

MÁSTER EN TRATAMIENTO ESTADÍSTICO-COMPUTACIONAL DE LA INFORMACIÓN

TRABAJO FIN DE MÁSTER

Predicción de Tormentas Geomagnéticas con Técnicas de Soporte Vectorial en Regresión

Raquel García Marañón



UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN
&
UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS MATEMÁTICAS

Tutor: Francisco Javier Yáñez Gestoso

Madrid, septiembre 2024

Resumen

Este trabajo de fin de máster presenta un nuevo enfoque para la predicción de tormentas geomagnéticas utilizando máquinas de soporte vectorial de regresión (SVR) y compara los resultados con modelos híbridos de redes neuronales (Long Short-Term Memory + Perceptrón multicapa). Se discuten los mecanismos de generación de tormentas geomagnéticas y se detallan las metodologías de medición y evaluación, con énfasis en el índice Dst. Además, se analizan varios modelos predictivos, centrándose en la predicción 2, 4 y 6 horas antes de una tormenta geomagnética. Los resultados muestran resultados competitivos en comparación con la LSTM+MLP y con estudios previos. Se concluye que el uso de SVR y, aún más, los modelos híbridos ofrecen una herramienta precisa para la predicción de eventos geomagnéticos.

Palabras clave

Tormentas geomagnéticas, Space Weather, Clima Espacial, Regresión, Predicción, Máquina de soporte vectorial de Regresión (SVR), Índice Dst, Viento solar, Reconexión magnética, Optimización de hiperparámetros, Validación cruzada K-Fold, Modelos híbridos, Modelos de aprendizaje automático, Series temporales, Procesos gaussianos (GP), Long Short-Term Memory (LSTM).

Abstract

This master's thesis presents a new approach for predicting geomagnetic storms using Support Vector Regression (SVR) and compares the results with hybrid neural network models (Long Short-Term Memory + Multilayer Perceptron). The mechanisms of geomagnetic storm generation are discussed, and measurement and evaluation methodologies are detailed, with an emphasis on the Dst index. Additionally, several predictive models are analyzed, focusing on predictions 2, 4, and 6 hours before a geomagnetic storm. The results show competitive performance compared to LSTM+MLP and previous studies. It is concluded that the use of SVR and, even more so, hybrid models, offer an accurate tool for predicting geomagnetic events.

Keywords

Geomagnetic storms, Space Weather, Space Climate, Regression, Prediction, Support Vector Regression (SVR), Dst Index, Solar Wind, Magnetic Reconnection, Hyperparameter Optimization, K-Fold Cross-Validation, Hybrid Models, Machine Learning Models, Time Series, Gaussian Processes (GP), Long Short-Term Memory (LSTM).

Índice

Índice	I
1 Introducción	1
1.1 Contexto histórico	2
1.2 Mecanismos de Generación de Tormentas Geomagnéticas	3
1.3 Medición y Evaluación de las Tormentas Geomagnéticas	3
1.3.1 Índice Dst	3
1.3.2 Fases de una Tormenta Magnética	3
1.4 Efectos de las Tormentas Geomagnéticas	4
1.5 Modelos de Predicción de Tormentas Geomagnéticas	5
1.5.1 Modelos Basados en Redes Neuronales Artificiales (ANN)	5
1.5.2 Modelos Híbridos y Otros Enfoques	5
1.6 Obtención y Procesamiento de Datos	6
2 Marco Teórico y Herramientas Matemáticas	7
2.1 Series Temporales	7
2.2 Detección de Anomalías con Machine Learning	7
2.3 Análisis Exploratorio	8
2.4 Machine Learning	8
2.4.1 Aprendizaje Supervisado	8
2.4.2 Aprendizaje No Supervisado	9
3 Máquina de Vector Soporte de Regresión	10
3.1 El problema de optimización y las condiciones KKT	10
3.2 Máquina de soporte vectorial relajada	15
3.3 Extensiones no lineales: núcleos o <i>kernels</i>	16
3.4 Máquina de soporte vectorial de regresión	17
4 Caso de estudio	22
4.1 Motivación y objetivos	22
4.2 Selección y descarga de datos	22
4.3 Procesamiento de los datos	24
4.4 Análisis exploratorio	27
4.5 Proceso de Optimización	28
4.5.1 Modelo alternativo: LSTM-MLP	30
4.6 Resultados	31
4.6.1 Modelo SVR	32
4.6.2 Modelo LSTM-MLP vs. SVR	40
5 Conclusión y futuros avances	45

Bibliografía	47
A Definiciones: Divergencia y Rotacional	I
B Repositorio del proyecto	II
C Entrenamientos realizados en el modelo SVR	III

Capítulo 1

Introducción

Las tormentas geomagnéticas son fenómenos complejos pero fundamentales en el estudio del clima espacial o *space weather*, y su entendimiento es crucial para la protección de infraestructuras tecnológicas avanzadas.

[Gonzalez et al. \(1994\)](#) definen una tormenta geomagnética como un periodo durante el cual un campo eléctrico de convección interplanetaria, suficientemente intenso y prolongado, provoca una energización sustancial en el sistema magnetosfera-ionosfera. Esto lleva a la formación de un anillo de corriente intensificado lo suficientemente fuerte como para superar un umbral clave del índice cuantificador de tiempo de tormenta ([Dst](#)). De forma más sencilla, estas tormentas son perturbaciones en el campo magnético de la Tierra, causadas principalmente por interacciones con el viento solar y el campo magnético interplanetario ([IMF](#)).

Estas tormentas afectan gravemente a infraestructuras tecnológicas críticas, incluyendo redes de transmisión eléctrica, sistemas de comunicación y navegación, y operaciones satelitales. La inducción de corrientes geomagnéticas en líneas de transmisión eléctrica puede provocar sobrecargas y fallos en los sistemas de potencia, dando lugar a apagones extendidos. Además, las perturbaciones en el campo magnético pueden desorientar los sistemas de navegación basados en GPS, afectando tanto a la aviación como a la navegación marítima. En el ámbito espacial, el aumento de la radiación y las partículas energéticas durante estas tormentas pueden degradar los componentes electrónicos de los satélites y poner en riesgo la salud de los astronautas y/o las misiones no tripuladas.

Los índices geomagnéticos, como el índice cuantificador del tiempo de tormenta, o *disturbance storm time* ([Dst](#), [Kyoto Dst](#)), que registra perturbaciones del campo magnético terrestre, son esenciales para predecir estos eventos. Históricamente se ha utilizado este índice ([Sugiura, 1963](#)), datado de 1963 y con una resolución temporal de una hora, pasando posteriormente al Índice de Perturbación simétrica en el campo magnético terrestre horizontal en latitudes medias ([SYM-H](#)) para estudios de alta resolución, debido a su granularidad de un minuto.

La investigación ha evolucionado desde modelos lineales hacia métodos más complejos que integran redes neuronales y mecanismos de atención, aprovechando grandes volúmenes de datos y poder computacional avanzado. Los modelos actuales utilizan datos de satélites mono y multi-instrumento, incorporando variables de viento solar para mejorar la precisión de las predicciones.

La explicación física de este fenómeno de acoplamiento viento solar/magnetosfera sigue sin tener una modelización validada científicamente. Los primeros estudios sobre tormentas geomagnéticas se remontan a [Carrington \(1859\)](#). No obstante, esta problemática se identificó formalmente por primera vez en el trabajo de [Dessler and Parker \(1959\)](#) y se subrayó en la definición de tormenta solar propuesta por [Gonzalez et al. \(1994\)](#), que persiste hasta el estado del arte actual. Según [Borovsky \(2021\)](#), nuestro conocimiento físico sobre la interacción entre el viento solar y la magnetósfera, fundamental para predecir el clima espacial, es incompleto, careciendo de una comprensión detallada de cómo las variables del viento solar impulsan la

actividad geomagnética, lo que dificulta las predicciones precisas del comportamiento de la magnetósfera y el desarrollo de modelos.

Esto puede indicarnos el nivel de complejidad del fenómeno y, por tanto, la necesidad del uso de aprendizaje automático o *machine learning* ([ML](#)) y modelos computacionales para la prevención de problemas asociados con este fenómeno.

A continuación, se presenta un análisis del contexto histórico y el estado del arte en la comprensión y predicción de tormentas geomagnéticas, así como los mecanismos de generación, medición y evaluación de estas tormentas.

1.1. Contexto histórico

El estudio de fenómenos geomagnéticos comenzó en 1724 cuando George Graham observó variaciones en la declinación magnética. La importancia de estos estudios se destacó con la Tormenta Solar de Carrington en 1859, que causó auroras en latitudes inusuales y daños a sistemas telegráficos ([Carrington, 1859](#); [Boteler, 2006](#)). En el siglo XIX, se establecieron observatorios como el de Colaba en India y el de Kew en Inglaterra, proporcionando datos valiosos sobre la variabilidad del campo magnético terrestre.

Hannes Alfvén desarrolló la teoría de la magnetósfera en el siglo XX, incluyendo conceptos como las ondas Alfvén y la reconexión magnética, cruciales para comprender la interacción entre el viento solar y la magnetósfera ([Alfvén, 1942](#)). Misiones espaciales como Explorer 1 en 1958, que descubrió los cinturones de radiación de Van Allen ([Van Allen and Frank, 1959](#)), y misiones como ACE y DSCOVR, han sido esenciales en la recolección de datos sobre el viento solar y el campo magnético interplanetario desde el punto de Lagrange L1 ([Stone et al., 1998](#); [Burt et al., 2015](#)).

La dependencia moderna de tecnologías afectadas por el clima espacial ha intensificado la necesidad de predicciones precisas y la integración de datos de observación en modelos predictivos ([Schrijver et al., 2015](#); [Baker et al., 2018](#)). Estos avances han mejorado nuestra capacidad para predecir eventos geomagnéticos, aunque persisten desafíos significativos en la comprensión completa de estos fenómenos mediante modelos físicos precisos.

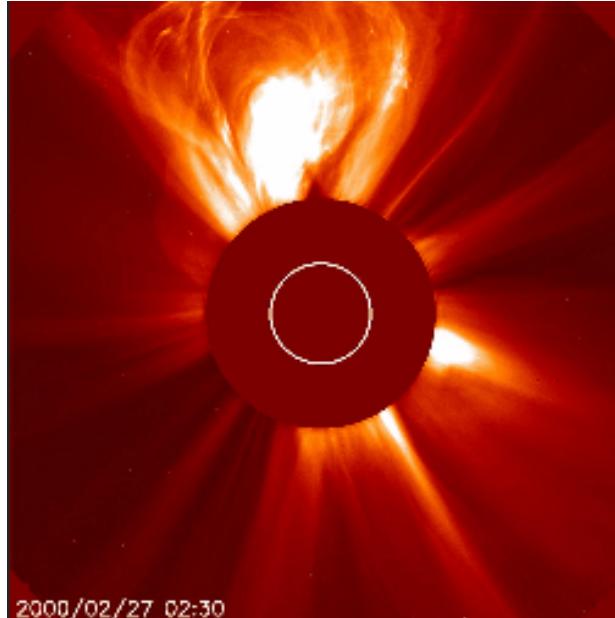


Figura 1.1: Eyección de masa coronal,
fuente: [SWL](#)

1.2. Mecanismos de Generación de Tormentas Geomagnéticas

Las tormentas geomagnéticas resultan de la interacción entre el viento solar y el campo magnético terrestre. Esta interacción puede inyectar partículas energéticas en las corrientes de plasma auroral y ecuatorial, perturbando el campo magnético (Burton et al., 1975). Las condiciones del viento solar que favorecen estas tormentas incluyen períodos de alta velocidad del viento solar y un campo magnético del viento solar dirigido hacia el sur. Esto puede desencadenar la reconexión magnética en la magnetosfera terrestre, intensificando las tormentas geomagnéticas. Incrementos en la densidad del plasma solar y estructuras interplanetarias, como las eyeciones de masa coronal (CME), también son críticos.

El proceso clave es la reconexión magnética, una interacción entre el campo magnético terrestre y el viento solar, que resulta en un intercambio de líneas de campo magnético y liberación de energía, acelerando partículas y corrientes de plasma (Birn et al., 2001). La reconexión magnética se describe mediante la ecuación de inducción magnética:

$$\text{rot } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (1.1)$$

Donde \mathbf{E} es el campo eléctrico y \mathbf{B} es el campo magnético. Esta ecuación describe cómo las variaciones en el campo magnético generan campos eléctricos y viceversa, permitiendo la transferencia de energía entre el viento solar y la magnetosfera durante la reconexión magnética. La definición de divergencia (div) y rotacional (rot) se encuentra en el Apéndice A.

1.3. Medición y Evaluación de las Tormentas Geomagnéticas

1.3.1. Índice Dst

El índice Dst se utiliza como medida de la actividad geomagnética y es fundamental en la predicción de tormentas geomagnéticas. Se calcula a partir de la componente horizontal del campo magnético terrestre y proporciona una medida de la intensidad de la tormenta, expresada en nanoteslas (nT).

1.3.2. Fases de una Tormenta Magnética

Una tormenta magnética típica presenta tres fases principales según las variaciones de Dst:

- **Caída Repentina:** Corresponde al inicio de la tormenta, caracterizada por una disminución brusca en el valor de Dst. Esta fase es inducida por el impacto inicial del viento solar comprimido o de una CME en la magnetosfera terrestre, conocido como choque de adelante.
- **Estado Excitado:** Durante esta fase, el valor de Dst permanece elevado mientras la corriente anular se intensifica. La energía y el momento transferidos a la magnetosfera durante la caída repentina continúan alimentando la actividad geomagnética (Gonzalez et al., 1994).

- **Recuperación:** Una vez que la componente z del campo magnético interplanetario (**IMF**) cambia de dirección, la corriente anular comienza a recuperarse y vuelve a su nivel normal. Esta fase puede durar desde varias horas hasta días, dependiendo de la magnitud de la tormenta.

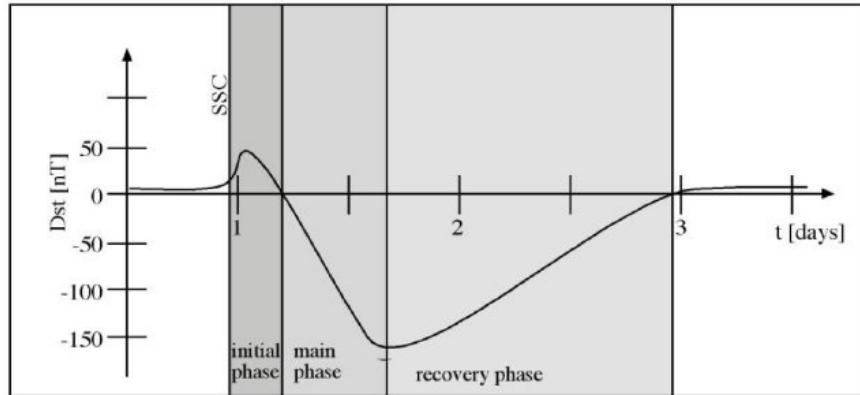


Figura 1.2: Fases de una tormenta geomagnética ([Kasran et al., 2018](#))

La evolución temporal de **Dst** se puede describir mediante un conjunto de ecuaciones diferenciales que modelan la interacción entre las corrientes en la magnetosfera y la actividad solar:

$$\frac{d(\text{Dst})}{dt} = Q(t) - \frac{\text{Dst}}{\tau} \quad (1.2)$$

Donde $Q(t)$ representa la tasa de inyección de energía en la corriente anular y τ es el tiempo de decaimiento de la corriente anular. Esta ecuación, derivada del modelo de Burton-McPherron-Russell ([Burton et al., 1975](#)), permite capturar la dinámica de las tormentas geomagnéticas, incluyendo la fase de caída, excitación y recuperación.

1.4. Efectos de las Tormentas Geomagnéticas

- **Impacto en Comunicación y Navegación:** Las tormentas geomagnéticas afectan la propagación de ondas de radio en la ionosfera, interrumpiendo comunicaciones (**HF**) y navegación (**GNSS**), induciendo errores en señales de posicionamiento ([Skone et al., 2001; Kintner et al., 2007](#)).
- **Efecto en Redes Eléctricas:** Las corrientes geomagnéticas inducidas (**GIC**) en redes eléctricas pueden causar sobrecargas y fallos en transformadores debido a variaciones rápidas del campo geomagnético ([Boteler et al., 1998](#)).
- **Impacto en Satélites y Astronáutica:** Las tormentas geomagnéticas aumentan la resistencia atmosférica y dañan componentes electrónicos de satélites, afectando su operatividad y la salud de los tripulantes de misiones espaciales ([Baker, 2000; Fennell et al., 2001](#)).

Las tormentas geomagnéticas han causado consecuencias adversas significativas. La tormenta más severa jamás registrada, la tormenta de Carrington del 1 y 2 de septiembre de 1859 ($\text{Dst} = -1760\text{nT}$) provocó auroras y fallos en sistemas telegráficos en Europa y América del Norte, causando incendios y pérdida de

comunicaciones (Lakhina et al., 2005); las tormentas del 29 al 31 de octubre de 2003 dañaron transformadores en EE.UU y Canadá (Kappenman, 2005) y la tormenta del 24 al 25 de octubre de 2011 afectó el sistema de navegación WAAS en EE.UU. (Datta-Barua et al., 2015). Más recientemente, la llamarada solar X1.1 del 23 de marzo de 2024 y la tormenta geomagnética del 12 de mayo de 2024 destacaron la amenaza continua de estos eventos. De hecho, en la actualidad nos encontramos en un periodo de aumento de la actividad solar debido a la periodicidad del ciclo solar de 11 años, por lo que cabe prever el aumento de estos eventos en los próximos años.

1.5. Modelos de Predicción de Tormentas Geomagnéticas

El objetivo de los modelos de predicción de tormentas solares es ser capaces de realizar la predicción de las fases inicial y principal con la suficiente antelación. Para ello, los modelos desarrollados han de ser capaces de detectar la pequeña subida del Índice Dst seguida de la notable bajada de su valor. Para realizar esta predicción de Dst (variable dependiente) se utilizan variables de viento solar, del campo magnético interplanetario (IMF) y del plasma como variables predictoras.

1.5.1. Modelos Basados en Redes Neuronales Artificiales (ANN)

Los modelos basados en ANN son efectivos para predecir Dst, utilizando datos históricos del viento solar y actividad geomagnética. Una ANN típica incluye capas de entrada, ocultas y de salida, con neuronas conectadas a todas las neuronas de la siguiente capa.

Para mejorar la precisión, se emplean técnicas de preprocesamiento como normalización y Análisis de Componentes principales (PCA), y arquitecturas avanzadas como Redes Neuronales Recurrentes (RNN) y Redes Neuronales convolucionales (CNN) para capturar relaciones temporales y espaciales en los datos del viento solar (Jagadeesh et al., 2020).

1.5.2. Modelos Híbridos y Otros Enfoques

Otros enfoques incluyen modelos híbridos que combinan técnicas físicas con Redes Neuronales Artificiales (ANN), y procesos Gaussianos para pronósticos probabilísticos, mejorando la precisión de las predicciones geomagnéticas (Bala and Reiff, 2012b; Chandorkar et al., 2017; Gruet et al., 2018).

Un ejemplo de modelo híbrido es usar simulaciones de MHD para generar datos sintéticos que entrena un modelo de ANN, incorporando conocimientos físicos fundamentales (Valdivia et al., 2013). Métodos como el filtro de Kalman y técnicas de asimilación de datos combinan observaciones en tiempo real con modelos numéricos para actualizar predicciones continuamente, mejorando la precisión de las alertas geomagnéticas (Schrijver et al., 2015).

El filtro de Kalman combina predicciones de un modelo dinámico con mediciones ruidosas para producir estimaciones más precisas del estado del sistema, ajustando las predicciones del modelo según observaciones del viento solar y el campo geomagnético.

1.6. Obtención y Procesamiento de Datos

La predicción y monitorización de tormentas geomagnéticas dependen de datos de misiones espaciales y observatorios terrestres, clasificados en *near real time* ([NRT](#)) para pronósticos inmediatos y *non-time critical* ([NTC](#)) para análisis científicos.

Misiones como [ACE](#) y [DSCOVR](#), ubicadas en [L1](#), son esenciales para recolectar datos del viento solar y el campo magnético interplanetario. [ACE](#), lanzada en 1997, estudia la composición y propiedades del viento solar, rayos cósmicos y otras partículas energéticas ([Stone et al., 1998](#)). [DSCOVR](#), lanzada en 2015, proporciona datos en tiempo real sobre el viento solar y el campo magnético interplanetario ([Burt et al., 2015](#)).

Los datos [NRT](#) se procesan en tiempo real para generar alertas y pronósticos, esenciales para la monitorización continua y respuesta rápida a eventos geomagnéticos. Para asegurar la calidad y precisión de los datos científicos, se realiza un reprocesado y validación posterior, generando los datos [NTC](#), fundamentales para estudios detallados y mejora de modelos predictivos.

Es importante distinguir entre datos provisionales y definitivos. Los datos provisionales son útiles para análisis preliminares y alertas rápidas, pero pueden contener errores. Los datos definitivos, completamente validados y reprocesados, se utilizan en estudios científicos y desarrollo de modelos predictivos, como es el caso de este estudio.

Capítulo 2

Marco Teórico y Herramientas Matemáticas

2.1. Series Temporales

Las series temporales son secuencias de datos, medidas típicamente a intervalos regulares de tiempo. Son fundamentales en muchos campos, como economía, finanzas, ciencias ambientales, y más, para analizar tendencias temporales, ciclos, o para predecir eventos futuros basados en datos históricos. Las herramientas estadísticas utilizadas en el análisis de series temporales incluyen:

- **Descomposición de series temporales:** Separa una serie temporal en componentes como tendencia, estacionalidad y ruido.
- **Modelos ARIMA (AutoRegressive Integrated Moving Average):** Utilizados para modelar y predecir datos de series temporales.
- **Suavizado Exponencial:** Incluye métodos como el Suavizado Exponencial Simple, Suavizado Exponencial Doble y Holt-Winters.
- **Máquina de Vectores de Soporte (SVM):** Las máquinas de Vectores de Soporte y la Máquina de Vectores de Soporte de Regresión son técnicas de aprendizaje automático utilizadas para tareas de clasificación y regresión, respectivamente. La **SVM** busca encontrar el hiperplano que mejor separe las clases en un espacio de alta dimensionalidad, mientras que la Máquina de Vector Soporte en Regresión (**SVR**) se enfoca en predecir valores continuos minimizando el error de predicción dentro de un margen de tolerancia específico.

A pesar de que este estudio no se centra en un análisis de series temporales, como se ha dicho anteriormente, es importante explicar las herramientas comunes utilizadas en su análisis, puesto que los datos empleados en este estudio poseen características de series temporales.

2.2. Detección de Anomalías con Machine Learning

La detección de anomalías es una técnica de aprendizaje automático (**ML**) utilizada para identificar patrones inusuales que no se ajustan al comportamiento esperado, llamados anomalías. Estas pueden indicar problemas críticos, como fallos en máquinas, fraude en transacciones bancarias o fallos de seguridad. Los modelos

de **ML** permiten automatizar y escalar la detección de anomalías aplicando métodos como el *clustering*, la detección de *outliers* o el uso de redes neuronales para identificar estas anomalías en grandes conjuntos de datos.

2.3. Análisis Exploratorio

El análisis exploratorio de datos (**EDA**) permite analizar datos mediante técnicas estadísticas y de visualización. Las etapas clave en EDA incluyen:

1. **Selección de Datos:** Consiste en elegir los datos relevantes para el análisis. La selección adecuada de los datos es crucial para asegurar la validez y la relevancia de los resultados obtenidos.
2. **Preprocesado de Datos:** El preprocesado de datos implica limpiar y preparar los datos para el análisis. Esto incluye la gestión de datos faltantes, la detección y corrección de *outliers*, y la agregación de datos.
3. **Análisis Exploratorio Visual:** El análisis visual de los datos a través de gráficos y diagramas ayuda a identificar patrones y relaciones que no son evidentes a partir de los datos brutos.
4. **Estudio de Correlación entre las Variables:** El análisis de correlación ayuda a identificar relaciones lineales entre las variables. Coeficientes de correlación como el de Pearson o el de Spearman son comúnmente utilizados para este propósito.

2.4. Machine Learning

El **ML** es un subcampo de la inteligencia artificial que se enfoca en el desarrollo de algoritmos que pueden aprender de los datos y hacer predicciones o tomar decisiones sin ser explícitamente programados para realizar una tarea específica. Estos algoritmos mejoran su rendimiento a medida que se exponen a más datos. Dentro del aprendizaje automático, distinguimos dos tipos principales de aprendizaje:

2.4.1. Aprendizaje Supervisado

El aprendizaje supervisado implica entrenar un modelo en un conjunto de datos que incluye las entradas y las salidas deseadas. El modelo aprende a asociar las entradas con las salidas correspondientes y puede generalizar esta relación para predecir las salidas para datos nuevos no vistos. Las técnicas de aprendizaje supervisado incluyen:

- **Regresión Lineal:** Utilizada para predecir un valor numérico continuo.
- **Regresión Logística:** Utilizada para problemas de clasificación binaria.
- **Árboles de Decisión:** Utilizados tanto para clasificación como para regresión.
- **Redes Neuronales Artificiales (**ANN**):** Modelos inspirados en el funcionamiento del cerebro humano, utilizados para tareas complejas de predicción y clasificación.

2.4.2. Aprendizaje No Supervisado

El aprendizaje no supervisado se utiliza con datos que no tienen etiquetas asignadas y el objetivo es inferir la estructura inherente presente dentro del conjunto de datos. Los usos comunes incluyen la agrupación de datos en clústeres y la reducción de dimensionalidad. Las técnicas de aprendizaje no supervisado incluyen:

- **Análisis de Componentes Principales (PCA)**: Utilizado para la reducción de dimensionalidad.
- **K-Means**: Algoritmo de agrupamiento que partitiona los datos en K clústeres.
- **Redes Neuronales Auto-Encoders**: Utilizadas para la reducción de dimensionalidad y la detección de anomalías.

Capítulo 3

Máquina de Vector Soporte de Regresión

Nota: El contenido teórico de este capítulo está basado en gran parte en los apuntes de Técnicas de Optimización del Máster TECI 2022/2023 de la UCM y la UPM ([Yáñez Gestoso, 2022](#)).

Las [SVM](#) y las [SVR](#) constituyen métodos de aprendizaje supervisado empleados para clasificación y regresión, respectivamente. Ambas técnicas se basan en problemas de optimización con restricciones de desigualdad, diseñados para encontrar el hiperplano que mejor divide las clases en el caso de [SVM](#), o que mejor se ajusta a los datos en el caso de [SVR](#), dentro de un margen de tolerancia especificado.

Esto se realiza mapeando los datos a un espacio de mayor dimensionalidad donde el margen entre las clases o los errores de predicción son minimizados. La optimización implica restricciones de desigualdad que aseguran que las instancias de datos se clasifiquen con un error dentro de un margen especificado del hiperplano, penalizando las correspondientes variables de holgura para manejar casos no separables linealmente. Este enfoque es parte de lo que se conoce como minimización del riesgo estructural ([SRM](#)), basado en la teoría de aprendizaje estadístico o *VC theory* propuesta por Vapnik y Chervonenkis ([Vapnik and Chervonenkis, 1974](#)), y popularizada en implementaciones de [SVM](#) por Vladimir Vapnik y Corinna Cortes en 1995 ([Cortes and Vapnik, 1995](#)).

3.1. El problema de optimización y las condiciones KKT

Sea un conjunto de datos clasificados

$$\{(x_1^i, x_2^i, \dots, x_n^i, y^i) \in \mathbb{R}^n \times \{-1, +1\}, i \in \{1, 2, \dots, m\}\} \quad (3.1)$$

que representa n características numéricas de m objetos, cuya clasificación binaria se conoce: el objeto i -ésimo es de clase 1 si $y^i = 1$ o es de la clase -1 si $y^i = -1$.

El problema de clasificación se plantea al intentar proponer un modelo matemático a partir de los datos anteriores de forma que si se presenta un objeto con características $(\bar{x}_1, \dots, \bar{x}_n)$ le asigne la clase 1 ó -1 considerando la información previa de la base de datos.

A partir de este problema de clasificación primaria, se ha de determinar un hiperplano de \mathbb{R}^n definido por las ecuaciones

$$H = \{x \in \mathbb{R}^n \mid \mathbf{w}^T \mathbf{x} = b\} \quad (3.2)$$

Observación Se supone que tal hiperplano H existe al estar los puntos de \mathbb{R}^n separados linealmente, lo que se puede conseguir con la elección adecuada.

La anchura de la franja separadora es la distancia (euclídea) entre los hiperplanos H^+ y H^- , $d(H^+, H^-)$; sin embargo, para conseguir una notación más compacta, se tratará de minimizar el cuadrado de la distancia $d(H, H^+)^2$, la mitad de la anchura de la franja al cuadrado.

El objetivo es identificar el vector \mathbf{w} y el escalar b que minimiza:

$$d(H, H^+)^2 = \min_{\mathbf{x} \in H, \mathbf{x} \in X^+} \left\{ \sum_{j=1}^n (x_j^+ - x_j)^2 \right\} \quad (3.3)$$

Considerando que $\sum_{j=1}^n w_j x_j = b$ y $\sum_{j=1}^n w_j x_j^+ = b + 1$, si se introduce el vector $\mathbf{d} \in \mathbb{R}^n$ verificando $x_j^+ = x_j + d_j \forall j \in \{1, 2, \dots, n\}$, la función a minimizar es:

$$\min_{\mathbf{d}} \sum_{j=1}^n d_j^2 \quad (3.4)$$

sujeto a

$$\sum_{j=1}^n w_j d_j = 1 \quad (3.5)$$

Este problema se resuelve introduciendo el lagrangiano

$$L(d, \alpha) = \sum_{j=1}^n d_j^2 + \alpha \left(\sum_{j=1}^n w_j d_j - 1 \right) \quad (3.6)$$

Al derivar e igualar a 0 se concluye:

$$d_j = \frac{w_j}{\sum_{j=1}^n w_j^2} \quad \forall j \in \{1, 2, \dots, n\} \quad (3.7)$$

$$d(H, H^+)^2 = \sum_{j=1}^n d_j^2 = \frac{1}{\sum_{j=1}^n w_j^2} = \frac{1}{\|\mathbf{w}\|^2} \quad (3.8)$$

En consecuencia, la anchura de la semirranja separadora es:

$$d(H, H^+) = \frac{1}{\|\mathbf{w}\|} \quad (3.9)$$

Añado las restricciones de separación de los objetos dependiendo de las clases $+1$ y -1 como

$$y^i (\mathbf{w}^T \mathbf{x}^i - b) \geq 1 \quad \forall i \in \{1, 2, \dots, m\} \quad (3.10)$$

El problema de optimización primal busca encontrar un hiperplano que separe dos clases de la mejor forma posible, minimizando la norma del vector de pesos w y maximizando el margen. Aquí, w es el vector de pesos del hiperplano, b es el sesgo, x_i son los vectores de características y y_i son las etiquetas de clase de los ejemplos de entrenamiento, con $y_i \in \{-1, 1\}$.

No obstante, por comodidad en la notación, el objetivo será maximizar el doble de esta anchura al cuadrado, por lo que el problema de optimización se plantea de la siguiente forma:

$$\begin{aligned} & \underset{w,b}{\max} \frac{2}{\|w\|^2} \\ & \text{sujeto a } y^i(w^T x^i - b) \geq 1 \quad \forall i \end{aligned} \tag{3.11}$$

Este problema es equivalente al de minimizar el inverso de la función objetivo anterior, se denomina problema primal:

$$P \left\{ \begin{array}{l} \min_{w,b} \frac{1}{2} \|w\|^2 \\ \text{sujeto a } y^i(w^T x^i - b) \geq 1 \quad \forall i \in \{1, 2, \dots, m\} \end{array} \right. \tag{3.12}$$

Una vez identificada la función objetivo óptima $z = \frac{1}{2} \|w\|^2$, la anchura de la franja (H^+, H^-) se determina por:

$$d(H^+, H^-) = 2d(H, H^+) = \frac{2}{\|w\|} = \frac{2}{\sqrt{2z}} = \sqrt{\frac{2}{z}} \tag{3.13}$$

El problema de clasificación se resuelve de la siguiente forma para un vector $\mathbf{x} \in \mathbb{R}^n$:

$$\text{Si } \mathbf{w}^\top \mathbf{x} \geq b \implies \text{clase}(\mathbf{x}) = +1 \tag{3.14}$$

$$\text{Si } \mathbf{w}^\top \mathbf{x} \leq b \implies \text{clase}(\mathbf{x}) = -1 \tag{3.15}$$

No obstante, es mejor plantear el problema dual del anterior y resolverlo; esta aparente complicación permitirá ampliar el modelo a problemas de clasificación no separables y modelos separables no lineales pero sí separables con una transformación de los datos por medio de una función kernel. Para introducir el problema dual se necesita el lagrangiano:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \alpha_i (1 - y^i(w^T x^i - b)) \tag{3.16}$$

Las ecuaciones del gradiente se traducen en:

$$\frac{\partial L}{\partial w_j} = 0 \iff w_j = \sum_{i=1}^m \alpha_i y^i x_j^i \quad \forall j \in \{1, 2, \dots, n\} \tag{3.17}$$

$$\frac{\partial L}{\partial b} = 0 \iff \sum_{i=1}^m \alpha_i y^i = 0 \quad (3.18)$$

Las ecuaciones de ortogonalidad asociadas a las condiciones KKT son:

$$\alpha_i (1 - y^i (w^t x^i - b)) = 0 \quad \forall i = 1, \dots, m \quad (3.19)$$

Esto implica que si $\alpha_i > 0$ entonces $y^i (w^t x^i - b) = 1$, indicando que el vector i -ésimo pertenece a uno de los hiperplanos H^+ (si $y^i = 1$ y, en consecuencia, $w^t x^i = b + 1$) o H^- (si $y^i = -1$ y $w^t x^i = b - 1$).

Un vector x^i del conjunto de datos será un vector soporte si $\alpha_i > 0$; en caso contrario, $\alpha_i = 0$ y está incluido en el exterior de la franja separadora perteneciendo a la clase $+1$ si $w^t x^i > b + 1$ o a la clase -1 si $w^t x^i < b - 1$.

En el interior de la franja ($-1 < w^t x^i < b + 1$) no hay ningún punto y se asignará a la clase $+1$ si $w^t x^i > b$ o a la clase -1 si $w^t x^i < b$. En el caso $w^t x^i = b$ se pueden asignar indistintamente a una clase u otra.

El problema dual de P es el siguiente:

$$\mathcal{D}\{ \begin{array}{l} \max_{\alpha} H(\alpha) \\ \text{sujeto a } \alpha_i \geq 0 \quad \forall i \in \{1, 2, \dots, m\} \end{array} \} \quad (3.20)$$

siendo

$$H(\alpha) = \min_{w, b} L(w, b, \alpha) \quad (3.21)$$

y

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \alpha_i (1 - y^i (w^t x^i - b)) \quad (3.22)$$

Al ser la función objetivo y las asociadas a las restricciones del problema P convexas, se verifica no sólo el teorema débil de dualidad sino también el fuerte. En consecuencia, si (\bar{w}, \bar{b}) es la solución óptima de P que alcanza la función objetivo $\gamma = \frac{1}{2} \|\bar{w}\|^2$ se verifica:

$$\max_{\alpha} \min_{w, b} L(w, b, \alpha) \leq \max_{\alpha} L(\bar{w}, \bar{b}, \alpha) = \frac{1}{2} \|\bar{w}\|^2 = \gamma \quad (3.23)$$

A continuación, y basándose en las condiciones de KKT, se expresa la función objetivo del problema dual D de una forma más asequible para resolver el problema del hiperplano separador.

$$H(\alpha) = \min_{w, b} \left\{ \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \alpha_i (1 - y^i (w^t x^i - b)) \right\} = \dots \quad (3.24)$$

$$\dots = \min_{w,b} \left\{ \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \alpha_i - \sum_{i=1}^m \alpha_i y^i w^t x^i + b \sum_{i=1}^m \alpha_i y^i \right\} \quad (3.25)$$

Hay dos opciones:

- $\sum_{i=1}^m \alpha_i y^i \neq 0$, en este caso, el problema \mathcal{D} es no acotado, pues se puede aumentar o disminuir el valor de la variable b arbitrariamente. En este caso, el problema P no tiene solución factible.
- $\sum_{i=1}^m \alpha_i y^i = 0$, en este caso, al sustituir (condiciones KKT)

$$w = \sum_{i=1}^m \alpha_i y^i x^i \quad (3.26)$$

se tiene que la función objetivo es

$$H(\alpha) = \min_{w,b} L(w,b,\alpha) = \left\{ \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^i y^j (x^i)^t x^j \right\} \quad (3.27)$$

El problema del hiperplano separador, en su formulación dual, queda de la siguiente forma:

$$\mathcal{D} \left\{ \begin{array}{l} \max_{\alpha \in \mathbb{R}^m} \left\{ \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^i y^j (x^i)^t x^j \right\} \\ \text{sujeto a } \sum_{i=1}^m \alpha_i y^i = 0 \\ \alpha_i \geq 0 \quad \forall i \in \{1, 2, \dots, m\} \end{array} \right. \quad (3.28)$$

A partir de la solución $\bar{\alpha}$ de este problema se determinan el vector w :

$$w = \sum_{i=1}^m \bar{\alpha}_i y^i x^i \quad (3.29)$$

El escalar \bar{b} se calcula identificando dos vectores soporte:

$$\blacksquare x^k: \quad \bar{\alpha}_k > 0 \quad \Rightarrow \quad y^k = -1 \quad \Rightarrow \quad w^t x^k = \bar{b} - 1 \quad (3.30)$$

A partir de estos dos vectores soporte se calcula el escalar óptimo \bar{b} :

$$\bar{b} = \frac{1}{2} (w^t x^k + w^t x^k) \quad (3.31)$$

El problema de clasificación por medio de la solución del problema dual \mathcal{D} se resuelve de la siguiente forma para un vector $x \in \mathbb{R}^n$:

- Si $\sum_{i=1}^m \alpha_i \cdot y^i \cdot \langle x^i, x \rangle \geq b \Rightarrow \text{clase}(x) = +1$
- Si $\sum_{i=1}^m \alpha_i \cdot y^i \cdot \langle x^i, x \rangle \leq b \Rightarrow \text{clase}(x) = -1$

3.2. Máquina de soporte vectorial relajada

En algunos casos no es posible separar perfectamente las clases con un hiperplano debido a la naturaleza de los datos o al ruido en los mismos. Para manejar esta situación, se introduce una versión relajada del **SVM**, conocida como **SVM** con margen suave o **SVM** relajada.

La **SVM** relajada permite que algunos puntos de entrenamiento estén dentro del margen o incluso en el lado incorrecto del hiperplano, penalizando estos errores mediante variables de holgura $\xi_i \geq 0$ y el costo asociado a las variables de holgura:

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i \\ & \text{sujeto a } y_i(\mathbf{w}^T \mathbf{x}_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \quad (3.32)$$

donde $C \geq 0$ es un parámetro de regularización que controla el trade-off entre la maximización del margen y la penalización por las violaciones del margen.

Esta **SVM** relajada permite encontrar un hiperplano de separación con margen más suave, mejorando la adaptación a casos no lineales en comparación a su forma clásica. El problema dual en este caso se formula como:

$$\begin{aligned} & \max_{\alpha \geq 0} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ & \text{sujeto a } \sum_{i=1}^m \alpha_i y_i = 0, 0 \leq \alpha_i \leq C \end{aligned} \quad (3.33)$$

donde α son los multiplicadores de Lagrange, con m representando el número total de ejemplos de entrenamiento y C siendo un parámetro de regularización.

El problema de clasificación en este caso se resuelve de la misma forma que en el caso estricto, la única diferencia es la identificación de los vectores soporte, necesarios para identificar el sesgo b .

Al igual que en el caso estricto, el escalar óptimo \bar{b} se determina identificando dos vectores soporte, si hubiera varios se puede elegir cualquiera de ellos: x^h y x^k en cada uno de los hiperplanos H^+ y H^- respectivamente; es decir, verificando

$$0 < \alpha_h < C \quad y^h = 1 \quad (3.34)$$

y

$$0 < \alpha_k < C \quad y^k = -1 \quad (3.35)$$

El escalar \bar{b} se determina según:

$$\bar{b} = \frac{1}{2} (\mathbf{w}^T \mathbf{x}^h + \mathbf{w}^T \mathbf{x}^k) \quad (3.36)$$

3.3. Extensiones no lineales: núcleos o *kernels*

En problemas de clasificación, los datos no siempre son linealmente separables. Para resolver esto, se usan funciones *kernel* (3.37) que permiten mapear los datos a un espacio de mayor dimensión, donde pueden ser linealmente separables. El objetivo es resolver el problema en este nuevo espacio sin calcular explícitamente la transformación Φ .

Dada una función kernel, Φ , se mapean los datos originales a un espacio de mayor dimensión:

$$\Phi : \mathbb{R}^n \rightarrow \mathbb{H} \quad (3.37)$$

En este nuevo espacio \mathbb{H} , los datos pueden ser linealmente separables. El objetivo es trabajar en \mathbb{H} sin calcular explícitamente Φ , utilizando una función kernel K para calcular los productos internos en \mathbb{H} desde los datos originales:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \quad (3.38)$$

El kernel lineal analizado en las subsecciones anteriores sería *sera* $\Phi(x) = x$, la función identidad, de forma que la función kernel es el producto escalar de x^i y x^j :

$$K(x^i, x^j) = \langle x^i, x^j \rangle \quad (3.39)$$

Funciones de Kernel útiles pueden ser las siguientes:

Funciones de Kernel útiles pueden ser las siguientes:

- Función de kernel polinómica: $K(\mathbf{x}^i, \mathbf{x}^j) = (\langle \mathbf{x}^i, \mathbf{x}^j \rangle + \gamma)^g$, siendo g el grado del polinomio (si $g = 1$ y $\gamma = 0$, es el modelo lineal).
- Función de kernel gaussiana o de base radial (**RB**F): $K(\mathbf{x}^i, \mathbf{x}^j) = \exp(-\gamma \langle \mathbf{x}^i - \mathbf{x}^j, \mathbf{x}^i - \mathbf{x}^j \rangle)$,
- Función de kernel sigmoidal: $K(\mathbf{x}^i, \mathbf{x}^j) = \tanh(-\gamma \langle \mathbf{x}^i, \mathbf{x}^j \rangle + \delta)$

El problema de optimización (dual) relajado, utilizando cualquier función de kernel sería el siguiente:

$$\begin{aligned} & \max_{\alpha \geq 0} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ & \text{sujeto a } \sum_{i=1}^m \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \end{aligned} \quad (3.40)$$

Al igual que en el caso lineal, con la formulación dual, los vectores soporte son aquellos que verifican:

$$0 < \alpha_h < C \quad y^h = 1 \quad (3.41)$$

$$0 < \alpha_k < C \quad y^k = -1 \quad (3.42)$$

El sesgo se determina a partir de estos vectores soporte:

$$\bar{b} = \frac{1}{2} \left(\sum_{i=1}^m \alpha_i y^i K(\mathbf{x}^i, \mathbf{x}^h) + \sum_{i=1}^m \alpha_i y^i K(\mathbf{x}^i, \mathbf{x}^k) \right) \quad (3.43)$$

Y el problema de clasificación de un vector $\bar{\mathbf{x}}$ es el siguiente:

$$\begin{cases} \sum_{i=1}^m \alpha_i y^i K(\mathbf{x}^i, \bar{\mathbf{x}}) > \bar{b} & \text{Clase de } \bar{\mathbf{x}} \text{ es } +1 \\ \sum_{i=1}^m \alpha_i y^i K(\mathbf{x}^i, \bar{\mathbf{x}}) < \bar{b} & \text{Clase de } \bar{\mathbf{x}} \text{ es } -1 \end{cases} \quad (3.44)$$

3.4. Máquina de soporte vectorial de regresión

El modelo de Máquina de Soporte Vectorial en regresión ([SVR](#)) fue introducido por Vapnik en 1996 ([Drucker et al. \(1996\)](#)).

Se parte de un conjunto de m puntos

$$\{(x_1^i, x_2^i, \dots, x_n^i, y^i) \in \mathbb{R}^{n+1}, \forall i \in \{1, 2, \dots, m\}\} \quad (3.45)$$

que es el conjunto de entrenamiento para ajustar el modelo; el objetivo es ajustar una función lineal:

$$f(x) = (w_1 x_1 + \dots + w_n x_n) + b = \langle w, x \rangle + b \quad (3.46)$$

de forma que dado un nuevo punto $\bar{x} \in \mathbb{R}^n$, la función anterior permita determinar el valor asociado $\bar{y} = f(\bar{x})$.

En este caso, la función f depende de los valores $w \in \mathbb{R}^n$ y $b \in \mathbb{R}$, por lo que este ajuste se resuelve identificando estos dos parámetros. Se tiene así definida la variable de decisión del problema de optimización.

A diferencia de la clasificación, donde se penalizan los datos mal clasificados con el modelo soft, en la regresión se penalizarán los datos cuyo valor y^i se aleje lo suficiente del valor predicho por la función $f(x^i)$.

Se define para ello una función de pérdida ϵ -sensible, que es una función lineal con una zona insensible, en la que el error es nulo:

$$L_\epsilon(y, f(x)) = \begin{cases} 0 & \text{si } |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon & \text{en otro caso} \end{cases} \quad (3.47)$$

Formalmente, se introduce para cada uno de los datos de entrenamiento dos variables (ξ_i^+, ξ_i^-) para cuantificar los distintos errores. La variable ξ_i^+ es positiva cuando la predicción $f(x^i)$ es mayor que el valor real y^i en una cantidad mayor que ϵ ; en cualquier otro caso, su valor es cero. Análogamente, la variable ξ_i^- es

positiva cuando la predicción $f(x^i)$ es menor que el valor real y^i en una cantidad mayor que ϵ ; en cualquier otro caso, su valor es cero.

En la figura 3.1 se muestra la relación entre las variables de holgura (ξ_i^+, ξ_i^-) asociadas a datos que quedan fuera de la zona tubular ϵ -insensible.

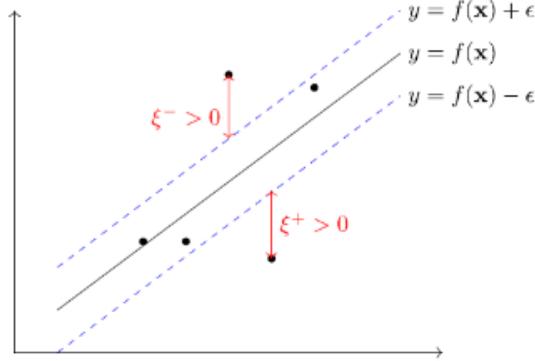


Figura 3.1: Representación gráfica de la zona ϵ -insensible

Así, el modelo del SVM en la regresión se basa en dos ideas:

- Establecer un margen de tolerancia de la predicción.
- Determinar una función que penalice los datos que se salgan del margen anterior.

El margen de tolerancia se identifica por el parámetro ϵ . Se define así una franja de amplitud 2ϵ que recogerá el dato (\mathbf{x}, y) como bien estimado si $f(\mathbf{x}) - \epsilon \leq y \leq f(\mathbf{x}) + \epsilon$.

La función que penaliza los datos mal estimados (fuera de la franja anterior) se basa en el modelo de la máquina de soporte vectorial, es decir, se trata de que la franja tenga una anchura máxima, lo que equivale a minimizar $\frac{1}{2}\|\mathbf{w}\|^2$.

A continuación, se formaliza el problema anterior como un problema de programación cuadrática, será la formulación primal para recoger la relación entre la variable de decisión y la función objetivo. Será un modelo análogo al de la máquina de soporte vectorial relajado, incluyendo el parámetro $C > 0$ penalizando las dos variables de holgura ξ_i^+ (asociada a los puntos por debajo de la franja) y ξ_i^- (asociada a los puntos por encima de la franja).

El modelo es el siguiente:

$$\begin{aligned} & \min \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i^+ + \xi_i^-) \\ & \text{sujeto a } (\mathbf{w}^T \mathbf{x}^i + b) - y^i \leq \epsilon + \xi_i^+ \\ & \quad y^i - (\mathbf{w}^T \mathbf{x}^i + b) \leq \epsilon + \xi_i^- \\ & \quad \xi_i^+, \xi_i^- \geq 0 \end{aligned} \tag{3.48}$$

donde w es el vector de pesos del hiperplano, b es el sesgo (*bias*) del hiperplano, x_i representa los vectores de características de los ejemplos de entrenamiento, y_i es la etiqueta de clase de los ejemplos de entrenamiento, donde $y_i \in \{-1, +1\}$, y ξ_i^+ y ξ_i^- son variables de holgura que permiten que algunas muestras violen las restricciones dadas por el hiperplano con un coste controlado por el hiperparámetro C . La constante ϵ mide el margen de insensibilidad a la pérdida para los errores de predicción.

OBSERVACIÓN: $y^i \in \mathbb{R}$ no es una variable categórica, como en la [SVM](#).

Para la formulación del problema dual, hay que considerar las variables no negativas $(\alpha_i^+, \alpha_i^-, \beta_i^+, \beta_i^-)$ asociadas a las anteriores desigualdades.

El lagrangiano sería:

$$L_C(w, b, \xi^+, \xi^-) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i^+ + \xi_i^-) + \sum_{i=1}^m \alpha_i^+ (w^T x^i + b - y^i - \epsilon - \xi_i^+) + \dots \quad (3.49)$$

$$\dots + \sum_{i=1}^m \alpha_i^- (y^i - w^T x^i - b - \epsilon - \xi_i^-) + \sum_{i=1}^m \beta_i^+ (-\xi_i^+) + \sum_{i=1}^m \beta_i^- (-\xi_i^-) \quad (3.50)$$

Aplicando las condiciones de KKT, se obtiene:

$$\frac{\partial L_C}{\partial w} = 0 \iff w = \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) x^i \quad (3.51)$$

$$\frac{\partial L_C}{\partial b} = 0 \iff \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) = 0 \quad (3.52)$$

$$\frac{\partial L_C}{\partial \xi_i^+} = 0 \iff C - \alpha_i^+ - \beta_i^+ = 0 \quad \forall i \in \{1, 2, \dots, m\} \quad (3.53)$$

$$\frac{\partial L_C}{\partial \xi_i^-} = 0 \iff C - \alpha_i^- - \beta_i^- = 0 \quad \forall i \in \{1, 2, \dots, m\} \quad (3.54)$$

Al ser las variables $\beta_i^+, \beta_i^- \geq 0$ se concluye:

$$0 \leq \alpha_i^+, \alpha_i^- \leq C \quad \forall i \in \{1, 2, \dots, m\} \quad (3.55)$$

Simplificando algunas expresiones, al final queda el problema dual:

$$D_C^R \left\{ \begin{array}{l} \text{máx}_{\alpha^+, \alpha^- \in \mathbb{R}^m} \left\{ \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) y^i - \epsilon \sum_{i=1}^m (\alpha_i^+ + \alpha_i^-) - \dots \right. \right. \\ \left. \left. \dots - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m (\alpha_i^+ - \alpha_i^-) \langle x^i, x^j \rangle (\alpha_j^+ - \alpha_j^-) \right\} \right. \\ \text{sujeto a} \\ \left. \left. \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) = 0 \right. \right. \\ \left. \left. 0 \leq \alpha_i^+, \alpha_i^- \leq C \quad \forall i \in \{1, 2, \dots, m\} \right. \right. \end{array} \right\} \quad (3.56)$$

Siendo el regresor asociado a la función lineal:

$$f(x) = \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) \langle x^i, x \rangle + b = w^T x + b \quad (3.57)$$

Al ser

$$w = (\alpha_1^+ - \alpha_1^-, \dots, \alpha_m^+ - \alpha_m^-) \begin{pmatrix} x_1^1 & \dots & x_n^1 \\ \vdots & \ddots & \vdots \\ x_1^m & \dots & x_n^m \end{pmatrix} \quad (3.58)$$

El escalar b es determinado a partir de algún dato que esté en alguno de los hiperplanos de la franja ϵ -sensible, por ejemplo, para algún índice $i \in \{1, 2, \dots, m\}$ se verifica $0 < \alpha_i^+ < C$, entonces:

- Al ser $\alpha_i^+ > 0$, la desigualdad $(\langle w, x^i \rangle + b) - y^i - \epsilon - \xi_i^+ \leq 0$ se alcanza en el límite, quedando $(\langle w, x^i \rangle + b) - y^i - \epsilon = \xi_i^+ = 0$.
- Al ser $\alpha_i^+ < C$, implica que $\beta_i^+ > 0$ y no hay error $\xi_i^+ = 0$.

Se obtiene, por tanto, que:

$$(\langle w, x^i \rangle + b) - y^i - \epsilon = 0$$

y x^i es un *vector soporte* que permite calcular el escalar b :

$$b = y_i - \langle w, x^i \rangle + \epsilon$$

Análogamente se podría haber calculado a partir de un punto x^j verificando $0 < \alpha_j^- < C$, en cuyo caso, el escalar b se obtendría de la siguiente forma:

$$b = y_j - \langle w, x^j \rangle - \epsilon$$

En el caso no lineal, con una función de kernel ($K_{\{i,j\}} = K(x^i, x^j)$), el modelo de programación dual sería:

$$\max_{\alpha^+, \alpha^- \in \mathbb{R}^m} \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) y^i - \epsilon \sum_{i=1}^m (\alpha_i^+ + \alpha_i^-) - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m (\alpha_i^+ - \alpha_i^-) K_{ij} (\alpha_j^+ - \alpha_j^-) \quad (3.59)$$

sujeto a

$$\sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) = 0 \quad (3.60)$$

$$0 \leq \alpha_i^+, \alpha_i^- \leq C \quad \forall i \in \{1, 2, \dots, m\} \quad (3.61)$$

donde $(\alpha_i^+ - \alpha_i^-)$ son los multiplicadores de Lagrange asociados con las restricciones de desigualdad del problema primal, K

La función regresora final es:

$$f(x) = \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) K(x_i, x) + b \quad (3.62)$$

donde b se determina usando un vector soporte x_i tal que $0 < \alpha_i^+ < C$ o $0 < \alpha_i^- < C$. Estos vectores soporte son fundamentales para definir el hiperplano de decisión y el margen de la SVR.

Por tanto, en **SVR**, la optimización se centra en encontrar los multiplicadores de Lagrange α_i^+ y α_i^- que maximizan la capacidad predictiva de la función regresora final $f(x)$, ajustando el modelo para minimizar la pérdida de predicción bajo las restricciones puestas por ϵ y C .

Capítulo 4

Caso de estudio

4.1. Motivación y objetivos

La motivación de este estudio radica en la creciente dependencia de infraestructuras tecnológicas críticas que son vulnerables a los efectos de las tormentas geomagnéticas (Sección 1.4). Desarrollar modelos predictivos más precisos y robustos mejorará nuestra capacidad para anticipar estos eventos y contribuirá a la protección y conservación de estas infraestructuras esenciales. No obstante, la mayoría de modelos basados en [ANN](#) carecen de interpretabilidad por su naturaleza de caja negra, y además son muy intensivos computacionalmente, lo que tiene un impacto económico y de sostenibilidad considerable. Por ese motivo, en este trabajo se propone un modelo más conservador conocido históricamente por su capacidad de predicción. El objetivo es evaluar su capacidad para competir en todos estos aspectos con estos modelos, destacando que puede ser más sostenible que los complejos modelos de [ANN](#) y, sobre todo, mucho más transparente e interpretable.

Este enfoque se centra en la utilización de [SVR](#) para predecir tormentas geomagnéticas mediante la predicción del Índice [Dst](#) (variable objetivo). Para ello, se emplearán datos históricos del viento solar y el índice [Dst](#), con el objetivo de desarrollar un modelo que pueda predecir estos eventos con antelación de 2, 4 y 6 horas.

Para contrastar estos resultados, se realizará una comparación exhaustiva entre el modelo [SVR](#) y un modelo alternativo basado en una red neuronal que combina bloques de Memoria a Largo Corto Plazo ([LSTM](#), por el inglés *Long Short-Term Memory*) y Perceptrón Multicapa ([MLP](#)), de ahora en adelante [LSTM-MLP](#).

4.2. Selección y descarga de datos

El conjunto de datos utilizado es el [OMNI2_H0_MRG1HR](#)¹, el cual proviene de las sondas *Advanced Composition Explorer* ([ACE](#)) y *Deep Space Climate Observatory* ([DSCOVR](#)) de la [NASA](#), y fue obtenido de [CDAweb](#).

Este recurso proporciona datos combinados omnidireccionales de [ACE](#) y [DSCOVR](#) con una resolución temporal de una hora, incluyendo el campo magnético interplanetario ([IMF](#)), datos de plasma, flujos de protones energéticos, además de índices solares y magnéticos. Aunque existe una versión con mayor resolución temporal (5 minutos), el [OMNI_HRO_5MIN](#), se optó por la resolución horaria por dos razones: la falta de

¹OMNI Combined, Definitive, Hourly [IMF](#) and Plasma Data, and Energetic Proton Fluxes, Time-Shifted to the Nose of the Earth's Bow Shock, plus Solar and Magnetic Indices - J.H. King, N. Papitashvili (ADNET, [NASA](#) GSFC).

datos [Dst](#) con resolución menor a una hora y la intención de reducir el requerimiento computacional del proyecto.

Los datos seleccionados abarcan el período del 14 de enero de 2001 al 31 de diciembre de 2016, coincidiendo con el conjunto de tormentas descrito en la [Tabla 1](#) de [Gruet et al. \(2018\)](#).

Se escogen las variables que influyen directamente en el índice [Dst](#) y que no son magnitudes derivadas, con la excepción de [E_field](#). Inicialmente, se eligen 14 variables (descritas en la [Tabla 4.1](#)), a las que se incluyen la variable temporal ([Datetime](#)) y las desviaciones de las variables seleccionadas que están disponibles. En la Sección [4.3](#), se descartan las variables [Bmag](#) y [AP](#). Durante el entrenamiento del modelo (Sección [4.5](#)), se ignoran las variables de identificación ([ID_IMF](#) y [ID_plasma](#)), las desviaciones ([dev_*](#)) y la variable temporal ([Datetime](#)), resultando en un total de 10 variables predictoras, que se muestran en las variables sin asterisco en la [Tabla 4.1](#).

Variable	Descripción
Dst (nT)	Representa el cambio en el campo magnético terrestre causado por tormentas solares. Especifica la perturbación del campo magnético en la magnetosfera terrestre debido a la interacción con el viento solar. El valor de la variable Dst en el momento actual se utiliza para predecir el valor de la variable Dst a <i>lookforward</i> horas. Por lo tanto, Dst es tanto la variable independiente como la variable objetivo.
Bmag (nT) *se descarta *por correlación	Representa la magnitud del campo magnético. Este es un indicador clave del estado magnético en el entorno espacial.
Bx (nT)	Componente del campo magnético a lo largo del eje X del sistema de coordenadas. Coincidente entre ambos sistemas de coordenadas elegidos ² .
By_gse (nT) y Bz_gse (nT) ²	Componentes del campo magnético en el sistema de coordenadas Geocéntrico Solar Eclíptico.
By_gsm (nT) y Bz_gsm (nT) ²	Componentes del campo magnético en el sistema de coordenadas Geocéntrico Solar Magnético.
P_density (cm ⁻³)	Densidad del plasma, importante para entender las condiciones del medio interplanetario.
AP *se descarta *por nulos	Es el ratio alfa/protón, que define el cociente entre la cantidad de partículas alpha y protones en el medio espacial.
E_field (mV/m)	Campo eléctrico derivado, relevante para estudios de interacciones magnéticas y eléctricas en el espacio. Se trata de una variable derivada, pero se considera altamente relevante en la predicción de tormentas geomagnéticas, y se dejará como parte del conjunto de datos.
plasma_T (K)	Temperatura del plasma, crucial para analizar las propiedades termodinámicas del medio espacial.
plasma_V (km/s)	Velocidad del plasma, que influye en la dinámica del viento solar y su interacción con la magnetosfera terrestre.

²La selección del campo magnético en dos sistemas de coordenadas se debe a la conclusión obtenida por ([Kumar and Raizada, 2009](#)), donde [Dst](#) parece tener una relación más marcada con la componente *z* del campo magnético para coordenadas geocéntrico solar magnéticas que para su análoga en coordenadas geocéntrico solar eclípticas, demostrando una posible discrepancia entre dichas medidas y su capacidad de predicción del índice objetivo. Ha de tenerse en cuenta que según el mismo estudio, la componente *B_z* no predice adecuadamente el índice [Dst](#) en ningún caso, y que esto se debe principalmente a las diferencias en la orientación de estos sistemas de coordenadas respecto al campo magnético terrestre. Más información en [NASA \(2024\)](#).

Variable	Descripción
Datetime *no utilizados	Marca temporal de los datos, esencial para correlacionar eventos y análisis temporal.
Id.s de satélite ID_IMF/ID_plasma *no utilizados	Utilizados para asegurar la consistencia en la recopilación de datos.

Tabla 4.1: Variables seleccionadas

4.3. Procesamiento de los datos

Antes de comenzar el procesamiento, se realiza un [EDA](#) sobre los datos descargados. Esto permite comprender los datos y sus características para enfocar el análisis y el procesamiento.

Estos datos se componen de 122712 registros y 22 columnas, incluyendo las variables de la Tabla 4.1 y sus desviaciones.

Las variables, como la magnitud del campo magnético (**Bmag**) y la densidad del plasma (**P_density**), muestran una amplia variabilidad, indicando fluctuaciones significativas en las condiciones solares. La variable **Dst**, que mide la actividad geomagnética, varía tomando un mínimo de -422. Se identifican datos faltantes en varias columnas, como **P_density** y **AP**. La presencia de datos incompletos indica que se requieren métodos de imputación en el procesamiento para mantener la precisión del modelo. Este [EDA](#) se puede ver en detalle en `01_01_EDA_storms_data_source.ipynb` y en el documento con las observaciones en [Anexos/01_EDA_source.pdf](#) (Apéndice B).

Para comenzar con el procesamiento, se realiza la normalización de los datos. Para ello, se utilizará el método *Standard Scaler*. Este es el método utilizado en este campo de estudio, puesto que no limita los datos a un rango específico, sino que mantiene la distribución original de los datos sin sesgarlos hacia un límite particular. En los datos de viento solar esto es una característica importante, ya que los datos extremos son información importante que permite predecir estos fenómenos. Este procedimiento se puede ver en mayor detalle en los `notebooks 03_0*_vDef_SVR_*h_ahead*.ipynb` (Apéndice B).

A continuación se realiza el tratamiento de nulos, esencial en una [SVR](#) porque los valores nulos pueden distorsionar el ajuste del modelo, llevando a predicciones inexactas o la imposibilidad de entrenar el modelo³. En este caso, se eliminarán las variables que contengan más de un 10 % del total de valores nulos, y se interpolarán linealmente los nulos en el resto de las variables.

Tras este análisis, que se muestra en la Tabla 4.2, se descarta la variable **AP**; es decir, el ratio alfa/protón.

Se interpolan el resto de datos de forma lineal con un *fillforward* y un *fillbackward*; es decir, los valores nulos en los extremos se rellenan utilizando los valores no nulos más cercanos anteriores y posteriores respectivamente. Los valores nulos intermedios se interpolan.

En este caso no se realiza tratamiento de *outliers* puesto que el objetivo del trabajo es predecir datos extremos y poder trabajar con datos ruidosos, por lo que se trata de conseguir que el modelo sea capaz de trabajar con ellos. Además, en estos datos los *outliers* pueden ser los puntos de interés que queremos identificar y analizar, ya que representan desviaciones significativas del comportamiento esperado y pueden significar períodos de tormenta.

³En caso de utilizar |SVR| de |scikit-learn|, la existencia de valores nulos en los datos impide el entrenamiento del modelo

Variable	Porcentaje nulos	Variable	Porcentaje nulos
ID_IMF	0.00	dev_Bz	0.00
ID_plasma	0.10	P_density	2.77
Bmag	0.00	dev_P_density	2.77
dev_Bmag	0.00	AP	11.98
Bx	0.00	dev_AP	11.98
By_gse	0.00	E_field	0.10
Bz_gse	0.00	plasma_T	2.40
By_gsm	0.00	dev_plasma_T	2.40
Bz_gsm	0.00	plasma_V	0.10
dev_Bx	0.00	Dst	0.00
dev_By	0.00	Datetime	0.00

Tabla 4.2: Porcentaje de valores nulos por variable en la selección de datos de tormenta

Para obtener un conjunto de datos con un alcance suficiente para probar el modelo sin exceder la extensión del trabajo, se filtra el conjunto de datos seleccionando únicamente los eventos de tormentas solares. En este caso, se han identificado 49 tormentas solares ocurridas entre 2001 y 2014, seleccionadas directamente de la Tabla 1 de [Gruet et al. \(2018\)](#). Las tormentas se consideran con un margen anterior y posterior de 5 días, ya que su duración estimada es de entre uno y tres días si se incluyen todas sus fases, de forma que este margen asegura la recopilación de todos los datos relevantes.

Este enfoque provoca solapamientos entre algunas tormentas. En caso de solapamiento, se ha decidido combinar las tormentas en una sola, ya que es común que varias tormentas sean parte de un mismo evento. Las tormentas geomagnéticas pueden originarse de [CME](#) muy cercanas en el tiempo, dado que ocurren durante períodos de alta actividad solar. Además, esto evita que los modelos entrenen múltiples veces con el mismo periodo temporal, evitando errores en el entrenamiento y sobreajuste del modelo. Esto reduce las 49 tormentas originales a un conjunto de 42 eventos, que se pueden observar en la Tabla 4.4.

Se muestra la distribución de los mínimos del Índice por tormenta en la Figura 4.1. La gravedad de estas tormentas se define mediante el convenio extendido en el estudio del clima espacial, que se puede observar en la Tabla 4.3.

Categoría	Rango de Dst (nT)
Calma	Dst superior a -30
Leve	Dst entre -30 y -50
Moderada	Dst entre -50 y -100
Fuerte	Dst entre -100 y -200
Severa	Dst entre -200 y -300
Muy severa	Dst inferior a -300

Tabla 4.3: Clasificación de la intensidad de las tormentas geomagnéticas basada en el índice [Dst](#)

Se realiza la selección de datos de entrenamiento, *test* y validación. Se obtiene una división estándar de 80% para entrenamiento, 20% para *test* y, para el modelo alternativo [LSTM-MLP](#), se obtiene un 10% para validación del *split* de entrenamiento. Como el 20% de los datos correspondería a 8,4 tormentas, se seleccionan un total de 9 tormentas para el *split* de prueba. Por tanto, los porcentajes reales ajustados son:

storm_index	date_start	date_end	min_DST	storm_index	date_start	date_end	min_DST
1	2001-03-15 13:00:00	2001-03-25 12:00:00	-149.0	22	2004-03-30 00:00:00	2004-04-08 23:00:00	-117.0
2	2001-03-26 08:00:00	2001-04-05 07:00:00	-387.0	23	2004-07-18 02:00:00	2004-08-01 12:00:00	-170.0
3	2001-04-13 06:00:00	2001-04-27 14:00:00	-114.0	24	2004-08-25 22:00:00	2004-09-04 21:00:00	-129.0
4	2001-08-12 21:00:00	2001-08-22 20:00:00	-105.0	25	2004-11-07 10:00:00	2004-11-17 09:00:00	-374.0
5	2001-09-26 08:00:00	2001-10-06 07:00:00	-166.0	26	2005-01-17 05:00:00	2005-01-27 04:00:00	-103.0
6	2001-10-16 21:00:00	2001-11-02 10:00:00	-187.0	27	2005-05-03 18:00:00	2005-05-13 17:00:00	-110.0
7	2002-03-19 09:00:00	2002-03-29 08:00:00	-100.0	28	2005-05-25 13:00:00	2005-06-04 12:00:00	-113.0
8	2002-04-13 07:00:00	2002-04-25 07:00:00	-149.0	29	2005-06-08 00:00:00	2005-06-17 23:00:00	-106.0
9	2002-05-06 19:00:00	2002-05-16 18:00:00	-110.0	30	2005-08-26 19:00:00	2005-09-05 18:00:00	-122.0
10	2002-05-18 17:00:00	2002-05-28 16:00:00	-109.0	31	2006-04-09 09:00:00	2006-04-19 08:00:00	-98.0
11	2002-07-28 05:00:00	2002-08-07 04:00:00	-102.0	32	2006-12-10 07:00:00	2006-12-20 06:00:00	-162.0
12	2002-08-30 05:00:00	2002-09-12 23:00:00	-181.0	33	2011-09-21 23:00:00	2011-10-01 22:00:00	-118.0
13	2002-09-26 16:00:00	2002-10-06 15:00:00	-176.0	34	2011-10-20 01:00:00	2011-10-30 00:00:00	-147.0
14	2002-11-16 10:00:00	2002-11-26 09:00:00	-128.0	35	2012-03-04 08:00:00	2012-03-14 07:00:00	-145.0
15	2003-05-24 23:00:00	2003-06-03 22:00:00	-144.0	36	2012-04-19 04:00:00	2012-04-29 03:00:00	-120.0
16	2003-06-13 09:00:00	2003-06-23 08:00:00	-141.0	37	2012-07-10 16:00:00	2012-07-20 15:00:00	-139.0
17	2003-07-07 05:00:00	2003-07-17 04:00:00	-105.0	38	2012-09-26 04:00:00	2012-10-14 07:00:00	-122.0
18	2003-08-13 15:00:00	2003-08-23 14:00:00	-148.0	39	2012-11-09 07:00:00	2012-11-19 06:00:00	-108.0
19	2003-11-15 20:00:00	2003-11-25 19:00:00	-422.0	40	2013-03-12 20:00:00	2013-03-22 19:00:00	-132.0
20	2004-01-17 13:00:00	2004-01-27 12:00:00	-130.0	41	2013-05-27 08:00:00	2013-06-06 07:00:00	-124.0
21	2004-02-06 17:00:00	2004-02-16 16:00:00	-93.0	42	2014-02-14 08:00:00	2014-02-24 07:00:00	-119.0

Tabla 4.4: Selección de tormentas obtenida de [Gruet et al. \(2018\)](#) combinada en períodos solapados

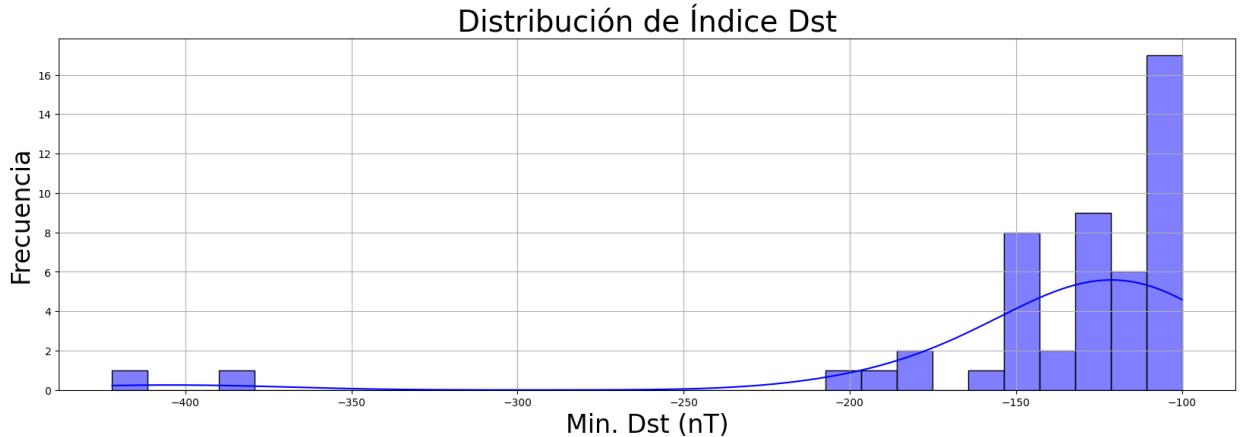


Figura 4.1: Distribución de Dst mínimo en tormentas de 4.4

78,57 % train, 21,43 % test para [SVR](#); 64,3 % train, 21,43 % test y 14,28 % validation en caso de [LSTM-MLP](#).

Para tener en cuenta el contexto temporal, se obtiene una **ventana temporal deslizante de los datos**. El parámetro *lookback* determina cuánto del pasado se considerará para hacer una predicción. Al utilizar varios puntos de datos anteriores como características, el modelo puede identificar patrones o tendencias en los datos. El *lookforward* determina cuánto adelante en el tiempo se desea predecir. En la práctica, esto significa que, para cada conjunto de datos basados en el *lookback*, intentamos predecir el valor de la variable objetivo varios pasos más adelante (*lookforward*).

En este caso, se está modelando un sistema donde los eventos pasados son significativos durante un tiempo considerable pero de magnitud desconocida. Se ha probado la función con valores de *lookback* de 12 y 24. Aunque un *lookback* mayor puede proporcionar más contexto histórico y potencialmente mejorar la precisión del modelo, también incrementa la carga computacional y el uso de memoria. Debido a estas limitaciones, se ha optado por utilizar un *lookback* de 12 en la implementación final y gran parte de los intentos previos para mantener un equilibrio entre precisión y eficiencia computacional.

En este estudio se va a realizar la predicción a 2, 4 y 6 horas. Por tanto, los valores de *lookforward* serán de 2, 4 y 6.

La información detallada sobre este proceso se puede ver en la función `create_window_df_nn` del Apéndice B.

4.4. Análisis exploratorio

Análisis descriptivo numérico El análisis descriptivo de los datos seleccionados y procesados ayuda a extraer varias observaciones relevantes sobre las **variables** medidas. Se ha obtenido un número considerable de observaciones (10786), lo que permite un entrenamiento adecuado del modelo, aunque se ha reducido el conjunto de datos para optimizar los recursos computacionales (la selección es $\approx 8,9\%$ de los datos originales).

Las medias de las **variables** son cercanas a cero, sugiriendo una distribución simétrica dada por el *Standard Scaler*. Sin embargo, los valores mínimos y máximos indican la presencia de valores extremos, especialmente en **Bz_gse** y **E_field**, con mínimos de -9,69 y -12,00 respectivamente, y máximos de 8,02 y 11,91. Esto concuerda con la ocurrencia de eventos extremos. La desviación estándar de todas las **variables** es aproximadamente 1, debido a la normalización de los datos. Los percentiles muestran que el 25 % de los valores de **By_gse** están por debajo de -0,57, mientras que el 75 % están por encima de 0,55, indicando una dispersión considerable. **plasma_T** presenta un valor máximo muy alto (35,99) comparado con su rango intercuartílico, lo que podría indicar una notable variabilidad en la temperatura del plasma.

Los histogramas muestran que la mayoría de las **variables**, como las componentes del campo magnético, tienen distribuciones normales centradas en cero, lo que indica simetría y variabilidad equilibrada. **P_density** y **plasma_T** presentan distribuciones sesgadas con colas largas a la derecha, sugiriendo eventos raros de alta magnitud, posiblemente las tormentas geomagnéticas que se quieren predecir. **E_field** muestra una dispersión significativa alrededor de cero, reflejando grandes variaciones. **plasma_V** está sesgada hacia bajas velocidades con algunas altas. **Dst** tiene una distribución sesgada hacia valores negativos, consistente con la presencia de tormentas geomagnéticas significativas.

Estudio de correlación Una de las correlaciones más destacadas es entre **By_gse** y **By_gsm** con un coeficiente de 0,96, lo que sugiere una alta relación entre estas componentes del campo magnético en diferentes sistemas de coordenadas. De manera similar, **Bz_gse** y **Bz_gsm** muestran una correlación de 0,93. A pesar de la fuerte correlación entre las componentes del campo magnético en los dos sistemas de coordenadas, se decide mantener ambos debido a las conclusiones de Kumar and Raizada (2009), que destaca la capacidad de predicción de **Bz_gse** que no comparte **Bz_gsm**. La correlación más fuerte entre variables es la que existe entre **B_mag** y las componentes del campo. En este caso, ya que se consideran todas las componentes en el estudio, se elimina **B_mag** por duplicidad de información.

Otra correlación notable es entre **E_field** y **plasma_V** (0,47), indicando que un aumento en el campo eléctrico está asociado con un incremento en la velocidad del plasma. Las correlaciones negativas fuertes también son de interés: **E_field** y **Bz_gse** tienen una correlación de -0,90, y **E_field** y **Bz_gsm** muestran -0,97, sugiriendo que un aumento en el campo eléctrico está asociado con una disminución en las componentes z del campo magnético en ambos sistemas de referencia.

Otras correlaciones moderadas incluyen la relación entre **plasma_T** y **plasma_V** (0,47), indicando que la temperatura del plasma y su velocidad están parcialmente correlacionadas. Las correlaciones más débiles, como

las entre `Dst` y las componentes del campo magnético, aunque menos significativas, pueden proporcionar información valiosa sobre la interacción entre el campo magnético y las condiciones geomagnéticas.

El análisis del mapa de calor de correlaciones revela varias relaciones significativas entre las variables del campo magnético y del plasma, indicando interdependencias que podrían influir en la dinámica del entorno espacial. Es importante considerar cómo la multicolinealidad (correlación alta entre variables predictoras) puede afectar el desempeño del modelo. No obstante, un estudio relevante concluyó que la multicolinealidad no afecta la precisión de las predicciones en los modelos de regresión lineal cuando el criterio es la precisión de predicción, lo que sugiere que se puede ignorar la multicolinealidad en este contexto (Morris and Lieberman, 2018)⁴.

Análisis de densidad de cada evento Durante eventos de tormentas geomagnéticas, las componentes del campo magnético $B\{x/y/z\}_gs\{m/e\}$ presentan distribuciones normales centradas en cero, indicando que estas **variables** tienden a fluctuar alrededor de un valor medio. Las distribuciones `deP_density` y la `T_plasma` tienen colas largas, sugiriendo la presencia de valores extremos durante algunas tormentas. El `E_field` muestra una gran variabilidad, reflejando la complejidad y la dificultad para predecir este fenómeno debido a las fluctuaciones a priori impredecibles. El resultado más interesante es que la velocidad del plasma parece presentar una distribución multimodal, lo que podría significar la presencia de diferentes *modos de plasma* durante las tormentas⁵. Esto se puede ver en la Figura 4.2.

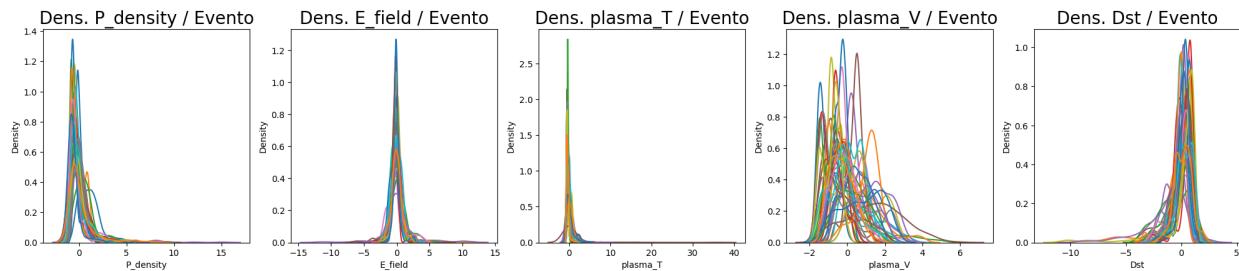


Figura 4.2: Densidad de cada variable por evento. Se han seleccionado únicamente las cinco variables más interesantes.

Lo mencionado anteriormente es solo un pequeño resumen de las observaciones y conclusiones más relevantes encontradas en el [EDA](#). Los resultados numéricos, gráficas y observaciones se encuentran detalladamente en el proceso que puede verse en `02_01_EDA_datos_procesados.ipynb` (Apéndice B).

4.5. Proceso de Optimización

Como se detalla en el Capítulo 1, la [SVR](#) define un hiperplano en un espacio de alta dimensionalidad mediante el uso de un kernel (por ejemplo, lineal, polinómico o RBF) para mapear los datos de entrada no lineales a un espacio lineal. Los hiperparámetros como C , ϵ y γ son cruciales para el rendimiento del modelo, y por ello la optimización de los mismos es el objetivo principal.

Durante la optimización con [Optuna](#)⁶, se definen estos hiperparámetros y se evalúa el modelo entrenándolo

⁴teniendo en cuenta que todos los modelos de la Sección 4.6 fueron de *kernel* lineal.

⁵Es decir, que pueden existir distintos estados o fuentes de plasma dependiendo de las condiciones geomagnéticas de la tormenta.

⁶Optuna es una librería de Python que utiliza métodos de búsqueda y optimización para encontrar la combinación de hiperparámetros que minimice la métrica de validación elegida.

y calculando el [MSE](#) en el conjunto de prueba. Para optimizar un modelo SVR utilizando Optuna, primero se definen los hiperparámetros a optimizar. Estos hiperparámetros se seleccionan de los valores predefinidos. Se implementa una función objetivo que entrena el modelo y calcula el [MSE](#) sobre un conjunto de prueba (Figura 4.3).

El espacio de hiperparámetros es explorado inicialmente mediante un *sampler* aleatorio: `RandomSampler()` de Optuna. Este método selecciona parámetros al azar, siguiendo una distribución especificada para cada hiperparámetro. Este enfoque, que no depende del historial de ensayos anteriores, realizó un total de 20 ensayos aleatorios para obtener una primera estimación de los resultados.

La exploración mediante un *sampler* en parrilla (*grid*) evalúa sistemáticamente combinaciones de hiperparámetros como C , γ , ϵ , y *degree*. Esta estrategia de búsqueda garantiza una exploración exhaustiva del espacio de hiperparámetros, aunque es computacionalmente intensa. Se realiza con múltiples combinaciones de parámetros, llegando finalmente a los modelos con los hiperparámetros definidos en la Tabla 4.5, que recorren localizaciones prometedoras del espacio de hiperparámetros.

Debido a los requerimientos de una [SVR](#), se observan colapsos de memoria durante el entrenamiento debido a entrenamientos individuales ineficientes que no convergen, lo que lleva a reiniciar el *kernel* y ajustar el número máximo de iteraciones (initialmente `max_iter=10000000` en [SVR](#) de `scikit-learn`) y el tiempo máximo de optimización (initialmente `timeout=1200` en `study.optimize` de [Optuna](#)). Tras múltiples configuraciones, se opta por una versión personalizada en la que se limita el tiempo máximo por iteración a 120s, cancelando el ensayo actual pero permitiendo la continuación de otros ensayos mediante la observación del estado del *kernel*, a diferencia del límite de tiempo `time_out` que detendría toda la optimización.

Validación Cruzada La validación cruzada implica dividir el conjunto de datos en múltiples pliegues, entrenando el modelo en algunos y validándolo en otros, lo que reduce la varianza en las estimaciones de rendimiento, ayuda a prevenir el sobreajuste y proporciona una estimación más precisa del desempeño del modelo en datos no vistos anteriormente.

En la validación Cruzada *K-Fold*, el conjunto se divide en K partes, entrenando el modelo repetidamente y utilizando cada vez un subconjunto diferente como prueba. Cada una de las K iteraciones utiliza una combinación diferente de pliegues para entrenamiento y prueba.

En este caso, la aplicación de K-folds se hace dentro del proceso de optimización con validación cruzada de 3 pliegues ($K = 3$). De esta manera, se garantiza que el proceso de optimización considere tanto la variabilidad en el rendimiento del modelo como su capacidad para generalizar a nuevos datos.

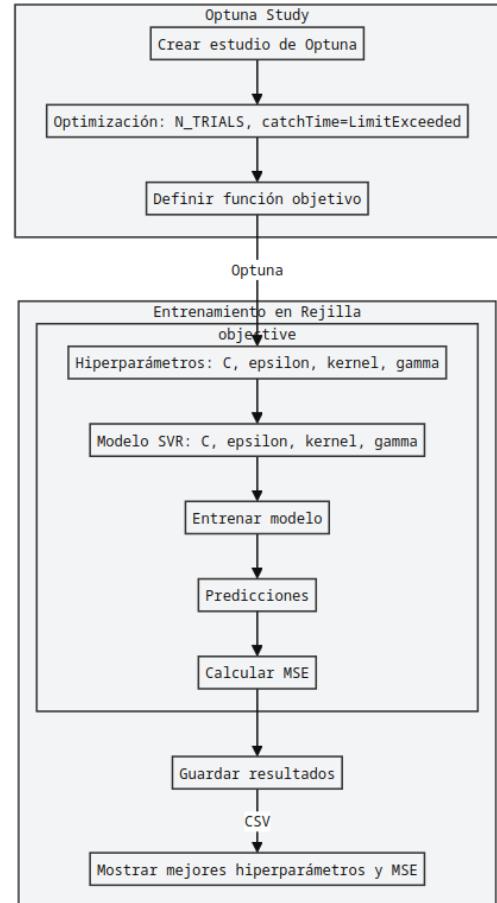


Figura 4.3: Arquitectura [SVR](#) (optimización con [Optuna \(2024\)](#))

Modelo y configuración final

<i>kernel</i>	Parámetro	Valores
Linear	C epsilon	[0,001; 0,005; 0,027; 0,139; 0,720; 3,728; 19,307; 100,000] [0,0001; 0,001; 0,01; 0,1; 1,0; 10,0]
Poly	C gamma epsilon degree coef0	[0,001; 1,000; 1000,000] [<i>scale</i>] [0,01; 0,316; 10,0] [3; 4; 5] [0,1; 1,0]
RBF	C gamma epsilon	[0,001; 0,016; 0,251; 3,981; 63,096; 1000,000] [<i>scale; auto</i>] [0,0001; 0,004; 0,215; 10,0]

Tabla 4.5: Parámetros de los diferentes tipos de *kernels* usados en el SVR

donde

- **scale:** $\gamma = \frac{1}{n \cdot \sigma^2}$, donde n es el número de variables predictoras y σ^2 es la varianza de las características.
- **auto:** $\gamma = \frac{1}{n}$, donde n es el número de variables predictoras.

El *kernel* sigmoide se descartó previamente puesto que no encajaba con el objetivo y los resultados eran insuficientes.

4.5.1. Modelo alternativo: LSTM-MLP

Se presenta un modelo que combina bloques **LSTM** y **MLP**, donde el **LSTM** procesa datos secuenciales, capturando dependencias a largo plazo y manteniendo memoria de contextos anteriores, mientras que el **MLP** toma las representaciones del **LSTM**, realiza la extracción de características y lleva a cabo la predicción final.

En una **LSTM**, los datos de entrada secuenciales se procesan paso a paso. En cada paso temporal, la **LSTM** decide qué información mantener y qué información descartar, y utiliza esta información para hacer predicciones. La idea detrás del uso de dos capas **LSTM** en lugar de una es permitir que el modelo aprenda representaciones más complejas y abstractas de los datos de entrada. La primera capa **LSTM** toma la secuencia de entrada y aprende características iniciales de esta secuencia. La salida de la primera capa **LSTM** se convierte en la entrada para la segunda capa **LSTM**. Esta segunda capa puede captar patrones más abstractos y de mayor nivel en los datos, basándose en las características aprendidas por la primera capa. El estado final de la segunda **LSTM** (`state_h_2`) se pasa a un **MLP**, que consiste en capas densas y capas de *dropout*. Esto permite transformar la salida de la **LSTM** en la predicción final.

Teniendo en cuenta que este es un modelo sencillo utilizado principalmente para comparar diferentes acercamientos al problema, se ha elegido un modelo sencillo que se pueda beneficiar de la aceleración por GPU ofrecida por cuDNN. Los parámetros elegidos para el modelo se pueden ver en la Figura 4.4. Se realiza un total de 150 entrenamientos por cada *lookforward*, que se pueden ver en detalle el repositorio (Apéndice B)

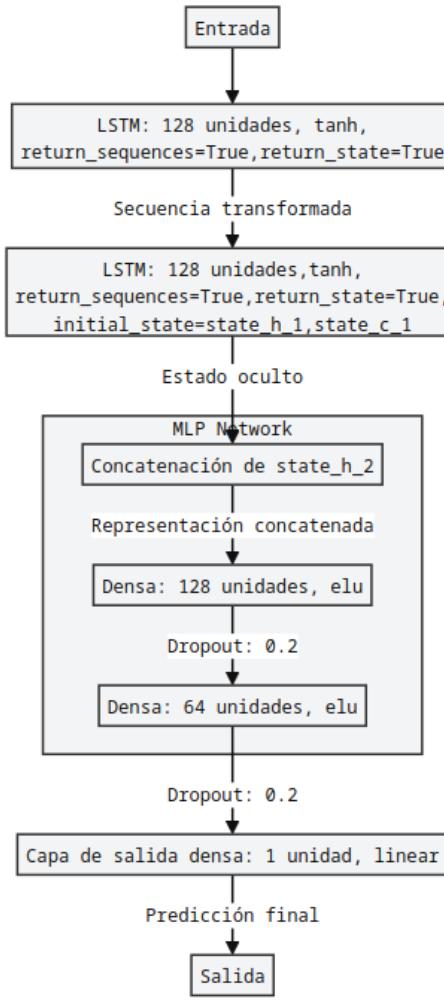


Figura 4.4: Arquitectura LSTM-MLP

4.6. Resultados

En esta sección se muestran los resultados de los modelos a 2, 4 y 6h de [SVR](#). Se presentan las métricas del mejor modelo obtenido para cada *lookforward*, así como el rendimiento medio de cada *kernel* para todos los entrenamientos realizados. Posteriormente, se muestra de forma visual y se calcularán de forma específica las métricas para dos tormentas: la más grave y la que mejor se ajusta (de *test*), obtenidas de la Tabla 4.4.

También se van a presentar los resultados del modelo alternativo [LSTM-MLP](#), y se van a comparar con el modelo principal del trabajo.

Finalmente, se realizará un comentario sobre la dificultad a la hora de comparar de forma consistente los resultados de este trabajo y los obtenidos en los estudios de referencia [Gruet et al. \(2018\)](#).

4.6.1. Modelo SVR

Modelo de regresión a 2h

Para este conjunto de datos y el horizonte de predicción de 2 horas, el modelo lineal es el más eficaz en términos de precisión y eficiencia. Su **MSE** es 0,0894 y sus hiperparámetros son: **kernel**: `linear`, **C**: 0.7197 y **epsilon**: 0.1

Al tratarse del kernel lineal, se puede deducir más fácilmente la ecuación de la regresión con las características más importantes. En este caso, se ha obtenido la ecuación completa. A continuación, se ha utilizado la Eliminación Recursiva de Características con Validación Cruzada (**RFECV**), una técnica de selección de características que elimina recursivamente las características menos importantes. En este caso, se ha determinado el número óptimo de características a retener mediante validación cruzada. Esto ha permitido seleccionar 6 variables y el intercepto (Ecuación 4.1).

$$\begin{aligned} \text{Dst}(t + 2h) = & -8,86 \times 10^{-4} \cdot \text{E_field}(t - 3h) \\ & + 5,77 \times 10^{-4} \cdot \text{Bz_gse}(t - 2h) \\ & - 1,20 \times 10^{-3} \cdot \text{E_field}(t - 2h) \\ & + 6,90 \times 10^{-5} \cdot \text{Dst}(t - 2h) \\ & + 7,35 \times 10^{-5} \cdot \text{Dst}(t - 1h) \\ & + 1,03 \times 10^{-4} \cdot \text{Dst}(t) \\ & - 0,9870 \end{aligned} \quad (4.1)$$

En total, la ecuación tendría 121 términos: 120 características y el intercepto. Esto se debe a que la **SVR** requiere que las secuencias de datos se aplaten (`.flatten`) en una sola fila de características, perdiendo la estructura temporal explícita de los datos. Esto hace que 10 variables predictoras por 12 horas de *lookback* generen 120 características que representan todas las observaciones en las 12 horas. Dado que las características están correlacionadas temporalmente, reducir a solo 6 características no captura adecuadamente la dinámica de los datos y no se recomienda. El resultado de los coeficientes se muestra de forma completa en la Tabla 4.6. Se puede ver el proceso detallado de obtención de las ecuaciones en `06_01_INTERPRETABILIDAD_SVR.ipynb` (Apéndice B).

Se puede ver en detalle la obtención de la ecuación en `06_01_INTERPRETABILIDAD_SVR.ipynb` (Apéndice B)

Al observar el promedio de **MSE** por **kernel** (Tabla 4.7), se puede ver que el **kernel** lineal presenta el MSE más bajo entre los tres, lo que sugiere que, en promedio, es el modelo más preciso para las predicciones a 2 horas. Además, es el que requiere menos parámetros y, por lo tanto, una rejilla con menos entrenamientos en total. El **kernel RBF** tiene un **MSE** ligeramente superior al del **kernel** lineal, pero sigue siendo significativamente menor que el del **kernel** polinómico. Además, los entrenamientos con este **kernel** fueron muy rápidos (detalles en Apéndice C). Por otro lado, el **kernel** polinómico muestra el **MSE** más alto, indicando que su desempeño es considerablemente peor en comparación.

En todos los **kernels**, la duración de las pruebas tiende a aumentar con valores más altos de **C**, lo cual es consistente con la naturaleza de la regularización en **SVM** y **SVR**, donde valores más altos de **C** pueden llevar a modelos más complejos que requieren más tiempo para entrenar. Se ha de destacar que los **NaN** en **value** se deben a una elección personal. Si un ensayo dura más de 2 minutos, se fuerza artificialmente

Var	Coef (2h)	Coef (4h)	Coef (6h)	Var	Coef (2h)	Coef (4h)	Coef (6h)
Bx(t-11h)	-4.52e-06	4.78e-05	4.78e-05	Bx(t-5h)	-8.07e-06	-7.47e-06	-7.47e-06
By_gse(t-11h)	-2.47e-05	3.62e-05	3.62e-05	By_gse(t-5h)	-7.50e-05	-6.61e-05	-6.61e-05
Bz_gse(t-11h)	6.11e-06	2.28e-05	2.28e-05	Bz_gse(t-5h)	-4.79e-06	-1.16e-05	-1.16e-05
By_gsm(t-11h)	-2.26e-05	3.63e-05	3.63e-05	By_gsm(t-5h)	-7.64e-05	-8.22e-05	-8.22e-05
Bz_gsm(t-11h)	1.87e-05	5.96e-05	5.96e-05	Bz_gsm(t-5h)	3.39e-05	8.49e-05	8.49e-05
P_density(t-11h)	-6.01e-06	7.87e-08	7.87e-08	P_density(t-5h)	4.47e-06	-4.12e-05	-4.12e-05
E_field(t-11h)	-3.92e-05	-1.10e-04	-1.10e-04	E_field(t-5h)	-6.16e-05	-1.26e-04	-1.26e-04
plasma_T(t-11h)	-5.06e-10	-1.26e-09	-1.26e-09	plasma_T(t-5h)	-2.16e-10	-2.15e-09	-2.15e-09
plasma_V(t-11h)	-1.40e-06	-1.59e-06	-1.59e-06	plasma_V(t-5h)	-2.45e-06	-5.23e-06	-5.23e-06
Dst(t-11h)	3.16e-06	3.69e-06	3.69e-06	Dst(t-5h)	7.55e-06	1.11e-05	1.11e-05
Bx(t-10h)	4.12e-06	3.45e-05	3.45e-05	Bx(t-4h)	-1.46e-05	-4.70e-05	-4.70e-05
By_gse(t-10h)	-3.18e-05	7.18e-06	7.18e-06	By_gse(t-4h)	-9.11e-05	-9.49e-05	-9.49e-05
Bz_gse(t-10h)	-3.93e-06	4.12e-05	4.12e-05	Bz_gse(t-4h)	7.14e-05	-1.29e-04	-1.29e-04
By_gsm(t-10h)	-3.45e-05	7.52e-06	7.52e-06	By_gsm(t-4h)	-7.34e-05	-1.35e-04	-1.35e-04
Bz_gsm(t-10h)	1.34e-05	7.75e-05	7.75e-05	Bz_gsm(t-4h)	1.08e-04	-3.57e-06	-3.57e-06
P_density(t-10h)	-9.47e-06	5.00e-05	5.00e-05	P_density(t-4h)	-6.54e-06	3.55e-05	3.55e-05
E_field(t-10h)	-2.62e-05	-1.39e-04	-1.39e-04	E_field(t-4h)	-2.40e-04	5.49e-05	5.49e-05
plasma_T(t-10h)	-4.28e-10	-9.78e-10	-9.78e-10	plasma_T(t-4h)	-2.10e-10	-1.94e-09	-1.94e-09
plasma_V(t-10h)	-1.50e-06	-1.67e-06	-1.67e-06	plasma_V(t-4h)	-2.11e-06	-5.75e-06	-5.75e-06
Dst(t-10h)	4.04e-06	5.46e-06	5.46e-06	Dst(t-4h)	8.09e-06	1.64e-05	1.64e-05
Bx(t-9h)	1.12e-06	6.68e-05	6.68e-05	Bx(t-3h)	-5.98e-06	-3.41e-07	-3.41e-07
By_gse(t-9h)	-3.76e-05	-4.03e-05	-4.03e-05	By_gse(t-3h)	-7.56e-05	-6.61e-05	-6.61e-05
Bz_gse(t-9h)	-8.19e-06	7.43e-05	7.43e-05	Bz_gse(t-3h)	7.65e-05	-1.17e-04	-1.17e-04
By_gsm(t-9h)	-4.02e-05	-4.34e-05	-4.34e-05	By_gsm(t-3h)	-6.00e-05	-1.00e-04	-1.00e-04
Bz_gsm(t-9h)	1.48e-05	9.82e-05	9.82e-05	Bz_gsm(t-3h)	1.05e-04	2.13e-05	2.13e-05
P_density(t-9h)	-1.03e-05	9.23e-05	9.23e-05	P_density(t-3h)	-1.74e-05	7.93e-05	7.93e-05
E_field(t-9h)	-2.64e-05	-1.79e-04	-1.79e-04	E_field(t-3h)	-2.42e-04	-1.53e-05	-1.53e-05
plasma_T(t-9h)	-3.72e-10	-8.47e-10	-8.47e-10	plasma_T(t-3h)	-4.03e-10	-1.52e-09	-1.52e-09
plasma_V(t-9h)	-1.62e-06	-1.60e-06	-1.60e-06	plasma_V(t-3h)	-2.13e-06	-5.73e-06	-5.73e-06
Dst(t-9h)	4.53e-06	7.74e-06	7.74e-06	Dst(t-3h)	1.24e-05	1.99e-05	1.99e-05
Bx(t-8h)	-1.53e-06	8.15e-05	8.15e-05	Bx(t-2h)	-8.49e-06	-5.75e-05	-5.75e-05
By_gse(t-8h)	-6.19e-05	-4.04e-05	-4.04e-05	By_gse(t-2h)	-5.72e-05	-1.19e-04	-1.19e-04
Bz_gse(t-8h)	-1.23e-05	3.22e-05	3.22e-05	Bz_gse(t-2h)	1.26e-04	4.30e-06	4.30e-06
By_gsm(t-8h)	-5.90e-05	-3.15e-05	-3.15e-05	By_gsm(t-2h)	-3.12e-05	-8.91e-05	-8.91e-05
Bz_gsm(t-8h)	1.99e-05	4.26e-05	4.26e-05	Bz_gsm(t-2h)	1.40e-04	1.52e-04	1.52e-04
P_density(t-8h)	-6.04e-06	7.85e-05	7.85e-05	P_density(t-2h)	4.07e-06	3.17e-05	3.17e-05
E_field(t-8h)	-3.31e-05	-8.24e-05	-8.24e-05	E_field(t-2h)	-3.17e-04	-3.03e-04	-3.03e-04
plasma_T(t-8h)	-7.20e-10	-2.98e-10	-2.98e-10	plasma_T(t-2h)	-3.23e-10	-5.32e-10	-5.32e-10
plasma_V(t-8h)	-1.99e-06	-1.73e-06	-1.73e-06	plasma_V(t-2h)	-2.04e-06	-4.51e-06	-4.51e-06
Dst(t-8h)	4.03e-06	9.73e-06	9.73e-06	Dst(t-2h)	1.62e-05	1.80e-05	1.80e-05
Bx(t-7h)	-1.13e-05	2.68e-05	2.68e-05	Bx(t-1h)	-5.34e-06	-1.09e-04	-1.09e-04
By_gse(t-7h)	-7.81e-05	-2.57e-05	-2.57e-05	By_gse(t-1h)	-3.70e-05	-1.02e-04	-1.02e-04
Bz_gse(t-7h)	-1.20e-05	-3.29e-07	-3.29e-07	Bz_gse(t-1h)	1.47e-04	8.87e-05	8.87e-05
By_gsm(t-7h)	-7.46e-05	-2.76e-05	-2.76e-05	By_gsm(t-1h)	-9.26e-06	-6.57e-05	-6.57e-05
Bz_gsm(t-7h)	2.66e-05	3.45e-05	3.45e-05	Bz_gsm(t-1h)	1.50e-04	2.13e-04	2.13e-04
P_density(t-7h)	-2.62e-05	2.64e-05	2.64e-05	P_density(t-1h)	5.70e-06	-3.04e-05	-3.04e-05
E_field(t-7h)	-5.09e-05	-1.47e-05	-1.47e-05	E_field(t-1h)	-3.39e-04	-4.33e-04	-4.33e-04
plasma_T(t-7h)	-8.98e-10	1.45e-10	1.45e-10	plasma_T(t-1h)	-1.12e-10	2.29e-11	2.29e-11
plasma_V(t-7h)	-2.44e-06	-1.88e-06	-1.88e-06	plasma_V(t-1h)	-1.97e-06	-4.70e-06	-4.70e-06
Dst(t-7h)	5.73e-06	9.30e-06	9.30e-06	Dst(t-1h)	1.70e-05	2.46e-05	2.46e-05
Bx(t-6h)	-1.15e-05	-1.20e-05	-1.20e-05	Bx(t)	2.79e-05	-1.19e-04	-1.19e-04
By_gse(t-6h)	-7.48e-05	-5.35e-05	-5.35e-05	By_gse(t)	-2.51e-05	-8.27e-05	-8.27e-05
Bz_gse(t-6h)	-2.21e-05	3.72e-05	3.72e-05	Bz_gse(t)	1.53e-04	3.18e-04	3.18e-04
By_gsm(t-6h)	-7.76e-05	-7.13e-05	-7.13e-05	By_gsm(t)	2.30e-06	-1.12e-05	-1.12e-05
Bz_gsm(t-6h)	1.80e-05	8.67e-05	8.67e-05	Bz_gsm(t)	1.52e-04	3.92e-04	3.92e-04
P_density(t-6h)	-1.32e-05	2.52e-05	2.52e-05	P_density(t)	1.69e-07	5.63e-06	5.63e-06
E_field(t-6h)	-2.13e-05	-1.26e-04	-1.26e-04	E_field(t)	-3.24e-04	-8.11e-04	-8.11e-04
plasma_T(t-6h)	-4.47e-10	-2.31e-09	-2.31e-09	plasma_T(t)	-1.84e-10	4.46e-10	4.46e-10
plasma_V(t-6h)	-2.37e-06	-3.62e-06	-3.62e-06	plasma_V(t)	-1.70e-06	-4.77e-06	-4.77e-06
Dst(t-6h)	7.88e-06	7.88e-06	7.88e-06	Dst(t)	2.06e-05	3.39e-05	3.39e-05

Tabla 4.6: Coeficientes de las ecuaciones de Regresión a 2h, 4h y 6h resultantes de la **SVR**.

<i>kernel</i>	MSE	Nº iteraciones
Linear	3.5076	36
RBF	4.2784	47
Poly	45.7375	51

Tabla 4.7: Error cuadrático medio promedio para cada *kernel* a 2h.

una excepción que pasa al siguiente ensayo devolviendo como valor del intento un NaN. Esto permite que el entrenamiento tenga una duración máxima determinada y no se inviertan incorrectamente los esfuerzos computacionales. A priori, esto podría dar lugar al descarte de un modelo mejor, pero la mejora se considera despreciable en favor de reducir al máximo el coste computacional.

La información completa sobre las iteraciones, los parámetros y el tiempo de ejecución se pueden encontrar en C ([RESULTADOS](#) en Apéndice B). Este anexo incluye la información completa de esta sección y las secciones 4.6.1 y 4.6.1.

En la tabla 4.8 y la figura 4.5, se muestran los resultados y métricas obtenidas para el conjunto completo de datos, las tormentas definidas en la Tabla 4.4.

Métrica	RMSE	MSE	MAE	R ²	MedAE	Var. expl.	Max error
Valor	11.11	123.35	6.11	0.9225	3.87	0.92	193.33

Tabla 4.8: Métricas de la predicción a 2h mediante [SVR](#) lineal para el conjunto completo de tormentas.

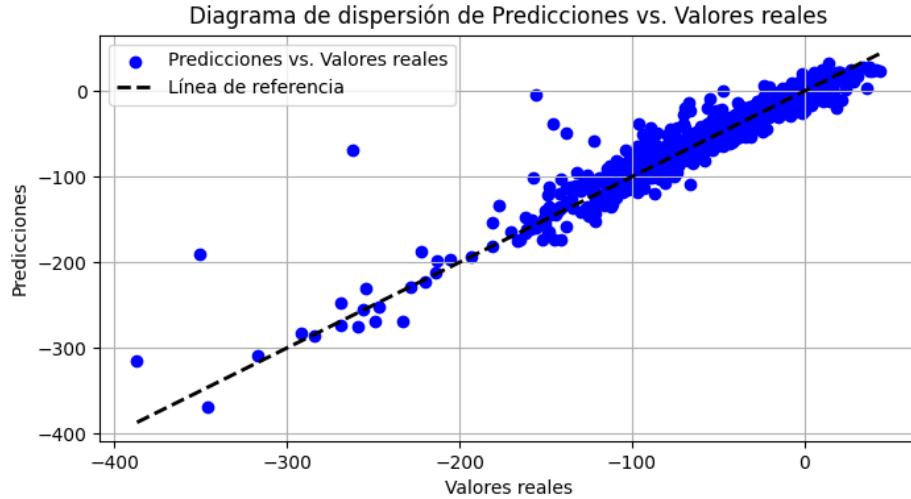


Figura 4.5: Diagrama de dispersión de predicciones contra valores reales de todas las tormentas en la predicción a 4h ([SVR](#))

A continuación, se muestran dos resultados para ilustrar la capacidad predictiva de este modelo. Primero, se muestra la tormenta nº 2, que es la tormenta más severa del grupo de prueba (Figura 4.6).

Como segundo ejemplo, se muestra la tormenta con mejor ajuste a los valores reales: la tormenta nº 31 (Figura 4.7). Se puede intuir que esta es la tormenta mejor predicha porque es una de las tormentas menos severas del grupo, lo cual hace disminuir el error en las métricas. No obstante, se puede observar en ambas figuras (4.6 y 4.7) que existe una *sombra* entre los valores predichos y los valores reales. Esto indica de forma clara que el modelo predice los valores de forma adecuada, pero siempre con un retraso considerable. Es

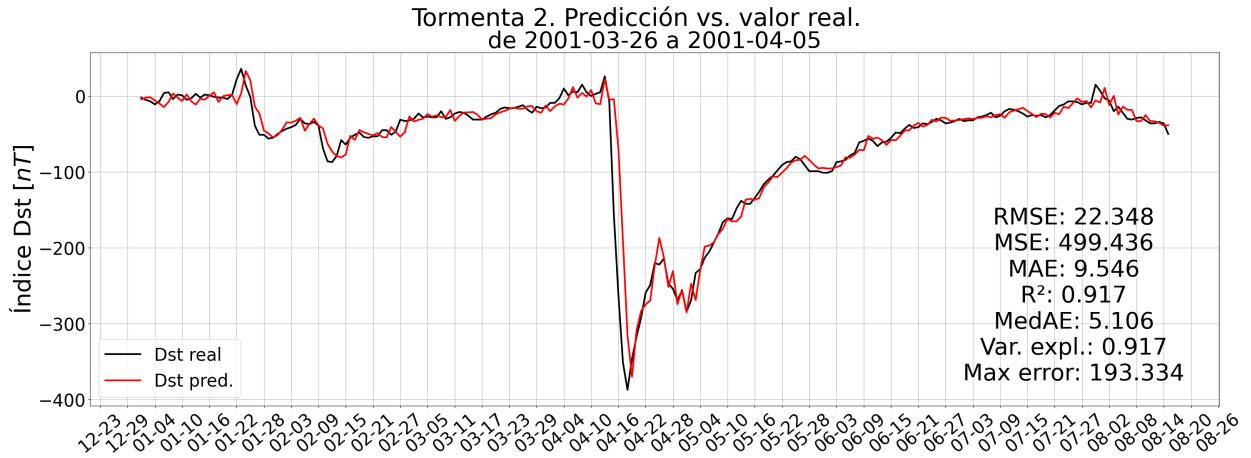


Figura 4.6: Tormenta 2: Valores predichos vs. valores reales (SVR) a 2h

decir, si quisieramos predecir una tormenta a 2h vista, no sería capaz de predecirla adecuadamente puesto que *avisaría* del fenómeno mucho más tarde. Esto es un resultado encontrado en la mayoría de estudios de Clima Espacial centrados en predicción de tormentas.

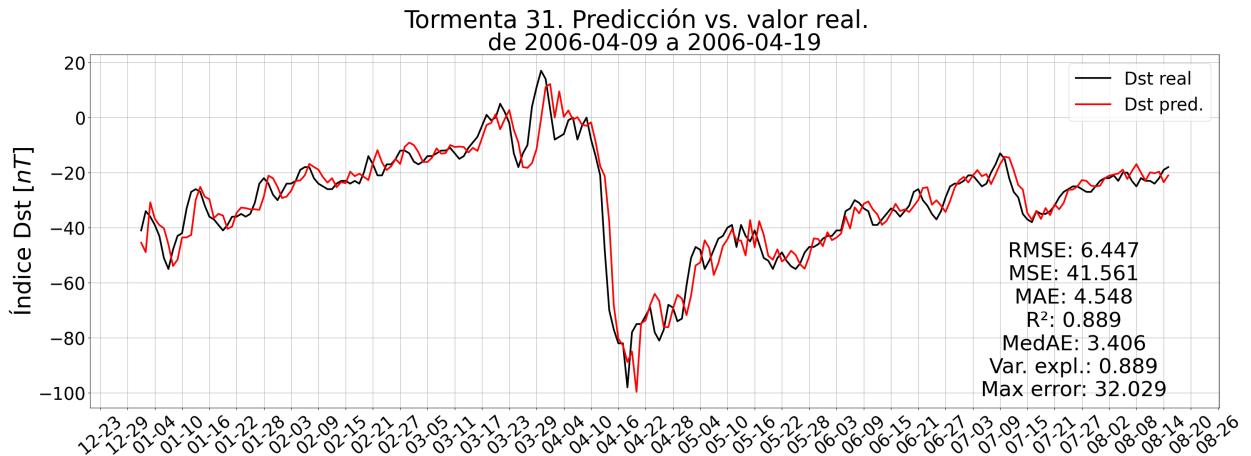


Figura 4.7: Tormenta 31: Valores predichos vs. valores reales (SVR) a 2h

Modelo de regresión a 4h

Para este conjunto de datos y el horizonte de predicción de 4 horas, el modelo lineal es el más eficaz en términos de precisión y eficiencia. Su MSE es 0,2573 y sus hiperparámetros son: *kernel*: linear, C: 0.7197 y *epsilon*: 0.1.

Al tratarse de nuevo del kernel lineal, se puede mostrar la ecuación de la regresión con las características más importantes. Se obtienen análogamente al apartado anterior y da como resultado la Ecuación 4.2, con 6 características y el intercepto.

$$\begin{aligned}
Dst(t + 4h) = & -8,86e - 04 \cdot E_field(t - 3h) \\
& + 5,77e - 04 \cdot Bz_gse(t - 2h) \\
& - 1,20e - 03 \cdot E_field(t - 2h) \\
& + 6,90e - 05 \cdot Dst(t - 2h) \\
& + 7,35e - 05 \cdot Dst(t - 1h) \\
& + 1,03e - 04 \cdot Dst(t) \\
& - 0,9870
\end{aligned} \tag{4.2}$$

Al observar el promedio de **MSE** por *kernel* (Tabla 4.9), se puede ver que el *kernel RBF* presenta el **MSE** más bajo entre los tres, lo que sugiere que, en promedio, es el modelo más preciso para las predicciones a 4 horas. Sin embargo, es importante destacar que el mejor modelo obtenido individualmente es de *kernel* lineal.

<i>kernel</i>	MSE	Nº iteraciones
Linear	5.7601	22
RBF	5.5090	36
Poly	21.4601	39

Tabla 4.9: Error cuadrático medio promedio para cada *kernel* a 4h.

Se realiza una gráfica de dispersión y el cálculo de las métricas para todo el conjunto de tormenta (Figura 4.8 y Tabla 4.10, respectivamente) para mostrar los resultados generales del modelo.

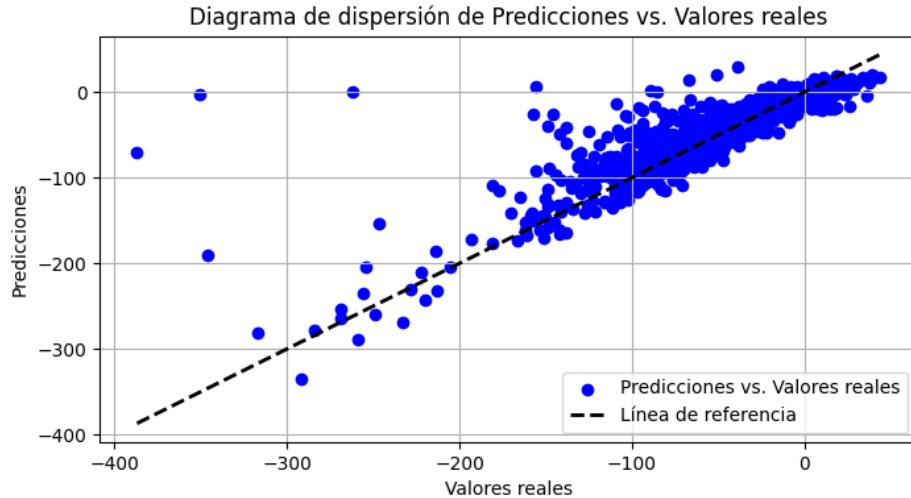


Figura 4.8: Diagrama de dispersión de predicciones contra valores reales de todas las tormentas en la predicción a 4h (**SVR**)

Métrica	RMSE	MSE	MAE	R ²	MedAE	Var. expl.	Max error
Valor	18.85	355.24	9.41	0.7778	5.66	0.78	348.96

Tabla 4.10: Métricas de la predicción a 4h mediante SVR lineal para el conjunto completo de tormentas.

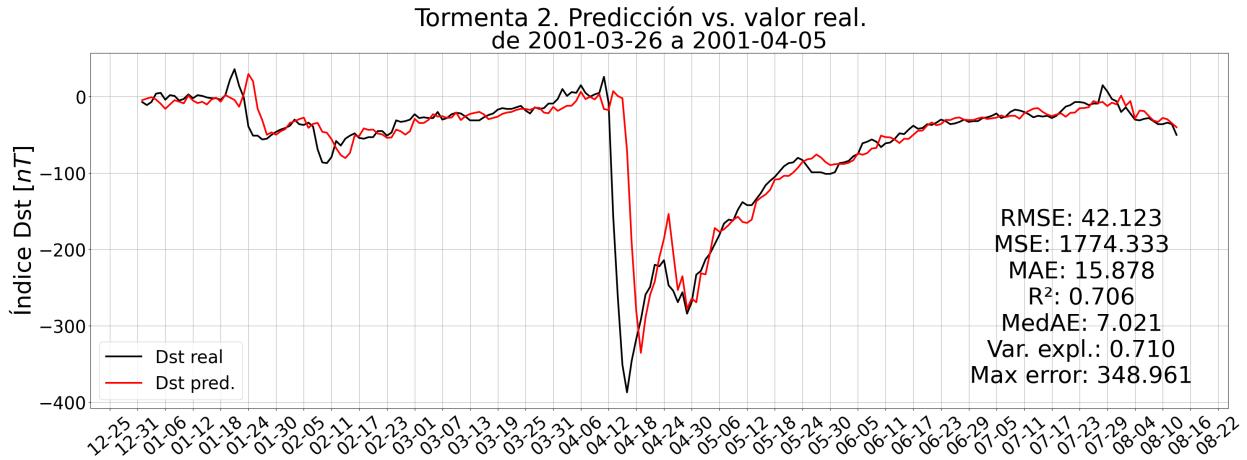


Figura 4.9: Tormenta 2: Valores predichos vs. valores reales ([SVR](#)) a 4h

De nuevo, se muestran las tormentas nº2 y nº31 para evaluar una muestra de los resultados (Figuras 4.9 y 4.10). Se puede observar que los resultados empeoran notablemente en comparación con los resultados a 2 horas. En este caso, la sombra se acentúa, llegando aún más tarde. Esto también cabía esperar y se observa en otros estudios como [Gruet et al. \(2018\)](#).

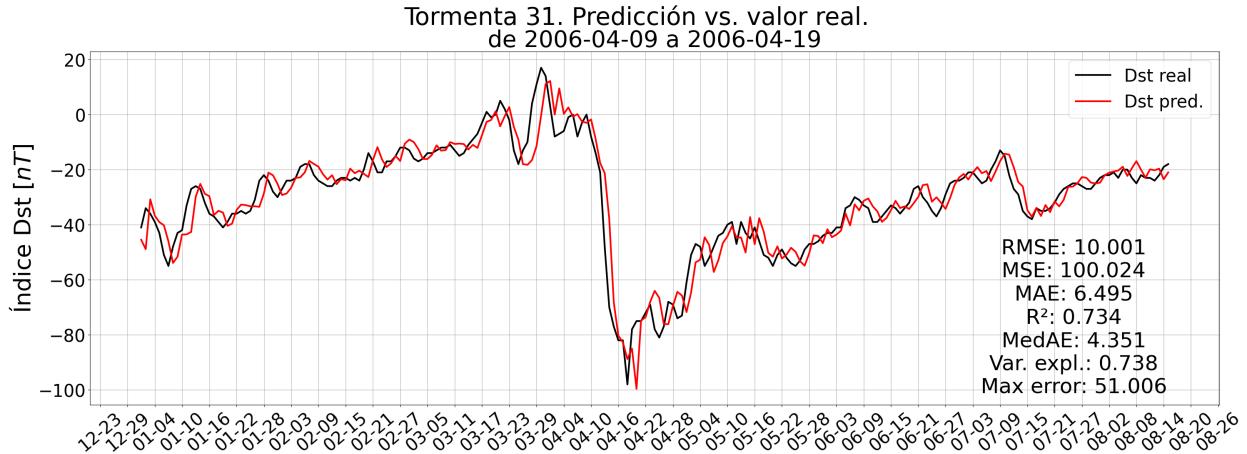


Figura 4.10: Tormenta 31: Valores predichos vs. valores reales ([SVR](#)) a 4h

De igual forma que en la predicción a 2h, se obtiene que el *kernel* lineal es significativamente mejor y se observa una sombra considerable en ambas tormentas, lo cual indica que se predicen con cierto retraso. Además, si se comparan las predicciones a 2h y 4h, se puede advertir que esta sombra es aún más prominente; es decir, que la predicción se realiza, proporcionalmente y de forma absoluta, más tarde.

Modelo de regresión a 6h

Para este conjunto de datos y el horizonte de predicción de 6 horas, el modelo lineal es de nuevo el más eficaz en términos de precisión y eficiencia. Su [MSE](#) es 0,4018 y sus hiperparámetros son: *kernel* : `linear`, *C*: 0.7197, *epsilon*: 0.1.

Al ser kernel lineal, se puede mostrar la ecuación de la regresión con las características más importantes. Se obtienen análogamente a los apartados anteriores y da como resultado la Ecuación 4.3, con 6 características y el intercepto.

$$\begin{aligned}
 Dst(t + 6h) = & -8,86e - 04 \cdot E_field(t - 3h) \\
 & + 5,77e - 04 \cdot Bz_gse(t - 2h) \\
 & - 1,20e - 03 \cdot E_field(t - 2h) \\
 & + 6,90e - 05 \cdot Dst(t - 2h) \\
 & + 7,35e - 05 \cdot Dst(t - 1h) \\
 & + 1,03e - 04 \cdot Dst(t) \\
 & - 0,9870
 \end{aligned} \tag{4.3}$$

<i>kernel</i>	MSE	Nº iteraciones
Linear	5.3953	27
RBF	6.9295	28
Poly	14.6611	37

Tabla 4.11: Error cuadrático medio promedio para cada *kernel* a 6h.

Al observar el promedio de MSE por *kernel* (Tabla 4.11), se puede ver que la justificación es análoga a la sección 4.6.1: el *kernel* lineal presenta el MSE más bajo entre los tres, por lo que, en promedio, es el modelo más preciso para las predicciones a 6 horas.

Se muestra el diagrama de dispersión de la predicción a 6h para el conjunto completo de tormentas, así como sus métricas (Figura 4.11 y Tabla 4.12).

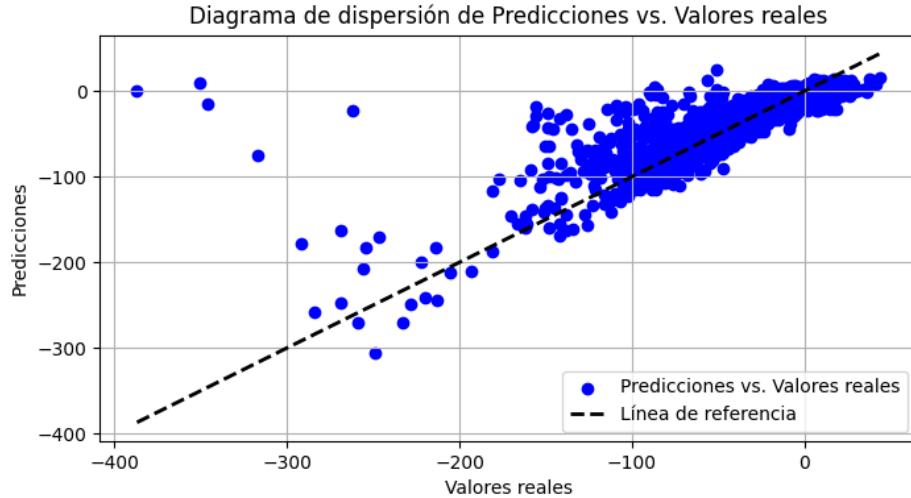


Figura 4.11: Diagrama de dispersión de predicciones contra valores reales de todas las tormentas en la predicción a 6h (SVR)

Finalmente, se grafican las tormentas n°2 y n°31 (Tabla 4.4) para evaluar una muestra de los resultados (Figuras 4.12 y 4.13).

Métrica	RMSE	MSE	MAE	R ²	MedAE	Var. expl.	Max error
Valor	23.55	554.80	11.67998	0.6541	6.60	0.66	387.14

Tabla 4.12: Métricas de la predicción a 6h mediante [SVR](#) lineal para el conjunto completo de tormentas.

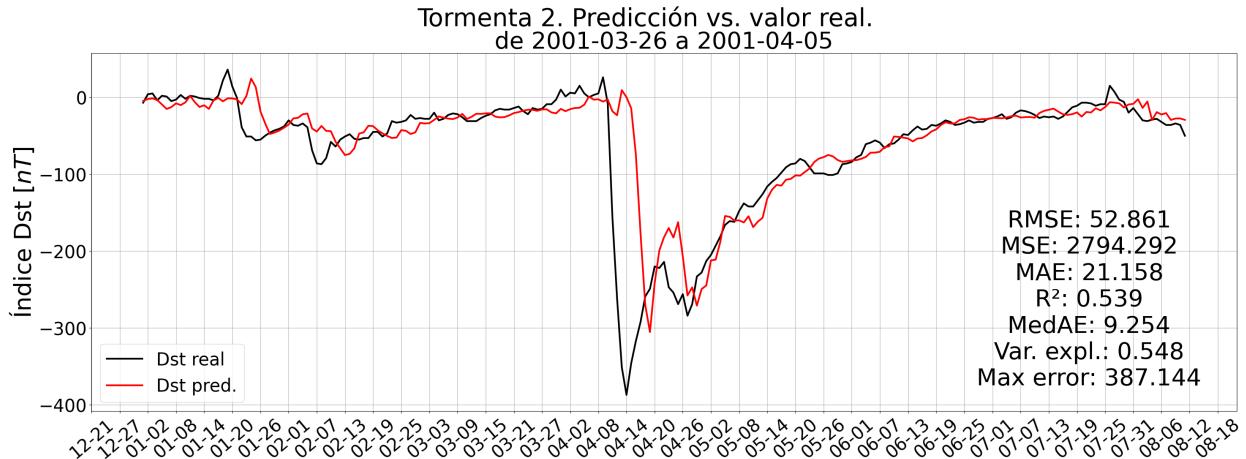


Figura 4.12: Tormenta 2: Valores predichos vs. valores reales ([SVR](#)) a 6h

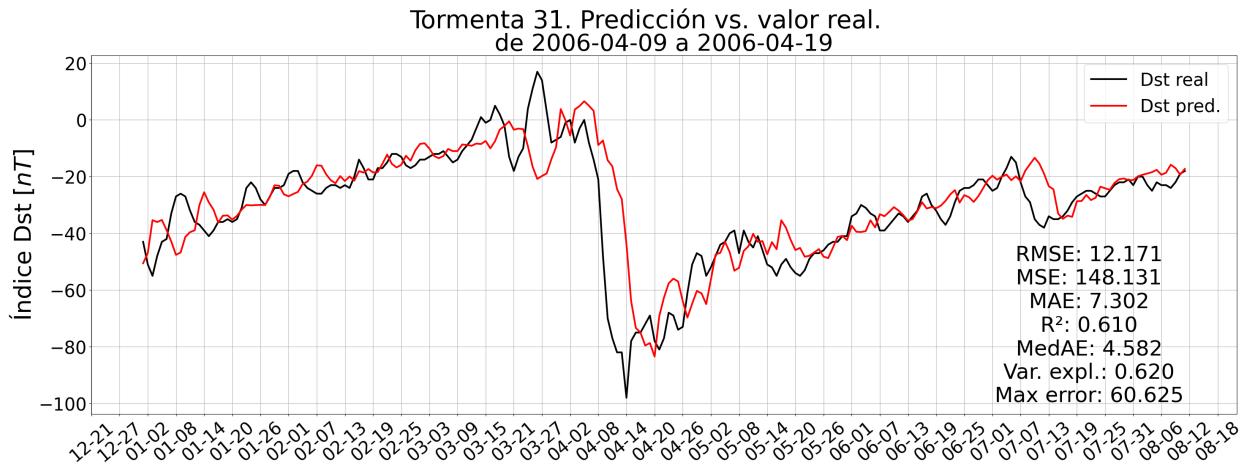


Figura 4.13: Tormenta 31: Valores predichos vs. valores reales ([SVR](#)) a 6h

La conclusión es similar a las Secciones anteriores de Predicción a 2h y 4h, viendo que se incrementa aún más la sombra de la predicción, como cabía esperar.

Observación

Comparando estos resultados con los reportados por ([Gruet et al., 2018](#)) y con los modelos utilizados para comparar en esta misma publicación: [Lazzús et al. \(2017\)](#), [Bala and Reiff \(2012a\)](#) y [Wu and Lundstedt \(1997\)](#), se observa lo siguiente:

Se observa que no se ha mejorado ninguno de los modelos de la bibliografía. No obstante, se puede justificar

Tiempo	Nuestro (2024)	Gruet et al. 2018 (no GPS)	Lazzús et al. 2017	Bala & Reiff 2012	Wu & Lundstedt 1997
$t + 2h$	11.11	6.65	7.05	-	16.3
$t + 4h$	18.85	8.86	10.44	-	19.9
$t + 6h$	23.55	10.24	13.09	11.09	20.8

Tabla 4.13: Comparación del RMSE con estudios previos

que esta comparación no es consistente.

La selección de tormentas y descarte de datos en calma realizados en este trabajo muy probablemente ha perjudicado al modelo desarrollado en este estudio, ayudando a Gruet et al. (2018) y demás bibliografía a superar el comportamiento de los modelos reportados por este trabajo.

Se ha de tener en cuenta que disponibilidad de datos más recientes y variados también podría mejorar la precisión de los modelos. Debido a que se quería realizar una comparación lo más justa posible, la selección de datos en este trabajo es un subconjunto de los datos seleccionados en el estudio realizado por Gruet et al. (2018), por lo que este motivo puede descartarse.

Sin embargo, se ha estudiado que el uso de datos en calma para la predicción reduce significativamente el error del modelo completo, inclinando las evidencias a que el uso de períodos de calma en predicción de tormentas geomagnéticas *falsea* la capacidad de predicción del modelo en eventos de tormenta (Borovsky and Steinberg, 2006; Stepanova et al., 2008)

4.6.2. Modelo LSTM-MLP vs. SVR

En esta sección, se realiza una comparación sencilla entre el rendimiento del modelo principal (SVR) y el modelo de comparación (LSTM-MLP). El estudio se extenderá más en la comparación del modelo a 2h por limitación en el alcance del proyecto. En la predicción a 4h y 6h se presentarán los resultados de forma más concisa y se asegura que siguen la misma lógica y conclusiones. Se puede ver en mayor detalle en el repositorio (Apéndice B).

Modelo a 2h

Teniendo en cuenta el conjunto total de todas las tormentas, se obtienen la Figura 4.14 y la Tabla 4.14. En comparación con la dispersión de todas las tormentas a 2h con el SVR (Tabla 4.8), se puede observar que el RMSE es menor para la SVR, así como el resto de métricas salvo por el Error Máximo. En este caso, las métricas podrían ser confusas o no representar del todo bien la capacidad de predicción del modelo. Por esto, se compara a su vez el diagrama de dispersión. En este caso, queda claro que la SVR es superior a la predicción dada por LSTM-MLP, ya que las predicciones se ajustan mejor a los valores reales.

Métrica	RMSE	MSE	MAE	R ²	MedAE	Varianza explicada	Max error
Valor	11.32	128.20	6.22	0.9195	3.91	0.9196	152.47

Tabla 4.14: Métricas de la predicción a 2h para el conjunto completo de tormentas.

De forma más concreta, se muestra la predicción para la tormenta más severa del split de *test*: la tormenta nº 2 de la Tabla 4.4, que se dio entre el 26 de marzo de 2001 y el 05 de abril de 2001. Se compara con los

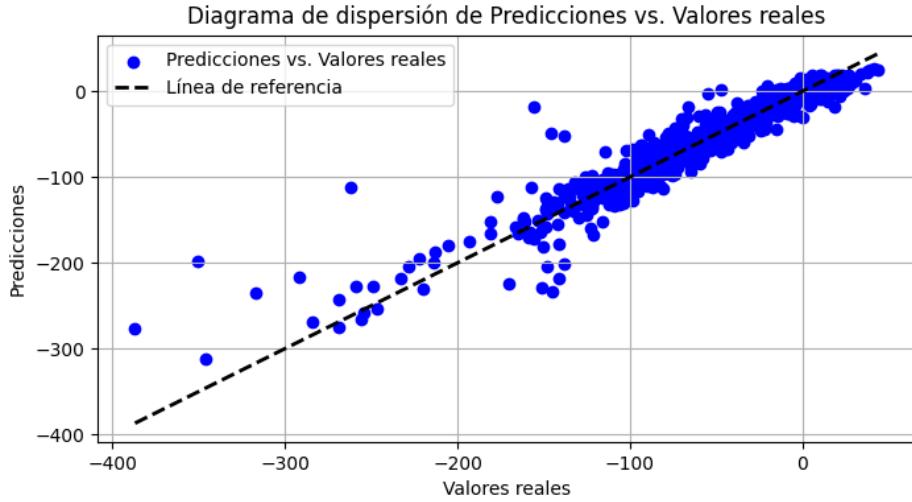


Figura 4.14: Diagrama de dispersión de predicciones contra valores reales de todas las tormentas en la predicción a 2 horas (**LSTM-MLP**)

resultados obtenidos en la **SVR** (Sección 4.6.1) en la Figura 4.15 y la Tabla 4.15.

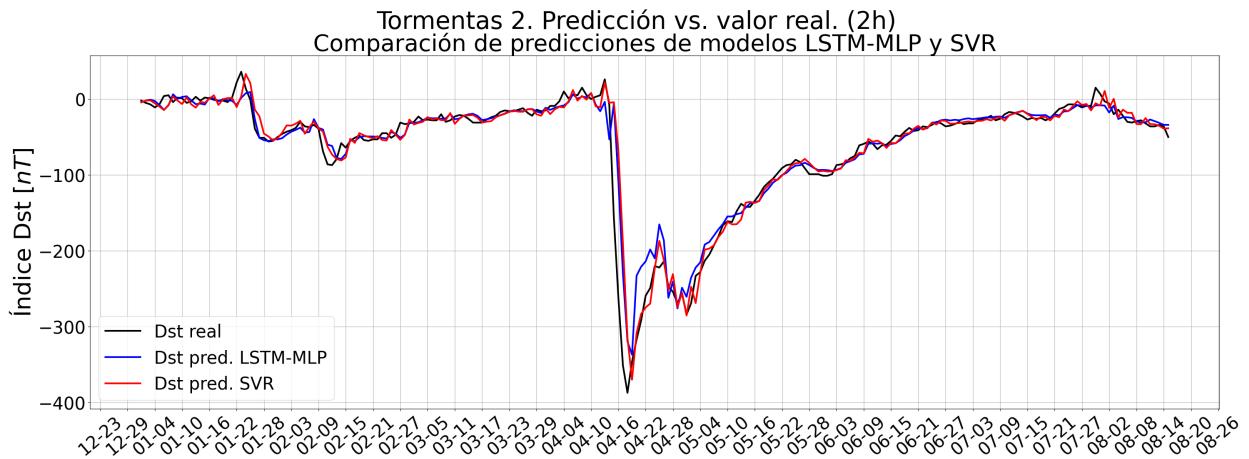


Figura 4.15: Comparación entre valores reales de Dst y valores predichos por **SVR** y **LSTM** a 2h en la tormenta 2.

Esta figura muestra la capacidad de predicción de ambos modelos. Se puede ver claramente que en este caso, el modelo **SVR** predice mejor la tormenta. No obstante, se ha de tener en cuenta que esto es un ejemplo y no puede generalizarse a todas las tormentas. Se pueden apreciar los resultados en más detalle en el repositorio (Apéndice B).

Predicción a 4h y 6h

En esta sección, se muestra únicamente un resumen de la comparación de los resultados de **SVR** y de **LSTM-MLP** a 4 y 6h. Se pueden ver los resultados de forma completa en el repositorio (Apéndice B)

Métrica	LSTM	SVR
RMSE	21.635	22.348
MSE	468.073	499.436
MAE	10.206	9.546
R ²	0.922	0.917
MedAE	5.410	5.106
Var. expl.	0.924	0.917
Max error	153.092	193.334

Tabla 4.15: Comparación de métricas entre LSTM y SVR a 2h, tormenta 2.

Las métricas obtenidas para todo el conjunto de tormentas a 4h y 6h se muestran en las tablas 4.16 y 4.17. La representación gráfica se muestra en las Figuras 4.16. Comparando estos resultados con sus equivalentes en la SVR , se advierte que para la predicción a 4h, el modelo LSTM es notablemente mejor que el SVR. No obstante, para la predicción a 6h, el modelo SVR es mejor que el LSTM. Por tanto, se puede decir con seguridad que la elección del modelo depende del problema en concreto, y que a rasgos generales los dos modelos tienen un rendimiento similar en lo que a métricas se refiere. Esto es habitual en los modelos de predicción de tormentas solares, ya que las características de las series temporales y la naturaleza de los datos pueden variar considerablemente en diferentes horizontes de predicción. Además, la complejidad y capacidad de captura de patrones temporales de los modelos LSTM y SVR se complementan entre sí, por lo que es difícil discriminar qué modelo es mejor de forma categórica.

Métrica	RMSE	MSE	MAE	R ²	MedAE	Var. exp.	Max error
Valor	17.76	315.51	9.60	0.8026	5.98	0.80	328.83

Tabla 4.16: Métricas de la predicción a 4h para el conjunto completo de tormentas. (LSTM)

Métrica	RMSE	MSE	MAE	R ²	MedAE	Var. exp.	Max error
Valor	22.30	497.43	12.20	0.6899	7.28	0.69	327.97

Tabla 4.17: Métricas de la predicción a 6h para el conjunto completo de tormentas. (LSTM)

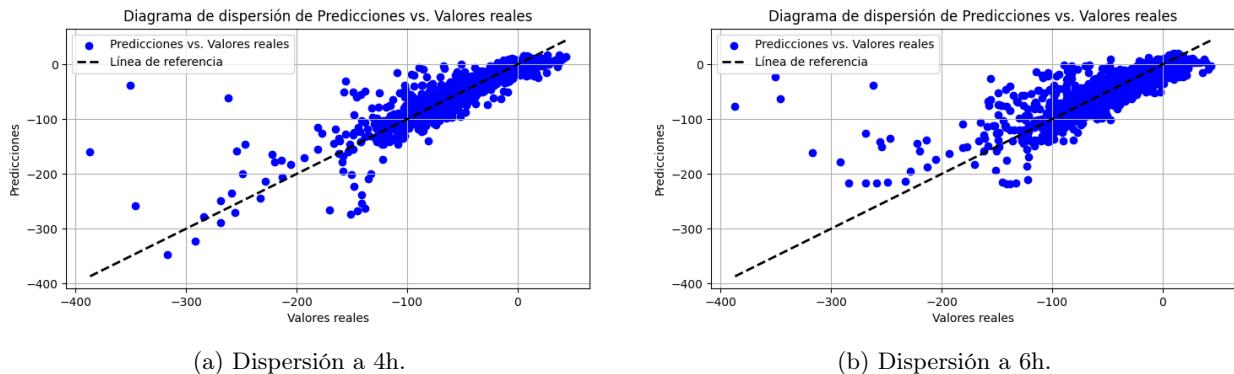


Figura 4.16: Diagramas de dispersión de predicciones contra valores reales en diferentes horizontes de predicción. (LSTM-MLP)

A continuación, se muestra la comparación gráfica de los resultados de SVR y LSTM-MLP a 4 y 6h para la tormenta n° 2.

Primero, se muestran los resultados a 4h en la Tabla 4.18 y la Figura 4.17.

Métrica	LSTM	SVR
RMSE	21.635	22.348
MSE	468.073	499.436
MAE	10.206	9.546
R ²	0.922	0.917
MedAE	5.410	5.106
Var. expl.	0.924	0.917
Max error	153.092	193.334

Tabla 4.18: Comparación de métricas entre LSTM y SVR a 4h.

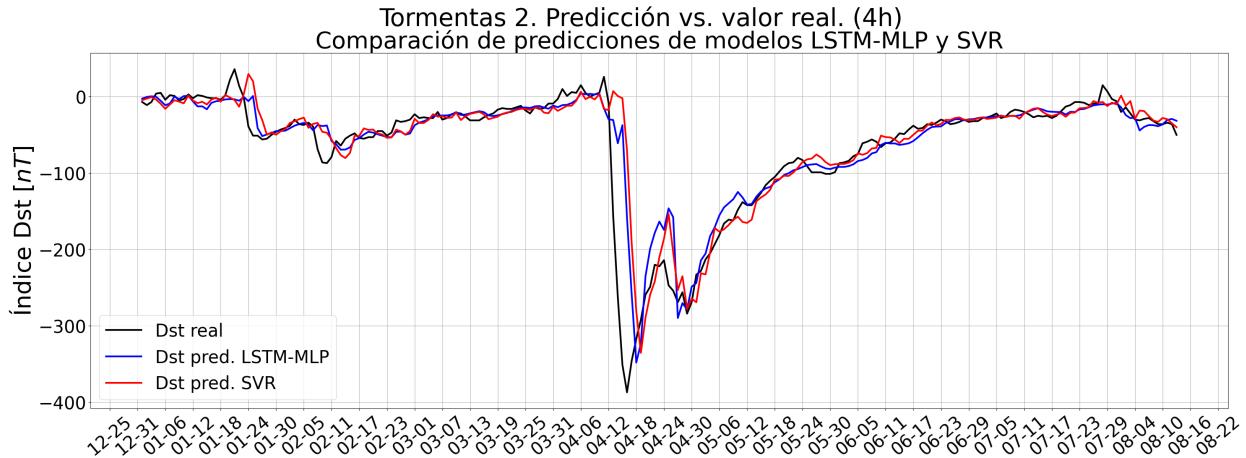


Figura 4.17: Comparación entre valores reales de Dst y valores predichos por SVR y LSTM a 4h en la tormenta 2.

A continuación, se muestran los resultados a 6h en la Tabla 4.19 y la Figura 4.18.

Métrica	LSTM	SVR
RMSE	21.635	22.348
MSE	468.073	499.436
MAE	10.206	9.546
R ²	0.922	0.917
MedAE	5.410	5.106
Var. expl.	0.924	0.917
Max error	153.092	193.334

Tabla 4.19: Comparación de métricas entre LSTM y SVR a 6h

Se puede observar que, en estos casos, la LSTM mejora notablemente a la SVR. No obstante, esto se debe a la selección de la tormenta. Por exemplificar este razonamiento, se muestra un caso favorable para la SVR: en la Tabla 4.20 y la Figura 4.19 se muestra la comparación de rendimiento para la tormenta nº 12 de la Tabla 4.4.

En este caso, todas las métricas del modelo SVR superan al rendimiento del modelo LSTM-MLP. De nuevo, estos resultados muestran que el rendimiento del modelo depende del problema específico, por lo que no se puede decir que un modelo sea, de forma general, mejor que otro para la detección de tormentas geomagnéticas. No obstante, se puede argumentar que el SVR es un modelo más sencillo y, a priori y si se

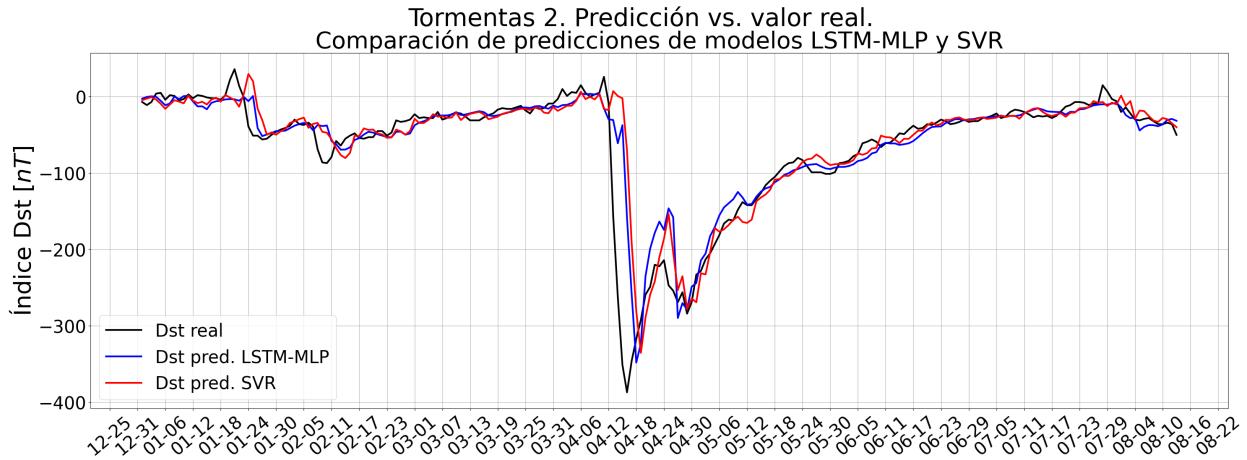


Figura 4.18: Comparación entre valores reales de **Dst** y valores predichos por **SVR** y **LSTM** a 6h en la tormenta 2.

Métrica	LSTM	SVR
RMSE	21.364	16.104
MSE	456.427	259.335
MAE	11.375	9.943
R ²	0.731	0.847
MedAE	5.742	6.283
Var. expl.	0.742	0.853
Max error	125.381	84.868

Tabla 4.20: Comparación de métricas entre LSTM y SVR

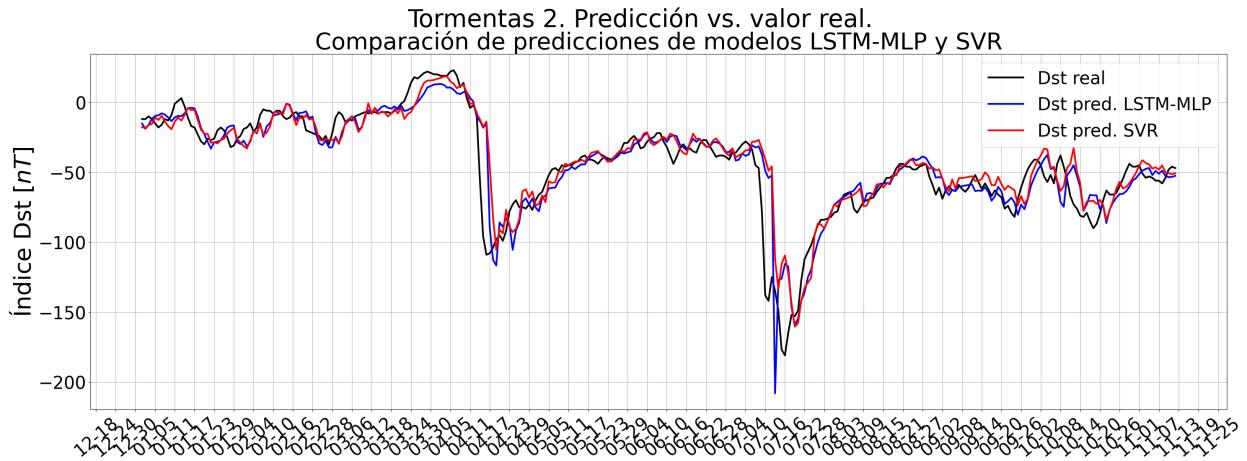


Figura 4.19: Comparación entre valores reales de **Dst** y valores predichos por **SVR** y **LSTM** a 6h en la tormenta 12.

realiza bien el ajuste de hiperparámetros, menos costoso computacionalmente. Otra de las posibles razones por las que una **SVR** podría ser una opción a considerar es la explicabilidad o interpretabilidad. No obstante, tiene ciertas desventajas que se desarrollan en las conclusiones del estudio (Capítulo 5).

Capítulo 5

Conclusión y futuros avances

El presente trabajo se ha centrado en la predicción de tormentas geomagnéticas utilizando **SVR** y comparando su rendimiento con un modelo alternativo basado en redes neuronales (**LSTM-MLP**). Los resultados obtenidos permiten extraer las siguientes conclusiones:

En cuanto al rendimiento, el modelo **SVR** demuestra ser eficaz en la predicción de tormentas geomagnéticas, proporcionando resultados precisos y relativamente consistentes en los diferentes horizontes temporales evaluados. Las métricas de evaluación **MSE**, **MAE** y **R²**, muestran valores favorables que destacan la capacidad predictiva del modelo, aunque su validez tiene limitaciones debidas a la naturaleza extrema de las tormentas. El modelo **SVR** logra capturar las variaciones en el índice **Dst** de manera efectiva y en línea con los resultados actuales.

Se puede observar en los resultados (Sección 4.6.1) que todos los *kernels* óptimos son lineales. En este caso, cabría preguntarse si la Regresión Lineal sería una mejor opción, dada su sencillez y poco requerimiento computacional. No obstante, la **SVR** con kernel lineal y la regresión lineal tienen diferencias clave: la **SVR** es más robusta frente a puntos atípicos debido a su enfoque en maximizar el margen utilizando solo los vectores de soporte, mientras que la regresión lineal minimiza el error cuadrático medio, haciéndola más susceptible a outliers. Además, la **SVR** incorpora regularización explícita mediante el parámetro C , lo que ayuda a evitar el sobreajuste y proporciona un control directo sobre la complejidad del modelo. Esto hace que la **SVR** con kernel lineal sea una opción atractiva cuando se necesitan predicciones más robustas y generalizables en presencia de datos ruidosos o outliers, por lo que sigue siendo la mejor opción en este caso.

En cuanto a las librerías utilizadas, **optuna** es poco intuitiva y presenta muchos fallos, además de tener una documentación complicada. Se recomienda usar alternativas más robustas y mejor documentadas, como **GridSearchCV** de **scikit-learn**, que es más sencilla y soportada para la búsqueda de hiperparámetros. En este proyecto no se cambió debido al estado avanzado del mismo.

Al comparar los resultados del modelo **SVR** con estudios previos, se observa que no se ha superado significativamente el rendimiento de algunos modelos de la bibliografía, como el de [Gruet et al. \(2018\)](#) y otros. Sin embargo, esta comparación no es completamente consistente debido a diferencias en la selección de tormentas y períodos de datos en calma, estos últimos descartados en el presente trabajo, lo que probablemente ha influido en los resultados obtenidos.

La comparación entre el modelo **SVR** y el modelo **LSTM-MLP** muestra que **SVR** ofrece un rendimiento similar en precisión y estabilidad. Aunque las redes neuronales como **LSTM-MLP** capturan patrones complejos en datos secuenciales, su naturaleza de *caja negra* dificulta su interpretabilidad. Por esta razón, una de las principales ventajas del modelo **SVR** es su capacidad de interpretabilidad debido a su naturaleza transparente. El **SVR** permite una mayor comprensión de los mecanismos subyacentes en la predicción de tormentas geomagnéticas. Esto es crucial para la validación científica y la implementación práctica, lo que

podría ser un aspecto diferencial de este modelo frente a otros para estudios futuros o ampliaciones del presente trabajo. La interpretabilidad facilita la identificación de las variables más influyentes en la predicción del índice **Dst**, lo que puede ayudar a comprender mejor el fenómeno y sus causas. Ejemplos de ello son Üstün et al. (2007); Rosenbaum et al. (2011).

Por otra parte, la precisión y la fiabilidad de los modelos predictivos dependen en gran medida de la calidad y la cantidad de datos disponibles. En este estudio, se ha utilizado un conjunto de datos históricos de las sondas **ACE** y **DSCOVR**, lo que ha permitido capturar una amplia gama de condiciones geomagnéticas. Como se explica en la Sección 1, los datos de satélite son limitados y difícilmente ampliables en el corto plazo. En este caso es importante considerar que la integración de datos adicionales, como observaciones en tiempo real y modelos físicos, podría mejorar aún más la capacidad predictiva de los modelos **SVR**, pero la implementación de estos datos podría ser difícil.

A pesar de los resultados prometedores, es importante señalar que el actual estado del arte se encuentra en un estado avanzado, y que hay modelos muy potentes que son capaces de predecir tormentas geomagnéticas de forma muy precisa, como los modelos desarrollados por **SeNMEs**, el Servicio Español de Clima Espacial (o, en inglés, *Spanish Space Weather Service*). Esto podría significar que la capacidad de predicción las máquinas de soporte vectorial de regresión (**SVR**) puede no ser suficiente para los continuos avances en la modelización del clima espacial. Específicamente, aunque las **SVR** han demostrado ser eficaces en diversas aplicaciones, su capacidad para procesar grandes volúmenes de datos en tiempo real y su inhabilidad de incorporar dinámicamente nuevas observaciones y ajustar sus predicciones en consecuencia, puede quedar rezagada en comparación con métodos más avanzados como las redes neuronales profundas y los modelos híbridos descritos en la Sección 1.5.2.

Finalmente, cabría destacar la vital importancia de continuar investigando teóricamente los mecanismos físicos subyacentes de las tormentas geomagnéticas para mejorar la interpretación y la validación de los modelos predictivos. Para este propósito, el modelo basado en **SVR** resulta más prometedor. En las tres ecuaciones dadas, las variables involucradas son las mismas: **E_field**, **Bz_gse** y **Dst**. Los signos y magnitudes de sus coeficientes son consistentes en todas las ecuaciones, con pequeñas variaciones debidas a los diferentes tiempos de pronóstico (2h, 4h y 6h). Esto sugiere que los impactos de estas variables sobre **Dst** son similares en diferentes horizontes temporales, aunque la influencia exacta de cada variable puede cambiar ligeramente con el tiempo. El análisis de la consistencia en las variables y sus coeficientes puede ser muy útil para la mejora de los modelos físicos. Primero, la estabilidad en los signos y magnitudes de los coeficientes confirma las relaciones subyacentes entre las variables. En posibles versiones mejoradas de este modelo, la comprensión de las variables predictoras y cómo afectan a la predicción podrían ser valores claves para el desarrollo del modelo o modelos físicos que expliquen el fenómeno.

En resumen, el uso de Máquinas de Soporte Vectorial para Regresión (**SVR**) ha demostrado ser una metodología efectiva para la predicción de tormentas geomagnéticas, ofreciendo una combinación de precisión e interpretabilidad que supera a los modelos basados en redes neuronales en ciertos aspectos y es limitado y superado en otros aspectos. La continuación de esta línea de investigación y la integración de otros datos o metodologías mixtas queda en un estado de incertidumbre, ya que los resultados obtenidos no son del todo concluyentes. Aunque el **SVR** ofrece ciertos beneficios teóricos y prácticos, no queda del todo claro que sea la herramienta más adecuada para la predicción de tormentas geomagnéticas. Basándonos en el análisis comparativo realizado, nos decantamos hacia la conclusión de que, en términos generales, el **SVR** no es la solución más eficiente y efectiva para esta problemática específica. No obstante, muestra resultados prometedores en el área de la interpretabilidad, lo que podría ayudar a la comprensión y modelización del fenómeno físico subyacente: el acoplamiento viento solar - magnetosfera.

Bibliografía

- Alfvén, H., 1942. Existence of electromagnetic-hydrodynamic waves. *Nature* 150, 405–406. doi:[10.1038/150405d0](https://doi.org/10.1038/150405d0).
- Baker, D.N., 2000. The occurrence of operational anomalies in spacecraft and their relationship to space weather. *IEEE Transactions on Plasma Science* 28, 2007–2016. doi:[10.1109/27.902228](https://doi.org/10.1109/27.902228).
- Baker, D.N., Balstad, R., Bodeau, J.M., Cameron, E., Fennell, J.F., Fisher, G.M., Forbes, K.F., Kintner, P.M., Leffler, L., Lewis, W.S., Reagan, J., Rusch, C., Savani, N.P., Schafer, J., Siscoe, G., Space, J., Green, J.L., Huba, J.D., Murtagh, W., Odenwald, S., Silvious, M., 2018. Space weather impacts in the united states: Major findings and key recommendations. *Space Weather* 16, e2017SW001805. doi:[10.1002/2017SW001805](https://doi.org/10.1002/2017SW001805).
- Bala, R., Reiff, P., 2012a. Improvements in short-term forecasting of geomagnetic activity. *Space Weather* 10. URL: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012SW000779>, doi:<https://doi.org/10.1029/2012SW000779>, arXiv:<https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012SW000779>.
- Bala, R., Reiff, P.H., 2012b. Improving predictions of geomagnetic activity by correcting the real-time solar wind with solar disk observations. *Space Weather* 10, S06001. doi:[10.1029/2012SW000794](https://doi.org/10.1029/2012SW000794).
- Birn, J., Drake, J.F., Shay, M.A., Rogers, B.N., Denton, R.E., Hesse, M., Kuznetsova, M., Ma, Z.W., Bhattacharjee, A., Otto, A., Pritchett, P.L., 2001. Geospace environmental modeling (gem) magnetic reconnection challenge. *Journal of Geophysical Research* 106, 3715–3720. doi:[10.1029/1999JA900449](https://doi.org/10.1029/1999JA900449).
- Borovsky, J., Steinberg, J., 2006. The calm before the storm in cir/magnetosphere interactions. *Journal of Geophysical research* doi:[10.1029/2005JA011397](https://doi.org/10.1029/2005JA011397).
- Borovsky, J.E., 2021. Is Our Understanding of Solar-Wind/ Magnetosphere Coupling Satisfactory? *Frontiers in Astronomy and Space Sciences* 8. doi:[10.3389/fspas.2021.634073](https://doi.org/10.3389/fspas.2021.634073).
- Boteler, D.H., 2006. The super storms of august/september 1859 and their effects on the telegraph system. *Advances in Space Research* 38, 159–172. doi:[10.1016/j.asr.2006.01.013](https://doi.org/10.1016/j.asr.2006.01.013).
- Boteler, D.H., Pirjola, R.J., Nevanlinna, H., 1998. The effects of geomagnetic disturbances on electrical systems at the earth's surface. *Advances in Space Research* 22, 17–27. doi:[10.1016/S0273-1177\(97\)01096-X](https://doi.org/10.1016/S0273-1177(97)01096-X).
- Burt, R., Smith, C.W., Skoug, R.M., Steinberg, J.T., Smith, S., 2015. Noaa's dscovr mission: Monitoring solar wind for space weather. *Space Weather* 13, 768–769. doi:[10.1002/2015SW001299](https://doi.org/10.1002/2015SW001299).
- Burton, R.K., McPherron, R.L., Russell, C.T., 1975. An empirical relationship between interplanetary conditions and dst. *Journal of Geophysical Research* 80, 4204–4214. doi:[10.1029/JA080i031p04204](https://doi.org/10.1029/JA080i031p04204).
- Carrington, R.C., 1859. Description of a singular appearance seen in the sun on september 1, 1859. *Monthly Notices of the Royal Astronomical Society* 20, 13–15. doi:[10.1093/mnras/20.1.13](https://doi.org/10.1093/mnras/20.1.13).
- CDAweb, NASA., . Coordinated Data Analysis Web (CDAWeb). <https://cdaweb.gsfc.nasa.gov/index.html>.
- Chandorkar, M., Camporeale, E., Wing, S., 2017. Probabilistic forecasting of the disturbance storm time index: An autoregressive Gaussian process approach. *Space Weather* doi:[10.1002/2017SW001627](https://doi.org/10.1002/2017SW001627).
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach Learning* 20, 273–297. doi:[10.1007/BF00994018](https://doi.org/10.1007/BF00994018).
- Datta-Barua, S., Su, Y., Deshpande, K., Bust, G., Hampton, D., Crowley, G., 2015. First light from a kilometer-baseline Scintillation Auroral GPS Array.pdf. *Geophysical Research Letters* 42, 3639–3646. doi:[10.1002/2015GL063556](https://doi.org/10.1002/2015GL063556).
- Dessler, A., Parker, E., 1959. Hydromagnetic theory of geomagnetic storms. *Journal of Geophysical Research* 64, 2239–2252. URL: https://consensus.app/papers/hydromagnetic-theory-storms-dessler/bbc637c9e99451dcaaa907e48815ed2a/?utm_source=chatgpt, doi:[10.1029/JZ064I012P02239](https://doi.org/10.1029/JZ064I012P02239).

- Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1996. Support vector regression machines, in: Mozer, M., Jordan, M., Petsche, T. (Eds.), Advances in Neural Information Processing Systems, MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.
- Fennell, J.F., Roeder, J.L., Blake, J.B., Spence, H.E., 2001. Comparison of crres inner zone electron data with nasa ae-8 radiation belt model. IEEE Transactions on Nuclear Science 48, 2028–2037. doi:[10.1109/23.983166](https://doi.org/10.1109/23.983166).
- Gonzalez, W.D., Joselyn, J.A., Kamide, Y., Kroehl, H.W., Rostoker, G., Tsurutani, B.T., Vasyliunas, V.M., 1994. What is a geomagnetic storm? Journal of Geophysical Research 99, 5771–5792. doi:[10.1029/93JA02867](https://doi.org/10.1029/93JA02867).
- Gruet, M., Chandorkar, M., Sicard, A., Camporeale, E., 2018. Multiple hours ahead forecast of the Dst index using a combination of Long Short-Term Memory neural network and Gaussian Process. Space Weather 16, 1882–1896. doi:[10.1029/2018SW001898](https://doi.org/10.1029/2018SW001898).
- Jagadeesh, P., Kumar, S., Seemakurthi, P., Kant, K., 2020. Prediction of geomagnetic storms using deep learning algorithms: a comparison study. Annales Geophysicae 38, 881–893. doi:[10.5194/angeo-38-881-2020](https://doi.org/10.5194/angeo-38-881-2020).
- Kappenman, J.G., 2005. An overview of the impulsive geomagnetic field disturbances and power grid impacts associated with the violent Sun-Earth connection events of 29–31 October 2003 and a comparative evaluation with other contemporary storms. Space Weather 3. doi:[10.1029/2004SW000128](https://doi.org/10.1029/2004SW000128).
- Kasran, F.A.M., Jusoh, M.H., Rahim, S.A.E.A., Abdullah, N., 2018. Geomagnetically Induced Currents (GICs) in Equatorial Region, in: 2018 IEEE 8th International Conference on System Engineering and Technology (ICSET), IEEE, Bandung. pp. 112–117. doi:[10.1109/ICSEngT.2018.8606391](https://doi.org/10.1109/ICSEngT.2018.8606391).
- Kintner, P.M., Humphreys, T., Hinks, J., 2007. Gnss and ionospheric scintillation: How to survive the next solar maximum. Inside GNSS 2, 22–30.
- Kumar, S., Raizada, A., 2009. Effect of solar features and interplanetary parameters on geomagnetosphere during solar cycle-23. Pramana 71, 1353–1366. doi:[10.1007/s12043-008-0189-7](https://doi.org/10.1007/s12043-008-0189-7).
- Lakhina, G.S., Alex, S., Tsurutani, B.T., Gonzalez, W.D., 2005. Research on Historical Records of Geomagnetic Storms. Proceedings of the International Astronomical Union 2004, 3–15. doi:[10.1017/S1743921305000074](https://doi.org/10.1017/S1743921305000074).
- Lazzús, J.A., Vega, P., Rojas, P., Saltate, I., 2017. Forecasting the dst index using a swarm-optimized neural network. Space Weather 15, 1068–1089. URL: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017SW001608>, doi:<https://doi.org/10.1002/2017SW001608>, arXiv:<https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017SW001608>.
- Morris, J.D., Lieberman, M.G., 2018. Multicollinearity's effect on regression prediction accuracy with real data structures. General Linear Model Journal 44, 29–34. doi:[10.31523/GLMJ.044001.004](https://doi.org/10.31523/GLMJ.044001.004).
- NASA, US., 2024. GSE vs. GSM - NASA - SPDF - Satellite Situation Center Web (SSCWeb). Satellite Situation Center Web .
- Optuna, 2024. Optuna: A hyperparameter optimization framework. <https://optuna.org/>. [Accedido: 31-jul-2024].
- Rosenbaum, L., Hinselmann, G., Jahn, A., Zell, A., 2011. Interpreting linear support vector machine models with heat map molecule coloring. Journal of Cheminformatics 3, 11 – 11. doi:[10.1186/1758-2946-3-11](https://doi.org/10.1186/1758-2946-3-11).
- Schrijver, C.J., Kauristie, K., Aylward, A.D., Denardini, C.M., Gibson, S.E., Glover, A., Gopalswamy, N., Grande, M., Hapgood, M., Heynderickx, D., Jakowski, N., Kalegaev, V.V., Lapenta, G., Linker, J.A., Liu, S., Mandrini, C.H., Mann, I.R., Nagatsuma, T., Nandy, D., Obara, T., Paul O'Brien, T., Onsager, T., Opgenoorth, H.J., Terkildsen, M., Valladares, C.E., Vilmer, N., 2015. Understanding space weather to shield society: A global road map for 2015–2025 commissioned by cospar and ilws. Advances in Space Research 55, 2745–2807. doi:[10.1016/j.asr.2015.03.023](https://doi.org/10.1016/j.asr.2015.03.023).
- Skone, S., Knudsen, K., De Jong, M., 2001. Limitations in gps receiver tracking performance under ionospheric scintillation conditions. Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy 26, 613–621. doi:[10.1016/S1464-1895\(01\)00054-5](https://doi.org/10.1016/S1464-1895(01)00054-5).

- Stepanova, M., Antonova, E., Munos-Uribe, F.A., Gordo, S.L.G., Torres-Sanchez, M.V., 2008. Prediction of geomagnetic storm using neural networks: Comparison of the efficiency of the satellite and ground-based input parameters. *Journal of Physics: Conference Series* 134, 012041. doi:[10.1088/1742-6596/134/1/012041](https://doi.org/10.1088/1742-6596/134/1/012041).
- Stone, E.C., Frandsen, A.M.A., Mewaldt, R.A., Christian, E.R., Margolies, D., Ormes, J.F., Snow, F., 1998. The advanced composition explorer. *Space Science Reviews* 86, 1–22. doi:[10.1023/A:1005082526237](https://doi.org/10.1023/A:1005082526237).
- Sugiura, M., 1963. Hourly values of equatorial dst for the igy: Hourly values for magnetic storm variation for International Geophysical Year. Technical Memorandum (TM) NASA-TM-X-55238, X-611-63-131. NASA Goddard Space Flight Center. URL: <https://ntrs.nasa.gov/api/citations/19650020355/downloads/19650020355.pdf>. accession Number: 65N29956, Distribution Limits: Public.
- Valdivia, J.A., Saavedra, F., Rogan, J., Munoz, V., Toledo, B., 2013. Improving geomagnetic storm predictions by combining magnetohydrodynamic simulations and neural networks. *Space Weather* 11, 71–84. doi:[10.1002/swe.20027](https://doi.org/10.1002/swe.20027).
- Van Allen, J.A., Frank, L.A., 1959. Radiation around the earth to a radial distance of 107,400 km. *Nature* 183, 430–434. doi:[10.1038/183430a0](https://doi.org/10.1038/183430a0).
- Vapnik, V., Chervonenkis, A.Ya., 1974. Teoriya raspoznavaniya obrazov [Theory of Pattern Recognition]. Nauka (‘), Moscú, Rusia.
- Wu, J.G., Lundstedt, H., 1997. Neural network modeling of solar wind-magnetosphere interaction. *Journal of Geophysical Research: Space Physics* 102, 14457–14466. URL: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97JA01081>, doi:<https://doi.org/10.1029/97JA01081>, arXiv:<https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/97JA01081>.
- Yáñez Gestoso, F.J., 2022. Apuntes de técnicas de optimización del máster teci 2022/2023 de la ucm y la upm. Documento interno.
- Üstün, B., Melssen, W., Buydens, L., 2007. Visualisation and interpretation of support vector regression models. *Analytica Chimica Acta* 595, 299–309. URL: <https://www.sciencedirect.com/science/article/pii/S0003267007004904>, doi:<https://doi.org/10.1016/j.aca.2007.03.023>. papers presented at the 10th International Conference on Chemometrics in Analytical Chemistry.

Apéndice A

Definiciones: Divergencia y Rotacional

En física, la divergencia y el rotacional son operadores diferenciales que describen ciertos aspectos de campos vectoriales.

La **divergencia** es un operador diferencial que se aplica a un campo vectorial en el espacio tridimensional. Dado un campo vectorial \mathbf{A} , la divergencia de \mathbf{A} , denotada como $\text{div } \mathbf{A}$ o $\nabla \cdot \mathbf{A}$, mide la magnitud de una fuente o sumidero en un punto dado. Matemáticamente, se define como:

$$\text{div } \mathbf{A} = \nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}$$

donde \mathbf{A} es un campo vectorial.

Usualmente en física la divergencia se expresa con la notación $\nabla \cdot$, pero debido a la posible confusión que podría suponer con un producto escalar, se utilizará la notación div .

El **rotacional** (o **curl** en inglés) es un operador diferencial que se aplica a un campo vectorial en el espacio tridimensional. Dado un campo vectorial \mathbf{A} , el rotacional de \mathbf{A} , denotado como $\text{rot } \mathbf{A}$ o $\nabla \times \mathbf{A}$, mide la tendencia de los vectores del campo a girar alrededor de un punto. Matemáticamente, se define como:

$$\text{rot } \mathbf{A} = \nabla \times \mathbf{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right)$$

Usualmente en física el rotacional se expresa con la notación $\nabla \times$, pero debido a la posible confusión que podría suponer con un producto vectorial, se utilizará la notación rot .

Apéndice B

Repositorio del proyecto

El repositorio del proyecto se encuentra disponible en el siguiente enlace: [GitHub - raqgmar /tsa4dst](https://github.com/raqgmar/tsa4dst).

```
00_download_data.ipynb
01_01_EDA_storms_data_source.ipynb
02_EDA_datos_procesados.ipynb
03_01_vDef_SVR_2h_ahead.ipynb
03_02_vDEF_SVR_4h_ahead.ipynb
03_03.1_vDef_SVR_6h_ahead_TRAIN.ipynb
03_03.2_vDef_SVR_6h_ahead_RESULTS.ipynb
04_01_LSTM-MLP_2h.ipynb
04_02_LSTM-MLP_4h.ipynb
04_03_LSTM-MLP_6h.ipynb
05_analisis_resultados_y_graficas_completar.ipynb
06_01_INTERPRETABILIDAD_SVR.ipynb
Anexos (*)
RESULTADOS
    SVR_*h_res_by_storm (*)
    NN_*h_res_by_storm (*)
TFM_data
    all_storms.csv
    historical_storms_gruet2018.csv
    OMNI2_HO_MRGIHR.csv
PRED_2h
    DEF_kaisa_kernel_LINEAR_pred2H_20240720.db
    DEF_kaisa_kernel_POLY_pred2H_20240720.db
    DEF_kaisa_kernel_RBF_pred2H_20240720.db
PRED_4h (*)
PRED_6h (*)
```

Apéndice C

Entrenamientos realizados en el modelo SVR

En `start` y `end` se ha eliminado la fecha `2024-07-20` ya que es la misma en todos los casos. También se ha eliminado `0 days` de `duration` ya que siempre el entrenamiento dura menos de un día.

number	- MSE	start	stop	duration	C	epsilon	kernel	state
1	0.0911	15:03:59.800140	15:04:03.531936	00:00:03.731796	0.0268	0.1	linear	COMPLETE
2	0.0926	15:04:03.667575	15:04:10.055316	00:00:06.387741	0.0268	0.0001	linear	COMPLETE
3	nan	15:04:10.094557	15:06:10.214667	00:02:00.120110	3.7276	0.01	linear	FAIL
4	0.1135	15:06:10.245238	15:06:12.680456	00:00:02.435218	0.001	0.0001	linear	COMPLETE
5	0.152	15:06:12.737667	15:06:13.746160	00:00:01.008493	0.7197	1.0	linear	COMPLETE
6	0.0911	15:06:13.805262	15:07:34.849784	00:01:21.044522	0.7197	0.01	linear	COMPLETE
7	0.1346	15:07:34.893620	15:07:35.194705	00:00:00.301085	0.1389	1.0	linear	COMPLETE
8	0.0915	15:07:35.238574	15:09:05.439751	00:01:30.201177	0.7197	0.0001	linear	COMPLETE
9	0.0917	15:09:05.499238	15:09:26.754600	00:00:21.255362	0.1389	0.0001	linear	COMPLETE
10	nan	15:09:26.815590	15:11:27.079791	00:02:00.264201	3.7276	0.0001	linear	FAIL
11	15.382	15:11:27.119896	15:11:27.249314	00:00:00.129418	0.1389	10.0	linear	COMPLETE
12	15.382	15:11:27.295410	15:11:27.435232	00:00:00.139822	0.001	10.0	linear	COMPLETE
13	0.0962	15:11:27.475779	15:11:31.395699	00:00:03.919920	0.0052	0.0001	linear	COMPLETE
14	nan	15:11:31.599343	15:13:31.855038	00:02:00.255695	3.7276	0.1	linear	FAIL
15	0.0894	15:13:31.899096	15:14:14.925133	00:00:43.026037	0.7197	0.1	linear	COMPLETE
16	nan	15:14:14.979742	15:16:15.348523	00:02:00.368781	19.307	0.001	linear	FAIL
17	0.0957	15:16:15.423809	15:16:17.320363	00:00:01.896554	0.0052	0.1	linear	COMPLETE
18	0.0896	15:16:17.386904	15:16:27.975146	00:00:10.588242	0.1389	0.1	linear	COMPLETE
19	nan	15:16:28.055276	15:18:28.246874	00:02:00.191598	19.307	0.0001	linear	FAIL
20	nan	15:18:28.295204	15:20:28.475462	00:02:00.180258	100.0	0.0001	linear	FAIL

21	0.0921	15:20:28.573262	15:20:35.404702	00:00:06.831440	0.0268	0.01	linear	COMPLETE
22	nan	15:20:35.462309	15:22:35.668712	00:02:00.206403	19.307	0.01	linear	FAIL
23	15.382	15:22:35.715787	15:22:35.984462	00:00:00.268675	0.7197	10.0	linear	COMPLETE
24	15.382	15:22:36.065113	15:22:36.274443	00:00:00.209330	0.0052	10.0	linear	COMPLETE
25	0.0926	15:22:36.341258	15:22:43.756626	00:00:07.415368	0.0268	0.001	linear	COMPLETE
26	0.16	15:22:43.843796	15:22:47.341030	00:00:03.497234	3.7276	1.0	linear	COMPLETE
27	0.1475	15:22:47.410811	15:22:47.821346	00:00:00.410535	0.0052	1.0	linear	COMPLETE
28	0.1135	15:22:47.890725	15:22:51.115666	00:00:03.224941	0.001	0.01	linear	COMPLETE
29	nan	15:22:51.193394	15:24:51.432405	00:02:00.239011	19.307	0.1	linear	FAIL
30	0.1653	15:24:51.483024	15:26:28.978899	00:01:37.495875	100.0	1.0	linear	COMPLETE
31	15.382	15:26:29.040284	15:26:29.256757	00:00:00.216473	100.0	10.0	linear	COMPLETE
32	nan	15:26:29.318597	15:28:29.512524	00:02:00.193927	3.7276	0.001	linear	FAIL
33	0.1135	15:28:29.559550	15:28:32.755569	00:00:03.196019	0.001	0.001	linear	COMPLETE
34	0.1127	15:28:32.795227	15:28:34.565038	00:00:01.769811	0.001	0.1	linear	COMPLETE
35	nan	15:28:34.615548	15:30:34.770300	00:02:00.154752	100.0	0.01	linear	FAIL
36	0.1248	15:30:34.818215	15:30:35.128663	00:00:00.310448	0.0268	1.0	linear	COMPLETE
37	0.1643	15:30:35.189855	15:30:52.040684	00:00:16.850829	19.307	1.0	linear	COMPLETE
38	0.0962	15:30:52.106449	15:30:57.166720	00:00:05.060271	0.0052	0.001	linear	COMPLETE
39	nan	15:30:57.228642	15:32:57.397862	00:02:00.169220	100.0	0.1	linear	FAIL
40	0.0964	15:32:57.460547	15:33:02.768590	00:00:05.308043	0.0052	0.01	linear	COMPLETE
41	0.0917	15:33:02.851903	15:33:32.194213	00:00:29.342310	0.1389	0.001	linear	COMPLETE
42	15.382	15:33:32.271207	15:33:32.485617	00:00:00.214410	3.7276	10.0	linear	COMPLETE
43	0.2344	15:33:32.559153	15:33:32.888519	00:00:00.329366	0.001	1.0	linear	COMPLETE
44	0.0916	15:33:32.952615	15:35:25.043703	00:01:52.091088	0.7197	0.001	linear	COMPLETE
45	nan	15:35:25.115035	15:37:25.303836	00:02:00.188801	100.0	0.001	linear	FAIL
46	15.382	15:37:25.349529	15:37:25.569301	00:00:00.219772	0.0268	10.0	linear	COMPLETE
47	0.0912	15:37:25.638532	15:37:53.130473	00:00:27.491941	0.1389	0.01	linear	COMPLETE

Tabla C.1: Grid de optuna: SVR, predicción a 2h. Kernel LINEAR.

number	- MSE	start	stop	duration	C	epsilon	kernel	state
1	15.382	15:53:27.662836	15:53:27.854636	00:00:00.191800	63.0957	10.0	rbf	COMPLETE
2	0.9051	15:53:27.898524	15:53:32.014654	00:00:04.116130	0.001	0.0001	rbf	COMPLETE
3	15.382	15:53:32.287992	15:53:32.440584	00:00:00.152592	3.9811	10.0	rbf	COMPLETE
4	0.2933	15:53:32.481839	15:53:34.269733	00:00:01.787894	3.9811	0.2154	rbf	COMPLETE
5	15.382	15:53:34.356176	15:53:34.506967	00:00:00.150791	0.0158	10.0	rbf	COMPLETE

<

6	15.382	15:53:34.549065	15:53:34.704557	00:00:00.155492	0.0158	10.0	rbf	COMPLETE
7	0.378	15:53:34.744252	15:53:38.433700	00:00:03.689448	0.2512	0.0046	rbf	COMPLETE
8	0.3779	15:53:38.490032	15:53:39.384444	00:00:00.894412	0.2512	0.2154	rbf	COMPLETE
9	0.9179	15:53:39.437662	15:53:41.509995	00:00:02.072333	0.001	0.2154	rbf	COMPLETE
10	0.3747	15:53:41.548919	15:53:45.376152	00:00:03.827233	0.2512	0.0046	rbf	COMPLETE
11	15.382	15:53:45.470878	15:53:45.617902	00:00:00.147024	0.2512	10.0	rbf	COMPLETE
12	0.3471	15:53:45.662442	15:54:25.855254	00:00:40.192812	63.0957	0.0001	rbf	COMPLETE
13	0.2894	15:54:25.908820	15:54:27.439567	00:00:01.530747	63.0957	0.2154	rbf	COMPLETE
14	0.4342	15:54:27.496848	15:55:54.240557	00:01:26.743709	1000.0	0.0001	rbf	COMPLETE
15	0.5435	15:55:54.444911	15:55:57.498228	00:00:03.053317	0.0158	0.0001	rbf	COMPLETE
16	0.3115	15:55:57.551557	15:55:59.457824	00:00:01.906267	1000.0	0.2154	rbf	COMPLETE
17	0.378	15:55:59.505431	15:56:03.279322	00:00:03.773891	0.2512	0.0001	rbf	COMPLETE
18	0.9073	15:56:03.468254	15:56:06.636331	00:00:03.168077	0.001	0.0001	rbf	COMPLETE
19	0.4281	15:56:06.679378	15:57:12.222667	00:01:05.543289	1000.0	0.0046	rbf	COMPLETE
20	0.2963	15:57:12.299958	15:57:20.625674	00:00:08.325716	3.9811	0.0001	rbf	COMPLETE
21	0.905	15:57:20.683772	15:57:23.671254	00:00:02.987482	0.001	0.0046	rbf	COMPLETE
22	0.3439	15:57:23.734091	15:58:01.912950	00:00:38.178859	63.0957	0.0001	rbf	COMPLETE
23	0.3747	15:58:01.974531	15:58:05.447114	00:00:03.472583	0.2512	0.0001	rbf	COMPLETE
24	15.382	15:58:05.497485	15:58:05.654872	00:00:00.157387	3.9811	10.0	rbf	COMPLETE
25	0.3746	15:58:05.701577	15:58:06.638345	00:00:00.936768	0.2512	0.2154	rbf	COMPLETE
26	0.2926	15:58:06.692371	15:58:08.469929	00:00:01.777558	63.0957	0.2154	rbf	COMPLETE
27	15.382	15:58:08.533898	15:58:08.693333	00:00:00.159435	0.001	10.0	rbf	COMPLETE
28	0.3092	15:58:08.739834	15:58:11.199419	00:00:02.459585	1000.0	0.2154	rbf	COMPLETE
29	0.9079	15:58:11.293494	15:58:14.776898	00:00:03.483404	0.001	0.0046	rbf	COMPLETE
30	0.5433	15:58:14.944927	15:58:17.963440	00:00:03.018513	0.0158	0.0046	rbf	COMPLETE
31	0.4279	15:58:18.007803	15:59:20.430722	00:01:02.422919	1000.0	0.0046	rbf	COMPLETE
32	0.2998	15:59:20.488181	15:59:29.209866	00:00:08.721685	3.9811	0.0001	rbf	COMPLETE
33	0.3451	15:59:29.445453	15:59:59.443069	00:00:29.997616	63.0957	0.0046	rbf	COMPLETE
34	0.5404	15:59:59.504363	16:00:00.786969	00:00:01.282606	0.0158	0.2154	rbf	COMPLETE
35	0.2961	16:00:00.828143	16:00:07.658042	00:00:06.829899	3.9811	0.0046	rbf	COMPLETE
36	15.382	16:00:07.711668	16:00:07.862734	00:00:00.151066	0.001	10.0	rbf	COMPLETE
37	15.382	16:00:07.908863	16:00:08.063087	00:00:00.154224	63.0957	10.0	rbf	COMPLETE
38	0.5394	16:00:08.107563	16:00:11.434938	00:00:03.327375	0.0158	0.0001	rbf	COMPLETE
39	0.539	16:00:11.485107	16:00:14.468782	00:00:02.983675	0.0158	0.0046	rbf	COMPLETE
40	0.2996	16:00:14.518884	16:00:21.650454	00:00:07.131570	3.9811	0.0046	rbf	COMPLETE
41	0.2897	16:00:21.705718	16:00:22.806011	00:00:01.100293	3.9811	0.2154	rbf	COMPLETE

42	0.4338	16:00:23.396360	16:01:42.026369	00:01:18.630009	1000.0	0.0001	rbf	COMPLETE
43	0.3422	16:01:42.076370	16:02:12.394127	00:00:30.317757	63.0957	0.0046	rbf	COMPLETE
44	15.382	16:02:12.468803	16:02:12.641451	00:00:00.172648	1000.0	10.0	rbf	COMPLETE
45	0.9155	16:02:12.685334	16:02:14.606727	00:00:01.921393	0.001	0.2154	rbf	COMPLETE
46	15.382	16:02:14.649981	16:02:14.819716	00:00:00.169735	0.2512	10.0	rbf	COMPLETE
47	15.382	16:02:14.861963	16:02:15.028522	00:00:00.166559	1000.0	10.0	rbf	COMPLETE

Tabla C.2: Grid de optuna: SVR, predicción a 2h. Kernel RBF.

number	- MSE	start	stop	duration	C	epsilon	kernel	state
1	0.2352	16:03:38.797090	16:03:41.451413	00:00:02.654323	0.001	0.01	poly	COMPLETE
2	15.382	16:03:41.508117	16:03:41.691311	00:00:00.183194	1.0	10.0	poly	COMPLETE
3	15.382	16:03:41.733103	16:03:41.928381	00:00:00.195278	1.0	10.0	poly	COMPLETE
4	176.3987	16:03:41.976892	16:03:43.656572	00:00:01.679680	1.0	0.3162	poly	COMPLETE
5	0.6554	16:03:43.714793	16:03:46.238300	00:00:02.523507	0.001	0.01	poly	COMPLETE
6	0.6183	16:03:46.289517	16:03:47.037387	00:00:00.747870	0.001	0.3162	poly	COMPLETE
7	15.382	16:03:47.098357	16:03:47.294213	00:00:00.195856	1.0	10.0	poly	COMPLETE
8	15.382	16:03:47.340460	16:03:47.525369	00:00:00.184909	0.001	10.0	poly	COMPLETE
9	208.8975	16:03:47.568305	16:03:53.153121	00:00:05.584816	1.0	0.01	poly	COMPLETE
10	nan	16:03:53.209240	16:05:53.449471	00:02:00.240231	1000.0	0.01	poly	FAIL
11	0.699	16:05:53.489266	16:05:56.170453	00:00:02.681187	0.001	0.01	poly	COMPLETE
12	2.4442	16:05:56.239468	16:05:58.422461	00:00:02.182993	1000.0	0.3162	poly	COMPLETE
13	15.382	16:05:58.486417	16:05:58.669589	00:00:00.183172	1000.0	10.0	poly	COMPLETE
14	0.6614	16:05:58.713703	16:06:00.496297	00:00:01.782594	0.001	0.3162	poly	COMPLETE
15	nan	16:06:00.539098	16:08:00.723388	00:02:00.184290	1000.0	0.01	poly	FAIL
16	344.1611	16:08:00.760297	16:08:02.798343	00:00:02.038046	1000.0	0.3162	poly	COMPLETE
17	46.5123	16:08:02.977873	16:10:01.454377	00:01:58.476504	1000.0	0.01	poly	COMPLETE
18	261.6895	16:10:01.513365	16:10:52.436627	00:00:50.923262	1000.0	0.01	poly	COMPLETE
19	15.382	16:10:52.491905	16:10:53.202066	00:00:00.710161	1.0	10.0	poly	COMPLETE
20	nan	16:10:53.296093	16:12:53.563138	00:02:00.267045	1000.0	0.01	poly	FAIL
21	0.6971	16:12:53.609411	16:12:55.438598	00:00:01.829187	0.001	0.3162	poly	COMPLETE
22	15.382	16:12:55.497760	16:12:55.709022	00:00:00.211262	0.001	10.0	poly	COMPLETE
23	0.243	16:12:55.755636	16:12:56.749681	00:00:00.994045	0.001	0.3162	poly	COMPLETE
24	0.3945	16:12:56.795695	16:12:57.751005	00:00:00.955310	1.0	0.3162	poly	COMPLETE
25	14.3218	16:12:57.795111	16:13:21.655102	00:00:23.859991	1.0	0.01	poly	COMPLETE
26	2.5293	16:13:21.708732	16:13:24.684209	00:00:02.975477	1000.0	0.3162	poly	COMPLETE

27	2.9229	16:13:24.729141	16:13:31.458630	00:00:06.729489	1.0	0.01	poly	COMPLETE
28	4.8188	16:13:31.503072	16:13:50.694886	00:00:19.191814	1.0	0.01	poly	COMPLETE
29	15.382	16:13:50.743622	16:13:50.936241	00:00:00.192619	0.001	10.0	poly	COMPLETE
30	0.5542	16:13:50.980784	16:13:53.952899	00:00:02.972115	0.001	0.01	poly	COMPLETE
31	15.382	16:13:53.994896	16:13:54.199135	00:00:00.204239	1.0	10.0	poly	COMPLETE
32	14.6355	16:13:54.242600	16:13:56.249580	00:00:02.006980	1000.0	0.3162	poly	COMPLETE
33	15.382	16:13:56.305309	16:13:56.505030	00:00:00.199721	1000.0	10.0	poly	COMPLETE
34	15.382	16:13:56.548841	16:13:56.740041	00:00:00.191200	0.001	10.0	poly	COMPLETE
35	0.7674	16:13:56.781358	16:13:57.688341	00:00:00.906983	1.0	0.3162	poly	COMPLETE
36	1.5442	16:13:57.734940	16:14:00.380522	00:00:02.645582	0.001	0.01	poly	COMPLETE
37	6.3043	16:14:00.434645	16:14:01.580021	00:00:01.145376	1.0	0.3162	poly	COMPLETE
38	1.4689	16:14:01.627537	16:14:03.532750	00:00:01.905213	0.001	0.3162	poly	COMPLETE
39	15.382	16:14:03.674391	16:14:03.869700	00:00:00.195309	1000.0	10.0	poly	COMPLETE
40	0.2142	16:14:03.917538	16:14:06.441606	00:00:02.524068	0.001	0.01	poly	COMPLETE
41	15.382	16:14:06.498457	16:14:06.776592	00:00:00.278135	1000.0	10.0	poly	COMPLETE
42	15.382	16:14:06.824107	16:14:07.008421	00:00:00.184314	0.001	10.0	poly	COMPLETE
43	0.2946	16:14:07.054548	16:14:07.957129	00:00:00.902581	1.0	0.3162	poly	COMPLETE
44	15.382	16:14:08.015022	16:14:08.206718	00:00:00.191696	1.0	10.0	poly	COMPLETE
45	0.2571	16:14:08.250231	16:14:09.090008	00:00:00.839777	0.001	0.3162	poly	COMPLETE
46	2.6722	16:14:09.141513	16:14:10.325576	00:00:01.184063	1.0	0.3162	poly	COMPLETE
47	939.0376	16:14:10.385372	16:14:44.478217	00:00:34.092845	1000.0	0.01	poly	COMPLETE
48	15.8197	16:14:44.534245	16:14:51.350512	00:00:06.816267	1.0	0.01	poly	COMPLETE
49	0.9499	16:14:51.468169	16:15:06.345479	00:00:14.877310	1.0	0.01	poly	COMPLETE
50	1.0714	16:15:06.407312	16:15:11.402538	00:00:04.995226	1000.0	0.3162	poly	COMPLETE
51	15.382	16:15:11.484184	16:15:11.677595	00:00:00.193411	0.001	10.0	poly	COMPLETE
52	15.382	16:15:11.724932	16:15:11.942766	00:00:00.217834	1000.0	10.0	poly	COMPLETE
53	1.245	16:15:11.987811	16:15:14.496429	00:00:02.508618	1000.0	0.3162	poly	COMPLETE

Tabla C.3: Grid de optuna: SVR, predicción a 2h. Kernel POLY.

number	- MSE	start	stop	duration	C	epsilon	kernel	state
1	0.2782	13:12:23.552645	13:12:36.337164	00:00:12.784519	0.001	0.1	linear	COMPLETE
2	nan	13:12:36.481753	13:14:36.694061	00:02:00.212308	3.7276	0.1	linear	FAIL
3	0.2573	13:14:36.772522	13:15:27.884128	00:00:51.111606	0.1389	0.1	linear	COMPLETE
4	15.3545	13:15:27.952013	13:15:28.168019	00:00:00.216006	0.7197	10.0	linear	COMPLETE
5	15.3545	13:15:28.236074	13:15:28.461712	00:00:00.225638	19.307	10.0	linear	COMPLETE

6	nan	13:15:28.523253	13:17:28.772133	00:02:00.248880	19.307	0.001	linear	FAIL
7	0.2803	13:17:28.866918	13:17:54.844185	00:00:25.977267	0.001	0.01	linear	COMPLETE
8	0.2596	13:17:54.966642	13:18:39.086462	00:00:44.119820	0.0268	0.01	linear	COMPLETE
9	nan	13:18:39.267195	13:20:39.618925	00:02:00.351730	19.307	0.1	linear	FAIL
10	nan	13:20:39.735742	13:22:40.129779	00:02:00.394037	0.7197	0.0001	linear	FAIL
11	15.3545	13:22:40.297844	13:22:41.119041	00:00:00.821197	0.0268	10.0	linear	COMPLETE
12	15.3545	13:22:41.458830	13:22:42.580611	00:00:01.121781	0.0052	10.0	linear	COMPLETE
13	nan	13:22:42.901466	13:24:43.639964	00:02:00.738498	3.7276	1.0	linear	FAIL
14	nan	13:24:43.805634	13:25:37.257860	00:00:53.452226	0.7197	0.001	linear	FAIL
15	nan	13:30:14.411294	13:32:14.777763	00:02:00.366469	100.0	1.0	linear	FAIL
16	nan	13:32:14.867445	13:34:15.173913	00:02:00.306468	19.307	0.1	linear	FAIL
17	nan	13:34:15.325652	13:36:15.785143	00:02:00.459491	100.0	0.1	linear	FAIL
18	nan	13:36:15.917236	13:38:16.407276	00:02:00.490040	19.307	0.0001	linear	FAIL
19	nan	13:38:16.538212	13:40:17.089058	00:02:00.550846	0.1389	0.01	linear	FAIL
20	0.2804	13:40:17.429282	13:41:09.135260	00:00:51.705978	0.001	0.001	linear	COMPLETE
21	0.264	13:41:09.354146	13:41:58.805434	00:00:49.451288	0.0052	0.001	linear	COMPLETE
22	nan	13:41:58.955774	13:43:59.388931	00:02:00.433157	0.1389	0.0001	linear	FAIL
23	0.2596	13:43:59.512851	13:45:27.482826	00:01:27.969975	0.0268	0.001	linear	COMPLETE
24	nan	13:45:27.739614	13:47:28.242846	00:02:00.503232	3.7276	1.0	linear	FAIL
25	nan	13:47:28.468858	13:49:29.468555	00:02:00.999697	0.1389	0.1	linear	FAIL
26	nan	13:49:29.643862	13:50:44.763936	00:01:15.120074	100.0	0.0001	linear	FAIL
27	nan	13:51:22.874954	13:53:23.221728	00:02:00.346774	100.0	0.0001	linear	FAIL
28	nan	13:53:23.313471	13:55:23.615962	00:02:00.302491	100.0	0.001	linear	FAIL
29	15.3545	13:55:23.718590	13:55:24.127113	00:00:00.408523	100.0	10.0	linear	COMPLETE
30	nan	13:55:24.247122	13:57:24.642725	00:02:00.395603	0.1389	0.001	linear	FAIL
31	0.2812	13:57:24.842826	13:57:37.245777	00:00:12.402951	0.7197	1.0	linear	COMPLETE
32	0.2597	13:57:37.376275	13:58:22.601805	00:00:45.225530	0.0268	0.001	linear	COMPLETE
33	15.3545	13:58:22.692187	13:58:23.046621	00:00:00.354434	19.307	10.0	linear	COMPLETE
34	0.2817	13:58:23.172804	14:00:06.467033	00:01:43.294229	3.7276	1.0	linear	COMPLETE
35	nan	14:00:06.689757	14:02:07.125238	00:02:00.435481	0.7197	0.01	linear	FAIL
36	15.3545	14:02:07.285817	14:02:07.895097	00:00:00.609280	0.1389	10.0	linear	COMPLETE
37	0.2641	14:02:08.132662	14:02:45.847038	00:00:37.714376	0.0052	0.001	linear	COMPLETE
38	nan	14:02:46.044864	14:04:46.513015	00:02:00.468151	100.0	0.01	linear	FAIL
39	0.2797	14:04:46.793213	14:04:54.108671	00:00:07.315458	0.1389	1.0	linear	COMPLETE
40	nan	14:04:54.238258	14:06:54.759005	00:02:00.520747	3.7276	0.0001	linear	FAIL
41	15.3545	14:06:55.015830	14:06:56.081953	00:00:01.066123	0.001	10.0	linear	COMPLETE

42	nan	14:06:56.508070	14:08:57.210791	00:02:00.702721	0.1389	0.0001	linear	FAIL
43	0.2645	14:08:57.484854	14:10:04.891768	00:01:07.406914	0.0052	0.01	linear	COMPLETE
44	nan	14:10:05.090779	14:12:05.655385	00:02:00.564606	0.1389	0.1	linear	FAIL
45	nan	14:12:05.843825	14:14:06.511841	00:02:00.668016	0.1389	0.01	linear	FAIL
46	nan	14:14:06.823738	14:16:07.500182	00:02:00.676444	100.0	1.0	linear	FAIL
47	nan	14:16:07.655146	14:18:08.213041	00:02:00.557895	3.7276	0.1	linear	FAIL

Tabla C.4: Grid de optuna: SVR, predicción a 4h. Kernel LINEAR.

number	- MSE	start	stop	duration	C	epsilon	kernel	state
1	15.3545	14:20:11.387624	14:20:12.176080	00:00:00.788456	0.2512	10.0	rbf	COMPLETE
2	0.4398	14:20:12.460769	14:21:39.113942	00:01:26.653173	0.2512	0.0001	rbf	COMPLETE
3	0.5881	14:21:39.354782	14:23:01.726003	00:01:22.371221	0.0158	0.0001	rbf	COMPLETE
4	0.9243	14:23:01.985750	14:24:26.343030	00:01:24.357280	0.001	0.0046	rbf	COMPLETE
5	0.9267	14:24:26.847831	14:25:49.043778	00:01:22.195947	0.001	0.0046	rbf	COMPLETE
6	15.3545	14:25:49.280687	14:25:50.015806	00:00:00.735119	63.0957	10.0	rbf	COMPLETE
7	0.395	14:25:50.283804	14:26:50.990124	00:01:00.706320	63.0957	0.2154	rbf	COMPLETE
8	15.3545	14:26:51.396785	14:26:53.779998	00:00:02.383213	1000.0	10.0	rbf	COMPLETE
9	nan	14:26:54.124765	14:28:54.793033	00:02:00.668268	1000.0	0.0001	rbf	FAIL
10	nan	14:28:54.964754	14:30:55.681186	00:02:00.716432	63.0957	0.0001	rbf	FAIL
11	15.3545	14:30:56.124608	14:30:57.813030	00:00:01.688422	0.2512	10.0	rbf	COMPLETE
12	15.3545	14:30:58.416806	14:30:59.776606	00:00:01.359800	0.0158	10.0	rbf	COMPLETE
13	0.357	14:31:00.116749	14:31:43.743956	00:00:43.627207	3.9811	0.2154	rbf	COMPLETE
14	0.4322	14:31:44.189821	14:32:16.791281	00:00:32.601460	0.2512	0.2154	rbf	COMPLETE
15	0.5923	14:32:17.071753	14:33:51.717038	00:01:34.645285	0.0158	0.0001	rbf	COMPLETE
16	0.4366	14:33:52.351814	14:35:31.873849	00:01:39.522035	0.2512	0.0001	rbf	COMPLETE
17	0.5923	14:35:32.169899	14:37:07.540788	00:01:35.370889	0.0158	0.0046	rbf	COMPLETE
18	15.3545	14:37:07.867957	14:37:08.676831	00:00:00.808874	0.0158	10.0	rbf	COMPLETE
19	nan	14:37:09.024545	14:39:10.572641	00:02:01.548096	3.9811	0.0046	rbf	FAIL
20	nan	14:39:10.951852	14:41:12.112994	00:02:01.161142	63.0957	0.0046	rbf	FAIL
21	15.3545	14:41:12.302782	14:41:13.191812	00:00:00.889030	0.001	10.0	rbf	COMPLETE
22	0.9266	14:41:13.531832	14:42:52.087508	00:01:38.555676	0.001	0.0001	rbf	COMPLETE
23	nan	14:42:52.673738	14:44:53.978050	00:02:01.304312	1000.0	0.0046	rbf	FAIL
24	0.9348	14:44:54.185921	14:46:04.657905	00:01:10.471984	0.001	0.2154	rbf	COMPLETE
25	15.3545	14:46:05.127857	14:46:06.851681	00:00:01.723824	0.001	10.0	rbf	COMPLETE
26	nan	14:46:07.517948	14:48:08.786985	00:02:01.269037	3.9811	0.0046	rbf	FAIL

x

27	0.924	14:48:09.032837	14:49:59.391936	00:01:50.359099	0.001	0.0001	rbf	COMPLETE
28	0.3605	14:49:59.752943	14:50:40.156077	00:00:40.403134	3.9811	0.2154	rbf	COMPLETE
29	15.3545	14:50:40.610800	14:50:42.176952	00:00:01.566152	63.0957	10.0	rbf	COMPLETE
30	0.4289	14:50:42.850824	14:51:17.342843	00:00:34.492019	0.2512	0.2154	rbf	COMPLETE
31	nan	14:51:17.844722	14:53:19.569483	00:02:01.724761	1000.0	0.0001	rbf	FAIL
32	0.4398	14:53:19.938931	14:55:08.649858	00:01:48.710927	0.2512	0.0046	rbf	COMPLETE
33	nan	14:55:08.995980	14:57:09.831687	00:02:00.835707	3.9811	0.0001	rbf	FAIL
34	nan	14:57:10.066912	14:59:11.175767	00:02:01.108855	3.9811	0.0001	rbf	FAIL
35	0.5754	14:59:11.646909	15:00:00.571900	00:00:48.924991	0.0158	0.2154	rbf	COMPLETE
36	0.932	15:00:00.902828	15:01:04.720855	00:01:03.818027	0.001	0.2154	rbf	COMPLETE
37	15.3545	15:01:05.036780	15:01:05.894240	00:00:00.857460	3.9811	10.0	rbf	COMPLETE
38	0.5791	15:01:06.185832	15:01:47.462969	00:00:41.277137	0.0158	0.2154	rbf	COMPLETE
39	0.4348	15:01:47.707614	15:03:33.215798	00:01:45.508184	1000.0	0.2154	rbf	COMPLETE
40	15.3545	15:03:33.748722	15:03:34.637777	00:00:00.889055	1000.0	10.0	rbf	COMPLETE
41	nan	15:03:34.949945	15:05:35.727939	00:02:00.777994	1000.0	0.0046	rbf	FAIL
42	0.3927	15:05:35.942822	15:06:52.964004	00:01:17.021182	63.0957	0.2154	rbf	COMPLETE
43	nan	15:06:53.350851	15:08:54.324939	00:02:00.974088	63.0957	0.0001	rbf	FAIL
44	0.4332	15:08:54.776123	15:10:45.001818	00:01:50.225695	1000.0	0.2154	rbf	COMPLETE
45	0.4365	15:10:45.570081	15:12:32.605048	00:01:47.034967	0.2512	0.0046	rbf	COMPLETE
46	15.3545	15:12:33.135898	15:12:34.904008	00:00:01.768110	3.9811	10.0	rbf	COMPLETE
47	0.5882	15:12:35.494821	15:14:21.096724	00:01:45.601903	0.0158	0.0046	rbf	COMPLETE

Tabla C.5: Grid de optuna: SVR, predicción a 4h. Kernel RBF.

number	- MSE	start	stop	duration	C	epsilon	kernel	state
1	2.0383	15:16:27.884913	15:17:23.107117	00:00:55.222204	1.0	0.3162	poly	COMPLETE
2	15.3545	15:17:23.532824	15:17:24.709808	00:00:01.176984	1.0	10.0	poly	COMPLETE
3	15.3545	15:17:25.029852	15:17:26.009978	00:00:00.980126	1000.0	10.0	poly	COMPLETE
4	5.3883	15:17:26.330709	15:18:14.346231	00:00:48.015522	1.0	0.3162	poly	COMPLETE
5	15.3545	15:18:14.681483	15:18:15.788041	00:00:01.106558	1000.0	10.0	poly	COMPLETE
6	nan	15:18:16.179069	15:20:17.237696	00:02:01.058627	1000.0	0.01	poly	FAIL
7	0.351	15:20:17.444838	15:21:43.497161	00:01:26.052323	0.001	0.01	poly	COMPLETE
8	nan	15:21:43.933640	15:23:45.023847	00:02:01.090207	1.0	0.01	poly	FAIL
9	nan	15:23:45.410614	15:25:47.064346	00:02:01.653732	1000.0	0.01	poly	FAIL
10	24.2708	15:25:47.313971	15:27:40.450990	00:01:53.137019	1000.0	0.3162	poly	COMPLETE
11	15.3545	15:27:41.095770	15:27:43.017731	00:00:01.921961	1.0	10.0	poly	COMPLETE

12	nan	15:27:43.632860	15:29:45.354774	00:02:01.721914	1000.0	0.01	poly	FAIL
13	nan	15:29:45.786734	15:31:47.656134	00:02:01.869400	1.0	0.01	poly	FAIL
14	15.3545	15:31:47.876771	15:31:49.221033	00:00:01.344262	1000.0	10.0	poly	COMPLETE
15	1.5731	15:31:49.687709	15:33:35.037029	00:01:45.349320	0.001	0.01	poly	COMPLETE
16	nan	15:33:35.740696	15:35:37.063806	00:02:01.323110	1000.0	0.3162	poly	FAIL
17	0.7392	15:35:37.386804	15:36:36.259157	00:00:58.872353	0.001	0.3162	poly	COMPLETE
18	15.3545	15:36:36.637830	15:36:37.810966	00:00:01.173136	0.001	10.0	poly	COMPLETE
19	0.3702	15:36:38.052002	15:37:11.962538	00:00:33.910536	0.001	0.3162	poly	COMPLETE
20	15.3545	15:37:12.306550	15:37:13.440058	00:00:01.133508	1.0	10.0	poly	COMPLETE
21	1.4808	15:37:13.770355	15:38:07.424738	00:00:53.654383	0.001	0.3162	poly	COMPLETE
22	0.3769	15:38:07.730383	15:38:31.126976	00:00:23.396593	0.001	0.3162	poly	COMPLETE
23	15.3545	15:38:31.471705	15:38:32.469794	00:00:00.998089	1000.0	10.0	poly	COMPLETE
24	6.8755	15:38:32.781819	15:39:04.173070	00:00:31.391251	1.0	0.3162	poly	COMPLETE
25	nan	15:39:04.618108	15:41:06.254095	00:02:01.635987	1000.0	0.01	poly	FAIL
26	nan	15:41:06.481607	15:43:07.531824	00:02:01.050217	1000.0	0.01	poly	FAIL
27	15.3545	15:43:07.787861	15:43:08.999801	00:00:01.211940	0.001	10.0	poly	COMPLETE
28	8.7544	15:43:09.362816	15:45:04.402818	00:01:55.040002	1000.0	0.3162	poly	COMPLETE
29	15.3545	15:45:04.718828	15:45:05.930724	00:00:01.211896	1.0	10.0	poly	COMPLETE
30	15.3545	15:45:06.262623	15:45:07.516741	00:00:01.254118	0.001	10.0	poly	COMPLETE
31	0.3787	15:45:07.903531	15:46:41.414035	00:01:33.510504	0.001	0.01	poly	COMPLETE
32	15.3545	15:46:41.833813	15:46:43.005990	00:00:01.172177	1000.0	10.0	poly	COMPLETE
33	0.7594	15:46:43.508750	15:48:16.565148	00:01:33.056398	0.001	0.01	poly	COMPLETE
34	360.6699	15:48:16.986982	15:49:38.734241	00:01:21.747259	1000.0	0.3162	poly	COMPLETE
35	nan	15:49:39.251257	15:51:41.105856	00:02:01.854599	1.0	0.01	poly	FAIL
36	15.3545	15:51:41.406844	15:51:42.679767	00:00:01.272923	0.001	10.0	poly	COMPLETE
37	nan	15:51:43.035702	15:53:44.498694	00:02:01.462992	1.0	0.01	poly	FAIL
38	15.3545	15:53:44.891835	15:53:46.300925	00:00:01.409090	0.001	10.0	poly	COMPLETE
39	141.3452	15:53:46.766883	15:54:42.622778	00:00:55.855895	1.0	0.3162	poly	COMPLETE
40	15.3545	15:54:42.900768	15:54:44.035625	00:00:01.134857	0.001	10.0	poly	COMPLETE
41	15.3545	15:54:44.361688	15:54:46.019970	00:00:01.658282	1.0	10.0	poly	COMPLETE
42	0.6554	15:54:46.589822	15:56:16.315811	00:01:29.725989	0.001	0.01	poly	COMPLETE
43	0.7349	15:56:16.620692	15:57:48.946482	00:01:32.325790	0.001	0.01	poly	COMPLETE
44	1.8086	15:57:49.305839	15:58:22.959863	00:00:33.654024	1.0	0.3162	poly	COMPLETE
45	0.6192	15:58:23.327529	15:59:10.103772	00:00:46.776243	1.0	0.3162	poly	COMPLETE
46	0.7578	15:59:10.442792	16:00:11.616743	00:01:01.173951	0.001	0.3162	poly	COMPLETE
47	15.3545	16:00:11.974720	16:00:13.244813	00:00:01.270093	1.0	10.0	poly	COMPLETE

48	0.6154	16:00:13.620958	16:00:39.783005	00:00:26.162047	0.001	0.3162	poly	COMPLETE
49	nan	16:00:40.320902	16:02:42.407905	00:02:02.087003	1.0	0.01	poly	FAIL
50	15.3545	16:02:42.888200	16:02:45.198135	00:00:02.309935	1000.0	10.0	poly	COMPLETE
51	nan	16:02:45.823623	16:04:47.818681	00:02:01.995058	1000.0	0.01	poly	FAIL
52	nan	16:04:48.139361	16:06:49.603638	00:02:01.464277	1.0	0.01	poly	FAIL
53	nan	16:06:49.870220	16:08:51.471719	00:02:01.601499	1000.0	0.3162	poly	FAIL

Tabla C.6: Grid de optuna: SVR, predicción a 4h. Kernel POLY.

number	- MSE	start	stop	duration	C	epsilon	kernel	state
1	0.4052	13:31:53.422718	13:32:55.620651	00:01:02.197933	0.1389	0.01	linear	COMPLETE
2	0.4256	13:32:55.733593	13:33:15.298446	00:00:19.564853	0.001	0.001	linear	COMPLETE
3	15.3291	13:33:15.370586	13:33:15.595243	00:00:00.224657	0.1389	10.0	linear	COMPLETE
4	nan	13:33:15.652062	13:35:15.841246	00:02:00.189184	100.0	0.0001	linear	FAIL
5	0.409	13:35:15.964449	13:35:33.909143	00:00:17.944694	0.0052	0.01	linear	COMPLETE
6	nan	13:35:33.998473	13:37:34.249090	00:02:00.250617	3.7276	0.1	linear	FAIL
7	0.4326	13:37:34.344890	13:38:02.003665	00:00:27.658775	0.7197	1.0	linear	COMPLETE
8	nan	13:38:02.217925	13:40:02.614964	00:02:00.397039	0.7197	0.1	linear	FAIL
9	0.4479	13:40:02.739785	13:40:06.851938	00:00:04.112153	0.0052	1.0	linear	COMPLETE
10	0.4249	13:40:07.166909	13:40:40.100154	00:00:32.933245	0.001	0.01	linear	COMPLETE
11	nan	13:40:40.237613	13:42:40.631011	00:02:00.393398	100.0	0.001	linear	FAIL
12	0.4058	13:42:40.818722	13:43:51.844939	00:01:11.026217	0.0268	0.001	linear	COMPLETE
13	nan	13:43:52.008178	13:45:52.342857	00:02:00.334679	3.7276	0.0001	linear	FAIL
14	nan	13:45:52.436781	13:47:52.769809	00:02:00.333028	3.7276	0.001	linear	FAIL
15	15.3291	13:47:52.860320	13:47:53.419805	00:00:00.559485	3.7276	10.0	linear	COMPLETE
16	0.5006	13:47:53.617839	13:47:56.623323	00:00:03.005484	0.001	1.0	linear	COMPLETE
17	nan	13:47:56.784331	13:49:57.223841	00:02:00.439510	100.0	1.0	linear	FAIL
18	0.4018	13:49:57.308174	13:51:06.516859	00:01:09.208685	0.0268	0.1	linear	COMPLETE
19	15.3291	13:51:06.755498	13:51:07.417818	00:00:00.662320	0.0052	10.0	linear	COMPLETE
20	nan	13:51:07.643889	13:52:47.929873	00:01:40.285984	0.1389	0.0001	linear	FAIL
21	15.3291	13:54:41.556257	13:54:41.774237	00:00:00.217980	0.7197	10.0	linear	COMPLETE
22	nan	13:54:41.833229	13:56:42.035334	00:02:00.202105	3.7276	0.001	linear	FAIL
23	nan	13:56:42.091306	13:58:42.299699	00:02:00.208393	100.0	0.01	linear	FAIL
24	nan	13:58:42.366834	14:00:42.631823	00:02:00.264989	0.7197	0.01	linear	FAIL
25	15.3291	14:00:42.708018	14:00:43.033700	00:00:00.325682	19.307	10.0	linear	COMPLETE
26	0.448	14:00:43.148559	14:00:44.874343	00:00:01.725784	0.0052	1.0	linear	COMPLETE

27	nan	14:00:44.969132	14:02:45.265810	00:02:00.296678	19.307	0.01	linear	FAIL
28	15.3291	14:02:45.373883	14:02:45.767784	00:00:00.393901	0.001	10.0	linear	COMPLETE
29	nan	14:02:45.902199	14:04:46.279927	00:02:00.377728	3.7276	0.0001	linear	FAIL
30	15.3291	14:04:46.412634	14:04:46.772978	00:00:00.360344	0.1389	10.0	linear	COMPLETE
31	nan	14:04:46.915413	14:06:47.434710	00:02:00.519297	3.7276	0.01	linear	FAIL
32	15.3291	14:06:47.768449	14:06:48.662808	00:00:00.894359	0.0268	10.0	linear	COMPLETE
33	15.3291	14:06:48.864799	14:06:49.362700	00:00:00.497901	3.7276	10.0	linear	COMPLETE
34	0.4099	14:06:49.546714	14:07:41.633882	00:00:52.087168	0.0052	0.0001	linear	COMPLETE
35	nan	14:07:41.796573	14:09:42.257980	00:02:00.461407	19.307	0.001	linear	FAIL
36	0.4258	14:09:42.499881	14:10:24.253872	00:00:41.753991	0.001	0.001	linear	COMPLETE
37	0.4066	14:10:24.700766	14:11:01.740827	00:00:37.040061	0.0052	0.1	linear	COMPLETE
38	nan	14:11:01.932661	14:13:02.337941	00:02:00.405280	100.0	0.001	linear	FAIL
39	nan	14:13:02.482780	14:15:02.936972	00:02:00.454192	0.1389	0.01	linear	FAIL
40	0.5	14:15:03.083580	14:15:06.437903	00:00:03.354323	0.001	1.0	linear	COMPLETE
41	0.422	14:15:06.660889	14:15:47.948808	00:00:41.287919	0.001	0.1	linear	COMPLETE
42	nan	14:15:48.152703	14:17:48.748836	00:02:00.596133	0.1389	0.001	linear	FAIL
43	0.4091	14:17:48.950850	14:18:54.776869	00:01:05.826019	0.0052	0.01	linear	COMPLETE
44	nan	14:18:55.030326	14:20:55.594826	00:02:00.564500	19.307	0.0001	linear	FAIL
45	0.4306	14:20:55.742942	14:21:09.053896	00:00:13.310954	0.1389	1.0	linear	COMPLETE
46	0.4058	14:21:09.503927	14:23:02.923720	00:01:53.419793	0.0268	0.001	linear	COMPLETE
47	nan	14:23:03.127212	14:25:03.926908	00:02:00.799696	100.0	0.1	linear	FAIL

Tabla C.7: Grid de optuna: SVR, predicción a 6h. Kernel LINEAR.

number	- MSE	start	stop	duration	C	epsilon	kernel	state
1	15.3291	14:27:08.579876	14:27:09.862859	00:00:01.282983	0.0158	10.0	rbf	COMPLETE
2	15.3291	14:27:10.249848	14:27:11.002879	00:00:00.753031	0.0158	10.0	rbf	COMPLETE
3	0.5195	14:27:11.302938	14:28:56.146854	00:01:44.843916	0.2512	0.0046	rbf	COMPLETE
4	nan	14:28:56.741824	14:30:58.057294	00:02:01.315470	1000.0	0.0001	rbf	FAIL
5	nan	14:30:58.425260	14:33:00.016022	00:02:01.590762	1000.0	0.0001	rbf	FAIL
6	nan	14:33:00.395725	14:35:01.931655	00:02:01.535930	1000.0	0.0046	rbf	FAIL
7	15.3291	14:35:02.417273	14:35:06.428932	00:00:04.011659	0.2512	10.0	rbf	COMPLETE
8	0.5026	14:35:06.907679	14:35:54.857715	00:00:47.950036	0.2512	0.2154	rbf	COMPLETE
9	0.4499	14:35:55.215795	14:37:01.137795	00:01:05.922000	3.9811	0.2154	rbf	COMPLETE
10	nan	14:37:01.522738	14:39:02.402481	00:02:00.879743	63.0957	0.0001	rbf	FAIL
11	nan	14:39:02.672913	14:41:03.666734	00:02:00.993821	0.001	0.0046	rbf	FAIL

12	15.3291	14:41:03.955072	14:41:05.018963	00:00:01.063891	0.001	10.0	rbf	COMPLETE
13	15.3291	14:41:05.464917	14:41:06.526960	00:00:01.062043	1000.0	10.0	rbf	COMPLETE
14	nan	14:41:06.945648	14:43:07.976996	00:02:01.031348	0.001	0.0001	rbf	FAIL
15	nan	14:43:08.480871	14:45:09.750244	00:02:01.269373	1000.0	0.2154	rbf	FAIL
16	nan	14:45:10.217958	14:47:11.965962	00:02:01.748004	0.0158	0.0046	rbf	FAIL
17	0.9589	14:47:12.245642	14:48:33.679823	00:01:21.434181	0.001	0.2154	rbf	COMPLETE
18	0.6341	14:48:34.381899	14:49:30.772857	00:00:56.390958	0.0158	0.2154	rbf	COMPLETE
19	nan	14:49:31.202919	14:51:32.204709	00:02:01.001790	0.001	0.0046	rbf	FAIL
20	nan	14:51:32.467835	14:53:33.203748	00:02:00.735913	63.0957	0.2154	rbf	FAIL
21	nan	14:53:33.464265	14:55:34.590125	00:02:01.125860	0.2512	0.0001	rbf	FAIL
22	15.3291	14:55:34.803558	14:55:35.857016	00:00:01.053458	0.2512	10.0	rbf	COMPLETE
23	0.5169	14:55:36.218978	14:57:35.380122	00:01:59.161144	0.2512	0.0046	rbf	COMPLETE
24	15.3291	14:57:36.139188	14:57:37.013760	00:00:00.874572	0.001	10.0	rbf	COMPLETE
25	0.4468	14:57:37.369721	14:58:32.453926	00:00:55.084205	3.9811	0.2154	rbf	COMPLETE
26	0.9542	14:58:33.199771	15:00:26.884246	00:01:53.684475	0.001	0.0001	rbf	COMPLETE
27	0.6466	15:00:27.243681	15:02:15.862110	00:01:48.618429	0.0158	0.0001	rbf	COMPLETE
28	nan	15:02:16.213836	15:04:17.348018	00:02:01.134182	63.0957	0.0046	rbf	FAIL
29	0.9612	15:04:17.819260	15:05:31.510549	00:01:13.691289	0.001	0.2154	rbf	COMPLETE
30	15.3291	15:05:31.833777	15:05:33.265010	00:00:01.431233	3.9811	10.0	rbf	COMPLETE
31	15.3291	15:05:33.860979	15:05:35.599118	00:00:01.738139	1000.0	10.0	rbf	COMPLETE
32	0.638	15:05:36.203239	15:06:23.647979	00:00:47.444740	0.0158	0.2154	rbf	COMPLETE
33	0.5053	15:06:23.963060	15:07:02.706924	00:00:38.743864	0.2512	0.2154	rbf	COMPLETE
34	nan	15:07:03.001783	15:09:03.749956	00:02:00.748173	3.9811	0.0046	rbf	FAIL
35	0.5172	15:09:03.971859	15:10:58.536770	00:01:54.564911	0.2512	0.0001	rbf	COMPLETE
36	0.5257	15:10:58.804771	15:12:27.430060	00:01:28.625289	63.0957	0.2154	rbf	COMPLETE
37	0.6496	15:12:27.686025	15:13:53.664020	00:01:25.977995	0.0158	0.0046	rbf	COMPLETE
38	nan	15:13:54.113122	15:15:55.297224	00:02:01.184102	1000.0	0.0046	rbf	FAIL
39	nan	15:15:55.632449	15:17:56.963938	00:02:01.331489	3.9811	0.0046	rbf	FAIL
40	15.3291	15:17:57.107692	15:17:57.755960	00:00:00.648268	3.9811	10.0	rbf	COMPLETE
41	nan	15:17:58.085814	15:19:58.903924	00:02:00.818110	1000.0	0.2154	rbf	FAIL
42	15.3291	15:19:59.101874	15:20:00.044884	00:00:00.943010	63.0957	10.0	rbf	COMPLETE
43	nan	15:20:00.413752	15:22:01.351830	00:02:00.938078	63.0957	0.0046	rbf	FAIL
44	nan	15:22:01.752941	15:24:02.841853	00:02:01.088912	3.9811	0.0001	rbf	FAIL
45	nan	15:24:03.160899	15:26:04.672812	00:02:01.511913	3.9811	0.0001	rbf	FAIL
46	0.6496	15:26:04.891917	15:27:55.864977	00:01:50.973060	0.0158	0.0001	rbf	COMPLETE
47	15.3291	15:27:56.479879	15:27:58.322937	00:00:01.843058	63.0957	10.0	rbf	COMPLETE

Tabla C.8: Grid de optuna: SVR, predicción a 6h. Kernel RBF.

number	- MSE	start	stop	duration	C	epsilon	kernel	state
1	0.4946	15:30:05.271441	15:30:35.796214	00:00:30.524773	0.001	0.3162	poly	COMPLETE
2	0.4805	15:30:36.181350	15:31:57.494839	00:01:21.313489	0.001	0.01	poly	COMPLETE
3	0.6667	15:31:57.796862	15:32:21.593994	00:00:23.797132	0.001	0.3162	poly	COMPLETE
4	15.3291	15:32:22.102606	15:32:23.997065	00:00:01.894459	0.001	10.0	poly	COMPLETE
5	15.3291	15:32:24.611768	15:32:25.963777	00:00:01.352009	1.0	10.0	poly	COMPLETE
6	nan	15:32:26.267696	15:34:27.204863	00:02:00.937167	1000.0	0.3162	poly	FAIL
7	15.3291	15:34:27.550898	15:34:29.594005	00:00:02.043107	1000.0	10.0	poly	COMPLETE
8	15.3291	15:34:30.267554	15:34:31.434749	00:00:01.167195	1.0	10.0	poly	COMPLETE
9	nan	15:34:31.773117	15:36:32.927848	00:02:01.154731	1000.0	0.3162	poly	FAIL
10	80.0153	15:36:33.351010	15:37:26.243812	00:00:52.892802	1.0	0.3162	poly	COMPLETE
11	2.8464	15:37:26.652856	15:38:29.268789	00:01:02.615933	1.0	0.3162	poly	COMPLETE
12	nan	15:38:29.579663	15:40:30.601248	00:02:01.021585	1000.0	0.01	poly	FAIL
13	15.3291	15:40:30.802913	15:40:31.948257	00:00:01.145344	1.0	10.0	poly	COMPLETE
14	nan	15:40:32.323403	15:42:33.466835	00:02:01.143432	1000.0	0.01	poly	FAIL
15	nan	15:42:33.912381	15:44:35.973839	00:02:02.061458	1000.0	0.01	poly	FAIL
16	0.4886	15:44:36.440849	15:45:06.299917	00:00:29.859068	0.001	0.3162	poly	COMPLETE
17	15.3291	15:45:07.001889	15:45:09.242342	00:00:02.240453	0.001	10.0	poly	COMPLETE
18	0.8195	15:45:09.900624	15:46:26.100010	00:01:16.199386	0.001	0.01	poly	COMPLETE
19	1.5927	15:46:26.677908	15:47:43.229755	00:01:16.551847	0.001	0.01	poly	COMPLETE
20	nan	15:47:43.553807	15:49:44.390719	00:02:00.836912	1000.0	0.01	poly	FAIL
21	7.064	15:49:44.655667	15:50:26.758838	00:00:42.103171	1.0	0.3162	poly	COMPLETE
22	155.7576	15:50:27.060705	15:52:00.026023	00:01:32.965318	1000.0	0.3162	poly	COMPLETE
23	nan	15:52:00.648925	15:54:01.587922	00:02:00.938997	1000.0	0.3162	poly	FAIL
24	15.3291	15:54:01.802914	15:54:03.756680	00:00:01.953766	0.001	10.0	poly	COMPLETE
25	15.3291	15:54:04.276896	15:54:06.500947	00:00:02.224051	1000.0	10.0	poly	COMPLETE
26	1.2254	15:54:07.075563	15:54:45.690840	00:00:38.615277	1.0	0.3162	poly	COMPLETE
27	0.5051	15:54:46.100739	15:56:16.788902	00:01:30.688163	0.001	0.01	poly	COMPLETE
28	15.3291	15:56:17.137697	15:56:18.297891	00:00:01.160194	1000.0	10.0	poly	COMPLETE
29	15.3291	15:56:18.623902	15:56:19.777804	00:00:01.153902	1000.0	10.0	poly	COMPLETE
30	15.3291	15:56:20.205531	15:56:21.958905	00:00:01.753374	1000.0	10.0	poly	COMPLETE
31	nan	15:56:22.594320	15:58:24.514009	00:02:01.919689	1.0	0.01	poly	FAIL
32	nan	15:58:24.809995	16:00:26.885125	00:02:02.075130	1.0	0.01	poly	FAIL

33	1.0874	16:00:27.055779	16:01:17.182856	00:00:50.127077	1.0	0.3162	poly	COMPLETE
34	15.3291	16:01:17.717729	16:01:19.486977	00:00:01.769248	1000.0	10.0	poly	COMPLETE
35	nan	16:01:20.026742	16:03:21.442995	00:02:01.416253	1.0	0.01	poly	FAIL
36	15.3291	16:03:21.825985	16:03:24.109710	00:00:02.283725	0.001	10.0	poly	COMPLETE
37	15.3291	16:03:24.506842	16:03:25.500914	00:00:00.994072	0.001	10.0	poly	COMPLETE
38	15.3291	16:03:25.800741	16:03:26.784861	00:00:00.984120	1.0	10.0	poly	COMPLETE
39	8.8108	16:03:27.112266	16:04:40.332259	00:01:13.219993	1.0	0.3162	poly	COMPLETE
40	0.732	16:04:40.644760	16:06:04.757810	00:01:24.113050	0.001	0.01	poly	COMPLETE
41	0.8181	16:06:05.078921	16:07:28.563791	00:01:23.484870	0.001	0.01	poly	COMPLETE
42	nan	16:07:28.885896	16:09:30.003961	00:02:01.118065	1000.0	0.3162	poly	FAIL
43	15.3291	16:09:30.389728	16:09:31.787632	00:00:01.397904	0.001	10.0	poly	COMPLETE
44	nan	16:09:32.107821	16:11:33.081983	00:02:00.974162	1.0	0.01	poly	FAIL
45	nan	16:11:33.247827	16:13:34.283909	00:02:01.036082	1000.0	0.01	poly	FAIL
46	0.8236	16:13:34.708873	16:14:22.536144	00:00:47.827271	0.001	0.3162	poly	COMPLETE
47	nan	16:14:22.761695	16:16:23.882755	00:02:01.121060	1.0	0.01	poly	FAIL
48	0.8178	16:16:24.105568	16:17:21.293843	00:00:57.188275	0.001	0.3162	poly	COMPLETE
49	nan	16:17:21.952726	16:19:23.989018	00:02:02.036292	1.0	0.01	poly	FAIL
50	nan	16:19:24.563193	16:21:26.215651	00:02:01.652458	1000.0	0.3162	poly	FAIL
51	1.4882	16:21:26.779939	16:22:30.206877	00:01:03.426938	0.001	0.3162	poly	COMPLETE
52	15.3291	16:22:30.565006	16:22:31.870719	00:00:01.305713	1.0	10.0	poly	COMPLETE
53	15.3291	16:22:32.119804	16:22:33.230698	00:00:01.110894	1.0	10.0	poly	COMPLETE

Tabla C.9: Grid de optuna: SVR, predicción a 6h. Kernel POLY.