000037	0.003788	0.999971
3	Regresión Lineal	2	0.000263	0.00005	0.00506	0.999842	0.000201	0.000073	0.007447	0.999804
4	MLP (50 épocas)	2	0.000263	0.00005	0.005114	0.999842	0.000202	0.000075	0.007656	0.999802
5	MLP (44 épocas)	3	0.000251	0.000054	0.005498	0.999856	0.000259	0.000082	0.008307	0.999676

## Input\_2

En la Tabla 9 observamos que para el conjunto de entradas Input\_2, los mejores modelos son los con MLP y Regresión Lineal

**Tabla 9.** Mejores resultados para Input\_2

Input_2			Vent	Train				Test			
#	Modelo			RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>
1	MLP (13 épocas)	1		0.000174	0.000024	0.002418	0.99993	0.000073	0.000031	0.003119	0.999974
2	Regresión Lineal	1	0.000174	0.000024	0.002415	0.99993	0.000073	0.000031	0.003113	0.999974	
3	MLP (18 épocas)	2	0.000264	0.000042	0.004291	0.99984	0.000195	0.000064	0.006507	0.999816	
4	Regresión Lineal	2	0.000264	0.000042	0.004304	0.99984	0.000195	0.000064	0.006527	0.999816	
5	MLP (13 épocas)	3	0.000252	0.000048	0.004971	0.999854	0.000255	0.000079	0.007996	0.999685	

## Input\_3

En la Tabla 10 observamos que para el conjunto de entradas Input\_3, los mejores modelos son los con MLP y Regresión Lineal

**Tabla 10.** Mejores resultados para Input\_3

Input_3			Vent	Train				Test			
#	Modelo			RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>
1	MLP (5 épocas)	1		0.000175	0.000016	0.001706	0.99993	0.000068	0.000013	0.001365	0.999977
2	Regresión Lineal	1	0.000175	0.000016	0.001702	0.99993	0.000068	0.000013	0.001361	0.999977	
3	MLP (6 épocas)	2	0.000266	0.00003	0.003068	0.999838	0.000192	0.000034	0.003427	0.999822	
4	Regresión Lineal	2	0.000266	0.00003	0.003069	0.999838	0.000192	0.000034	0.003427	0.999822	
5	MLP (9 épocas)	3	0.000255	0.000035	0.003572	0.999852	0.000253	0.000045	0.004561	0.999689	

Entre los conjuntos de entradas, se nota una ligera mejora en el Input\_3, tanto de las métricas, como por ejemplo de la cantidad de épocas necesarias para el entrenamiento.

## 6.2. Análisis por algoritmo

Para este análisis, seleccionamos los modelos con los 5 mejores resultados para cada uno de los algoritmos utilizados, independientemente de la entrada de datos utilizada.

Los modelos son ordenados del mejor al peor resultado.

### Regresión Lineal

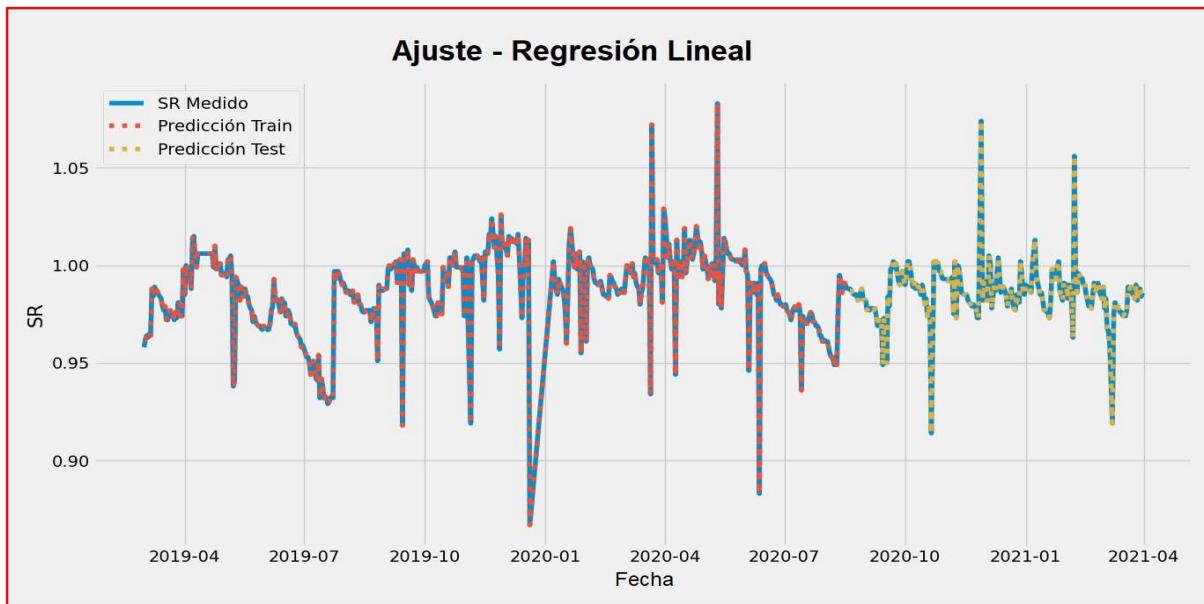
En la Tabla 11 observamos que los mejores modelos se los han obtenido con ventanas de 1 y 2 días. Además, el método de Regresión Lineal ha sido eficiente para los 3 conjuntos de datos, con su mejor resultado para el Input\_3 con RMSE de un 0.000068. Entre los mejores resultados, su media de RMSE es de un 0.0001212.

**Tabla 11.** Mejores resultados para el algoritmo de Regresión Lineal

#	Input	Vent	Train				Test			
			RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>
1	Input_3	1	0.000175	0.000016	0.001702	0.99993	0.000068	0.000013	0.001361	0.999977
2	Input_2	1	0.000174	0.000024	0.002415	0.99993	0.000073	0.000031	0.003113	0.999974
3	Input_1	1	0.000173	0.000029	0.002985	0.999931	0.000078	0.000037	0.003788	0.999971
4	Input_3	2	0.000266	0.00003	0.003069	0.999838	0.000192	0.000034	0.003427	0.999822
5	Input_2	2	0.000264	0.000042	0.004304	0.99984	0.000195	0.000064	0.006527	0.999816
<b>Media de las métricas:</b>			<b>0.0002104</b>	<b>0.0000282</b>	<b>0.002895</b>	<b>0.9998938</b>	<b>0.0001212</b>	<b>0.0000358</b>	<b>0.0036432</b>	<b>0.999912</b>

El RMSE medio obtenido con la Regresión Lineal ha sido de un 0.001022, y su peor resultado un 0.003456, para el conjunto Input\_1, con ventana de 14 días.

En la **Figura 21**, se expone la gráfica del ajuste para el mejor modelo obtenido. Se nota que el ajuste se aadecua muy bien. La línea azul, correspondiente al SR Medido (el *label* de nuestro conjunto de datos) está prácticamente cubierta por las líneas de las predicciones.



**Figura 21. Gráfico del ajuste del mejor modelo de Regresión Lineal**  
Elaboración propia

## Árbol de Decisión

En la Tabla 12 observamos que los 4 mejores resultados son para la entrada del Input\_3, con ventana de 1 día y variados *maxDepth*, lo que nos lleva a creer que, para esta configuración, la profundidad máxima del árbol no es tan relevante. Por esta razón, se presentan en esta tabla los 10 mejores resultados, con la finalidad de tener una muestra mayor de los resultados obtenidos.

Para los resultados de Árbol de Decisión, la predominancia es de mejores resultados para el Input\_3, aunque con alguna variedad de ventanas.

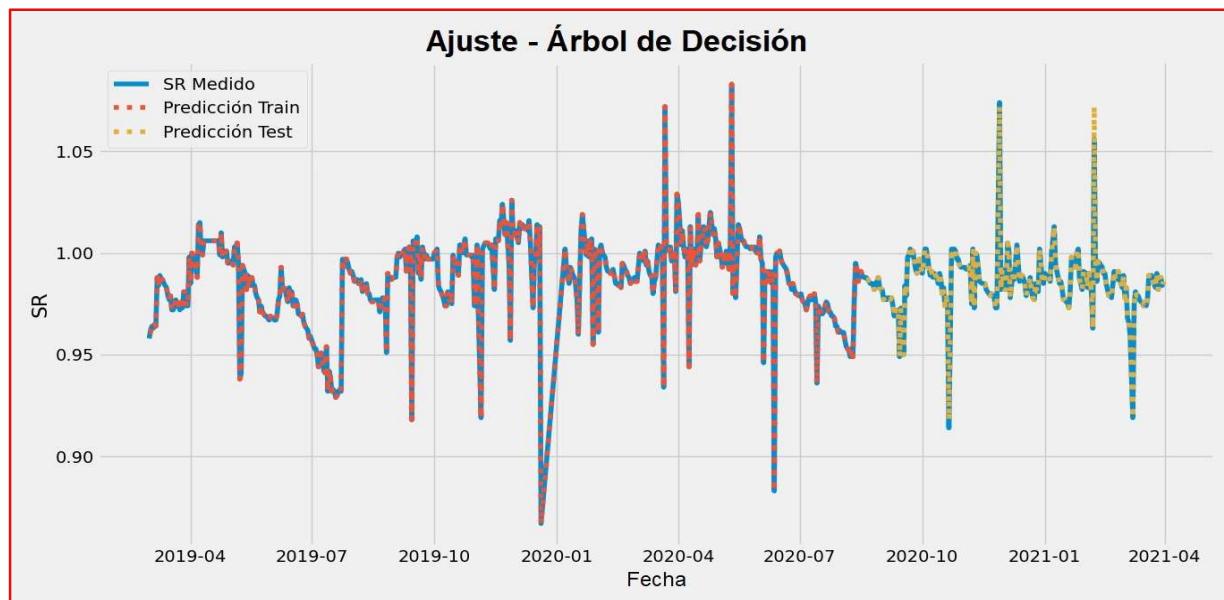
**Tabla 12.** Mejores resultados para el algoritmo de Árbol de Decisión

#	Input	Vent	max Depth	Train				Test			
				RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>
1	Input_3	1	10	0.000152	0.000012	0.001204	0.99995	0.001117	0.000114	0.011171	0.99396
2	Input_3	1	20	0.000152	0.000012	0.001204	0.99995	0.001117	0.000114	0.011171	0.99396
3	Input_3	1	25	0.000152	0.000012	0.001204	0.99995	0.001117	0.000114	0.011171	0.99396
4	Input_3	1	30	0.000152	0.000012	0.001204	0.99995	0.001117	0.000114	0.011171	0.99396
5	Input_3	1	7	0.000285	0.000116	0.011705	0.99981	0.001126	0.000138	0.013589	0.99386
6	Input_3	2	25	0.00019	0.000014	0.00147	0.99992	0.001131	0.000129	0.012689	0.9938
7	Input_3	3	20	0.000181	0.000015	0.001603	0.99993	0.001151	0.000152	0.01506	0.99358

8	Input_1	2	30	0	0	0	1	0.001154	0.000146	0.014419	0.99356
9	Input_2	1	20	0	0	0	1	0.001156	0.000164	0.016282	0.99353
10	Input_3	2	7	0.000319	0.000126	0.012819	0.99977	0.001162	0.000167	0.016602	0.99346
<b>Media de las métricas:</b>		<b>0.000158</b>	<b>0.000032</b>	<b>0.003241</b>	<b>0.999921</b>	<b>0.001135</b>	<b>0.000135</b>	<b>0.013333</b>	<b>0.993762</b>		

El RMSE medio obtenido con Árbol de Decisión ha sido de un 0.001135, y su mejor resultado un 0.001117, para el conjunto Input\_3, con ventana de 1 día, sin distinguir la profundidad.

En la **Figura 22**, se expone la gráfica del ajuste para el mejor modelo obtenido. Se ha utilizado la profundidad máxima =10 para el gráfico. Se nota que el ajuste se adapta muy bien. La línea azul, correspondiente al SR Medido (el *label* de nuestro conjunto de datos) está prácticamente cubierta por las líneas de las predicciones.



**Figura 22.** Gráfico del ajuste del mejor modelo con Árbol de Decisión

Elaboración propia

## Random Forest

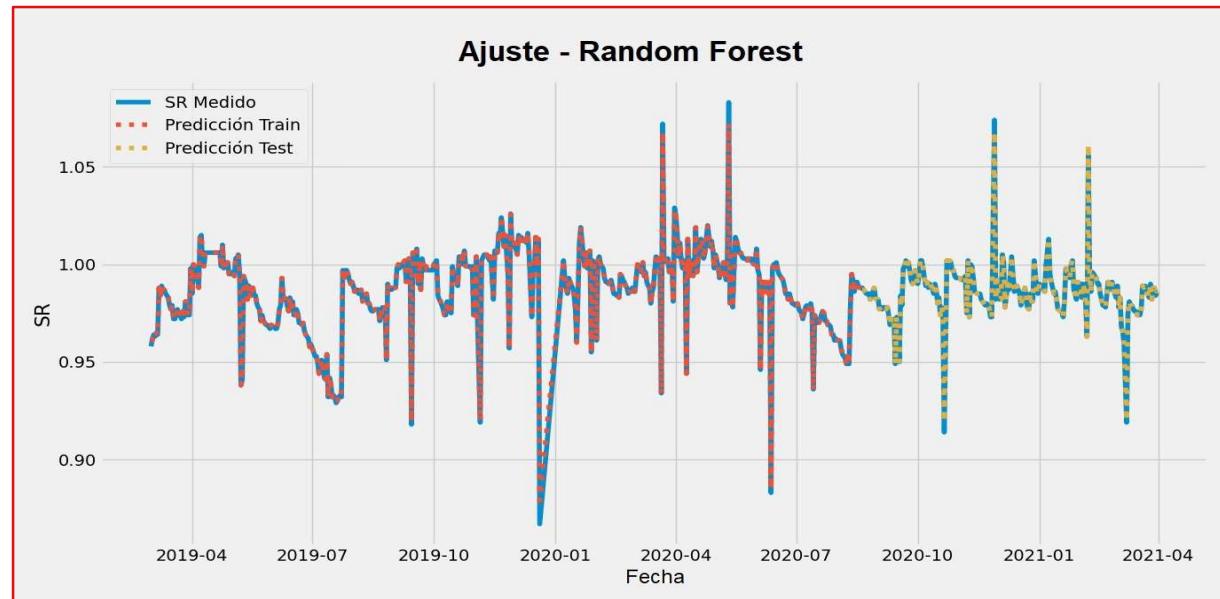
En los resultados de Random Forest (Tabla 13), el conjunto de datos Input\_3 se destaca entre las 5 primeras posiciones. También se identifica que la profundidad máxima es un hiperparámetro importante en estos modelos.

**Tabla 13.** Mejores resultados para el algoritmo Random Forest

#	Input	Vent	max Depth	Train				Test			
				RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>
1	Input_3	1	30	0.000624	0.000081	0.008318	0.999109	0.000576	0.000096	0.009762	0.998393
2	Input_3	3	20	0.000649	0.000112	0.011706	0.999036	0.000604	0.000123	0.012579	0.998233
3	Input_3	2	20	0.000912	0.000124	0.012961	0.998093	0.000613	0.000117	0.011862	0.99818
4	Input_3	1	20	0.000691	0.000087	0.008991	0.998907	0.000618	0.000102	0.010179	0.998149
5	Input_3	1	10	0.000637	0.000084	0.008673	0.999072	0.00062	0.000101	0.010146	0.998141
Media de las métricas:				0.000703	0.000098	0.010130	0.998843	0.000606	0.000108	0.010906	0.998219

El RMSE medio obtenido con Random Forest ha sido de un 0.000606, y su mejor resultado un 0.000576, para el conjunto Input\_3, con ventana de 1 día, y máxima profundidad de 30. Diferentemente de los resultados con Árbol de Decisión, en este caso los *maxDepth* más altos obtienen mejores resultados.

En la **Figura 23**, se expone la gráfica del ajuste para el mejor modelo obtenido. Se nota que el ajuste se aadecua muy bien. La línea azul, correspondiente al SR Medido (el *label* de nuestro conjunto de datos) está prácticamente cubierta por las líneas de las predicciones, con excepción de algunos pocos puntos en las extremidades superiores e inferiores.



**Figura 23.** Gráfico del ajuste del mejor modelo con Random Forest  
Elaboración propia

## MLP – Perceptrón Multicapa

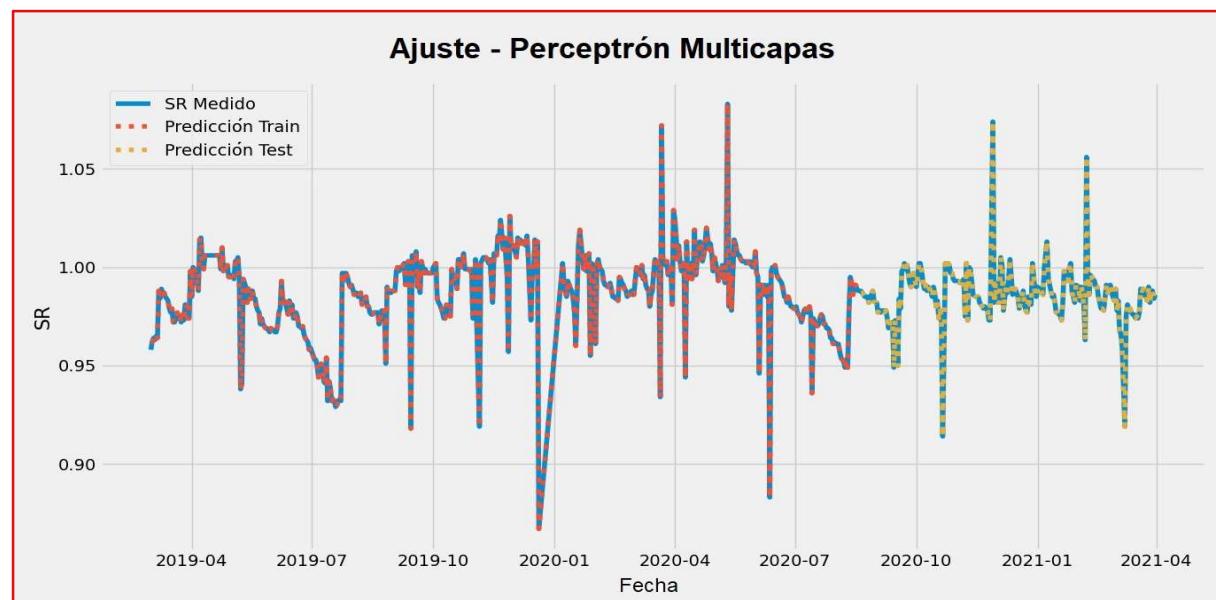
Entre los mejores resultados del método MLP (Tabla 14) se contemplan los tres conjuntos de datos, con mejores resultados para las ventanas de 1 y 2 días.

Se observa que el número de épocas de entrenamiento ha variado en un largo rango, de 5 a 43 épocas.

**Tabla 14.** Mejores resultados para el algoritmo MLP - Perceptrón Multicapa

#	Input	Vent	Train				Test				epoch
			RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>	
1	Input_3	1	0.000175	0.000016	0.001706	0.99993	0.000068	0.000013	0.001365	0.99998	5
2	Input_2	1	0.000174	0.000024	0.002418	0.99993	0.000073	0.000031	0.003119	0.99997	13
3	Input_1	1	0.000173	0.000029	0.002995	0.99993	0.000078	0.000036	0.003712	0.99997	43
4	Input_3	2	0.000266	0.00003	0.003068	0.99984	0.000192	0.000034	0.003427	0.99982	6
5	Input_2	2	0.000264	0.000042	0.004291	0.99984	0.000195	0.000064	0.006507	0.99982	18
<b>Media de las métricas:</b>			<b>0.00021</b>	<b>0.000028</b>	<b>0.002896</b>	<b>0.99989</b>	<b>0.000121</b>	<b>0.000036</b>	<b>0.003626</b>	<b>0.99991</b>	

El RMSE medio obtenido con MLP ha sido de un 0.000121, y su mejor resultado un 0.000068, para el conjunto Input\_3 con ventana de 1 día. En la **Figura 24**, se expone la gráfica del ajuste para el mejor modelo obtenido. Se nota que la línea azul, correspondiente al SR Medido (el *label* de nuestro conjunto de datos) está prácticamente cubierta por las líneas de las predicciones.



**Figura 24.** Gráfico del ajuste del mejor modelo con MLP  
Elaboración propia

## LSTM

Entre los mejores resultados del LSTM (Tabla 15) se contemplan los tres conjuntos de datos. Diferentemente de los algoritmos anteriores, aquí se ven entre los mejores resultados ventanas de 3 y 5 días.

Se observa que el número de épocas de entrenamiento ha variado en un largo rango, de 91 a 313 épocas, considerablemente más alto que en los entrenamientos de MLP.

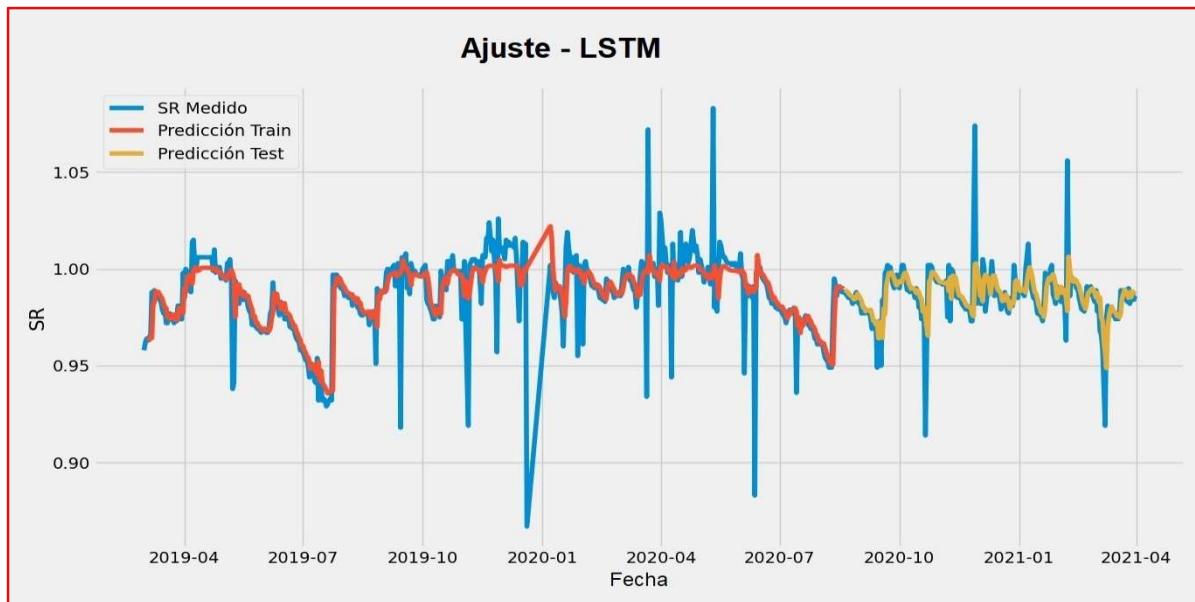
**Tabla 15.** Mejores resultados para el algoritmo LSTM

#	Input	Vent	Neur	Activ	Train				Test				epoch
					RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>	
1	Input_3	3	5	relu	0.014437	0.006798	0.695736	0.521137	0.012109	0.00639	0.646637	0.289877	91
2	Input_3	3	10	relu	0.01435	0.006765	0.692594	0.526915	0.012182	0.00632	0.639833	0.281291	84
3	Input_3	5	3	tanh	0.014595	0.00706	0.722919	0.510093	0.012211	0.00649	0.656496	0.277871	111
4	Input_1	3	3	sigmoid	0.015905	0.008107	0.830708	0.418833	0.012228	0.00607	0.613869	0.275792	105
5	Input_2	1	5	sigmoid	0.015871	0.00813	0.832358	0.422099	0.012232	0.00656	0.663857	0.275359	313
Media de las métricas:					0.015032	0.007372	0.754863	0.479815	0.012192	0.006364	0.644138	0.280038	

El RMSE medio obtenido con MLP ha sido de un 0.012192, y su mejor resultado un 0.012109, para el conjunto Input\_3 con ventana de 3 días, 5 neuronas en la capa oculta y método de activación “relu”.

Esos valores nos muestran un resultado bastante peor que los anteriores vistos, una sorpresa para un método tan potente. Una de las posibles explicaciones es que el conjunto de datos pueda ser demasiadamente pequeño para el entrenamiento del LSTM.

En la **Figura 25**, se expone la gráfica del ajuste para el mejor modelo obtenido. Se nota una grande diferencia con los gráficos anteriores. El ajuste del LSTM de manera general sigue la tendencia del SR Medido, pero los errores aquí son más expresivos que en los métodos anteriores.



**Figura 25.** Gráfico del ajuste del mejor modelo con LSTM

Elaboración propia

### 6.3. Análisis entre todos los experimentos

Para este análisis, seleccionamos los modelos con los 20 mejores resultados entre todos los modelos experimentados e independientemente de la entrada de datos utilizada. En total son 297 experimentos.

Nuestra medida principal sigue siendo el mejor RMSE en el Test y, para desempate se guía por la prioridad establecida para las métricas:

$$\text{RMSE} > \text{MAE} > \text{MAPE} > R^2$$

En la Tabla 16 se presentan estos resultados. Además, se incluyen las métricas de bondad de los dos modelos matemáticos utilizados para comparación en este estudio.

Los modelos son ordenados del mejor al peor resultado.

**Tabla 16.** Mejores resultados entre todos los experimentos realizados y modelos matemáticos

#	Modelo	Input	Vent	Train				Test			
				RMSE	MAE	MAPE	R <sup>2</sup>	RMSE	MAE	MAPE	R <sup>2</sup>
1	Regresión Lineal	Input_3	1	0.000175	0.000016	0.001702	0.99993	0.000068	0.000013	0.001361	0.999977
2	MLP (5 épocas)	Input_3	1	0.000175	0.000016	0.001706	0.99993	0.000068	0.000013	0.001365	0.999977
3	MLP (13 épocas)	Input_2	1	0.000174	0.000024	0.002418	0.99993	0.000073	0.000031	0.003119	0.999974
4	Regresión Lineal	Input_2	1	0.000174	0.000024	0.002415	0.99993	0.000073	0.000031	0.003113	0.999974
5	MLP (43 épocas)	Input_1	1	0.000173	0.000029	0.002995	0.999931	0.000078	0.000036	0.003712	0.999971
6	Regresión Lineal	Input_1	1	0.000173	0.000029	0.002985	0.999931	0.000078	0.000037	0.003788	0.999971
7	MLP (6 épocas)	Input_3	2	0.000266	0.00003	0.003068	0.999838	0.000192	0.000034	0.003427	0.999822
8	Regresión Lineal	Input_3	2	0.000266	0.00003	0.003069	0.999838	0.000192	0.000034	0.003427	0.999822
9	MLP (18 épocas)	Input_2	2	0.000264	0.000042	0.004291	0.99984	0.000195	0.000064	0.006507	0.999816
10	Regresión Lineal	Input_2	2	0.000264	0.000042	0.004304	0.99984	0.000195	0.000064	0.006527	0.999816
11	Regresión Lineal	Input_1	2	0.000263	0.00005	0.00506	0.999842	0.000201	0.000073	0.007447	0.999804
12	MLP (50 épocas)	Input_1	2	0.000263	0.00005	0.005114	0.999842	0.000202	0.000075	0.007656	0.999802
13	MLP (9 épocas)	Input_3	3	0.000255	0.000035	0.003572	0.999852	0.000253	0.000045	0.004561	0.999689
14	Regresión Lineal	Input_3	3	0.000255	0.000035	0.003571	0.999852	0.000253	0.000045	0.004561	0.999689
15	Regresión Lineal	Input_2	3	0.000252	0.000048	0.004968	0.999854	0.000255	0.000079	0.007992	0.999685
16	MLP (13 épocas)	Input_2	3	0.000252	0.000048	0.004971	0.999854	0.000255	0.000079	0.007996	0.999685
17	MLP (44 épocas)	Input_1	3	0.000251	0.000054	0.005498	0.999856	0.000259	0.000082	0.008307	0.999676
18	Regresión Lineal	Input_1	3	0.00025	0.000055	0.005589	0.999856	0.00026	0.000086	0.008768	0.999672
19	Random Forest (maxDepth=30)	Input_3	1	0.000624	0.000081	0.008318	0.999109	0.000576	0.000096	0.009762	0.998393
20	MLP (9 épocas)	Input_3	5	0.002068	0.000423	0.043483	0.990207	0.000577	0.000289	0.029185	0.998389

#	Modelo Matemático	RMSE	MAE	MAPE	R <sup>2</sup>
21	Coello y Boyle (Datos de Suelo)	0.038538	0.023643	2.420922	0.205066
22	You (Datos de Suelo)	0.019935	0.013792	1.418861	0.347129

Se observa que los resultados obtenidos son bastante buenos en relación a los modelos matemáticos actualmente utilizados.

Como los dos mejores resultados entre todos los experimentos realizados, se destacan la Regresión Lineal y el Perceptrón Multicapa, ambos utilizando el Input\_3 y con ventana de 1 día. Sus métricas son casi idénticas, con una pequeña ventaja de 0.000004 para el modelo de Regresión Lineal en el MAPE.

También observamos el tercer mejor modelo obtenido, es el Perceptrón Multicapa, utilizando el Input\_2 y con ventana de 1 día.

Analizando los resultados es posible inferir que los parámetros ambientales pueden suponer algún ruido para los modelos. Además, claramente se destacan las ventanas con menos días.

Se destaca que, para este problema en concreto, los modelos con Regresión Lineal y MLP son los que mejor funcionan. Solamente un modelo con Random Forest aparece en la posición 19. Los modelos con Árbol de Decisión y LSTM no aparecen en esta clasificación, sus mejores resultados aparecen en las posiciones 66 y 154, respectivamente.

## 7. PREDICCIONES SECUENCIALES

Teniendo en cuenta los tres mejores modelos seleccionados en el capítulo [6.3 Análisis Global de los Experimentos](#), en este presente capítulo se propone utilizarlos para hacer predicciones secuenciales del SR.

Las predicciones secuenciales son predicciones futuras en la serie temporal. O sea, se entrena el modelo con las ventanas de los valores medidos y se predice utilizando las salidas del modelo como entrada para la siguiente predicción y así sucesivamente hasta el horizonte temporal futuro que se desee.

En la **Figura 26**, se observa un ejemplo de cómo se construirían los datos de entrada para los modelos de la base de datos utilizada en este estudio para una predicción de 90 días secuenciales, con ventana de 2 días.

Fecha	SRIndex	Entrada: Ventanas (2 días)	Predicción	Validación
01/01/2019	SR0			SR0
02/01/2019	SR1			SR1
03/01/2019	SR2	V0 = SR0, SR1	y_train0	SR2
04/01/2019	SR3	V1 = SR1, SR2	y_train1	SR3
05/01/2019	SR4	V2 = SR2, SR3	y_train2	SR4
06/01/2019	SR5	V3 = SR3, SR4	y_train3	SR5
(...)	(...)	(...)	(...)	(...)
30/12/2020	SR650	V648 = SR648, SR649	y_train648	SR650
31/12/2020	SR651	V649 = SR649, SR650	y_train649	SR651
Datos de Test	01/01/2021	V_test0 = SR650, SR651	y_test0	SR652
	02/01/2021	V_test1 = SR651, y_test0	y_test1	SR653
	03/01/2021	V_test2 = y_test0, y_test1	y_test2	SR654
	04/01/2021	V_test3 = y_test1, y_test2	y_test3	SR655
	05/01/2021	V_test4 = y_test2, y_test3	y_test4	SR656
	06/01/2021	V_test5 = y_test3, y_test4	y_test5	SR657
	(...)	(...)	(...)	(...)
	31/03/2021	V_test5 = y_test88, y_test89	y_test89	SR741

**Figura 26. Ejemplo de entradas y salidas en una predicción secuencial de 90 días**  
Elaboración propia

La columna **SRIndex** contiene los datos de mediciones de SR para una determinada fecha, los mismo que serán utilizados para la validación de las predicciones. Para la entrada de los datos de entrenamiento, se utilizan los SR medidos en los 2 días anteriores (en razón de que se está utilizando una ventana de 2 días) y el modelo realiza una predicción.

Para la entrada de los datos de test, la ventana para la primera predicción es construida con los últimos SR medidos del conjunto de entrenamiento. Para las predicciones de los días siguientes, se empieza a utilizar las salidas del modelo asociadas a dichas predicciones como entrada al modelo. O sea, las entradas del modelo predictor están compuestas de las salidas que en tiempo va calculando.

Para cada uno de los mejores modelos obtenidos, se propone observar qué horizonte de tiempo futuro es más adecuado para realizar las predicciones. Han sido probados los siguientes horizontes futuros (días secuenciales): 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 20, 30, 45, 60, 75 y 90.

Como los dos mejores modelos utilizan como entrada el conjunto de datos [Input\\_3](#) (sólo con las ventanas deslizantes), será realizada la predicción con Regresión Lineal y MLP para este Input, que es compuesto solamente por las ventanas de los días anteriores, es decir, con los SR de días anteriores. Además, se realizará también la predicción con en [Input\\_2](#) (con las variables utilizadas en los modelos matemáticos y con las ventanas deslizantes) utilizando MLP, que corresponde al tercer mejor modelo obtenido.

Además, aunque los mejores resultados han sido obtenidos con ventana de 1 día, para estos experimentos de predicción secuencial, haremos las predicciones con modelos de ventanas de 1, 2 y 3 días.

## 7.1. Modelo con Regresión Lineal – [Input\\_3](#)

El modelo con Regresión Lineal y ventana de 1 día ha presentado el mejor resultado entre todos los experimentos anteriormente realizados.

En la Tabla 17, se presenta los resultados obtenidos en este experimento con la Regresión Lineal. Los modelos son ordenados del mejor al peor resultado. Donde se expone “N/A” para el  $R^2$ , es en razón de que Sklearn no calcula esta métrica para apenas 1 valor. Además, se advierte que el  $R^2$  no es preciso cuando es calculado para pocos valores.

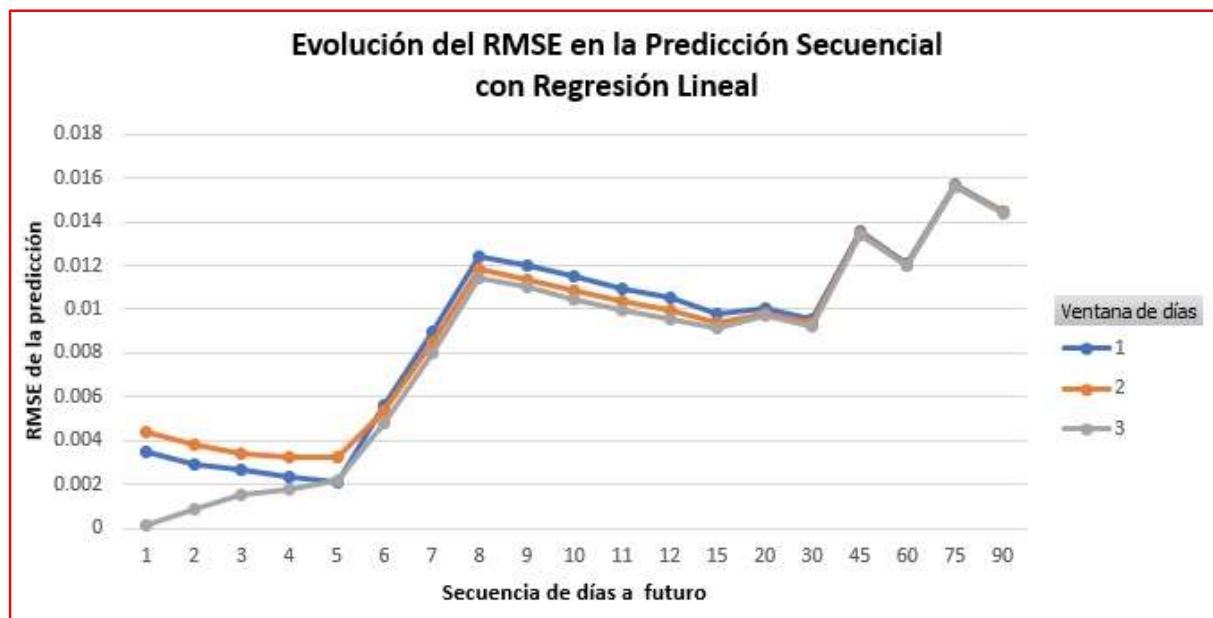
**Tabla 17.** Resultados de la predicción secuencial para el modelo con Regresión Lineal

#	días Secuenciales	Ventana	Predicción			
			RMSE	MAE	MAPE	$R^2$
1	1	1	0.0035	0.0035	0.353543	N/A
2	2	1	0.002934	0.002864	0.289416	-33.4301
3	3	1	0.00268	0.002603	0.263055	-31.313
4	4	1	0.002334	0.002077	0.209941	-5.836
5	5	1	0.002139	0.00187	0.189131	-1.334895
6	6	1	0.005629	0.003714	0.373144	-0.534779
7	7	1	0.009002	0.005958	0.595462	-0.619646
8	8	1	0.012443	0.008452	0.840774	-0.731482
9	9	1	0.012024	0.008392	0.835666	-0.81887
10	10	1	0.011474	0.007943	0.791515	-0.801886
11	11	1	0.010943	0.007303	0.727887	-0.707624
12	12	1	0.010495	0.006869	0.684959	-0.560104
13	15	1	0.009835	0.006581	0.658866	-0.194543
14	20	1	0.010059	0.00751	0.757897	-0.005062
15	30	1	0.00958	0.007427	0.748218	-0.057212
16	45	1	0.013579	0.008277	0.826609	-0.062277
17	60	1	0.012103	0.007214	0.722075	-0.028624
18	75	1	0.01564	0.009421	0.960985	-0.016414
19	90	1	0.014441	0.008573	0.874604	-0.02485
20	1	2	0.004402	0.004402	0.44461	N/A
21	2	2	0.003847	0.0038	0.383968	-58.199
22	3	2	0.003382	0.003257	0.329161	-50.471
23	4	2	0.003215	0.003105	0.31399	-11.9704
24	5	2	0.003267	0.003177	0.321479	-4.443925
25	6	2	0.005421	0.004496	0.452718	-0.423369
26	7	2	0.008454	0.006425	0.6435	-0.428415
27	8	2	0.011811	0.008723	0.869249	-0.559957
28	9	2	0.011381	0.008538	0.851406	-0.629427
29	10	2	0.010845	0.008007	0.798885	-0.609805
30	11	2	0.010341	0.007313	0.729665	-0.524821
31	12	2	0.009927	0.006914	0.690224	-0.395991
32	15	2	0.009383	0.006676	0.669068	-0.08739
33	20	2	0.009775	0.007623	0.769813	0.050893
34	30	2	0.009359	0.007486	0.754583	-0.008937
35	45	2	0.013465	0.008303	0.829494	-0.04446
36	60	2	0.012009	0.007243	0.725288	-0.012698
37	75	2	0.015604	0.009464	0.965507	-0.01169
38	90	2	0.014413	0.008612	0.878771	-0.02086
39	1	3	0.000142	0.000142	0.014352	N/A
40	2	3	0.000857	0.000673	0.067999	-1.935
41	3	3	0.001545	0.001243	0.125724	-9.735

42	4	3	0.001748	0.001495	0.151251	-2.8333
43	5	3	0.002214	0.001897	0.192113	-1.500972
44	6	3	0.004821	0.003368	0.338758	-0.125526
45	7	3	0.008021	0.005406	0.540634	-0.285894
46	8	3	0.011423	0.007775	0.773691	-0.459265
47	9	3	0.01099	0.007641	0.761083	-0.519558
48	10	3	0.010463	0.007153	0.712824	-0.498347
49	11	3	0.009976	0.006511	0.648863	-0.419092
50	12	3	0.009589	0.006216	0.619871	-0.302563
51	15	3	0.009145	0.006193	0.620554	-0.032935
52	20	3	0.00968	0.007329	0.740449	0.06915
53	30	3	0.009257	0.007265	0.732385	0.012912
54	45	3	0.013405	0.008144	0.813642	-0.035133
55	60	3	0.011962	0.007132	0.714178	-0.004769
56	75	3	0.015598	0.009393	0.958477	-0.010914
57	90	3	0.014412	0.008558	0.873456	-0.020745

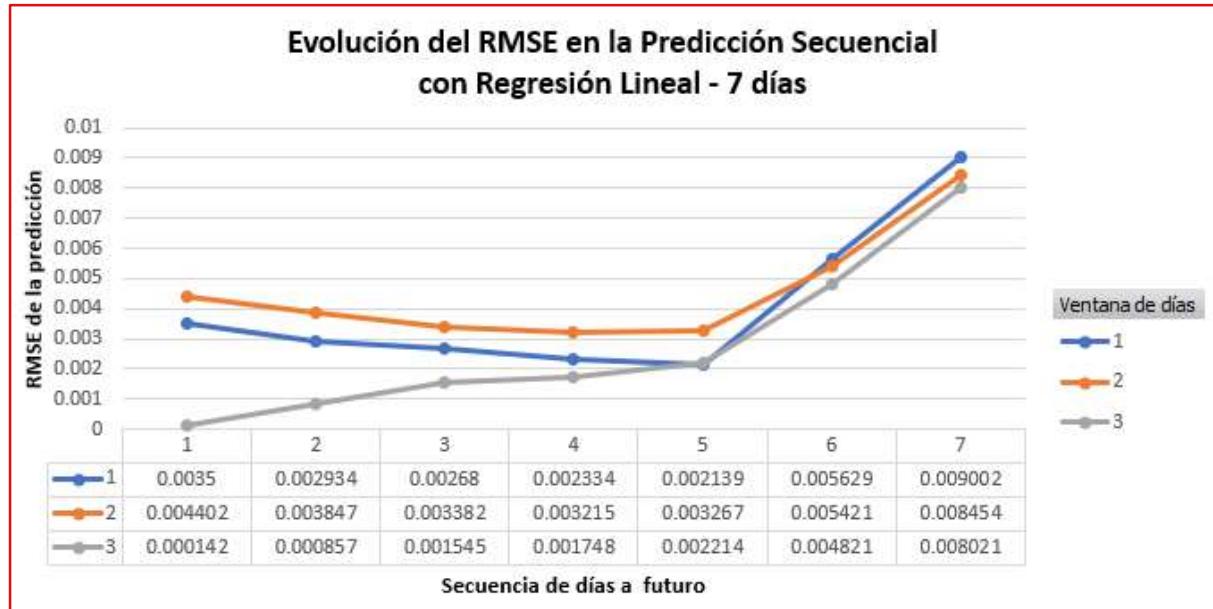
Los mejores resultados obtenidos son los que utilizan la ventana de 3 días para las predicciones de los 1, 2, 3 y 4 días subsecuentes, respectivamente.

En la **Figura 27**, se expone la evolución del RMSE en las predicciones secuenciales para hasta 90 días. Se observa que a partir del quinto día la métrica tiene un crecimiento más acentuado para todas las ventanas.



**Figura 27.** Gráfico de evolución: RMSE en la predicción secuencial de hasta 90 días con Regresión Lineal  
Elaboración propia

Para mejor visualización, en la **Figura 28**, se aproxima el gráfico restringiendo la secuencia de días para las predicciones de hasta 5 días. Además, se incluye la tabla de los datos del gráfico.



**Figura 28.** Gráfico de evolución: RMSE en la predicción secuencial de hasta 5 días con Regresión Lineal  
Elaboración propia

Es interesante observar como para las predicciones secuenciales la ventana de 3 días (línea gris) obtiene mejores resultados que la ventana de 1 día, que había sido la mejor clasificada entre los experimentos anteriores. Esta situación se queda evidente también cuando se analizan Modelo con Perceptrón Multicapa

## 7.2. Modelo con Perceptrón Multicapa – Input\_3

El Perceptrón Multicapa ha presentado el segundo mejor resultado entre todos los experimentos anteriormente realizados, utilizando el Input\_3.

Conforme mencionado anteriormente, el modelo ha sido construido con una capa oculta con 30 neuronas, solver “lbfgs” y función de activación “identity”.

Observamos estos datos en la Tabla 18, donde se exponen los resultados obtenidos en este experimento con MLP. Se les señalan en verde y en rojo, respectivamente los 5 mejores y los 5 peores valores de RMSE obtenidos

Los modelos son ordenados por orden de ejecución, primeramente, las secuencias para ventana 1, luego para la ventana 2 y finalmente para la ventana 3. Donde se expone “N/A” para el  $R^2$ , es en razón de que Sklearn no calcula esta métrica para apenas 1 valor. Además, se advierte que el  $R^2$  no es preciso cuando calculado para pocos valores.

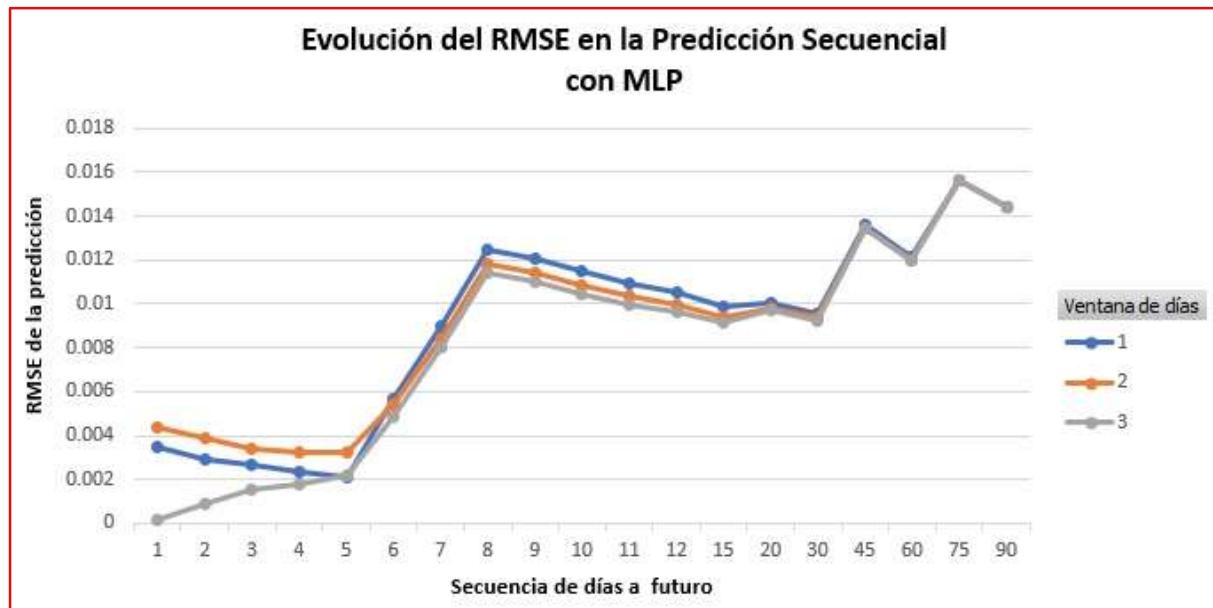
**Tabla 18.** Resultados de la predicción secuencial para el modelo con MLP

#	días Secuenciales	Ventana	Predicción			
			RMSE	MAE	MAPE	$R^2$
1	1	1	0.0035	0.0035	0.353542	N/A
2	2	1	0.002934	0.002864	0.289416	-33.433
3	3	1	0.00268	0.002603	0.263055	-31.313
4	4	1	0.002334	0.002077	0.209941	-5.836
5	5	1	0.002139	0.00187	0.189131	-1.334
6	6	1	0.005629	0.003714	0.373144	-0.534779
7	7	1	0.009002	0.005958	0.595464	-0.619652
8	8	1	0.012443	0.008452	0.840774	-0.731482
9	9	1	0.012024	0.008392	0.835666	-0.81887
10	10	1	0.011474	0.007943	0.791515	-0.801885
11	11	1	0.010943	0.007303	0.727887	-0.707624
12	12	1	0.010495	0.006869	0.684959	-0.560105
13	15	1	0.009835	0.006581	0.658866	-0.194544
14	20	1	0.010059	0.00751	0.757897	-0.005062
15	30	1	0.00958	0.007427	0.748218	-0.057212
16	45	1	0.013579	0.008277	0.826609	-0.062277
17	60	1	0.012103	0.007214	0.722075	-0.028624
18	75	1	0.01564	0.009421	0.960986	-0.016414
19	90	1	0.014441	0.008573	0.874604	-0.02485
20	1	2	0.004402	0.004402	0.44461	N/A
21	2	2	0.003847	0.0038	0.383967	-58.199
22	3	2	0.003382	0.003257	0.329161	-50.471
23	4	2	0.003215	0.003105	0.31399	-11.970
24	5	2	0.003266	0.003177	0.321474	-4.443
25	6	2	0.005421	0.004496	0.452719	-0.423367
26	7	2	0.008454	0.006425	0.643499	-0.428415
27	8	2	0.011811	0.008723	0.869249	-0.55995
28	9	2	0.011381	0.008538	0.851406	-0.629426
29	10	2	0.010845	0.008007	0.798885	-0.609805
30	11	2	0.010341	0.007313	0.729665	-0.524823
31	12	2	0.009927	0.006914	0.690223	-0.395987
32	15	2	0.009383	0.006676	0.669067	-0.08739
33	20	2	0.009775	0.007623	0.769813	0.050893
34	30	2	0.009359	0.007486	0.754583	-0.008936
35	45	2	0.013465	0.008303	0.829494	-0.044461
36	60	2	0.012009	0.007243	0.725288	-0.012698
37	75	2	0.015604	0.009464	0.965506	-0.01169
38	90	2	0.014413	0.008612	0.878772	-0.02086
39	1	3	0.000142	0.000142	0.014353	N/A
40	2	3	0.000857	0.000673	0.067999	-1.935
41	3	3	0.001545	0.001244	0.125761	-9.737

42	4	3	0.001748	0.001495	0.151252	-2.833
43	5	3	0.002214	0.001897	0.192114	-1.501
44	6	3	0.004821	0.003368	0.338758	-0.125526
45	7	3	0.008021	0.005406	0.540634	-0.285898
46	8	3	0.011423	0.007775	0.773691	-0.459265
47	9	3	0.01099	0.007641	0.761082	-0.519555
48	10	3	0.010463	0.007153	0.712825	-0.498349
49	11	3	0.009976	0.006511	0.648863	-0.419095
50	12	3	0.009589	0.006216	0.619871	-0.302562
51	15	3	0.009145	0.006193	0.620554	-0.032935
52	20	3	0.00968	0.007329	0.740448	0.06915
53	30	3	0.009257	0.007265	0.732384	0.012912
54	45	3	0.013405	0.008144	0.813642	-0.035137
55	60	3	0.011962	0.007132	0.714178	-0.004769
56	75	3	0.015598	0.009393	0.958482	-0.010918
57	90	3	0.014412	0.008558	0.873457	-0.020746

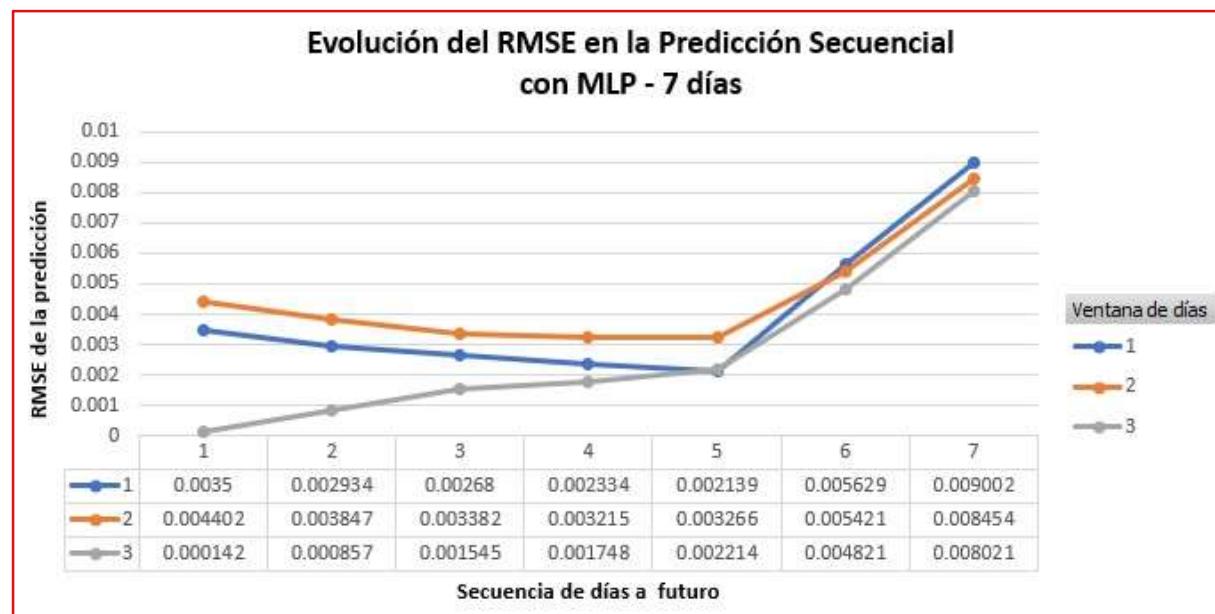
Los mejores resultados obtenidos son los que utilizan la ventana de 3 días para las predicciones de los 1, 2, 3 y 4 días subsecuentes, respectivamente. Y los peores para las secuencias más largas, conforme esperado.

En la **Figura 29**, se expone la evolución del RMSE en las predicciones secuenciales para hasta 90 días. Se observa que, de la misma manera que con la Regresión Lineal, a partir del quinto día la métrica tiene un crecimiento más acentuado para todas las ventanas.



**Figura 29.** Gráfico de evolución: RMSE en la predicción secuencial de hasta 90 días con Perceptrón Multicapa  
Elaboración propia

Para mejor visualización, en la **Figura 30**, se aproxima el gráfico restringiendo la secuencia para las predicciones de hasta 7 días. Además, se incluye la tabla de los datos del gráfico.



**Figura 30.** Gráfico de evolución: RMSE en la predicción secuencial de hasta 5 días con Perceptrón Multicapa  
Elaboración propia

Es interesante observar cómo, igual al resultado de Regresión Lineal, para las predicciones secuenciales la ventana de 3 días (línea gris) obtiene mejores resultados que la ventana de 1 día, que había sido la mejor clasificada entre los experimentos anteriores.

### 7.3. Modelo con Perceptrón Multicapa – Input\_2

El Perceptrón Multicapa utilizando el Input\_2, ha presentado el tercero mejor resultado entre todos los experimentos anteriormente realizados.

Conforme mencionado anteriormente, el modelo ha sido construido con una capa oculta con 30 neuronas, solver “lbfgs” y función de activación “identity”.

Observamos estos datos en la Tabla 19Tabla 18, donde se exponen los resultados obtenidos en este experimento con MLP. Se les señalan en verde y en rojo, respectivamente los 5 mejores y los 5 peores valores de RMSE obtenidos

Los modelos son ordenados por orden de ejecución, primeramente, las secuencias para ventana 1, luego para la ventana 2 y finalmente para la ventana 3. Donde se expone “N/A” para el  $R^2$ , es en razón de que Sklearn no calcula esta métrica para sólo 1 valor. Además, se advierte que el  $R^2$  no es preciso cuando calculado para pocos valores.

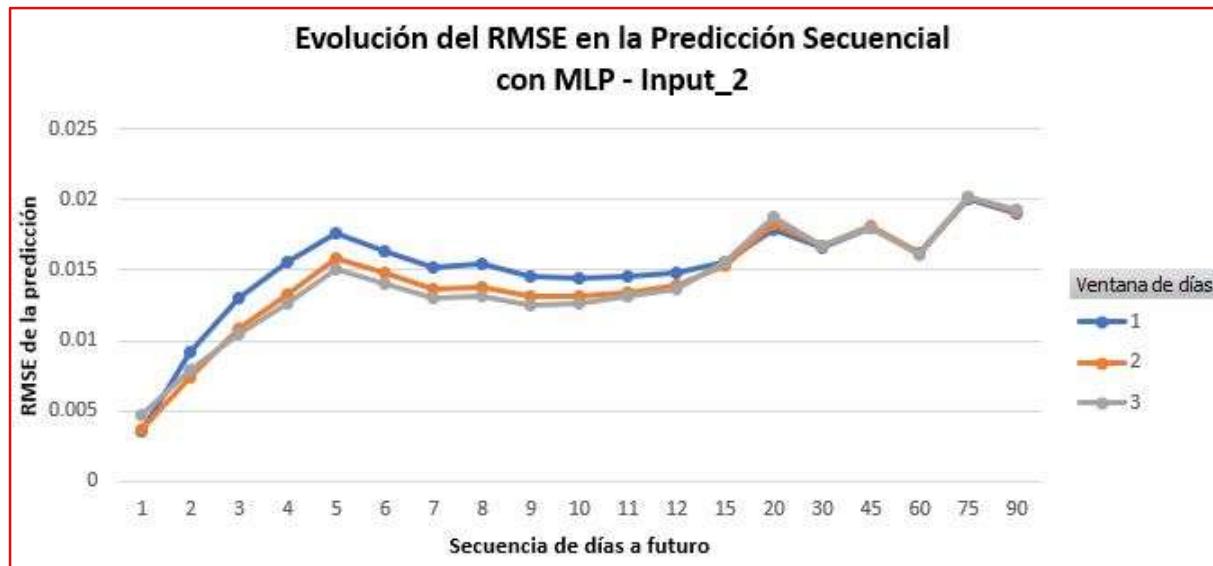
**Tabla 19.** Resultados de la predicción secuencial para el modelo con MLP con Input\_2

#	días Secuenciales	Ventana	Predicción			
			RMSE	MAE	MAPE	$R^2$
1	1	1	0.003519	0.003519	0.355499	N/A
2	2	1	0.009172	0.008002	0.80893	-335.522
3	3	1	0.013057	0.01151	1163694	-766.164
4	4	1	0.015561	0.013978	141404	-302.873
5	5	1	0.017589	0.015991	1618975	-156.836
6	6	1	0.016335	0.014552	1471768	-11.923
7	7	1	0.015198	0.013045	1318307	-3.616967
8	8	1	0.01548	0.01358	1367282	-1.679778
9	9	1	0.0146	0.012203	1228679	-1.681692
10	10	1	0.014362	0.012184	1226957	-1.823159
11	11	1	0.014571	0.012578	1267424	-2.027559
12	12	1	0.014781	0.012939	13049	-2.094545
13	15	1	0.015506	0.01387	140273	-1.969703
14	20	1	0.017928	0.016123	1638108	-2.192541
15	30	1	0.016649	0.014853	1504203	-2.1929
16	45	1	0.01802	0.014428	1451626	-0.870675
17	60	1	0.016189	0.012538	1263067	-0.840397
18	75	1	0.020026	0.015087	1538996	-0.666505
19	90	1	0.019011	0.014459	1474841	-0.776102
20	1	2	0.003654	0.003654	0.369072	N/A
21	2	2	0.007376	0.006712	0.678528	-216.642
22	3	2	0.010805	0.009655	0.976077	-524.414
23	4	2	0.013289	0.011959	120981	-220.619
24	5	2	0.015824	0.014239	1441646	-126.756
25	6	2	0.014784	0.013148	1329627	-9.585922
26	7	2	0.013712	0.011591	1171608	-2.75802
27	8	2	0.013833	0.011974	1205982	-1.139929
28	9	2	0.013101	0.01106	1113788	-1.159381
29	10	2	0.013087	0.01125	1133207	-1.34441
30	11	2	0.013396	0.011697	1178925	-1.559093
31	12	2	0.013858	0.012237	1234504	-1.720465
32	15	2	0.0153	0.013721	1388409	-1.891005
33	20	2	0.018256	0.016401	1667274	-2.310714
34	30	2	0.016711	0.014818	15016	-2.216622
35	45	2	0.0181	0.014502	1459669	-0.887309

36	60	2	0.016193	0.012466	1256093	-0.84131
37	75	2	0.020157	0.015138	1544653	-0.688394
38	90	2	0.019171	0.014553	148472	-0.806002
39	1	3	0.004677	0.004677	0.472413	N/A
40	2	3	0.007957	0.007456	0.753616	-252.229
41	3	3	0.010473	0.009713	0.98192	-492.560
42	4	3	0.012664	0.011704	1184009	-200.267
43	5	3	0.015079	0.013814	1398629	-115.003
44	6	3	0.014069	0.0127	1284352	-8.587772
45	7	3	0.013054	0.011214	1133499	-2.406181
46	8	3	0.013126	0.011515	1159869	-0.926796
47	9	3	0.012494	0.010807	108845	-0.963815
48	10	3	0.012684	0.011155	1123712	-1.202057
49	11	3	0.013095	0.011655	1174801	-1.445134
50	12	3	0.013666	0.012253	1236292	-1.645461
51	15	3	0.015519	0.014005	1417406	-1.974613
52	20	3	0.018745	0.01689	1717282	-2.490172
53	30	3	0.016699	0.01462	1482244	-2.212039
54	45	3	0.018014	0.014259	1435571	-0.869421
55	60	3	0.016077	0.012181	122759	-0.814893
56	75	3	0.020165	0.014967	1527684	-0.689621
57	90	3	0.019207	0.014438	1473515	-0.812896

En este caso, los mejores resultados obtenidos son los para las predicciones de los 1 y 2 subsecuentes, ya sin depender tanto de las ventanas. Además, los peores para las secuencias más largas, conforme lo esperado.

En la **Figura 31**, se expone la evolución del RMSE en las predicciones secuenciales para hasta 90 días. Se observa que, diferentemente de los modelos anteriores, aquí no hay una estabilidad inicial, y se ve el crecimiento del error de manera expresiva ya en los primeros días.



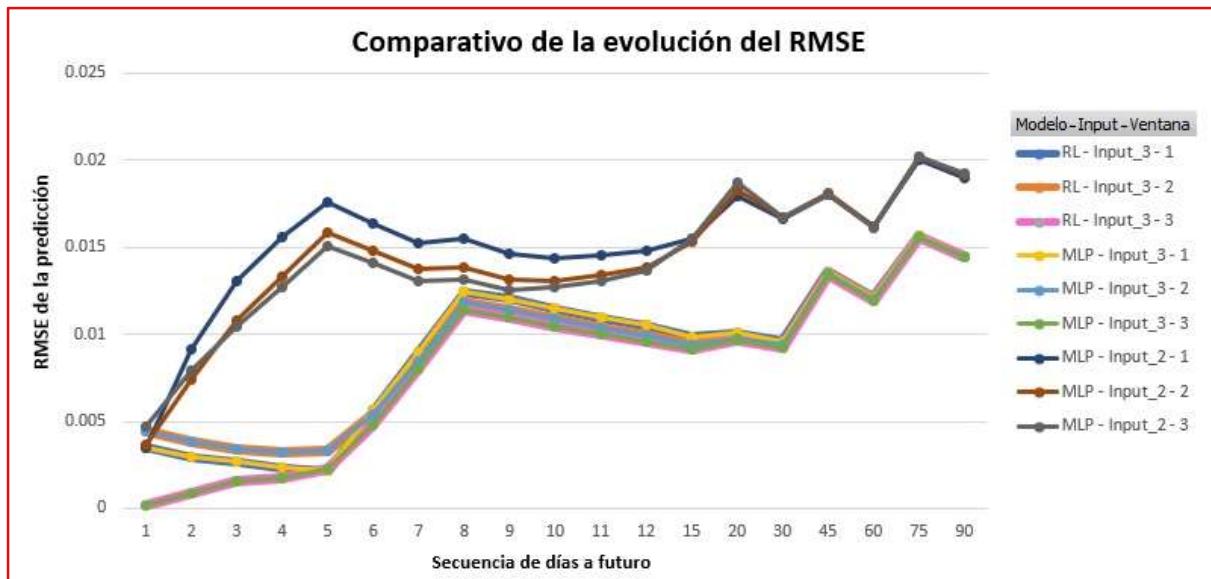
**Figura 31.** Gráfico de evolución: RMSE en la predicción secuencial de hasta 90 días con Perceptrón Multicapa para Input\_2

Elaboración propia

#### 7.4. Comparación entre las predicciones secuenciales

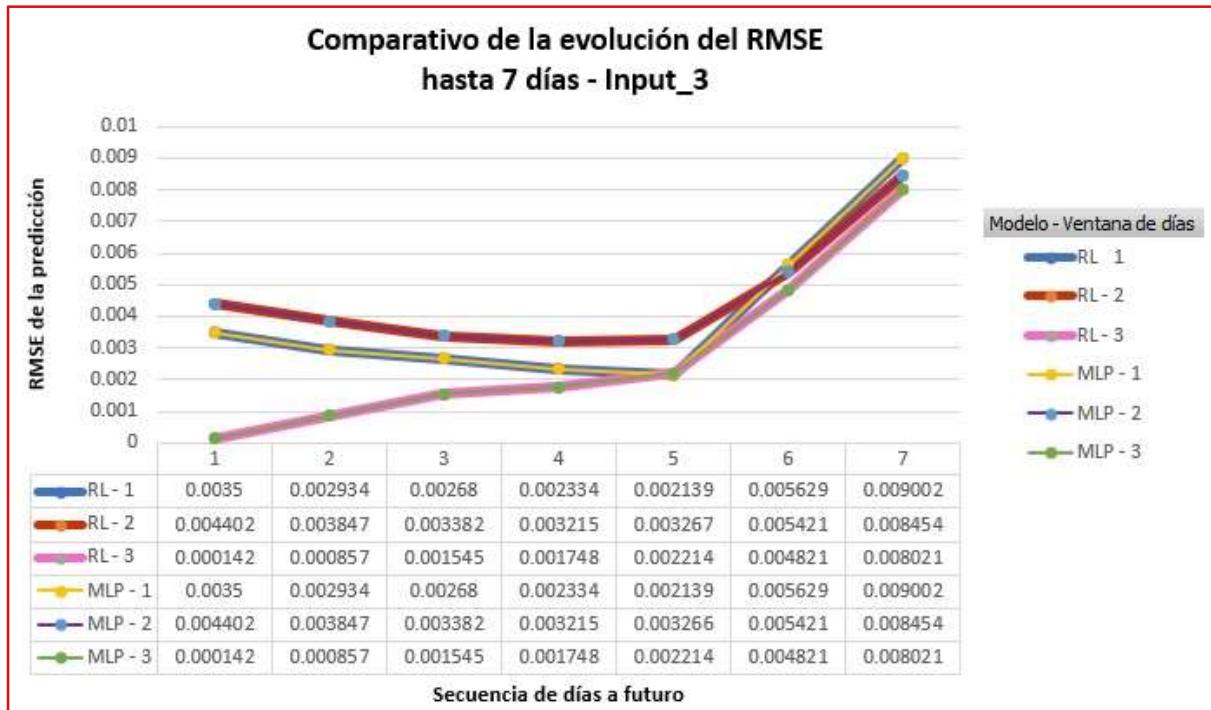
Se ha podido observar que los modelos de Regresión Lineal y MLP con Input\_3 en este caso tienen comportamientos bastante similares. Ya el modelo MLP con Input\_2 es menos eficaz.

En la **Figura 32** se presenta el gráfico comparativo de la evolución del RSME de cada uno de los modelos, con relación a los días a futuro de la predicción secuencial.



**Figura 32.** Gráfico comparativo entre modelos: predicciones secuenciales hasta 90 días  
Elaboración propia

Para mejor visualización, de la semejanza de los modelos que utilizan el Input\_3, en la **Figura 33**, se aproxima el gráfico restringiendo para las predicciones de hasta 7 días. Además, se incluye la tabla de los datos del gráfico para facilitar el análisis.



**Figura 33.** Gráfico comparativo entre modelos con Input\_3: predicciones secuenciales hasta 7 días  
Elaboración propia

Observamos en la Tabla 20, los 10 mejores resultados para cada uno de los algoritmos. Los modelos son ordenados del mejor al peor resultado. Además, se recuerda que donde se expone “N/A” para el  $R^2$ , es en razón de que Sklearn no calcula esta métrica para sólo 1 valor. Se advierte que el  $R^2$  no es preciso cuando calculado para pocos valores.

**Tabla 20.** Comparativo de resultados de la predicción secuencial

Perceptrón Multicapa - Input_3			Predicción				Regresión Lineal - Input_3						
#	días Sec	Ventana	RMSE	MAE	MAPE	$R^2$	#	días Sec	Ventana	RMSE	MAE	MAPE	$R^2$
1	1	3	0.00014	0.00014	0.01435	N/A	1	1	3	0.00014	0.00014	0.01435	-0.802
2	2	3	0.00086	0.00067	0.068	-1.935	2	2	3	0.00086	0.00067	0.068	-0.61
3	3	3	0.00155	0.00124	0.12576	-9.737	3	3	3	0.00155	0.00124	0.12572	-0.498
4	4	3	0.00175	0.0015	0.15125	-2.833	4	4	3	0.00175	0.0015	0.15125	-0.708
5	5	1	0.00214	0.00187	0.18913	-1.334	5	5	1	0.00214	0.00187	0.18913	-0.525
6	5	3	0.00221	0.0019	0.19211	-1.501	6	5	3	0.00221	0.0019	0.19211	-0.419
7	4	1	0.00233	0.00208	0.20994	-5.836	7	4	1	0.00233	0.00208	0.20994	-0.56
8	3	1	0.00268	0.0026	0.26306	-31.313	8	3	1	0.00268	0.0026	0.26306	-0.396
9	2	1	0.00293	0.00286	0.28942	-33.433	9	2	1	0.00293	0.00286	0.28942	-0.303
10	4	2	0.00322	0.00311	0.31399	-11.970	10	4	2	0.00322	0.00311	0.31399	-0.195

Perceptrón Multicapa - Input_2			Predicción			
#	días Sec	Ventana	RMSE	MAE	MAPE	$R^2$
1	1	1	0.00352	0.00352	0.3555	N/A
2	1	2	0.00365	0.00365	0.36907	N/A
3	1	3	0.00468	0.00468	0.47241	N/A
4	2	2	0.00738	0.00671	0.67853	-216.642
5	2	3	0.00796	0.00746	0.75362	-252.229
6	2	1	0.00917	0.008	0.80893	-335.522
7	3	3	0.01047	0.00971	0.98192	-492.560
8	3	2	0.01081	0.00966	0.97608	-524.414
9	9	3	0.01249	0.01081	108845	-0.96381
10	4	3	0.01266	0.0117	1184009	-200.267

Conforme lo esperado, una vez que en los análisis entre todos los algoritmos estos dos ya presentaban sus similitudes, los resultados así siguen. Se observa que los modelos que utilizan el Input\_3 presentan no sólo rendimientos semejantes si no que iguales. Las diferencias se observan en el  $R^2$ , pero se sabe que en este caso no es una métrica muy fiable.

Con este análisis se observa que, hasta el día 5, ambos modelos con Input\_3 tienen buenos resultados. Sin embargo, a partir del sexto día, el RMSE empieza a presentar una subida expresiva.

Ya el modelo MLP con el Input\_2, presenta un crecimiento del RMSE constante desde el inicio de las predicciones, eso puede significar que la utilización de los parámetros ambientales puede suponer algún ruido al predictor.

## 8. CONCLUSIONES

Como se ha visto anteriormente, los estudios del *soiling* están en pleno crecimiento debido a la importancia de mitigar sus efectos en la generación de energía solar. En este trabajo se ha podido aplicar técnicas de *machine learning* a este problema con resultados muy satisfactorios.

Observamos que, para las predicciones secuenciales, el RMSE medio obtenido para predicciones de hasta 5 días ha sido en torno de un 0.004 (media de los 10 mejores resultados obtenidos para cada modelo), frente a un error medio de un 0.026 de los modelos matemáticos utilizados. Además, si enfocamos sólo en los modelos con Input\_3, que al final obtuvieron los mejores resultados, se tiene un error medio de un 0.002.

Son resultados muy prometedores, así que, desde este punto de vista, mismo las predicciones a más largo plazo han presentado buenos resultados, con un RMSE de 0.009, 0.012 y 0.014 para las predicciones de 30, 60 y 90 días, respectivamente, utilizándose el Input\_3.

Además, hay que considerar el contexto de los datos obtenidos: la procedencia geográfica de las mediciones (ya que Jaén no es considerada una ubicación con severos niveles de *soiling*) y el hecho de que los modelos han sido entrenados con un conjunto de datos relativamente pequeño. Ambos puntos influyen negativamente en su capacidad de aprendizaje.

Desde este aspecto, los resultados obtenidos en este trabajo son bastante favorables en la práctica para el tema de la predicción del *soiling* ya que se obtiene bajos errores en una ubicación dónde justamente se espera que la predicción no sea tan precisa.

En relación al uso de los parámetros ambientales, más investigación se hace necesaria para certificar su influencia en los resultados de los modelos de este trabajo. No se ha podido ratificar si los mejores resultados son obtenidos sin el uso de esos datos por el hecho de que datos de satélite son más susceptibles a errores, o si realmente, se tornan ruido para el predictor al ser datos externos a la serie temporal de mediciones.

## 9. LÍNEAS DE FUTURO

A partir de los resultados obtenidos en los capítulos anteriores, se plantean nuevas metas de investigación. Algunas de estas cuestiones a resolver en el futuro serán:

- Construir modelos con datos ambientales de suelo y compararlos a los resultados obtenidos en este trabajo asociados a los datos de satélites, específicamente para la ubicación de Jaén.
- Validar los modelos construidos en este trabajo con mediciones de SR procedentes de otras ubicaciones, principalmente que sufran de manera más severa los efectos del *soiling* y con más mediciones disponibles.
- Extender la colaboración con el CEACTEMA/UJA para comparar los modelos propuestos en este trabajo con nuevos modelos matemáticos actualmente desarrollados por estos investigadores. En este aspecto se ha planteado la publicación de un artículo científico abordando el tema.
- Plantear la utilización de pronósticos meteorológicos futuros en las predicciones temporales como entrada externa a los modelos de *machine learning* desarrollados.

## 10. GLOSARIO

**Aerosoles:** es la mezcla formada por partículas sólidas o líquidas suspendidas en el aire. [38]

**CEACTEMA:** Centro de Estudios Avanzados en Ciencias de la Tierra, Energía y Medio Ambiente de la Universidad de Jaén.

**FV:** Fotovoltaico(a)

**Irradiancia:** la cantidad de energía que incide por la radiación solar en un período de tiempo determinado en un área.[39]

**NASA:** sigla para “*National Aeronautics and Space Administration*”. En libre traducción, la Administración Nacional de Aeronáutica y el Espacio de Estados Unidos. Es una agencia que tiene como enfoque la investigación y exploración del espacio.

**SR:** *Soiling Ratio*. Tasa de medición del nivel de suciedad de un módulo fotovoltaico.

**UJA:** Universidad de Jaén

## 11. BIBLIOGRAFÍA

- [1] H. Ritchie, «Electricity Mix», *Our World in Data*.  
<https://ourworldindata.org/electricity-mix> (accedido abr. 13, 2021).
- [2] J. A. Pomilio, «Energia elétrica e fontes renováveis», *Unicamp*, 2013.  
<https://www.dsce.fee.unicamp.br/~antenor/pdffiles/it744/cap2.pdf> (accedido abr. 15, 2021).
- [3] Red Electrica de España, «La demanda de energía eléctrica de España aumenta un 4,8 % en marzo», *Red Electrica de España*, 2021.  
<https://www.ree.es/es/sala-de-prensa/actualidad/nota-de-prensa/2021/04/la-demanda-de-energia-electrica-de-espana-aumenta-4-8-por-ciento-marzo> (accedido abr. 12, 2021).
- [4] J. G. Bessa, L. Micheli, F. Almonacid, y E. F. Fernández, «Monitoring photovoltaic soiling: assessment, challenges, and perspectives of current and potential strategies», *iScience*, vol. 24, n.º 3, p. 102165, mar. 2021, doi: 10.1016/j.isci.2021.102165.
- [5] M. Coello y L. Boyle, «Simple Model for Predicting Time Series Soiling of Photovoltaic Panels», *IEEE J. Photovoltaics*, vol. 9, n.º 5, pp. 1382-1387, 2019, doi: 10.1109/JPHOTOV.2019.2919628.
- [6] S. Toth, M. Hannigan, M. Vance, y M. Deceglie, «Predicting Photovoltaic Soiling From Air Quality Measurements», *IEEE J. Photovoltaics*, vol. 10, n.º 4, pp. 1142-1147, 2020, doi: 10.1109/JPHOTOV.2020.2983990.
- [7] S. You, Y. J. Lim, Y. Dai, y C. H. Wang, «On the temporal modelling of solar photovoltaic soiling: Energy and economic impacts in seven cities», *Appl. Energy*, vol. 228, pp. 1136-1146, oct. 2018, doi: 10.1016/j.apenergy.2018.07.020.
- [8] A. Younis y Y. Alhorrr, «Modeling of dust soiling effects on solar photovoltaic performance: A review», *Sol. Energy*, vol. 220, pp. 1074-1088, may 2021, doi: 10.1016/j.solener.2021.04.011.

- [9] B. Guo, W. Javed, S. Khan, B. Figgis, y T. Mirza, «Models for Prediction of Soiling-Caused Photovoltaic Power Output Degradation Based on Environmental Variables in Doha, Qatar». jun. 26, 2016, doi: 10.1115/ES2016-59390.
- [10] APPA, «¿Qué es la energía fotovoltaica?», *Asociación de Empresas de Energías Renovables*. <https://www.appa.es/appa-fotovoltaica/que-es-la-energia-fotovoltaica/> (accedido abr. 12, 2021).
- [11] J. A. Roca, «Las 20 mayores plantas fotovoltaicas del mundo: India manda en el ranking y España entra en el Top 20», *El Periódico de la Energía*, 2020. <https://elperiodicodelaenergia.com/las-10-mayores-plantas-fotovoltaicas-del-mundo/> (accedido abr. 15, 2021).
- [12] Portal Solar, «Energia Fotovoltaica», *Portal Solar*. <https://www.portalsolar.com.br/energia-fotovoltaica.html> (accedido abr. 14, 2021).
- [13] Portal Solar, «História e origem da Energia Solar», *Portal Solar*, 2016. <https://www.portalsolar.com.br/blog-solar/energia-solar/historia-origem-da-energia-solar.html> (accedido abr. 12, 2021).
- [14] A. Comerio, «Avaliação do impacto de sujidade de módulos fotovoltaicos na geração de energia elétrica», Universidade Federal do Espírito Santo, Vitória, 2019.
- [15] Our World in Data, «Solar PV module prices», *Our World in Data*, 2020. <https://ourworldindata.org/grapher/solar-pv-prices?time=earliest..latest> (accedido abr. 13, 2021).
- [16] J. G. Bessa, «Estudio de la influencia de la suciedad en módulos fotovoltaicos de concentración», Universidad de Jaén, 2018.
- [17] J. Bengoechea, M. Murillo, I. Sánchez, y A. R. Lagunas, «Soiling and abrasion losses for concentrator photovoltaics ARTICLES YOU MAY BE INTERESTED IN NoDustPV project: Development and testing of anti-soiling coatings AIP Conference Alternative technique for temperature control and automated dust

cleaning in CPV installations: A qualitative approach AIP Conference», *Proceedings*, vol. 2012, p. 20010, 2012, doi: 10.1063/1.5053531.

- [18] H. P. Garg, «Effect of dirt on transparent covers in flat-plate solar energy collectors», *Sol. Energy*, vol. 15, n.º 4, pp. 299-302, abr. 1974, doi: 10.1016/0038-092X(74)90019-X.
- [19] S. C. Silva Costa, A. S. A. Cardoso Diniz, V. A. Camatta Santana, M. Muller, L. Micheli, y L. L. Kazmerski, «Avaliação da sujidade em módulos fotovoltaicos em Minas Gerais, Brasil», Gramado, abr. 2018. Accedido: abr. 06, 2021. [En línea]. Disponible en:  
<https://anaiscbens.emnuvens.com.br/cbens/article/view/191>.
- [20] W. Javed, B. Guo, y B. Figgis, «Modeling of photovoltaic soiling loss as a function of environmental variables», *Sol. Energy*, vol. 157, pp. 397-407, nov. 2017, doi: 10.1016/j.solener.2017.08.046.
- [21] K. Chiteka, R. Arora, y S N Sridhara, «A method to predict solar photovoltaic soiling using artificial neural networks and multiple linear regression models», *Energy Syst.*, vol. 11, pp. 981-1002, 2020, doi: 10.1007/s12667-019-00348-w.
- [22] J. M. Carmona, P. Gupta, D. F. Lozano-García, A. Y. Vanoye, F. D. Yépez, y A. Mendoza, «Spatial and Temporal Distribution of PM2.5 Pollution over Northeastern Mexico: Application of MERRA-2 Reanalysis Datasets», *Remote Sens.*, vol. 12, n.º 14, p. 2286, jul. 2020, doi: 10.3390/rs12142286.
- [23] A. Skomedal y M. Deceglie, «Combined Estimation of Degradation and Soiling Losses in Photovoltaic Systems», *IEEE J. Photovoltaics*, vol. 10, p. 1788, 2020, doi: 10.1109/JPHOTOV.2020.3018219.
- [24] NASA, «MERRA-2», NASA. <https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/> (accedido abr. 19, 2021).
- [25] Global Modeling and Assimilation Office (GMAO), «MERRA-2 tavg\_2d\_flx\_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Surface Flux Diagnostics V5.12.4», *Goddard Earth Sciences Data and Information Services Center (GES DISC)*, 2015. doi:

10.5067/7MCPBJ41Y0K6 (accedido abr. 23, 2021).

- [26] Global Modeling and Assimilation Office (GMAO), «MERRA-2 tavg1\_2d\_aer\_Nx: 2d,1-Hourly,Time-averaged,Single-Level,Assimilation,Aerosol Diagnostics V5.12.4», *Goddard Earth Sciences Data and Information Services Center (GES DISC)*, 2015. doi: 10.5067/KLICLTZ8EM9D (accedido abr. 19, 2021).
- [27] Global Modeling and Assimilation Office (GMAO), «MERRA-2 inst3\_3d\_aer\_Nv: 3d,3-Hourly,Instantaneous,Model-Level,Assimilation,Aerosol Mixing Ratio V5.12.4», *Goddard Earth Sciences Data and Information Services Center (GES DISC)*, 2015. doi: 10.5067/LTVB4GPCOTK2 (accedido abr. 21, 2021).
- [28] «Soda - Solar Radiation Data». <http://www.soda-pro.com/es/home> (accedido abr. 23, 2021).
- [29] Global Modeling and Assimilation Office (GMAO), «MERRA-2 FAQ». <https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/FAQ/> (accedido abr. 21, 2021).
- [30] PRTR - España: Registro Estatal de Emisiones y Fuentes Contaminantes, «Partículas PM10». <https://prtr-es.es/Particulas-PM10,15673,11,2007.html> (accedido jun. 20, 2021).
- [31] J. Brownlee, «How to Check if Time Series Data is Stationary with Python», 2020. <https://machinelearningmastery.com/time-series-data-stationary-python/> (accedido may 28, 2021).
- [32] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *JMLR*, vol. 12, 2011, doi: <https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf>.
- [33] Python Software Foundation, «Python». <https://www.python.org/> (accedido may 25, 2021).
- [34] Spyder Website Contributors, «Spyder». <https://www.spyder-ide.org/> (accedido

may 25, 2021).

- [35] «Project Jupyter». <https://jupyter.org/> (accedido may 25, 2021).
- [36] F. Chollet, «Keras», *GitHub*, 2015. <https://keras.io>.
- [37] M. B. Kursa, A. Jankowski, y W. R. Rudnicki, «Boruta – A System for Feature Selection», *Fundam. Informaticae*, vol. 101, n.º 4, pp. 271 – 285, 2010, doi: 10.3233/FI-2010-288.
- [38] M. Santiago Aladro, «Estudio de la formación de aerosoles orgánicos secundarios en un modelo fotoquímico mediante experimentos en una cámara de simulación atmosférica», Universidad Complutense de Madrid, 2013.
- [39] «Aprende con Energía», *Ministerio de Energía y Fundación Chile*. <https://www.aprendeconenergia.cl/glosario/#a> (accedido abr. 19, 2021).
- [40] M. G. Bosilovich, R. Lucchesi, y M. Suarez, «MERRA-2: File Specification», *Global Modeling and Assimilation Office*, 2016. <https://gmao.gsfc.nasa.gov/pubs/docs/Bosilovich785.pdf> (accedido abr. 19, 2021).

## ANEXO I: Diccionario de datos de la base utilizada

### Descripción de las columnas de la base de datos

Datos unificados de las bases originales

Nombre del Campo	Tipo del Dato	Descripción*	Base original / Origen del campo**
Index	date	Fecha de la medición	Campo llave entre las bases
SR	float	Soiling Ratio. Valor medido del nivel de suciedad de la placa fotovoltaica	Atonometrics - CACTEMA, UJA
Temperature	float	Temperatura a 2 metros por encima del suelo. Medida en Kelvin (K)	Soda/M2T1NXFLX, MERRA-2, NASA
Relative humidity	float	Humedad relativa a 2 metros por encima del suelo (Medida en %)	Soda/M2T1NXFLX, MERRA-2, NASA
Pressure	float	Presión atmosférica a nivel del suelo (Medida en hPa)	Soda/M2T1NXFLX, MERRA-2, NASA
Wind speed	float	Velocidad del viento a 10 metros por encima del suelo (Medida en m/s)	Soda/M2T1NXFLX, MERRA-2, NASA
Wind direction	float	Dirección del viento a 10 metros por encima del suelo. 0 significa Norte, 90 Este, 180 Sur y 270 Oeste (Medida en grados)	Soda/M2T1NXFLX, MERRA-2, NASA
Rainfall	float	Precipitación o profundidad de lluvia en mm (Medida en kg/m2)	Soda/M2T1NXFLX, MERRA-2, NASA
Short-wave irradiation	float	Entrada de radiación de ondas cortas (Medida en Wh/m2)	Soda/M2T1NXFLX, MERRA-2, NASA
DUSMASS25	float	Dust Surface Mass Concentration - PM 2.5	M2T1NVAER, MERRA-2, NASA
AIRDENS	float	air density	M2I3NVAER, MERRA-2, NASA
BCPHILIC	float	Hydrophilic Black Carbon	M2I3NVAER, MERRA-2, NASA
BCPHOBIC	float	Hydrophobic Black Carbon	M2I3NVAER, MERRA-2, NASA
DU001	float	Dust Mixing Ratio (bin 001)	M2I3NVAER, MERRA-2, NASA
DU002	float	Dust Mixing Ratio (bin 002)	M2I3NVAER, MERRA-2, NASA
DU003	float	Dust Mixing Ratio (bin 003)	M2I3NVAER, MERRA-2, NASA
DU004	float	Dust Mixing Ratio (bin 004)	M2I3NVAER, MERRA-2, NASA
OCPHILIC	float	Hydrophilic Organic Carbon (Particulate Matter)	M2I3NVAER, MERRA-2, NASA
OCPHOBIC	float	Hydrophobic Organic Carbon (Particulate Matter)	M2I3NVAER, MERRA-2, NASA
SO4	float	Sulphate aerosol	M2I3NVAER, MERRA-2, NASA
SS001	float	Sea Salt Mixing Ratio (bin 001)	M2I3NVAER, MERRA-2, NASA
SS002	float	Sea Salt Mixing Ratio (bin 002)	M2I3NVAER, MERRA-2, NASA
SS003	float	Sea Salt Mixing Ratio (bin 003)	M2I3NVAER, MERRA-2, NASA
SS004	float	Sea Salt Mixing Ratio (bin 004)	M2I3NVAER, MERRA-2, NASA
PM10	float	Partículas aerosoles de diámetro entre 2,5 y 10 micrómetros, directamente relacionadas a la contaminación atmosférica	Cálculo propio [29]

\* La descripción de los campos de las bases M2T1NVAER y M2I3NVAER está en su idioma original para que se mantenga la fidelidad a su sentido, debido a sus características técnicas. [40]

\*\* Bases de datos: Atonometrics - Base de mediciones de niveles de soiling

Soda/M2T1NXFLX - Base de datos meteorológicos [25]

M2T1NVAER - Base de diagnósticos de aerosoles [26]

M2I3NVAER - Base de índices de aerosoles mezclados [27]