UIMP
Universidad Internacional
Menéndez Pelayo

Evaluación del impacto de las medidas de confinamiento sobre la propagación de la enfermedad COVID-19 mediante técnicas de aprendizaje automático

(Assessment of the impact of containment measures on the spread of COVID-19 disease using machine learning techniques)

Trabajo de Fin de Máster
para acceder al

MÁSTER EN CIENCIA DE DATOS

Autora: María Oyarzabal Alcain

Directora: María Castrillo Melguizo

Septiembre - 2020

Agradecimientos

Agradecimiento al proyecto "Impacto de las medidas de distanciamiento social sobre la expansión de la epidemia de COVID-19 en España" (Distancia COVID - CSICCOV19-039) en el cual participa el Grupo de Computación Avanzada y e-Ciencia y en el marco del cual se ha llevado a cabo este Trabajo Fin de Máster.

Resumen

El brote de la enfermedad COVID-19, causada por el virus SARS-CoV-2, ha afectado a más de 170 países, acumulando más de 25 millones de contagios y miles de muertes en todo el mundo. Siendo así una de las mayores pandemias mundiales vividas en la historia. No existe tratamiento efectivo contra la enfermedad, lo que ha hecho que los gobiernos adopten medidas de confinamiento para frenar la expansión del virus.

En este trabajo se aplican técnicas de aprendizaje automático para evaluar el impacto de las medidas de confinamiento sobre la propagación de la enfermedad en España. El objetivo es diseñar un modelo que pueda aplicarse sobre los datos a nivel provincial y que permita predecir la evolución de la pandemia en un periodo corto de plazo. En particular, se pretende estudiar qué medidas de distanciamiento social fueron más efectivas a la hora de frenar su expansión.

Se han calculado tanto el número básico reproductivo como el porcentaje de incremento de casos a partir del número de contagios por día. Posteriormente se han ajustado un modelo *random forest* y una red neuronal perceptrón multicapa y se ha comprobado cuál de los dos tiene mejor desempeño. Finalmente, se ha evaluado qué medidas de confinamiento tuvieron mayor impacto a la hora de frenar el virus creando nuevos escenarios y viendo como afectan a las variables objetivo.

Palabras clave: COVID-19, random forest, redes neuronales, aprendizaje automático.

Abstract

The outbreak of COVID-19 disease, caused by the SARS-CoV-2 virus, has affected more than 170 countries, accumulating more than 25 million infections and thousands of deaths worldwide. This is one of the largest global pandemics in history. There is no effective treatment against the disease, which has led governments to adopt containment measures to stop the spread of the virus.

In this paper, machine learning techniques are applied to assess the impact of containment measures on the spread of the disease in Spain. The objective is to design a unique model that can be applied to data by province and that will allow the prediction of the evolution of the pandemic in a short period of time. In particular, it is intended to study which social distancing measures were most effective in slowing its expansion.

Both the basic reproductive number and the percentage of increase in cases based on the number of infections per day have been calculated. Subsequently, a random forest model and a multi-layer perceptron neural network have been adjusted and it has been proven which of the two has better performance. Finally, it has been evaluated which containment measures had the greatest impact in stopping the virus by creating new scenarios and seeing how they affect the target variables.

Key words: COVID-19, random forest, neural networks, machine learning.

Índice general

Capítulo 1

Introducción

1.1. Motivación

La *enfermedad COVID-19*, también conocida como enfermedad por coronavirus, es una enfermedad infecciosa causada por el virus SARS-CoV-2.

Se detectó por primera vez en la ciudad de Wuhan, provincia de Hubei (China), a mediados de diciembre de 2019 [1]. El 30 de enero de 2020 la Organización Mundial de la Salud (OMS), declaró la enfermedad como una emergencia sanitaria de preocupación internacional debido a un número creciente de casos confirmados, no solo en China sino también en otros países [2]. El 11 de marzo la enfermedad se encontraba ya en más de 100 países del mundo y la OMS la reconoció como pandemia [3].

En España, se confirmó el primer caso a finales de febrero y el 9 de abril ya era el segundo país con más casos confirmados y el tercero en número de muertes. A partir del 3 de abril, el número de nuevos casos diarios se estabilizó e incluso comenzó a disminuir, a pesar de que el número de casos activos seguía aumentando [4].

La transmisión de persona a persona ocurre principalmente por contacto directo o a través de gotas propagadas por la tos o el estornudo de un individuo infectado. Además, se ha demostrado que el virus puede permanecer en superficies incluso días [5].

Los síntomas de la infección por COVID-19 aparecen después de un período de incubación de aproximadamente 5,2 días [6]. Los más comunes al comienzo de la enfermedad son fiebre, tos y fatiga, pero también hay casos en los que se presentan síntomas como disnea, dolor de cabeza y diarrea. Sin embargo, las personas infectadas también pueden ser asintomáticas, siendo esta una de las dificultades para frenar la expansión de la enfermedad en la población.

Al no existir tratamiento efectivo contra la enfermedad, los gobiernos adoptaron medidas de distanciamiento social para evitar la propagación de la misma. En España se decretó el estado de alarma el 14 de marzo de

2020, recogido en Boletin Oficial del Estado [7], para afrontar la situación de emergencia provocada por la enfermedad COVID-19. Con la declaración del estado de alarma se limitó la libertad de la circulación de las personas permitiendo únicamente realizar tareas como la compra de alimentos, asistencia a centros sanitarios o desplazamiento al lugar de trabajo. En general, únicamente se permitía la realización de tareas esenciales. Además, fueron suspendidas todas las actividades presenciales en todos los centros educativos.

Estas medidas fueron efectivas a la hora de controlar la pandemia, ya que se consiguió en un primer momento estabilizar el número de casos y posteriormente reducirlo. No obstante, esta situación produjo graves consecuencias en ámbitos como la economía y la sociedad. Todos los sectores de la economía mundial se vieron afectados por esta crisis. En concreto, países cuya economía depende más del turismo, como España, se vieron más afectados. Además, se ha demostrado en [8] que las medidas de aislamiento tienen efectos psicológicos negativos en las personas.

Por lo tanto, estas medidas no se podían mantener durante un largo periodo de tiempo. Una vez que la pandemia estuviera controlada, el levantamiento de las medidas debía realizarse de forma que se evitara un repunte de la enfermedad. Para ello, es de gran utilidad conocer qué medidas y con qué intensidad tenían un mayor impacto sobre la propagación de la enfermedad. Igualmente, en el caso de un rebrote que implique la necesidad de restablecer medidas de distanciamiento, el conocimiento de este impacto permitiría priorizar aquellas medidas que tuvieron un mayor efecto sobre la detención de la propagación.

Sin embargo, no es fácil distinguir qué medidas fueron las más efectivas a la hora de frenar la expansión de la enfermedad. Todas las medidas de distanciamiento social fueron tomadas prácticamente al mismo tiempo, por lo que es difícil distinguir cuáles causaron más efecto. Además, al ser una enfermedad que no se había detectado antes, se está investigando constantemente cuáles son sus características, como por ejemplo los síntomas, las vías de transmisión y el tiempo de incubación. Por otro lado, aunque los datos del número de contagios y del número de muertes son datos públicos, se pueden encontrar dificultades al recogerlos. Muchos casos no fueron contabilizados o fueron contabilizados en días posteriores. Todo ello, dificulta la comprensión de la evolución de la pandemia.

A lo largo de los años, la ciencia epidemiológica ha tenido un gran peso a la hora de modelizar agentes infecciosos. Los modelos epidemiológicos son representados a través de ecuaciones diferenciales. El modelo base es el modelo SIR, que divide la población en tres grupos de acuerdo a su estado de salud como susceptibles (S), infectados (I) y recuperados (R) [9]. A partir de este modelo se incorporan diferentes estados como vacunados (V) y en cuarentena (Q) en función de las características más importantes de la infección. En el caso de la enfermedad por coronavirus, se toman los

grupos de susceptibles, expuestos, infectados, en cuarentena, hospitalizados, recuperados y fallecidos [10].

Hoy en día, las técnicas de aprendizaje automático son capaces de conseguir un gran desempeño en tareas de predicción. Por esto, en este estudio se proponen diferentes técnicas de aprendizaje automático para la evaluación de la propagación de la enfermedad.

1.2. Objetivo

El objetivo de este trabajo es, por tanto, evaluar el impacto de las medidas de distanciamiento social adoptadas en la propagación de la enfermedad COVID-19 en España mediante técnicas de aprendizaje automático. Para ello, los datos utilizados están tomados por provincias y se pretende diseñar un modelo único que pueda aplicarse sobre los datos de cada provincia y que permita predecir cómo evolucionará la pandemia en un periodo a corto plazo en función de las medidas de distanciamiento adoptadas. La memoria se estructura de la siguiente manera:

En el Capítulo 1, se han presentado tanto la motivación como los objetivos del trabajo.

En el Capítulo 2, se describen las fuentes de datos utilizadas, se explica el proceso de preprocesado de los datos y se detallan ciertas propiedades de las series temporales que ayudarán a la descripción y comprensión de los datos.

En el Capítulo 3, se presenta la metodología empleada para realizar el estudio. Para ello, primero se explican los modelos que se utilizan, luego el entrenamiento de dichos modelos y finalmente como se evaluarán los mismos.

En el Capítulo 4, se especifican los modelos empleados, la evaluación de los mismos y se muestran los resultados obtenidos.

Finalmente, en el Capítulo 5, se recogen las conclusiones y dificultades presentadas a lo largo del trabajo y se plantean posibles mejoras para trabajos futuros.

Para la realización del trabajo se ha utilizado principalmente el lenguaje de programación *Python*. Las librerías utilizadas se detallarán a lo largo de la memoria cuando se considere relevante.

Capítulo 2

Análisis de datos

2.1. Descripción

En primer lugar se van a describir los dos conjuntos de datos utilizados, ambos elaborados por el grupo de Computación Avanzada y e-Ciencia del Instituto de Física de Cantabria. El primero de ellos se encuentra disponible en [11] y corresponde al número de contagios y muertes por provincia y día. Dispone de un total de 2477 observaciones tomadas desde el 27 de febrero de 2020, que fue cuando se empezaron a detectar los primeros casos, hasta el 18 de abril de 2020. A partir del 18 de abril, en algunas comunidades y provincias no se diferencian entre los test rápidos y las pruebas PCR [12] que se realizan a la población para detectar la enfermedad, por lo que los datos no pueden ser tratados de la misma manera. Estos datos han sido tomados diariamente a través de los portales de los gobiernos autonómicos y consultando notas de prensa diarias. Destacar que para los primeros días no existen datos de todas las provincias. El conjunto de datos consta de 9 variables que se explican a continuación y a partir de las cuales se han obtenido las variables objetivo que se detallarán posteriormente:

- **Date:** Fecha de publicación de los datos.
- **Id:** Número de identificación de la comunidad autónoma.
- **Region:** Nombre de la comunidad autónoma.
- **Pid:** Número de identificación de la provincia.
- **Province:** Nombre de la provincia.
- **Cases acc:** Número de casos infectados acumulados de la enfermedad COVID-19 hasta la fecha.
- **Cases new:** Número de casos nuevos infectados de COVID-19 registrados en la fecha dada.

- **Deaths acc:** Número de muertes por enfermedad COVID-19 acumuladas hasta la fecha.
- **Deaths new:** Número de muertes nuevas por enfermedad COVID-19 registradas en la fecha dada.

El segundo conjunto de datos recoge distintas variables de cada provincia en situación normal, es decir, cuando no existe ninguna medida de confinamiento. En el artículo [13], se establece que el contacto social se produce principalmente en el hogar, en la escuela, en el lugar de trabajo y en la comunidad en general. Por lo tanto, se han tenido en cuenta variables de estos cuatro ámbitos y además se han añadido datos demográficos y de turismo para crear el conjunto de datos. El conjunto de datos consta de un total de 17 variables para cada provincia, recogidas de los portales de *INEbase* [14] y *EDUCAbase* [15], que se explican a continuación y que se utilizarán como variables predictoras:

- **Población total:** Población total de la provincia.
- **Densidad:** Densidad de población de la provincia.
- **Población hombres:** Población masculina de la provincia.
- **Población >60:** Población mayor a 60 años.
- **Población hogar >4 personas:** Población que vive en hogares con más de 4 personas.
- **Población matriculada no uni:** Población estudiante no matriculada en enseñanzas universitarias.
- **Población matriculada uni:** Población estudiante matriculada en enseñanzas universitarias.
- **Ocupados agricultura:** Población ocupada en el sector de la agricultura.
- **Ocupados industria:** Población ocupada en el sector de la industria.
- **Ocupados construcción:** Población ocupada en el sector de la construcción.
- **Ocupados servicios:** Población ocupada en servicios.
- **Transporte en metro:** Viajeros transportados en metro.
- **Transporte autobús urbano:** Viajeros transportados en autobús urbano.

- **Población residencias mayores:** Población que vive en residencias de personas mayores.
- **Viajeros residencia España:** Viajeros alojados en hoteles con residencia en España.
- **Viajeros residencia extranjero:** Viajeros alojados en hoteles con residencia en el extranjero.
- **Movilidad laboral:** Movilidad laboral por provincia.

2.2. Preprocesado

Como se ha mencionado anteriormente, el primer conjunto de datos contiene el número de contagios y muertes por provincia y día de las 52 provincias españolas. En la comunidad catalana, no se han notificado las muertes por día clasificadas por provincias, por lo que dichos valores son desconocidos (NA). En este trabajo únicamente se utilizan el número de contagios por día para el cálculo de las variables objetivo, por lo que estos valores faltantes se han sustituido por valores nulos, ya que no afectan al estudio.

Además, para la provincia de Barcelona, no se habían registrado los datos para dos días posteriores al primer dato tomado. Para facilitar los cálculos, se han asignado valores nulos al número de contagios de estos dos días, por ser la diferencia del número de casos acumulados entre los días anteriores y los siguientes pequeña.

Los datos han sido recogidos en las fechas que se han notificado, sin embargo, esto no coincide con el día en el que se producen los contagios. Según un informe del Ministerio de Sanidad [16], se estima que el retraso que hay entre el día que ocurre el contagio y el día que se notifica es de 9 días. Esto es importante a la hora de evaluar si las medidas de confinamiento han influido en la reducción de la propagación del virus. Por lo tanto, se han trasladado las observaciones a nueve días antes, teniendo así datos desde el 18 de febrero de 2020 hasta el 9 de abril de 2020.

Por otro lado, los primeros casos que se registran son casos muy aislados, por lo que a la hora de evaluar como evoluciona la propagación del virus se han tenido en cuenta únicamente los días en los que el número de casos acumulados por provincia es al menos 12. El conjunto de datos se reduce entonces a 1990 observaciones.

Para poder asignar a cada día registrado las variables predictoras correspondientes en función de la provincia y del día, se explican a continuación las distintas fases de confinamiento dadas en España. El estado de alarma se decretó el 14 de marzo del 2020, provocando así la primera fase de confinamiento. Aunque el cierre de instituciones educativas se adelantó en La Rioja al 11 de marzo del 2020, en Madrid y Álava al 12 de marzo del 2020

y en el resto de España al 13 de marzo del 2020. El 29 de marzo del 2020 se endurecieron las medidas, únicamente se permitía la práctica de trabajos esenciales. En la siguiente tabla, se recogen los porcentajes a los que se redujeron cada una de las variables predictoras en función de la fase de confinamiento. Estos porcentajes también han sido estimados por el grupo de Computación Avanzada y e-Ciencia del Instituto de Física de Cantabria en algunos casos mediante la información obtenida en prensa.

	Base	Fin de semana	Confinamiento 14/03/2020	Confinamiento intensificado 29/03/2020
Poblacion total	100	100	100	100
Densidad	100	100	100	100
Poblacion hombres	100	100	100	100
Poblacion >60	100	100	100	100
Poblacion hogar >4 personas	100	100	100	100
Poblacion matriculada no uni	100	0	0	0
Poblacion matriculada uni	100	0	0	0
Ocupados agricultura	100	100	90	90
Ocupados industria	100	40	80	20
Ocupados construccion	100	0	80	10
Ocupados servicios	100	40	40	20
Transporte en metro	100	100	15	10
Transporte autobús urbano	100	100	10	10
Población residencias mayores	100	100	100	100
Viajeros residencia España	100	100	0	0
Viajeros residencia extranjero	100	100	40	0
Movilidad laboral	100	100	60	40

Tabla 2.1: Porcentajes a los que se reducen cada variable en función de la fase de confinamiento.

Entonces, teniendo en cuenta los porcentajes de la Tabla 2.1, a cada observación del conjunto de datos que recoge tanto el número de contagios como de muertes por día, se le ha asignado las variables predictoras de su provincia con su correspondiente reducción aplicada en función de la fecha. De esta manera se consigue un único conjunto de datos con toda la información necesaria.

2.3. Variables objetivo

Las variables objetivo consideradas en el estudio son el *número básico reproductivo* y el *porcentaje de incremento de casos*.

2.3.1. Número básico reproductivo

El *número básico reproductivo*, $\mathcal{R}_0$, es el número esperado de contagios secundarios generados por un individuo infectado [17, 18]. En concreto,

$$\mathcal{R}_0 = \tau \cdot \overline{c} \cdot d$$

donde τ es la probabilidad de infección dado que ha habido contacto entre un individuo susceptible e infectado, $\overline{c}$ es la tasa media de contacto entre individuos susceptibles e infectados y d es la duración del riesgo de contagio [19].

Si $\mathcal{R}_0 < 1$, cada individuo infectado generará menos de un nuevo contagiado. Por lo tanto, el número de nuevos casos contagiados disminuirá con el tiempo y el brote terminará por sí solo.

Si $\mathcal{R}_0 = 1$, cada individuo infectado generará un nuevo infectado. Por lo tanto, el número de casos contagiados se mantendrá estable.

Si $\mathcal{R}_0 > 1$, cada individuo infectado generará más de un nuevo infectado. Por lo tanto, el número de nuevos casos contagiados aumentará y podrá producirse un brote.

El valor de $\mathcal{R}_0$ se aplica únicamente si todos los individuos de una población son completamente vulnerables a una enfermedad. Es decir, cuando se cumplen las siguientes condiciones:

- Nadie ha sido vacunado.
- Nadie ha tenido la enfermedad antes.
- No hay forma de controlar la propagación de la enfermedad.

Para poder calcular el número básico reproductivo de la enfermedad, es necesario conocer la distribución del intervalo serial. El *intervalo serial* es el período de tiempo entre fases análogas de una enfermedad infecciosa en casos sucesivos de una cadena de infección. En el estudio [20] se estima que el intervalo serial del COVID-19 sigue una distribución de media 4,7 días y desviación estándar de 2,9 días.

El número básico reproductivo ha sido calculado utilizando el paquete *EpiEstim* de *R*. Se calcula tomando ventanas de tiempo, en este caso de días. Si la ventana es muy pequeña, el valor de $\mathcal{R}_0$ es más inestable debido a que el número de casos tiene saltos e irregularidades. En este caso, se han tomado ventanas de 7 días tal y como se recomienda en la documentación del paquete [21]. El valor obtenido para cada ventana se ha asignado al último día de la ventana, perdiendo así los datos para los 7 primeros días.

Entonces, el conjunto de datos resultante del cálculo del $\mathcal{R}_0$ consta de 1626 observaciones.

2.3.2. Porcentaje de incremento de casos

El *porcentaje de incremento de casos*, PI, es la proporción en la que aumentan el número de casos respecto al día anterior. Se calcula de la siguiente manera:

$$PI = \frac{\text{nº de nuevos casos}}{\text{nº total de casos}} \cdot 100$$

El conjunto de datos resultante del cálculo del porcentaje de incremento de casos consta de 1990 observaciones.

2.4. Series temporales

Una *serie temporal* es un conjunto de observaciones tomadas secuencialmente a lo largo del tiempo. Si el conjunto es continuo, se dice que la serie temporal es *continua*. Si el conjunto es discreto, se dice que la serie temporal es *discreta* [22].

El campo de aplicación de las series temporales es muy amplio. Se utilizan en ámbitos como la economía, la industria, la demografía, la medicina o la meteorología. En general, en cualquiera en el que la variable temporal sea necesaria para la comprensión del conjunto de datos observados.

Las series temporales tienen en particular que las observaciones futuras dependen de las anteriores, es decir, no son independientes. Por lo que es importante el orden de las mismas.

Uno de los principales objetivos de su estudio es la descripción de los datos, para así entender su comportamiento. Esta comprensión de los datos se realiza para predecir lo que va a ocurrir en tiempos futuros basándose en tiempos pasados.

En función del número de observaciones registradas por unidad de tiempo se pueden clasificar las series temporales en *univariantes* y *multivariantes*:

- **Serie temporal univariante:** Únicamente se observa una variable a lo largo del tiempo.
- **Serie temporal multivariante:** Se observan más de una variable a lo largo del tiempo.

Una característica importante de las series temporales es la *tendencia*. La *tendencia* es el comportamiento a lo largo del tiempo de la serie.

Otra característica importante de las series temporales es si son *estacionarias* o no. Se dice que una serie temporal es *estacionaria* si cumple que [23]:

- La varianza es finita.
- La media es constante y no depende del tiempo t.
- La covarianza en un intervalo de tiempo solo depende de la longitud del intervalo y no del momento en el que ocurre.

Es decir, una serie temporal es estacionaria si es estable. Gráficamente se puede observar que los valores oscilan alrededor de una media constante y que la varianza respecto a esa media también es constante. No presentan tendencias ni crecientes ni decrecientes.

El problema presentado es una serie temporal ya que se tienen las observaciones de las variables objetivo a lo largo de los días. Los conjuntos en los que toman valores dichas variables objetivo son conjuntos continuos, por lo que es una serie temporal continua.

Aunque se han considerado dos variables objetivo para resolver el problema, se han tomado como dos series temporales univariantes y no como una serie temporal multivariante. Esto se debe a que es más simple entender y trabajar con series temporales univariantes. Las series temporales multivariantes son más difíciles de modelar y muchos modelos clásicos del aprendizaje automático no funcionan bien.

2.4.1. Visualización

Visualizando la serie temporal se puede detectar si la tendencia es creciente o decreciente, si es estacionaria o no y si presenta valores atípicos o *outliers*.

A continuación se visualizan las variables objetivo de las primeras 6 provincias:

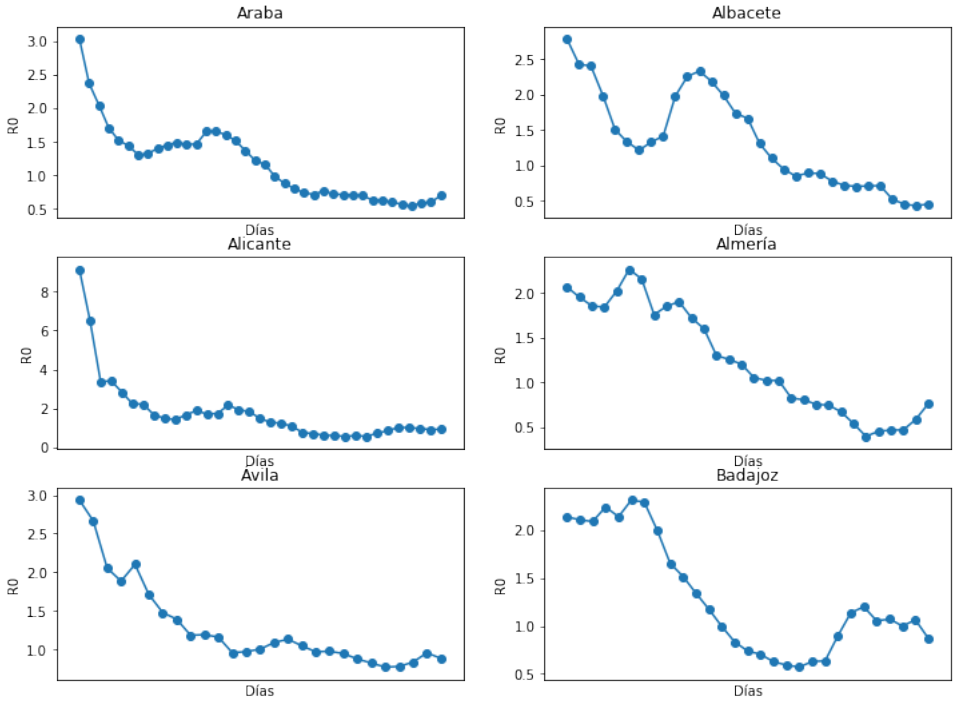

Figura 2.1: Variable objetivo $\mathcal{R}_0$ para las 6 primeras provincias.

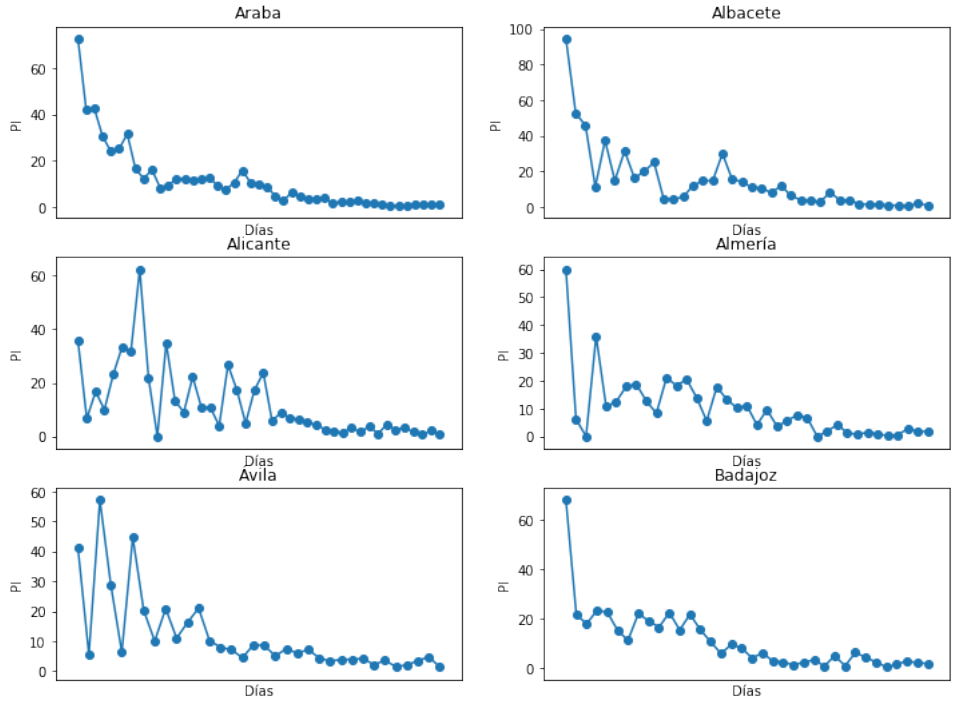

Figura 2.2: Variable objetivo PI para las 6 primeras provincias.

En la Figura 2.1 se puede observar que la tendencia de la serie del número básico reproductivo es generalmente descendente y que no es una serie temporal estacionaria.

En la Figura 2.2, en cambio, se puede ver que el porcentaje de incremento de casos presenta ciertos saltos e irregularidades. Una forma habitual de evitar estas irregularidades y poder observar la tendencia de la serie es aplicar técnicas de suavizado como las *medias móviles*.

La *media móvil simple* con n periodos, MMS, es la media de las n observaciones anteriores. Es decir, si $X_1, ..., X_n$ son las n observaciones anteriores, entonces,

$$MMS = \frac{X_1 + ... + X_n}{n}$$

Se ha calculado, por tanto, la media móvil del porcentaje de incrementos de casos, MPI, con periodos de 4 días para evitar los saltos. A continuación se puede observar la serie temporal suavizada:

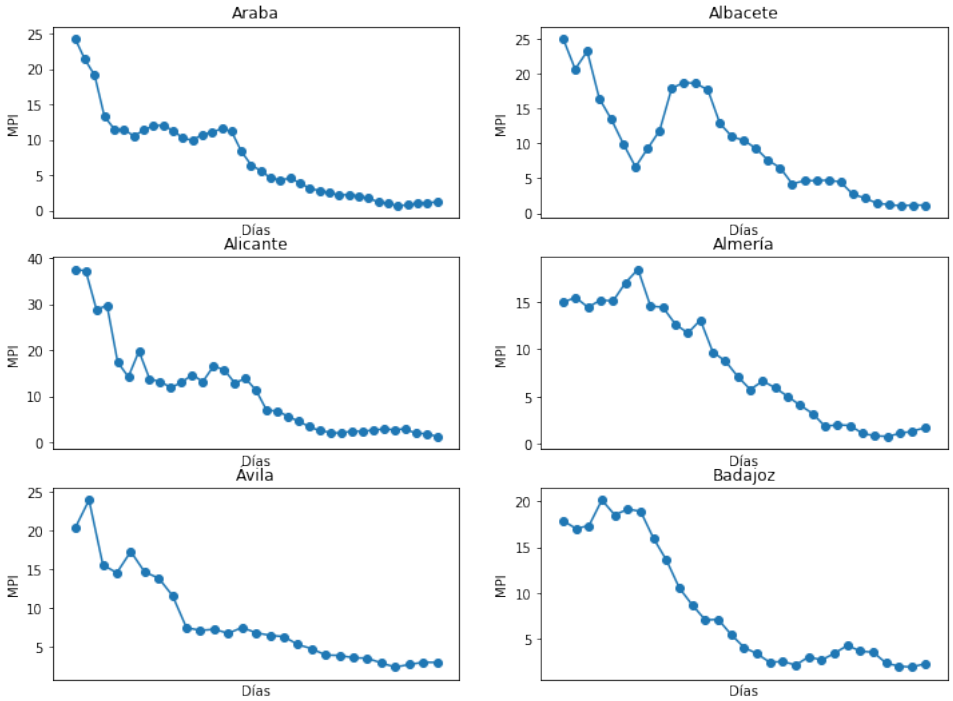

Figura 2.3: Media movil del MPI para las 6 primeras provincias.

En la Figura 2.3 se puede observar que la tendencia de la serie también es generalmente decreciente y que no es estacionaria.

Tras calcular la media móvil del porcentaje de incremento de casos, el conjunto de datos consta de 1834 observaciones. Sin embargo, se han eliminado los 4 primeros días para cada provincia para así tener los mismos días en ambas variables objetivo. Por tanto, se reduce a 1626 observaciones.

2.4.2. Método de ventanas deslizantes

El *método de ventanas deslizantes* es un método que convierte una serie temporal en un problema de aprendizaje supervisado, es decir, en un problema que consta de unas variables de entrada, X, y una variable de salida, y, y se utiliza un algoritmo para aprender la función de mapeo de la entrada a la salida [24].

Este método consiste en utilizar los pasos de tiempo anteriores como variables de entrada y los pasos siguientes como variables de salida. Para ello, se reestructura el conjunto de datos del problema de esta manera. El número de pasos anteriores y siguientes es un parámetro a elegir en función del problema.

En este caso, se han tomado 4 días anteriores para predecir el siguiente día. Es decir, se han utilizado las variables en los tiempos $t-1, t-2, t-3, t-4$

para predecir los valores en t. Estos valores han sido seleccionados después de realizar diferentes pruebas. Como es una serie temporal univariante que además consta de variables predictoras, se han tomado las variables predictoras y la variable objetivo en $t-1, t-2, t-3, t-4$ y las variables predictoras en t como variables de entrada para predecir lo que ocurre con la variable objetivo en t. A continuación, en la Figura 2.4, se muestra un esquema de las ventanas deslizantes empleadas donde X_i son las variables predictoras e y_i son las observaciones de la variable objetivo en el tiempo i con $i = \{1, 2, ..., N\}$:

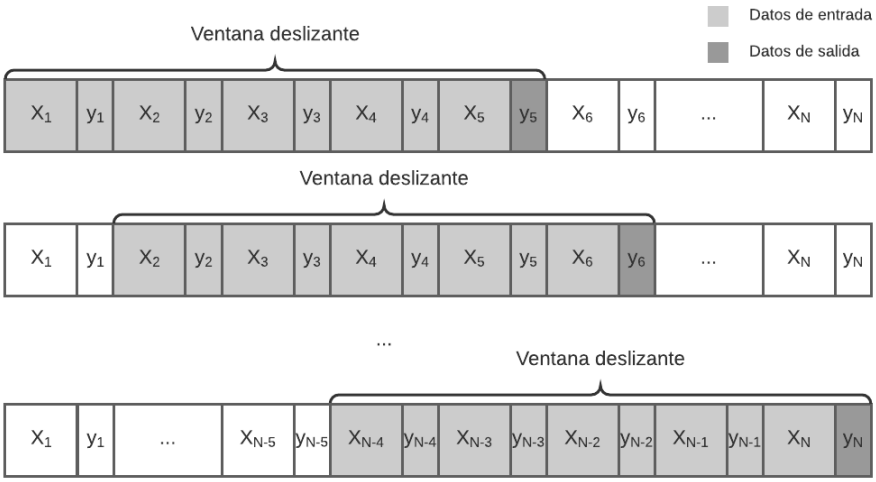

Figura 2.4: Esquema de las ventanas deslizantes empleadas.

Una vez realizada esta transformación, los conjuntos de datos se reducen a 1418 observaciones.

Capítulo 3

Metodología

Como se ha mencionado en el apartado 2.4.2, una vez aplicado el método de ventanas deslizantes al conjunto de datos, el problema planteado es un problema de aprendizaje supervisado, es decir, los modelos aprenden basándose en observaciones pasadas. Se utilizan muestras etiquetadas como datos de entrenamiento. Luego, para un conjunto de datos de prueba sin etiquetar, se desea estimar la variable objetivo.

Además, es un problema de regresión porque los conjuntos en los que toman valores las variables objetivo del problema son continuos.

Los modelos seleccionados para realizar el estudio son los *métodos basados en árboles*, en concreto, los modelos *random forest* y las *redes neuronales*, en concreto las *redes perceptrón multicapa*, que se explican a continuación.

Este capítulo está principalmente basado en [27, 28, 29].

3.1. Métodos basados en árboles

Los *métodos basados en árboles* son métodos que dividen el espacio de características en una serie de regiones simples. Para hacer una predicción de una observación dada, normalmente se utilizan la media o la moda de las observaciones de entrenamiento en la región a la que pertenece. Como el conjunto de reglas de división utilizadas para segmentar el espacio de predicción puede resumirse en un árbol, estos tipos de enfoques se conocen como métodos de *árboles de decisión*.

Los árboles de decisión pueden utilizarse tanto para problemas de clasificación como para problemas de regresión. Son muy intuitivos y fáciles de explicar gracias a que se pueden visualizar gráficamente. Sin embargo, también presentan ciertas dificultades. Cuando el árbol de decisión es muy profundo y por lo tanto muy complejo, puede perder capacidad de generalización en nuevas muestras que no han sido utilizadas durante el entrenamiento. Cuando esto sucede se dice que el modelo sobreajusta. Además, son poco robustos, es decir, un pequeño cambio en los datos puede causar un

gran cambio en el árbol estimado final.

Una manera de mejorar su capacidad predictiva es agregando muchos árboles. Las dos técnicas más utilizadas para el agregado de modelos simples son la técnica de *bagging* y la técnica de *boosting*.

Los *métodos de bagging* son métodos que utilizan los modelos simples en paralelo, es decir, crea los modelos de manera independiente para reducir el error promediando las salidas de estos.

Los *métodos de boosting* son métodos que utilizan los modelos simples secuencialmente, es decir, crea los modelos en función de la información dada por los modelos creados anteriormente. De esta manera, mejora el rendimiento dando importancia a los errores cometidos en modelos entrenados previamente.

3.1.1. Random forest

Random forest es una técnica de *bagging* en la que los modelos simples son árboles. Como se ha mencionado anteriormente, la idea principal de *bagging* es promediar muchos modelos con ruido pero poco sesgados, para reducir la varianza. Los árboles son candidatos ideales para ello, ya que pueden capturar estructuras complejas de interacción en los datos, y si crecen a suficiente profundidad, tienen un sesgo relativamente bajo.

Una característica del *random forest* es que al construir cada árbol de decisión, cada vez que se considera una división en un árbol, se elige una muestra aleatoria de m predictores como candidatos a la división del conjunto total de p predictores. Esto evita que los árboles sean similares entre sí y por lo tanto, evita que las predicciones de los árboles agregados estén altamente correlacionadas. Al promediar las predicciones se conseguirá una mayor reducción de la varianza si no están correlacionadas.

En las técnicas de *bagging*, los árboles de decisión se entrenan tomando un subconjunto de la muestra total de entrenamiento. Por lo que por cada muestra tomada de los datos de entrenamiento, quedarán muestras que no se incluyeron. Estas muestras se denominan *Out-Of-Bag (OOB)*. El error estimado de las muestras *Out-Of-Bag* proporciona una buena estimación del error que se puede esperar en un conjunto de evaluación nuevo, debido a que estas muestras no han sido incluidas en el conjunto de entrenamiento del modelo.

Una de la desventajas de los *random forest* es la pérdida de la interpretabilidad. Al construirse muchos árboles y obtener un promedio de las predicciones, no se puede visualizar el modelo tal y como ocurría con los árboles de decisión simples. Sin embargo, los algoritmos *random forest* devuelven una salida con la importancia de las variables predictoras. Esto ayuda a interpretar de los resultados.

Los principales hiperparámetros a ajustar son el número de estimadores, es decir, el número de árboles, la profundidad máxima del árbol, el número

mínimo de muestras necesarias para dividir un nodo, el número mínimo de muestras requerido para un nodo terminal, el número de predictores considerado en cada división del árbol y si cada árbol se entrena o no con la muestra de entrenamiento completa. Para ajustarlos, se han seleccionado varios valores para cada hiperparámetro y se ha entrenado un modelo para cada combinación de dichos valores, obteniendo así los valores para los cuales el modelo tiene mayor capacidad de generalización. Para determinar la capacidad de generalización, se crean particiones del conjunto de datos, de forma que se tiene un conjunto de datos para entrenar y otro para validar el modelo que se llamará *muestra de validación*. Realizar estas particiones reduce notablemente el conjunto utilizado para el aprendizaje del modelo. Para evitar esto, se utiliza la técnica de *validación cruzada*.

La técnica de *validación cruzada con k-fold* consiste en dividir la muestra en k particiones y realizar k iteraciones de tal manera que en cada una de ellas se considere como muestra de validación una de las particiones. Así, se garantiza que todos los datos del conjunto aparezcan en la muestra de validación y que las muestras de validación nunca se solapen. Para determinar el desempeño del modelo se realiza el promedio de las métricas de validación obtenidas de cada una de las iteraciones.

A continuación en la Figura 3.1 se muestra un esquema de un modelo *random forest* con N árboles:

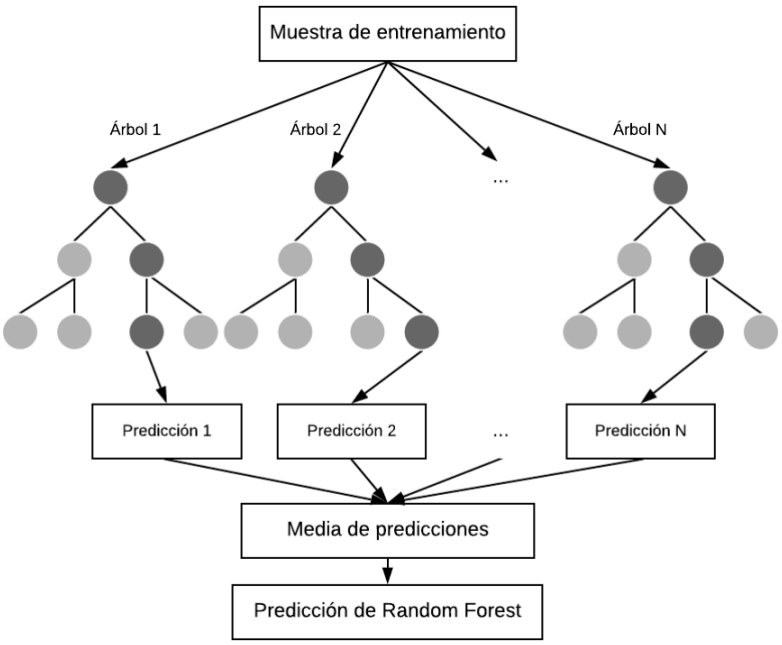

Figura 3.1: Esquema de un modelo *Random Forest* con N árboles.

Para la implementación del algoritmo *random forest* se ha utilizado la

librería *scikit-learn* de Python.

3.2. Redes neuronales

Las *redes neuronales artificiales* son técnicas populares de aprendizaje automático que simulan el mecanismo de aprendizaje de los organismos biológicos. Se aplican tanto a problemas de clasificación como de regresión.

El elemento fundamental de una red neuronal es una *neurona*. Su función es recibir los datos de entrada y realizar una suma ponderada por unos pesos para producir una salida.

Para evitar que el modelo solo sea capaz de aprender relaciones lineales, se añade una *función de activación*, que es una transformación no lineal que se aplica a las salidas de las neuronas. En este trabajo se ha utilizado la función de activación *ReLU* (*Rectified Linear Unit*) que se define de la siguiente manera:

$$f(x) = max(0, x)$$

La función *ReLU* no es derivable, en cambio una de las ventajas que presenta es que su derivada es constante en todos los puntos en los que está definiada, ya que para $x > 0$ su derivada es 1. Por esto, es eficiente computacionalmente.

Agrupando un conjunto de neuronas se forma una *capa*. A la primera capa se le llama capa de entrada y a la última capa de salida. A las capas situadas entre estas se les denomina capas ocultas. El número de neuronas de la capa de entrada viene dado por el número de variables predictoras consideradas en el problema. El número de neuronas de la capa de salida dependerá del tipo de problema, en este caso, al ser un problema de regresión con una etiqueta, tendrá una neuronas en la capa de salida.

El objetivo de una red neuronal es capturar las relaciones entre los datos y para ello ajusta los pesos que conectan las neuronas. Este proceso de ajustar los pesos se llama *entrenamiento*. El entrenamiento consta de dos fases: *propagación hacia delante* y *propagación hacia atrás* o *backpropagation*.

En la *propagación hacia delante*, el modelo obtiene una predicción para los datos de entrenamiento con los pesos actuales avanzando en la red desde la capa de entrada hasta la capa de salida. Una vez obtenidas dichas predicciones, se comparan con los valores reales para obtener el error. En la *propagación hacia atrás*, el error se propaga desde la capa de salida hacia todas las capas que contribuyen a la salida del modelo. A cada neurona se le asigna una fracción del error total, en función de su contribución a la salida. Este proceso se repite hasta que cada neurona tenga asignado un error. Una vez que se tienen dichos errores asignados, los pesos se van actualizando utilizando un algoritmo de optimización [26].

El algoritmo de optimización utilizado en este trabajo es el *optimizador Adam*. El *optimizador Adam* se basa en los métodos de descenso del gradiente que consisten en encontrar el mínimo local de la función de pérdida, es decir, de la función de error definida. Para ello, se actualizan los pesos en dirección contraria al gradiente de esta función. Un parámetro a ajustar en el descenso del gradiente es el *ratio de aprendizaje*, α, que determina el tamaño del paso en cada iteración mientras se mueve hacia un mínimo. La diferencia que presenta el optimizador *Adam* respecto al resto de métodos basados en el descenso del gradiente es que este calcula ratios de aprendizaje adaptativos individuales para cada uno de los parámetros [25].

Existen muchas arquitecturas para la configuración de las redes neuronales. Es difícil escoger el número de capas ocultas y el número de neuronas en cada una de ellas. Si el número de neuronas es muy pequeño, la red podría no tener la suficiente flexibilidad para capturar relaciones de los datos. En cambio, si el número de neuronas es muy alto, el modelo podría ser muy complejo y podría sobreajustar. Para evitar el sobreajuste se aplican distintas técnicas llamadas *técnicas de regularización*.

Una de las ténicas de regularización empleadas en estre trabajo es la técnica de *apagado de neuronas* o *dropout*. La técnica de *apagado de neuronas* es una técnica de *bagging*, explicada en el apartado 3.1, que consiste en añadir una capa que apague un número fijo de neuronas aleatorias durante el entrenamiento de la red. Para ello, se multiplican por 0 los valores de salida de dichas neuronas.

Es importante también configurar el tamaño de los lotes de entrenamiento. Cuando las muestras de entrenamiento son grandes y se toma el conjunto entero de datos para entrenar el modelo, se actualizan los pesos una vez recorrida la muestra completa. De esta manera, cada iteración puede tardar mucho tiempo. En cambio, si se actualizan los parámetros para cada muestra del conjunto, cada iteración será mucho más rápida pero puede dar problemas de convergencia en regiones grandes. Por lo tanto, es aconsejable dividir la muestra de entrenamiento en lotes de tamaño intermedio. Es recomendable que los tamaños de los lotes sean potencias de dos por como funciona el acceso a memoria en los ordenadores.

Por otro lado, es necesario escalar los datos para que las variables predictoras del problema tengan magnitudes comparables. De esta manera, se evitarán los sesgos durante el aprendizaje hacia las variables de mayor magnitud. Por lo tanto, es conveniente estandarizar las entradas de manera que estas tengan media cero y desviación estándar uno.

Las *redes perceptrón multicapa* son redes neuronales compuestas por al menos una capa oculta en las que la información únicamente avanza hacia delante, no admiten ciclos. Además, todas las neuronas de la capa anterior están conectadas con las neuronas de la siguiente capa. A continuación, en la Figura 3.2, se muestra un esquema de una red perceptrón multicapa, que es el tipo de red escogida para este trabajo:

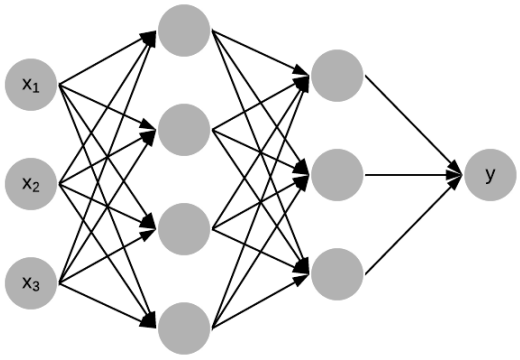

Figura 3.2: Esquema de una red perceptrón multicapa con dos capas ocultas.

Para la implementación de la red perceptrón multicapa se ha utilizado la librería *keras* de *Python*.

3.3. Entrenamiento

En el entrenamiento de los modelos se han considerado, tal y como se explica en el apartado 2.4.2, las muestras de entrenamiento de tal manera que se utilizan los 4 días anteriores para predecir lo que ocurre con la variable objetivo ese día.

El objetivo de los modelos de aprendizaje automático es que sean capaces de generalizar los resultados, es decir, que el modelo no pierda capacidad predictiva cuando la muestra es nueva y no ha sido utilizada en el entrenamiento. Para poder comprobar la capacidad de generalización del modelo, se utiliza una partición del conjunto de datos llamada *muestra de entrenamiento* o *train* para entrenar el modelo y otra partición del conjunto de datos llamada *muestra de evaluación* o *test* únicamente para la evaluación del modelo.

En este estudio la división de los datos se ha hecho de manera aleatoria utilizando el 80 % del conjunto total como muestra de entrenamiento y el 20 % restante como muestra de evaluación.

3.4. Evaluación

Una vez entrenados los modelos, se evalúa su rendimiento. Para ello, se comparan diferentes métricas en la muestra de entrenamiento como de evaluación.

Al ser un problema de regresión, las métricas elegidas para la evaluación son el error absoluto medio, el error cuadrático medio, la raíz del error cuadrático medio que se explican a continuación. Para ello, se denotan por

$y_1, y_2, ..., y_N$ los valores reales, por $\hat{y}_1, \hat{y}_2, ..., \hat{y}_N$ los valores predichos y por N el número total de observaciones.

El *error absoluto medio*, EAM, es la media de la diferencia absoluta entre el valor real y el valor predicho por el modelo. Se calcula de la siguiente manera:

$$EAM = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j|$$

El *error cuadrático medio*, ECM, es la media de la diferencia cuadrada entre el valor real y el valor predicho por el modelo de regresión. Viene dado por:

$$ECM = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

La *raíz del error cuadrático medio*, $RECM$, es la raíz cuadrada del error cuadrático medio, es decir, es la raíz cuadrada de la media de la diferencia cuadrada entre el valor real y el valor predicho por el modelo. Se define de la siguiente manera:

$$RECM = \sqrt{\frac{\sum_{j=1}^{N}(y_j - \hat{y}_j)^2}{N}}$$

El error absoluto medio es robusto a valores atípicos, por lo que no penaliza tanto los errores como lo puede hacer el error cuadrático medio, ya que este al elevar al cuadrado los errores penaliza incluso los errores más pequeños. La raíz del error cuadrático medio penaliza mucho los errores grandes.

Capítulo 4

Resultados

4.1. Configuración de modelos

Antes del entrenamiento de los modelos, es importante realizar una *selección de variables*. La *selección de variables* es el proceso a través del cual se elige un subconjunto de las variables predictoras que más contribuyen a la predicción de la variable objetivo del problema. De esta forma, se eliminan las variables irrelevantes que pueden empeorar el desempeño del modelo. Así, se consigue reducir el sobreajuste de los modelos, mejorar su rendimiento y reducir el tiempo de entrenamiento.

En las variables predictoras que se tienen en este problema, hay variables que se mantienen estáticas a lo largo del tiempo. Son variables que durante las fases de confinamiento se han mantenido iguales que en estado normal y se ha comprobado que los modelos resultantes incluyendo estas variables no varían respecto a los modelos entrenados eliminándolas. Por lo tanto, se han eliminado las variables *población total*, *densidad*, *población hombres*, *población >60*, *población hogar >4 personas* y *población residencias mayores*.

Entonces, la selección de variables resultante para el entrenamiento de los modelos consta de las 11 variables dinámicas que son: *población matriculada no uni*, *población matriculada uni*, *ocupados agricultura*, *ocupados industria*, *ocupados construcción*, *ocupados servicios*, *transporte en metro*, *transporte autobús urbano*, *viajeros residencia España*, *viajeros residencia extranjero* y *movilidad laboral*.

Una vez seleccionadas las variables, se ha aplicado el método de ventanas deslizantes tal y como se explica en el apartado 2.4.2 y las muestras resultantes constan de 59 variables de entrada y una de salida. Se han entrenado un modelo *random forest* y una red perceptrón multicapa para cada una de las variables objetivo. Para el ajuste de los hiperparámetros de los modelos random forest, tal y como se ha mencionado en el apartado 3.1.1, se han probado diferentes valores para cada uno de los hiperparámetros a ajustar, con una validación cruzada de 5 *folds* y finalmente se han entrenado los

siguientes modelos:

Variable objetivo $\mathcal{R}_0$:

- 300 árboles.
- No tiene profundidad máxima.
- El número mínimo de muestras para dividir un nodo es 3.
- El número mínimo de muestras requerido para un nodo terminal es 2.
- 45 predictores considerados en cada división del árbol.
- Cada árbol no se entrena con la muestra completa.

Variable objetivo MPI:

- 300 árboles.
- Profundidad máxima 12.
- El número mínimo de muestras para dividir un nodo es 3.
- El número mínimo de muestras requerido para un nodo terminal es 1.
- 45 predictores considerados en cada división del árbol.
- Cada árbol no se entrena con la muestra completa.

La arquitectura de la red neuronal perceptrón entrenada para las dos variables objetivo es la siguiente:

- Capa de entrada con 64 neuronas.
- Capa oculta de 256 neuronas con función de activación *ReLU*.
- Capa de apagado de neuronas con probabilidad de apagado de 0,1.
- Capa oculta de 128 neuronas con función de activación *ReLU*.
- Capa de apagado de neuronas con probabilidad de apagado de 0,1.
- Capa oculta de 64 neuronas con función de activación *ReLU*.
- Capa de apagado de neuronas con probabilidad de apagado de 0,1.
- Capa de salida con 1 neurona con función de activación *ReLU*.

Las redes han sido entrenadas durante 400 épocas, es decir, se recorre el conjunto de entrenamiento 400 veces. Antes de entrenar la red, el conjunto de datos ha sido estandarizado para obtener una media 0 y desviación estándar 1. El tamaño de los lotes de entrenamiento es de 64 y se ha empleado un optimizador *Adam* con parámetros de $\alpha = 0{,}0001$, $\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$ y $\epsilon = 10^{-8}$. En el apartado 3.3, se ha fijado que los modelos se entrenan con el 80 % del conjunto de datos. Sin embargo, para entrenar las redes y comprobar si sobreajustan se ha utilizado el 10 % de esa partición para validar el modelo.

4.2. Evaluación de modelos

Los modelos entrenados se han evaluado con las métricas mencionadas en el apartado 3.4 tanto en el conjunto de entrenamiento como en el de evaluación. Los resultados para los modelos *random forest* para cada una de las variables objetivo son las siguientes:

		EAM	**ECM**	**RECM**
$\mathcal{R}_0$	Entrenamiento	0,0492	0,0127	0,1125
	Evaluación	0,1137	0,0382	0,1954
MPI	Entrenamiento	0,3970	0,3912	0,6254
	Evaluación	1,0206	2,5698	1,6031

Tabla 4.1: Métricas obtenidas con *random forest* para cada variable objetivo en los conjuntos de entrenamiento y evaluación.

En la Tabla 4.1 se puede observar que el error en el conjunto de evaluación tanto para la variable objetivo $\mathcal{R}_0$ como para la variable objetivo MPI es superior al error en el conjunto de entrenamiento. Esto indica que los modelos pierden la capacidad de generalizar en muestras nuevas que no han sido utilizadas en el entrenamiento. Es decir, sufren ligeramente de sobreajuste.

Al comprobar las variables que más influyen en los modelos *random forest* en el valor de las variables de salida, es decir, las más importantes, coinciden en ambos casos con las variables objetivo en $t-1$, $t-2$, $t-3$ y $t-4$ en ese orden. Esto se debe a que tal y como se ha explicado en el apartado 2.4, las observaciones de una serie temporal no son independientes entre sí, sino que las observaciones pasadas condicionan las observaciones futuras.

Por otro lado, los resultados de las métricas obtenidas para las redes perceptrón multicapa entrenadas para cada una de las variables objetivo son las siguientes:

		EAM	**ECM**	**RECM**
$\mathcal{R}_0$	Entrenamiento	0,1042	0,0236	0,1537
	Evaluación	0,1184	0,0320	0,1789
MPI	Entrenamiento	0,7836	1,3367	1,1562
	Evaluación	0,9770	2,2129	1,4876

Tabla 4.2: Métricas obtenidas con la red perceptrón multicapa para cada variable objetivo en los conjuntos de entrenamiento y evaluación.

En la Tabla 4.2, se puede observar que las diferencias de los errores entre el conjunto de entrenamiento y de evaluación son menores que los obtenidos por lo modelos *random forest*, por lo que tienen mayor capacidad de generalización. A continuación se muestran los progresos de las funciones de

pérdida en los conjuntos de entrenamiento y validación para ambos modelos a lo largo de las épocas donde se ve que los valores que toma la función de pérdida en los conjuntos de entrenamiento y validación son similares:

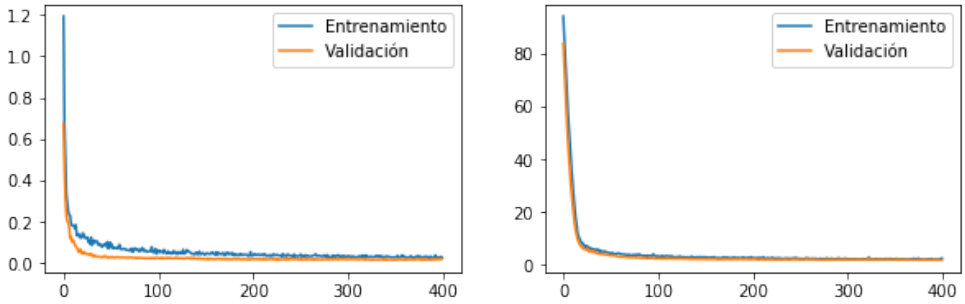

Figura 4.1: Evolución de las funciones de perdida. A la izquierda del modelo con variable objetivo $\mathcal{R}_0$ y a la derecha del modelo con variable objetivo MPI.

En las siguientes imágenes se muestran las predicciones de la curva de las primeras 6 provincias de $\mathcal{R}_0$ de ambos modelos:

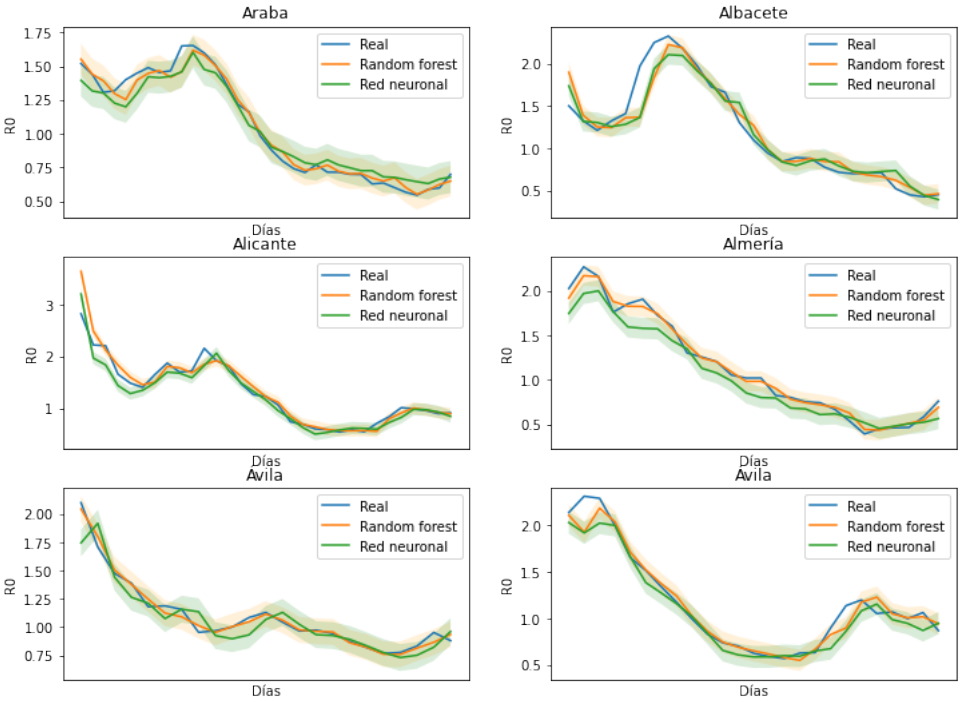

Figura 4.2: Predicción de $\mathcal{R}_0$ con ambos modelos para las 6 primeras provincias. En sombreado, la $predicción \pm EAM$ de cada modelo.

A continuación se muestran las predicciones de MPI de ambos modelos para las primeras 6 provincias:

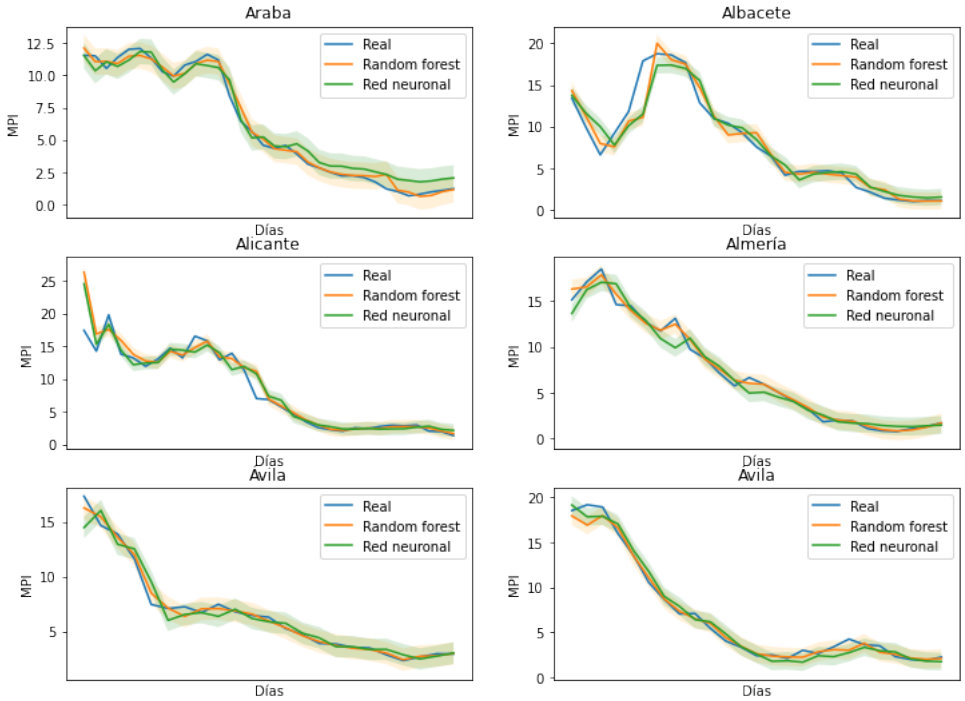

Figura 4.3: Predicción de MPI con ambos modelos para las 6 primeras provincias. En sombreado, la *predicción* $\pm$ EAM de cada modelo.

Destacar que en la predicciones que se muestran en las Figuras 4.2 y 4.3, algunos de los datos han sido incluidos en el entrenamiento y otros no.

Se ha comprobado que los modelos de las redes perceptrón multicapa sufren menor sobreajuste que los modelos *random forest*, por lo que se han utilizado estos modelos para la evaluación de las medidas de confinamiento. Además, se ha elegido la variable objetivo $\mathcal{R}_0$ ya que obtiene mayor capacidad de generalización por presentar menores diferencias entre los errores de los conjuntos de entrenamiento y evaluación. Para ello, se ha entrenado el modelo de nuevo con la muestra completa.

En la Tablas A.1 y A.2 se recogen los errores obtenidos con dicho modelo para cada una de las provincias y se puede observar que el error del modelo no es homogéneo en todas las provincias.

Provincias como Madrid y Barcelona obtienen errores altos en comparación con el resto. Esto puede darse porque son provincias con mucha población. Los valores de las variables predictoras en situación normal son más elevados en estas provincias. Por lo tanto, son casos en los que los datos son muy diferentes al resto de provincias y la red neuronal tienen menos

muestras similares de las que aprender. En cambio, para las provincias con tamaño de población intermedio hay más datos similares y el error obtenido para dichas provincias es menor.

Un ejemplo de este caso es Araba, que es una provincia de tamaño de población intermedio. Además, Araba fue una de las provincias que acumuló más casos en un primer momento, por lo que hay más días registrados para dicha provincia. y por tanto hay más datos en lo que se puede basar la red para aprender. Esto explica que el error obtenido para esta provincia sea más bajo que para el resto.

4.3. Evaluación de medidas de confinamiento

El objetivo del trabajo es, como se ha explicado anteriormente, evaluar las medidas de confinamiento que más han afectado a la propagación del virus por provincias. Para ello, se plantea modificar a partir del último día en el que hay datos registrados las medidas adoptadas. Se suponen nuevos escenarios en los que en cada uno de ellos se levanta una medida y se observa mediante los modelos entrenados como afectan los escenarios supuestos a las variables objetivo.

Los modelos predicen lo que va a ocurrir en un tiempo futuro de un día. Para poder evaluar las medidas, es interesante que puedan predecir a más días vista, para comprobar qué ocurre una vez que ya no haya datos registrados.

Por lo tanto, a partir del último día que se tienen los datos, se modifican las variables predictoras en función de los nuevos escenarios supuestos y se predice con los últimos 4 días registrados, $t-4, t-3, t-2, t-1$ lo que va a ocurrir con el siguiente día t en función de dichas variables predictoras. Una vez predicho el valor de la variable objetivo en t, se predice con los valores de las variables en $t-3, t-2, t-1, t$ el valor en $t+1$. Es decir, el valor en t se considera como el valor en $t-1$. Se realiza esto sucesivamente hasta llegar al número de días vista que se quiera obtener.

Para calcular el error que se comete prediciendo de esta manera, se ha hecho esta predicción para cada valor observado. Es decir, para cada uno de los datos registrados, se ha calculado la predicción en $t, t+1, t+2, t+3...$ actualizando el conjunto de entrada al modelo como se ha explicado. Una vez obtenidos dichos valores, se han comparado con los valores reales. En el esquema siguiente se muestra como para cada punto t de la curva de una provincia, se calcula la predicción en $t, t+1, t+2$ y $t+3$.

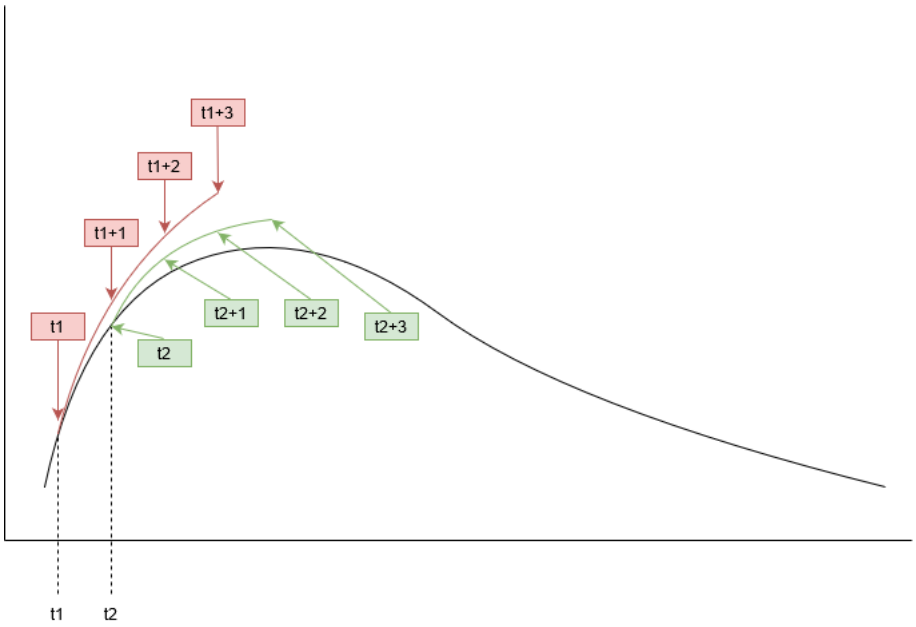

Figura 4.4: Esquema de predicción para cada punto t de la curva de los valores t, $t+1$, $t+2$ y $t+3$.

A continuación se muestran los valores de los errores obtenidos prediciendo la variable objetivo $\mathcal{R}_0$ a 6 días vista tal y como se ha explicado para la primera provincia:

	t	t+1	t+2	t+3	t+4	t+5
EAM	0,0446	0,0711	0,0965	0,1228	0,1411	0,1463
ECM	0,0037	0,0093	0,0154	0,0233	0,0300	0,0333
RECM	0,0609	0,0966	0,1240	0,1527	0,1731	0,1825

Tabla 4.3: Errores obtenidos prediciendo a 6 días vista tal y como se ha explicado para la primera provincia.

Se puede observar en la Tabla 4.3, que los errores van aumentando a medida que se predice a más días vista. Esto se debe a que la predicción se va realizando en base a predicciones anteriores y no a valores reales. Se ha considerado un máximo de 4 días vista, es decir, hasta $t+3$. Así, los valores en los que se basa para predecir no son todos predichos.

Para evaluar cuáles han sido las medidas que más impacto han tenido a la hora de frenar la propagación del virus, se han creado diferentes escenarios y se ha comprobado como han variado las variables objetivo en función del escenario propuesto. El último día en el que había datos observados es el 9 de abril de 2020, por lo que en cada uno de los escenarios considerados se levanta una medida de confinamiento distinta a partir de este día. Es decir,

se predice lo que ocurriría desde el 10 de abril hasta el 13 de abril. Para ello, se propone crear 4 escenarios distintos en los que en cada uno de ellos se levanta una medida, es decir, se restablece el valor de una de las variables predictoras a su valor en situación normal. La escenarios se han considerado de tal manera que en cada uno de ellos se levante una medida de distinto ámbito de los 4 mencionados en el apartado 2.1.

- *Escenario 1:* Se levanta la medida de que los estudiantes universitarios acudan a las aulas y se mantienen el resto de restricciones.
- *Escenario 2:* Se levanta la medida de que los ocupados en la industria retomen su actividad y se mantienen el resto de restricciones.
- *Escenario 3:* Se levanta la medida de aforo en los autobuses y se mantienen el resto de restricciones.
- *Escenario 4:* Se levanta la medida de que los residentes en España no puedan viajar dentro del estado y se mantienen el resto de restricciones.

En la siguiente tabla se recoge a qué porcentaje se reducen las variables predictoras en función de las medidas de confinamiento propuestas en cada escenario:

	Escenario 1	*Escenario 2*	*Escenario 3*	*Escenario 4*
Poblacion matriculada no uni	0	0	0	0
Poblacion matriculada uni	100	0	0	0
Ocupados agricultura	90	90	90	90
Ocupados industria	20	100	20	20
Ocupados construccion	10	10	10	10
Ocupados servicios	20	20	20	20
Transporte en metro	10	10	10	10
Transporte autobús urbano	10	10	100	10
Viajeros residencia España	0	0	0	100
Viajeros residencia extranjero	0	0	0	0
Movilidad laboral	40	40	40	40

Tabla 4.4: Porcentajes a los que se reducen cada variable en función de los escenarios.

En la Figura A.1, se visualizan las predicciones obtenidas para la variable $\mathcal{R}_0$ en función de los escenarios propuestos para 8 provincias de España. Además, se observan los valores reales para los 4 días anteriores para poder detectar cuál era la tendencia del valor de $\mathcal{R}_0$ hasta el momento en el que se aplican los cambios propuestos. Las provincias han sido seleccionadas por densidad siendo las 4 primeras las provincias menos densas y las 4 últimas las más densas.

Se puede observar en la imagen que en general, si se levanta una única medida en las provincias menos densas, el impacto es menor que en las más densas. En las provincias menos densas, el levantamiento de ninguna de las medidas parece que implique un repunte muy alto de la variable objetivo, en muchas de ellas la variable sigue la tendencia decreciente o constante que llevaba anteriormente.

Dos de las provincias de las 4 más densas son Ceuta y Melilla, pero hay que tener en cuenta que estas son casos extremos por ser ciudades autónomas. Las dos otras provincias más densas mostradas son Barcelona y Madrid. En estos casos se ve que el levantamiento de una medida si que hace aumentar notablemente el valor de la variable objetivo. Por ejemplo, el escenario 1, es decir, la vuelta de los universitarios a la aulas, afecta de manera significativa al aumento del valor de $\mathcal{R}_0$ en ambas.

En las Figuras A.2 y A.3, se pueden visualizar de la misma manera las predicciones obtenidas para la variable $\mathcal{R}_0$ en función de los escenarios propuestos y la tendencia de la curva hasta el momento para todas las provincias de España. Las provincias están ordenadas de izquierda a derecha y de arriba a abajo de menor a mayor población.

Se puede observar que en las provincias de mayor población una de las medidas que afecta a un mayor repunte del valor de $\mathcal{R}_0$ es el escenario 4, es decir, que se permita viajar dentro de España. Se puede ver que esta medida afecta especialmente a provincias como Girona, Tarragona, Granada Asturias, Baleares, Cádiz, Málaga y Alicante, que pueden ser provincias en las que el turismo nacional tenga gran peso.

Sin tener en cuenta los casos de Ceuta y Melilla, que como se ha mencionado son casos extremos, las provincias con menor población, al igual que las menos densas, sufren menos el levantamiento de una única medida. Las medidas de manera aislada no parecen tener un gran impacto a la hora de frenar la propagación del virus en estas provincias.

Las provincias de Madrid y Barcelona son las provincias que mayor población tienen en las 4 variables tomadas en cuenta. Por esto, se ve que el levantamiento de cualquiera de las 4 medidas hace que el aumento de $\mathcal{R}_0$ sea considerable en cualquiera de los 4 escenarios presentados.

La provincia de Valencia es la tercera provincia que mayor número de ocupados en industria tiene después de Barcelona y Madrid. Esto se ve reflejado en que el levantamiento de esa medida provoca el mayor repunte de los 4 escenarios en Valencia.

Hay provincias como la de Cantabria, que se puede observar que en un primer momento algunas medidas como el escenario 3 aumentan la variable objetivo y luego disminuye ligeramente en $t+3$. Esto puede indicar que aunque en un primer momento el levantamiento de esta medida pueda generar un pequeño repunte, dicha medida no tiene gran impacto en la propagación de la pandemia en Cantabria ya que eliminarla no genera un aumento notable de los casos.

En el caso de Zaragoza, que es la tercera provincia con mayor número de viajeros de autobús urbano después de Barcelona y Madrid, no es la medida que al eliminarla más afecta a esta provincia. Esto se contradice con lo se puede esperar en un primer momento. Puede ser causa de las limitaciones que puede presentar el modelo o del error que comete. Dichas limitaciones pueden ser debidas a los problemas presentados a lo largo del trabajo.

Tal y como se ha mostrado con la variable objetivo $\mathcal{R}_0$, se pueden utilizar los modelos para predecir como afectan los diferentes escenarios a cualquiera de las dos variables objetivo del problema a cada una de las provincias. Además, se pueden adoptar los escenarios deseados para ver como varían.

Capítulo 5

Conclusiones

En este trabajo se han desarrollado varios modelos para predecir la evolución de la enfermedad COVID-19 que se pueden aplicar a los datos de cada provincia. Para ello, se ha realizado el preprocesado de los datos necesario teniendo en cuenta las propiedades de la enfermedad y a partir del número de casos de contagios diarios se han calculado como variables objetivo del problema el número básico reproductivo y el porcentaje de incremento de casos. Se han estudiado propiedades de las series temporales necesarias para la comprensión de los datos.

Los modelos entrenados para el estudio son un modelo *random forest* y una red perceptrón multicapa para cada una de las variables objetivo. Los hiperparámetros de los modelos *random forest* se han ajustado mediante validación cruzada. Para el entrenamiento de los modelos se han utilizado el 80 % de los datos y para la evaluación el 20 % y una vez entrenados los modelos, se han evaluado teniendo en cuenta el error absoluto medio, el error cuadrático medio y la raíz del error cuadrático medio.

Las redes perceptrón multicapa en general han presentado un mejor desempeño que los modelos *random forest* ya que han obtenido mayor capacidad de generalización, es decir, menor sobreajuste. En concreto, la mayor capacidad de generalización ha sido obtenida por la red perceptrón multicapa para la variable objetivo del número básico reproductivo. Por esta razón, se ha utilizado este modelo entrenado con la muestra completa para evaluar cómo afectan las medidas de distanciamiento social en la propagación de la enfermedad.

Se han propuesto diferentes escenarios en los que en cada uno de ellos se levanta una medida de confinamiento diferente a partir del último día que se tienen datos para ver en qué casos aumentan más la variable objetivo. Esto podría realizarse para todos los escenarios posibles.

En general, se ha detectado que la restricción del turismo en territorio nacional afecta más a las provincias de mayor población. Además, se ha observado que para las provincias menos densas o con menor población el

levantamiento de las medidas de manera aislada no supone un repunte en la variable objetivo y que la curva sigue con la tendencia constante o descendiente que llevaba anteriormente. También se ha comprobado que en algunos casos, las provincias con magnitudes mayores en las variables restablecidas sufrían un mayor repunte del número reproductivo de casos. Sin embargo, también se han presentado casos en los que eso no ocurre, pudiendo ser esto una indicación de que el modelo tiene ciertas limitaciones.

A la hora de realizar el trabajo se han encontrado ciertas dificultades. Uno de los principales problemas ha sido que durante el inicio de la pandemia muchos casos no fueron detectados. En un principio no se conocían los síntomas que causaba la enfermedad y se confundían los casos con otras enfermedades. Además, los casos asintomáticos no fueron contabilizados. Por otro lado, en un primer momento se priorizó diagnosticar los casos más graves y no se realizaron las pruebas suficientes a casos más leves por no haber recursos necesarios. Por esto, se sabe que hay un inframuestreo en los datos. En cambio, no se sabe si este inframuestreo ha afectado igual a todas las provincias. Esto puede verse reflejado en que dos provincias con características similares podrían haber tenido una progresión similar de la pandemia y que no se vea plasmado en los datos.

Por otro lado, se han encontrado grandes diferencias entre los datos proporcionados por las diferentes comunidades autónomas. Algunas comunidades han publicado los datos diariamente en sus portales de datos de manera clara, accesible y distinguidos por las provincias que las componen. Los datos de otras comunidades, en cambio, solo se podían obtener consultando notas de prensa diarias lo que dificultaba la localización de los mismos. Además, presentaban en ocasiones datos faltantes y algunos de los datos no se asignaban a ninguna de sus provincias lo que ha hecho que se pierdan estos datos.

Se quiere destacar en este trabajo la importancia de que los datos en abierto cumplan los principios *FAIR*, es decir, que sean localizables, accesibles, interoperables y reutilizables para optimizar el estudio de los mismos.

Otra de las dificultades presentadas a lo largo del trabajo es el desconocimiento sobre la enfermedad COVID-19. Es una enfermedad que no ha existido antes y se está investigando constantemente sus características como el tiempo de incubación.

Como trabajo futuro se plantea obtener un conjunto de datos mayor en el que se recojan los datos a lo largo de más días, teniendo en cuenta tanto los test rápidos como las pruebas PCR. Además, se propone tener en cuenta todas los avances respecto a la enfermedad para la comprensión de los datos. Por último, se podría estudiar la aplicación de otros modelos como pueden ser las redes neuronales recurrentes, que son un tipo de red neuronal con un estado interno llamado memoria que permite tomar como entrada secuencias de datos y modelizar la variabilidad temporal entre ellos.

Apéndice A

Tablas y gráficas

	EAM	**ECM**	***RECM***
Araba	0,0437	0,0034	0,0586
Albacete	0,0967	0,0202	0.142
Alicante	0,0933	0,0159	0,1262
Almería	0,1091	0,0173	0,1316
Ávila	0,0786	0,0106	0,1030
Badajoz	0,0816	0,0127	0,1127
Balears, Illes	0,0921	0,0162	0,1274
Barcelona	0,1866	0,0576	0,2400
Burgos	0,0647	0,0068	0,0825
Cáceres	0,1190	0,0202	0,1420
Cádiz	0,0770	0,0098	0,0992
Castellón	0,0986	0,0180	0,1341
Ciudad Real	0,0723	0,0089	0,0942
Córdoba	0,1110	0,0157	0,1252
Coruña, A	0,0681	0,0086	0,0927
Cuenca	0,1506	0,0356	0,1887
Girona	0,0623	0,0057	0,0753
Granada	0,0529	0,0051	0,0715
Guadalajara	0,1016	0,0158	0,1259
Gipuzkoa	0,0560	0,0058	0,0760
Huelva	0,0821	0,0137	0,1168

Tabla A.1: Errores obtenidos por la red perceptrón multicapa entrenada con toda la muestra en la predicción de $\mathcal{R}_0$ para cada provincia de España (parte I).

	EAM	**ECM**	**RECM**
Huesca	0,1181	0,0197	0,1403
Jaén	0,0610	0,0053	0,0726
León	0,0733	0,0140	0,1182
Lleida	0,0603	0,0056	0,0746
Rioja, La	0,0964	0,0200	0,1415
Lugo	0,1466	0,0590	0,2429
Madrid	0,1842	0,0501	0,2239
Málaga	0,0709	0,0070	0,0837
Murcia	0,1159	0,0189	0,1373
Navarra	0,0452	0,0031	0,0556
Ourense	0,1219	0,0222	0,1491
Asturias	0,0569	0,0052	0,0722
Palencia	0,0572	0,0069	0,0829
Palmas, Las	0,0533	0,0051	0,0712
Pontevedra	0,0632	0,0062	0,0790
Salamanca	0,0761	0,0094	0,0968
Santa Cruz de Tenerife	0,0691	0,0087	0,0934
Cantabria	0,0757	0,0135	0,1164
Segovia	0,0689	0,0107	0,1032
Sevilla	0,0670	0,0064	0,0798
Soria	0,0979	0,0135	0,1163
Tarragona	0,0576	0,0051	0,0714
Teruel	0,1057	0,0170	0,1304
Toledo	0,0874	0,0113	0,1064
Valencia	0,1512	0,0622	0,2495
Valladolid	0,0745	0,0089	0,0946
Bizkaia	0,0488	0,0040	0,0632
Zamora	0,1446	0,0366	0,1912
Zaragoza	0,0612	0,0065	0,0805
Ceuta	0,1573	0,0372	0,1928
Melilla	0,1043	0,0166	0,1288

Tabla A.2: Errores obtenidos de la red perceptrón multicapa entrenada con toda la muestra en la predicción de $\mathcal{R}_0$ para cada provincia de España (parte II).

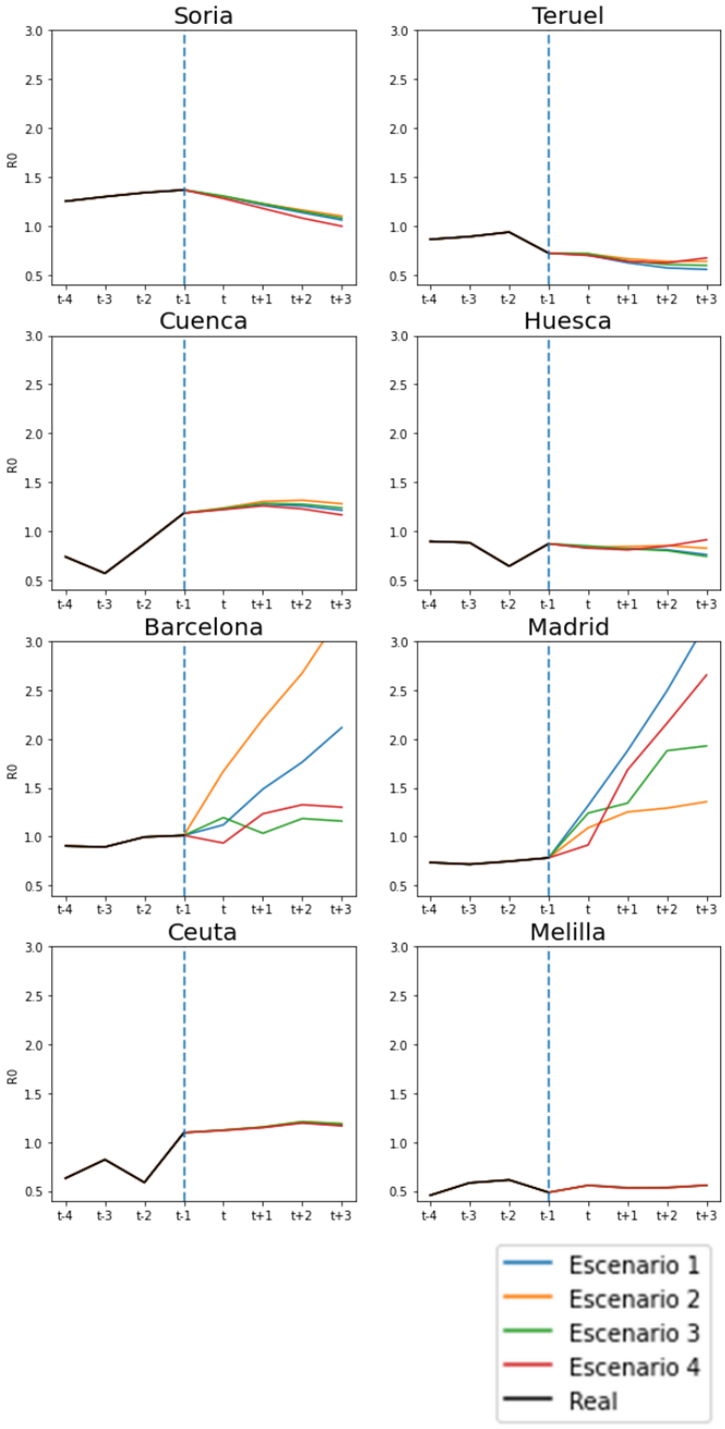

Figura A.1: Valores reales de $\mathcal{R}_0$ de los 4 días anteriores y predicciones a 4 días vista de las 4 provincias menos densas (parte de arriba) y más densas (parte de abajo) en función de los escenarios de la Tabla 4.4.

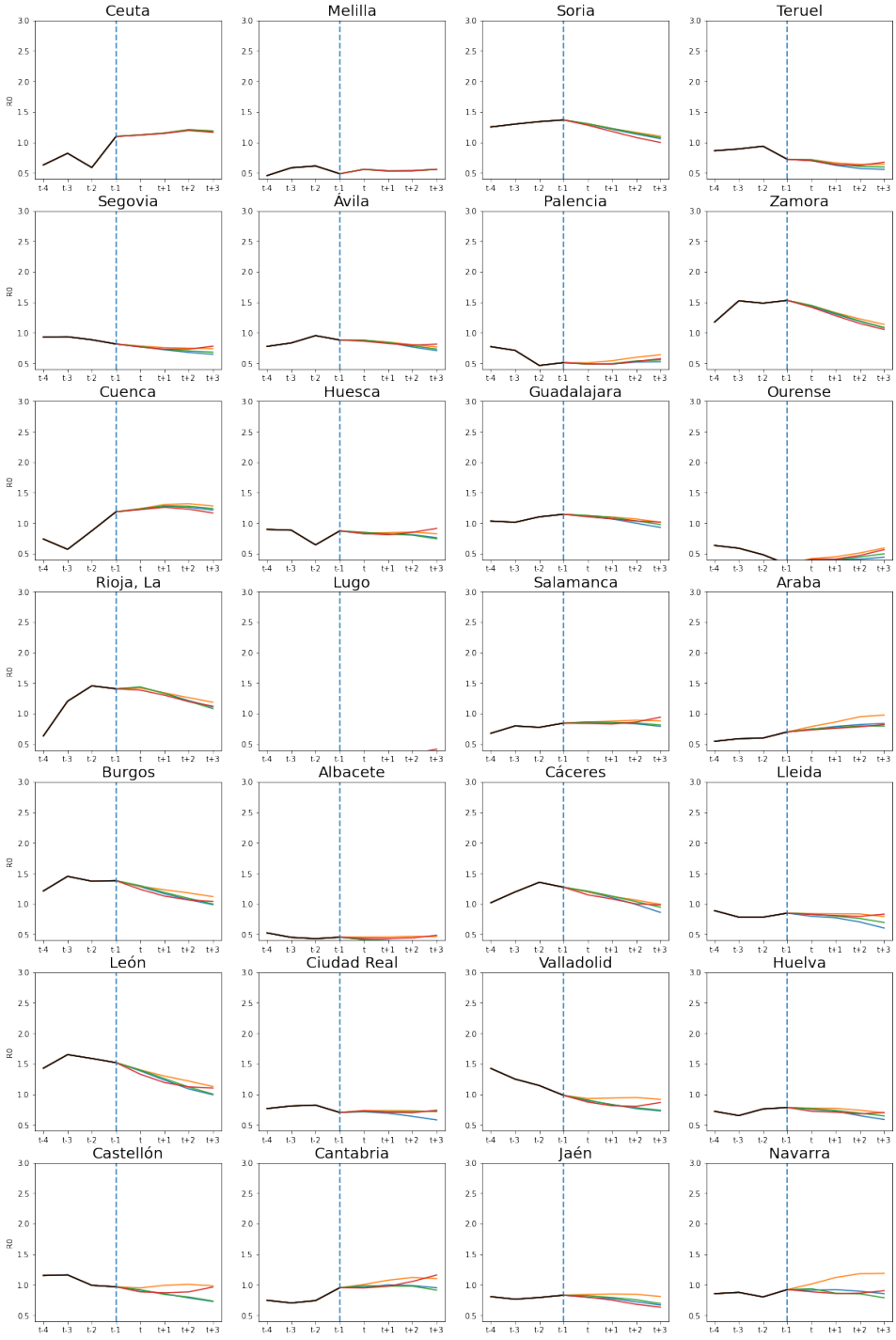

Figura A.2: Valores reales de $\mathcal{R}_0$ de los 4 días anteriores y predicciones a 4 días vista de todas las provincias de España en función de los escenarios de la Tabla 4.4. Las provincias se encuentran ordenadas de izquierda a derecha y de arriba a abajo de menor a mayor población (parte I).

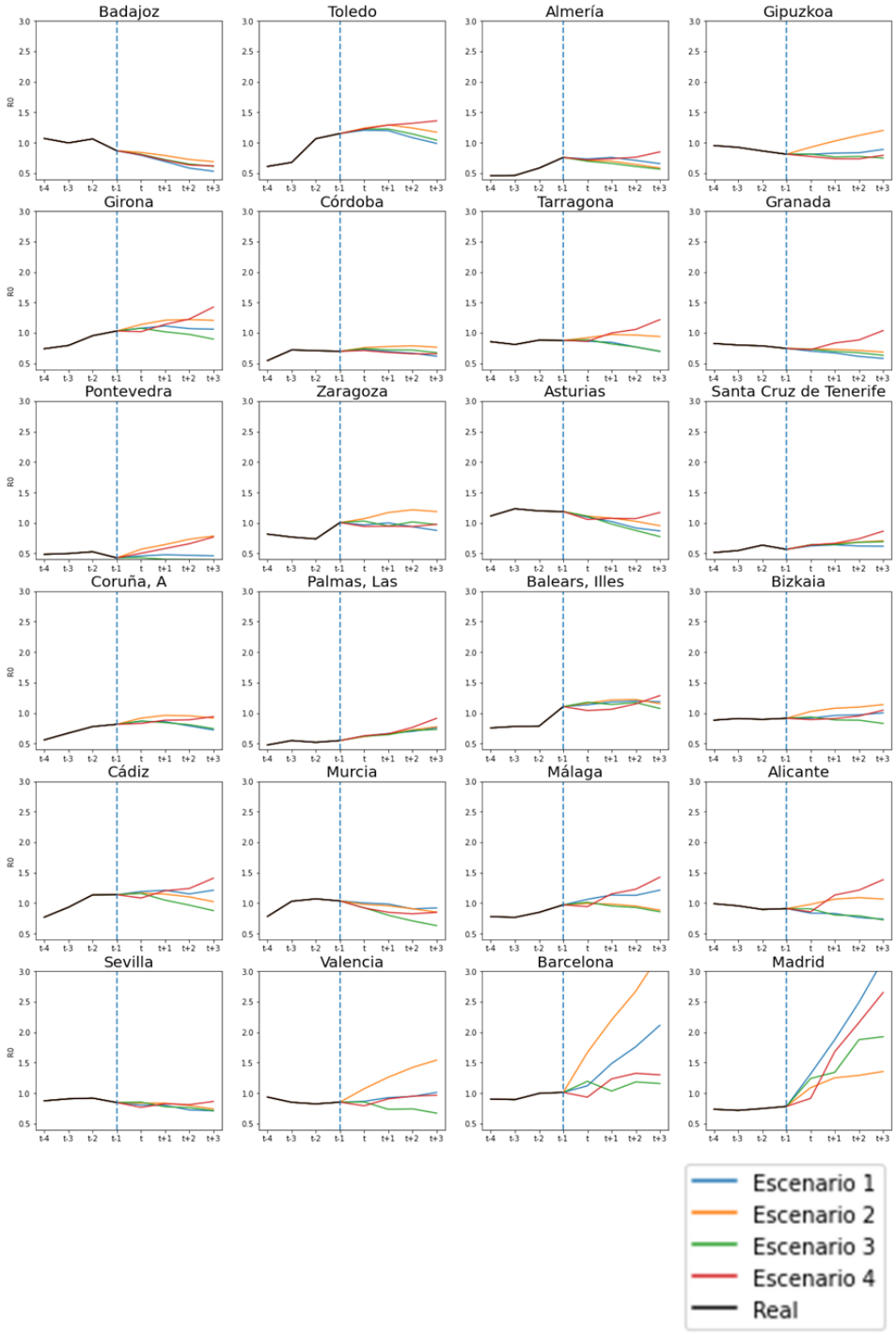

Figura A.3: Valores reales de $\mathcal{R}_0$ de los 4 días anteriores y predicciones a 4 días vista de todas las provincias de España en función de los escenarios de la Tabla 4.4. Las provincias se encuentran ordenadas de izquierda a derecha y de arriba a abajo de menor a mayor población (parte II).

Bibliografía

[1] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Niu, P. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*.

[2] Sifuentes-Rodríguez, E., & Palacios-Reyes, D. (2020). COVID-19: The outbreak caused by a new coronavirus. *Bol Med Hosp Infant Mex*, 77(2), 47-53.

[3] Organización Mundial de la Salud. (2020). *Timeline of WHO's response to COVID-19*. Recuperado de `https://www.who.int/news-room/detail/29-06-2020-covidtimeline`.

[4] Rodríguez-Rey, R., Garrido-Hernansaiz, H., & Collado, S. (2020). Psychological impact and associated factors during the initial stage of the coronavirus (COVID-19) pandemic among the general population in Spain. *Frontiers in psychology*, 11, 1540.f

[5] Van Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., Williamson, B. N., ... & Lloyd-Smith, J. O. (2020). Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *New England Journal of Medicine*, 382(16), 1564-1567.

[6] Rothan, H. A., & Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *Journal of autoimmunity*, 102433.

[7] Gobierno de España. (2020). *Real Decreto 463/2020, de 14 de marzo, por el que se declara el estado de alarma para la gestión de la situación de crisis sanitaria ocasionada por el COVID-19.* BOE-A-2020-3692). Madrid, España: Agencia Estatal Boletín Oficial del Estado Recuperado de `https://bit.ly/3bZDDnD`.

[8] Rubin, G. J., & Wessely, S. (2020). The psychological effects of quarantining a city. Bmj, 368.

[9] Siettos, C. I., & Russo, L. (2013). Mathematical modeling of infectious disease dynamics. *Virulence*, 4(4), 295-306.

[10] Ramos, A. M., Ferrándezb, M. R., Vela-Pérez, M., & Ivorra, B. (2020). A simple but complex enough θ-SIR type model to be used with COVID-19 real data. Application to the case of Italy.

[11] IFCA Advanced Computing and e-Science group. (2020). COVID-19 en España. *GitHub repository*. Recuperado de `https://github.com/IFCA/covid-19-spain/`.

[12] Instituto de Salud Carlos III. (2020). *Pruebas de diagnóstico del coronavirus: ¿qué es la PCR?, ¿qué son los test rápidos? ¿en qué se diferencian?*. Recuperado de `https://www.isciii.es/InformacionCiudadanos/DivulgacionCulturaCientifica/DivulgacionISCIII/Paginas/Divulgacion/COVID19_PCR_test.aspx#:~:text=La%20PCR%2C%20siglas%20en%20ingl%C3%A9s,material%20gen%C3%A9tico%20de%20un%20pat%C3%B3geno.`

[13] Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., ... & Dighe, A. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand.

[14] Instituto Nacional de Estadística. (2020). *INEbase*. Recuperado de `https://www.ine.es/`.

[15] Ministerio de Educación y Formación Profesional. (2020). Estadísticas de la educación. *EDUCAbase*. Recuperado de `http://www.educacionyfp.gob.es/servicios-al-ciudadano/estadisticas.html`.

[16] Ministerio de Sanidad, Consumo y Bienestar Social. (2020). *Actualización nº 116. Enfermedad por el coronavirus (COVID-19). 25.05.2020.* Recuperado de `https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/documentos/Actualizacion_116_COVID-19.pdf`.

[17] Flaxman, S., Mishra, S., Gandy, A., Unwin, H., Coupland, H., Mellan, T., ... & Schmit, N. (2020). Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries.

[18] Arenas, A., Cota, W., Gomez-Gardenes, J., Gomez, S., Granell, C., Matamalas, J. T., ... & Steinegger, B. (2020). Derivation of the effective reproduction number R for COVID-19 in relation to mobility restrictions and confinement. *medRxiv*.

[19] Jones, J. H. (2007). Notes on R0. *California: Department of Anthropological Sciences*, 323, 1-19.

[20] Centre for mathematical modelling of infectious disease. (2020). Temporal variation in transmission during the COVID-19 outbreak. *CMMID Repository*. Recuperado de `https://cmmid.github.io/topics/covid19/global-time-varying-transmission.html`.

[21] Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178(9), 1505-1512.

[22] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[23] Shumway, R. H., & Stoffer, D. S. (2000). Time series analysis and its applications. Studies In Informatics And Control, 9(4), 375-376.

[24] Brownlee, J. (2016). Time Series Forecasting as Supervised Learning. *Machine Learning Mastery*. Recuperado de `https://machinelearningmastery.com/time-series-forecasting-supervised-learning/#:~:text=The%20use%20of%20prior%20time,or%20size%20of%20the%20lag.`.

[25] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[26] Aggarwal, C. C. (2018). *Neural networks and deep learning*. Springer.

[27] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

[28] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

[29] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.