Universidad Internacional de La Rioja
Escuela Superior de Ingeniería y Tecnología - ESIT

Máster Universitario en Industria 4.0

Aplicación de Machine Learning para la previsión de las curvas de contagios de COVID-19 en Europa

Trabajo fin de estudio presentado por:	Natalia Puente Carreño
Tipo de trabajo:	Investigación
Director/a:	Arturo Peralta Martin-Palomino
Fecha:	15 de julio de 2021

Resumen

La epidemia de COVID-19 es uno de los grandes retos del siglo XXI, desde un punto de vista sanitario, social y económico. Poder realizar predicciones acertadas sobre su evolución nos permite anticipar qué acciones son necesarias a largo plazo. Este trabajo presenta un nuevo procedimiento en el que se utilizarán curvas de epidemias pasadas como base para la predicción de los contagios de COVID-19 en 5 países europeos, mediante el procesamiento de los datos y técnicas de machine learning.

Se ha logrado obtener resultados a largo plazo para los 5 países, con un error similar al obtenido en estudios anteriores, y permitiendo estimar el valor absoluto y la forma del final de la curva de contagios de COVID-19 para cada caso.

Se complementa el trabajo con el desarrollo de un panel de visualización para que el análisis de los resultados sea claro y sencillo.

Palabras clave: COVID-19, machine-learning, análisis de datos, visualización, predicción

Abstract

The COVID-19 epidemic is one of the great challenges of the 21st century, from a sanitary, social, and economic perspective. Making accurate predictions on the COVID-19 evolution allows us to anticipate what actions are necessary in the long term. This work presents a new procedure in which curves of past epidemics will be used as the basis for predicting COVID-19 infections in 5 European countries, through data processing and machine learning techniques.

Finally, long-term results have been obtained for the 5 countries, with an error similar to that obtained in previous studies, and estimating the absolute value and the shape of the end of the COVID-19 cases curve for each country.

The work is complemented with the development of a visualization panel so that the analysis of the results is simple and clear.

Keywords: COVID-19, machine-learning, data analysis, visualization, prediction

Índice de contenidos

1. Introducción 9
1.1. Motivación 9
1.2. Planteamiento del trabajo 11
1.3. Estructura del documento 11
2. Contexto y estado del arte 14
2.1. Descripción general del contexto del proyecto 14
2.2. Proyectos y tecnologías relacionados 15
2.3. Conclusiones sobre el estado del arte 18
3. Descripción general de la contribución del TFM 20
3.1. Objetivos 20
3.2. Metodología del trabajo 20
3.3. Descripción general de los componentes de la propuesta 21
3.3.1. Planteamiento de la propuesta y resultados esperados 21
3.3.2. Alcance y limitaciones 22
3.3.3. Listado de participantes 23
3.3.4. Tecnologías implicadas 23
3.3.5. Arquitectura, componentes e integración de tecnologías 23
3.3.6. Presupuesto y retorno esperado de la inversión 25
3.3.7. Planificación general 28
4. Desarrollo de la propuesta para el modelo de previsión de las curvas de COVID-19 30
4.1. Análisis preliminar de los datos 30
4.1.1. Recolección y análisis de los datos de COVID-19 31
4.1.2. Recolección y análisis de los datos de influenza estacional 32
4.1.3. Recolección y análisis de los datos de la influenza H1N1 de 2009: Gripe A 35

4.1.4. Transformación de los datos tipo fecha en “Microsoft Excel date serial numbers” 38

4.2. Previsión temporal preliminar y determinación del punto de madurez de las curvas de COVID-19 39

4.2.1. Ajuste manual de las curvas influenza A de Reino Unido, Alemania e Italia. 42

4.2.2. Localización del punto de madurez de las curvas y ajuste de la escala temporal 43

4.3. Algoritmos de Machine Learning para la transformación de la curva de referencia en los datos conocidos de COVID-19 47

4.3.1. Ajuste con el algoritmo Long-Short-Term Memory (LSTM) 48

4.3.2. Ajuste con el algoritmo Linear Regression (LR) 49

4.3.3. Ajuste con el algoritmo Least Absolute Shrinkage and Selection Operator (LASSO) 49

4.4. Resultados y análisis de los modelos obtenidos 50

4.4.1. Análisis de la exactitud de la predicción 53

4.4.2. Análisis del ajuste del modelo 54

4.4.3. Conclusiones globales 56

4.4.4. Recomendación por país del modelo a utilizar 57

4.4.5. Comparativa respecto a los modelos propuestos en el estado del arte 58

4.5. Dashboard para la visualización y el análisis de los resultados de predicción 61

4.5.1. Arquitectura de datos para los resultados 61

4.5.2. Dashboard de visualización de resultados 63

5. Conclusiones y trabajos futuros 69

5.1. Trabajos Futuros 71

Referencias bibliográficas 72

Anexo A. Tabla estática ‘model’ 75

Índice de figuras

Figura 1. Comparación de la eficacia de la predicción de los modelos según el RSME para el experimento de Tian, Luthra, y Zhang (2020) con datos normalizados por el millón de la población del país. 16

Figura 2. Flujo de almacenamiento y transformación de los datos para la propuesta. (Elaboración propia) 24

Figura 3. Cronograma de la planificación del proyecto. (Elaboración propia) 29

Figura 4. Visualización preliminar de los datos de casos acumulados de COVID-19 a fecha de 15 de mayo de 2021. (Elaboración propia) 32

Figura 5. Número total de positivos de influenza estacional semanales para EU5 desde el año 2010 al año 2019 según los datos de la OMS. (Elaboración propia) 34

Figura 6. Número total de positivos de influenza estacional acumulados por ciclo anual para EU5. (Elaboración propia) 35

Figura 7. Número total de positivos de influenza A (H1N1pdm09) semanales por ciclo anual para EU5. (Elaboración propia) 37

Figura 8. Número total de positivos de influenza A (H1N1pdm09) acumulados por ciclo anual para EU5. (Elaboración propia) 37

Figura 9. Comparativa de las curvas reales de contagios de ES a la aproximación por medio de una función logística. (Elaboración propia) 40

Figura 10. Previsión preliminar de los casos de COVID19 para EU5 basada en la aproximación a una curva logística. (Elaboración propia) 41

Figura 11. Ajuste logístico automático obtenido para los datos de H1N1 de UK, DE, y IT. (Elaboración propia) 42

Figura 12. Ajuste logístico manual para los datos de H1N1 de UK, DE, y IT. (Elaboración propia) 43

Figura 13. Representación gráfica del modelo logístico. (Espina Marconi, 1984) *(Imagen alterada)* 44

Figura 14. Comparativa de las curvas de referencia tras la transformación de su escala temporal y las curvas originales en el periodo equivalente. (Elaboración propia) 46

Figura 15. Predicción de la incidencia acumulada de COVID-19 para UK a partir del 23 de abril de 2021 para cada modelo estudiado. (Elaboración propia) 50

Figura 16. Predicción de la incidencia acumulada de COVID-19 para DE a partir del 21 de junio de 2021 para cada modelo estudiado. (Elaboración propia) 51

Figura 17. Predicción de la incidencia acumulada de COVID-19 para FR a partir del 21 de junio de 2021 para cada modelo estudiado. (Elaboración propia) 51

Figura 18. Predicción de la incidencia acumulada de COVID-19 para IT a partir del 21 de junio de 2021 para cada modelo estudiado. (Elaboración propia) 52

Figura 19. Predicción de la incidencia acumulada de COVID-19 para ES a partir del 23 de abril de 2021 para cada modelo estudiado. (Elaboración propia) 52

Figura 20. RSME de la predicción de cada modelo en el periodo de evaluación. (Elaboración propia) 54

Figura 21. RSME del ajuste de cada modelo en el periodo completo. (Elaboración propia)....55

Figura 22. Vista del eje ajustado para el RMSE de la predicción de cada modelo tras la etapa de experimentación. (Elaboración propia) 56

Figura 23. Arquitectura de las tablas de datos de los resultados. (Elaboración propia).....62

Figura 24. Página Overview del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia) 64

Figura 25. Demostración del funcionamiento de los filtros de la página Overview del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia).....65

Figura 26. Página Graph: Best Model del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia) 66

Figura 27. Página Graph: Model Deep Dive del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia).....67

Figura 28. Ejemplo de filtrado en la página Graph: Model Deep Dive del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia).....68

Índice de tablas

Tabla 1. Rendimiento de los modelos en la previsón futura de nuevos casos de infección confirmados según el experimento de Rustam y otros (2020). 18

Tabla 2. Desglose de los costes totales asociados a los recursos humanos. (Elaboración propia) 25

Tabla 3. Desglose de los costes totales asociados al hardware. (Elaboración propia) 26

Tabla 4. Desglose de los costes totales asociados al software. (Elaboración propia) 26

Tabla 5. Desglose de los costes totales asociados a bibliografía y datos. (Elaboración propia) 27

Tabla 6. Desglose de los costes totales asociados a los suministros. (Elaboración propia) 27

Tabla 7. Desglose de los costes totales asociados al proyecto. (Elaboración propia) 27

Tabla 8. Ingresos percibidos para el proyecto. (Elaboración propia) 28

Tabla 9. Parámetros del ajuste logístico de las curvas de influenza estacional de 2018/19. (Elaboración propia) 43

Tabla 10. Parámetros del ajuste logístico de las curvas de Influenza A H1N1. (Elaboración propia) 43

Tabla 11. RSME de la predicción de cada modelo en el periodo de evaluación. (Elaboración propia) 53

Tabla 12. RSME del ajuste de cada modelo en el periodo completo. (Elaboración propia) 55

Tabla 13. % RMSE de la media de casos considerados para las predicciones de los modelos. (Elaboración propia) 59

Tabla 14. % RMSE de la media de casos considerados para las predicciones de los modelos de Tian, Luthra, & Zhang (2020). *(Elaboración propia)* 60

Tabla 15. % RMSE de la media de casos considerados para las predicciones de los modelos de Rustam, y otros (2020). *(Elaboración propia)* 60

1. Introducción

Este capítulo plantea el problema que deberá solventar la propuesta. La solución presenta los objetivos que se esperan de ella, y además su contribución particular. Para ello, se divide el capítulo en la motivación para el desarrollo de la propuesta, el planteamiento global del proyecto, y finalmente, la estructura que sigue el documento.

1.1. Motivación

Uno de los grandes sucesos que ha marcado el comienzo de la década de 2020 es la pandemia de la COVID-19, que ha puesto en jaque a grandes potencias occidentales, provocando una crisis sanitaria, que posteriormente se ha extendido al ámbito político, social y económico.

La enfermedad del COVID-19, o enfermedad por coronavirus de 2019, es una enfermedad infecciosa provocada por el virus SARS-CoV-2, que fue inicialmente detectada por la Organización Mundial de la Salud (OMS) el 31 de diciembre de 2019. El cuadro sintomático de esta enfermedad abarca principalmente fatiga, fiebre, y diversas afecciones respiratorias. La OMS calcula que aproximadamente un 80% de las personas contagiadas que desarrollan síntomas se recuperan sin necesitar tratamiento hospitalario, sin embargo, un 15% experimenta una sintomatología grave, necesitando aporte adicional de oxígeno, y llegando a ser muy grave en el 5% restante, para los casos en los que son esenciales los cuidados intensivos. (World Health Organization, 2020).

Adicionalmente, el SARS-CoV-2 es un virus que se trasmite principalmente por vía aérea, a través de gotículas expulsadas por las vías respiratorias de personas infectadas. Existe una gran variabilidad en la tasa de contagios estimada en función de la población, sin embargo, en una población no controlada, éste valor generalmente se sitúa por encima de 1, siendo el rango de 1.4-2.5 el valor estimado por la OMS (Najafimehr, Mohamed Ali, Safari, Yousefifard, & Hosseini, 2020). Por tanto, la COVID-19 es considerada una enfermedad de gran transmisión, lo que implica que, sin medidas adicionales, los contagios tienden multiplicarse siguiendo un patrón exponencial. Tras un año desde que la OMS declarara la pandemia de la COVID-19, se han superado los 128 millones de casos, siendo 3.3 millones únicamente en España, representando casi un 7% de la población (El Mundo Gráficos, 2021). Actualmente

esta cifra llega a los 3.8 millones, y otros países como Francia, Reino Unido, Italia o Alemania acumulan 5.77, 4.8, 4.26, y 3.75 millones de casos respectivamente a fecha de 30 de junio de 2021.

La combinación de las complicaciones del cuadro sintomático en una parte significativa de los casos, y de la tasa de contagio del virus, provocó en los estadios iniciales del virus, una gran saturación de los recursos sanitarios, entre los que se incluye maquinaria, como respiradores, instalaciones, como camas UCI, o recursos humanos. De acuerdo con los datos del Instituto Nacional de estadística, la defunciones en España alcanzaron el valor de 20.800 durante la semana más trágica de la primera ola, que se corresponde con la primera semana de abril, mientras que el valor habitual para esta semana, de acuerdo con los datos históricos, suele encontrarse alrededor de 8.000 defunciones (Instituto Nacional de Estadística, s.f.).

Para prevenir el colapso sanitario, los gobiernos regulan las medidas de contención para detener la expansión del virus. Desde el mes de marzo de 2020, hasta aproximadamente el mes de junio, se impuso un confinamiento duro, que posteriormente derivó en medidas más laxas en términos económicos y movilidad, como la implementación de toques de queda, cierre de establecimiento de hostelería u ocio, imposición de distanciamiento social, o la prohibición de reunión de grupos numerosos.

Si bien es cierto que estas medidas han tenido efectos beneficiosos para frenar la expansión del virus, también han provocado graves consecuencias económicas y sociales. La principal secuela de las medidas de contención consiste en la limitación de la movilidad y del consumo, y la atenuación o paralización de ciertas actividades económicas, que provocó en Europa en 2020 una pérdida del 14.6% de las horas trabajadas, por motivos de desempleo, inactividad, o reducción de jornada (Jackson, Weiss, Schwarzenberg, & Nelson, 2020). Estas medidas también afectan a la población desde una perspectiva social, ya que limitan la movilidad de las personas, su comodidad, o su estilo de vida.

Por este motivo, es esencial realizar un seguimiento de la expansión del virus y poder anticipar la curva de incidencia futura, y tomar las decisiones con respecto a medidas contra el COVID-19 más acertadas, de manera que se mantenga un control efectivo sobre la expansión del virus, limitando lo menor posible la actividad económica y social. Adicionalmente, una previsión de incidencia permitirá conocer de antemano los recursos sanitarios que deberán

estar disponibles en cada momento, y evitar la saturación de estos durante las olas de la enfermedad.

1.2. Planteamiento del trabajo

Dado el impacto que tiene la pandemia de COVID-19 a nivel mundial, detallado en el apartado anterior, una de las grandes incógnitas desde el surgimiento de la enfermedad, es prever cuál será su evolución para predecir su impacto futuro y ajustar las medidas de contención necesarias en cada momento. Esto permitiría mantener el nivel de contagios bajo control, limitando los efectos sobre la economía e interacción social.

Para este Trabajo Fin de Máster, se desea desarrollar un modelo que pueda prever la evolución de la curva de contagios de COVID-19. La curva está demostrando tener un comportamiento distinto en función del país al que se refiera, por lo que se desea desarrollar un modelo que sea efectivo para diferentes países de los que se dispongan suficientes datos, o incluso que pueda extrapolado a epidemias futuras. Sin embargo, se plantea utilizar como ejemplo para el desarrollo del trabajo las curvas de COVID-19 de Reino unido, Alemania, Francia, Italia y España (UK, DE, FR, IT, y ES, en conjunto EU5) y realizar un análisis de los resultados que valide la implementación.

Finalmente, se propone utilizar como base la evolución de curvas de infección reales, como puede ser curvas de variantes especialmente infecciosas de gripe o variantes de la pandemia de gripe de 1918. Aunque la curva de previsión deberá seguir la evolución de los datos reales de COVID-19 recogidos hasta el momento.

1.3. Estructura del documento

La memoria del presente Trabajo Fin de Máster está compuesta de un total de cinco capítulos, encontrándose actualmente en el Capítulo 1. Introducción. Este capítulo contiene los elementos básicos para introducir la propuesta como la motivación, y el planteamiento del trabajo. Los cuatro capítulos restantes tienen los siguientes contenidos, y estructura dentro del documento:

- **2. Contexto y estado del arte:** en este capítulo se describe el contexto en el que se ubica el proyecto, y en este contexto se introducen antecedentes que resuelven problemas similares al que se ha planteado para el proyecto. Finaliza el capítulo con una reflexión sobre los elementos que resultarían útiles y las mejoras que lograría la nueva propuesta.
- **3. Descripción general de la contribución del TFM:** Este capítulo contiene los detalles generales de la propuesta, planteando los objetivos y la metodología del proyecto, y la descripción general de otros componentes relevantes.
- **4. Desarrollo de la propuesta para el modelo de previsión de las curvas de COVID-19:** En este capítulo se describe de manera específica el desarrollo de la propuesta. Este capítulo puede dividirse a su vez en cinco apartados que diferencian las etapas en el desarrollo:
    - **4.1. Análisis preliminar de los datos:** Este apartado contiene la descripción y visualización de las fuentes de datos utilizadas, realizando transformaciones básicas, y sacando pequeñas conclusiones iniciales.
    - **4.2. Previsión temporal preliminar y determinación del punto de madurez de las curvas de COVID-19**: Este apartado describe el grueso de la transformación de los datos, de manera que puedan utilizarse con técnicas de machine learning para obtener previsiones.
    - **4.3. Algoritmos de Machine Learning para la transformación de la curva de referencia en los datos conocidos de COVID-19:** Este apartado contiene las decisiones tomadas y el procedimiento utilizado a la hora de implementar técnicas de machine learning.
    - **4.4. Resultados y análisis de los modelos obtenidos**: Consiste en la presentación de los resultados que se han obtenido mediante las técnicas anteriores, y un análisis de éstos resultados., acompañados por recomendaciones y conclusiones.
    - **4.5. Dashboard para la visualización y el análisis de los resultados:** Describe las vista diseñadas para el panel de visualización. Adicionalmente, se describe la arquitectura de datos en las que se almacenan los resultados de manera que se puedan transmitir desde la salida del modelo construido, a la entrada del software de visualización.

- **5. Conclusiones y trabajos futuros:** En este capítulo se realiza a una valoración global del trabajo realizado, cuantificando el cumplimiento de diversos hitos intermedios, y del objetivo principal de la propuesta. Finalmente, se incluye un apartado final con trabajos futuros que podrían complementar, ampliar, o mejorar el proyecto.

Finalmente, en la parte final del documento se encuentra el conjunto de las referencias utilizadas en la memoria. Se incluye también un único Anexo que contiene información complementaria para apoyar la descripción del apartado 4.5 del documento.

2. Contexto y estado del arte

Este capítulo, se compone de tres partes fundamentales. La primera consiste en la descripción del contexto de proyecto que permite ubicar la terminología básica y las tecnologías implicadas en el mismo. Posteriormente, se incluye un conjunto de antecedentes que componen el estado del arte de proyecto, y son en el punto de partida de cara al desarrollo. Por tanto, se incluyen conclusiones para el estado del arte, que permiten analizar los elementos de utilidad para el proyecto, pero también los elementos diferenciadores que componen la innovación del mismo.

2.1. Descripción general del contexto del proyecto

El contexto del proyecto se puede definir desde dos perspectivas diferentes. En primer lugar, el proyecto está ubicado en un ámbito epidemiológico. De acuerdo con la definición de Baños, Brotons y Farré (1998):

"(…) (La epidemiología) es la disciplina que estudia la distribución de las enfermedades o los estados relacionados con la salud y sus determinantes en poblaciones específicas, y su aplicación en el control de los problemas de salud. (…)."

Para este proyecto se plantea un caso de pronóstico epidemiológico, en el que el objetivo es la predicción de la evolución futura de la enfermedad COVID-19, en poblaciones con diferentes características y elementos externos que puedan afectar a su comportamiento. El elemento sobre el que se aplica el foco de la investigación del proyecto es la denominada curva de contagios. Para el caso, se va a definir la curva de contagios de una enfermedad como la evolución temporal del número de contagios registrados de dicha enfermedad, pudiendo determinarse a partir de ella su tendencia y magnitud, local y global, y su periodicidad.

En este contexto, el proyecto alcanza una segunda dimensión de entendimiento, como el desarrollo de modelos de previsión y pronóstico, mediante técnicas de análisis predictivo y Machine Learning. Desde esta perspectiva, se puede definir los siguientes términos:

"*El análisis predictivo (…) describe una variedad de técnicas estadísticas y analíticas usadas para desarrollar modelos para predecir eventos y comportamientos futuros*." (Nyce, 2007).

Russell y Norvig (2010) definen inteligencia artificial como "el estudio de agentes que reciben percepciones del entorno y realizan acciones", similarmente determinan que un agente aprende "si mejora su rendimiento en tareas futuras tras realizar observaciones (…)." De manera análoga, las técnicas de machine permiten otorgar capacidad de aprendizaje a sistemas informáticos.

Definidas las dos áreas fundamentas en las que se ubica el presente proyecto, se considera que el interés principal del proyecto se encuentra en el desarrollo tecnológico del modelo. De manera que el modelo permita ser revisado desde un punto de vista de exactitud epidemiológica.

2.2. Proyectos y tecnologías relacionados

Desde que surgieron los primeros casos de COVID-19 y se evidenció la gravedad de la situación, también aparecieron los primeros modelos para prevenir las curvas de contagios. Dado que el interés por realizar investigaciones ligadas a la pandemia es relativamente reciente, en la mayor parte de los casos, los proyectos se han orientado a fases de experimentación inicial y con un ámbito principalmente académico. El propósito de estos proyectos suele ser la evaluación de diferentes técnicas de análisis de datos y la comprobación de su efectividad al recrear el comportamiento de la curva original.

Este es el caso de la investigación llevada a cabo por Tian, Luthra, y Zhang (2020). En ella se utilizan datos de 6 países para comprobar la eficacia en la predicción de casos de COVID-19 de 3 modelos de machine learning. Los países involucrados en el estudio son Alemania, Italia, Estados Unidos, Taiwán, Japón y Corea del Sur, cuyos datos de casos COVID-19 fueron recolectados del repositorio de GitHub oficial de la Johns Hopkins University (Dong, Du H, & Gardner).

Para contextualizar el proyecto, se procede a definir brevemente las tres técnicas de machine learning que se evaluaron:

- Hierarchical Bayes Model (HBM): Se trata de un modelo con base estadística, escrito en múltiples niveles, que estima una distribución utilizando métodos Bayesianos. Es conveniente cuando se dispone de información en diferentes niveles en términos de unidades observacionales, como en este caso, en el que las unidades informacionales

son cada uno de los países con una curva de casos diferente (Allenby, Rossi, & McCulloch, 2005).

- Long-Short-Term-Memory (LSTM): es una arquitectura de red neuronal recurrente capaz de aprender dependencias largas (Hochreiter & Schmidhuber, 1997), por lo que se trata de un método efectivo en series temporales. Como indican Tian, Luthra, y Zhang, frente al hierarchical bayes model no presupone antecedentes sobre la distribución estadística, por lo que es conveniente para sets de datos nuevos.
- Hidden Markov Model (HMM): modelos basados en modelos estadísticos Markov, que se basan en la identificación de estados ocultos dependientes de una variable temporal (Jurafsky & Martin, 2020).

Tian, Luthra, y Zhang realizan una comparativa entre los resultados de los modelos logrados con cada método en base a la raíz del error cuadrático medio (RMSE) para la predicción de los últimos días. El experimento obtuvo los resultados demostrados en la Figura 1, que concluyen que, para la mayor parte de países, el mejor comportamiento lo tiene el método LSTM RNN, mientras que el peor lo tiene el método Hidden Markov Model.

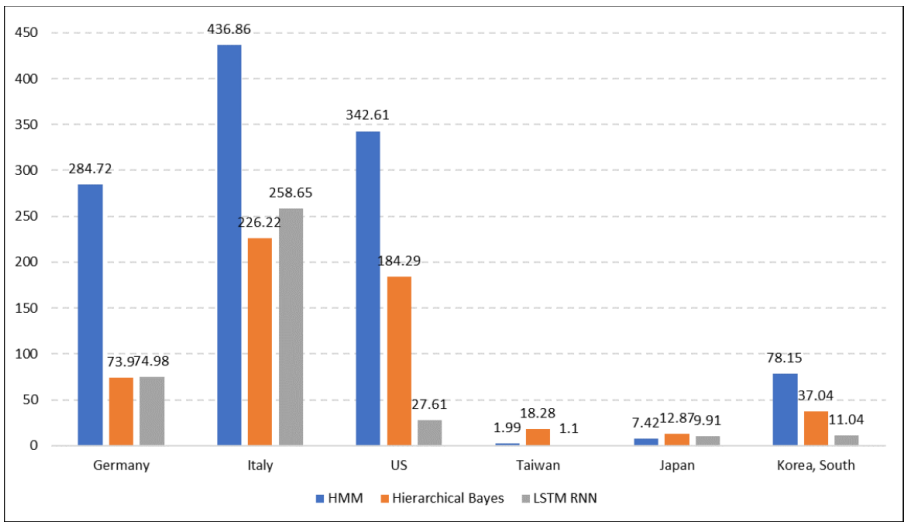

Figura 1. Comparación de la eficacia de la predicción de los modelos según el RSME para el experimento de Tian, Luthra, y Zhang (2020) con datos normalizados por el millón de la población del país.

Lee, Lei, y Mallick (2020) también estudian el uso de 3 modelos jerárquicos bayesianos, en este caso, utilizando los datos de 40 países del mismo repositorio de GitHub de la Johns Hopkins University (Dong, Du H, & Gardner). El primer modelo analiza individualmente cada

país sin asumir relaciones entre ellos, un segundo modelo jerárquico utiliza las trayectorias de los 40 países, y finalmente, el tercer modelo adiciona covariables específicas de cada país. Los resultados demostraron, en base al error cuadrático medio, que el modelo con peor comportamiento fue el primero, y aquel con mejor comportamiento fue el tercero.

Diferentes técnicas de aprendizaje supervisado fueron utilizadas por Rutam y otros (2020), para predecir los casos de COVID-19. En el proyecto se utilizó también el repositorio de la Johns Hopkins University (Dong, Du H, & Gardner), queriendo anticipar el número de casos positivos, el número de muertes, y el número de recuperaciones. Los algoritmos de aprendizaje elegidos para las previsiones entran todos en la clasificación de algoritmos de regresión que pueden ser utilizados con series temporales. En este caso se desea conocer la función de la curva de COVID-19 con respecto del eje temporal, y para ello estas técnicas utilizan el aprendizaje de comportamiento frente al tiempo de la parte inicial de la serie (datos de entrenamiento), para encontrar el modelo que mejor defina este comportamiento, y hacer predicciones con nuevos datos temporales. Las técnicas de aprendizaje valoradas en el experimento fueron las siguientes:

- Linear regression (LR): modelos basados en regresiones lineales en los que existe una componente dependiente y una componente independiente, tratando de identificar su relación para minimizar el error entre las observaciones y el resultado del modelo.
- Least absolute shrinkage and selection operator (LASSO): También pertenece al grupo de regresiones lineales, aunque en este caso se realiza una disminución de los valores extremos frente a los valores centrales, produciendo resultados potencialmente más estables y con menor error.
- Support vector machine (SVM): utilizado tanto en regresión y clasificación. Para el caso de la regresión el algoritmo puede utilizar transformaciones de Kernel para mapear los datos a un espacio de dimensión en el que se puede utilizar la regresión lineal.
- Exponential smoothing (ES): basa la previsión en el ponderado de datos previos decayendo según la antigüedad de estos datos.

Para evaluar los modelos producidos por cada una estas técnicas, Rustam y otros (2020) utilizan la valoración del coeficiente de determinación (R^2 score), del coeficiente de determinación ajustado (R^2 adjusted), del error cuadrático medio (MSE), del error absoluto medio (MAE), y de la raíz del error cuadrático medio (RMSE). Particularmente, para el caso de

la previsión de la curva de contagios, el mejor comportamiento se observa en los modelos LASSO y ES de acuerdo con las métricas observadas en la Tabla 1, y el peor comportamiento se le atribuye a la técnica SVM.

Model	R^2 Score	$R^2_{Adjusted}$	MSE	MAE	RMSE
LR	0.83	0.79	1472986504.96	30279.55	38390.51
LASSO	0.98	0.97	234489560.99	11693.97	15322.11
SVM	0.59	0.47	5760890969.30	60177.90	75911.28
ES	0.98	0.97	283201302.2	8867.43	16828.58

Tabla 1. Rendimiento de los modelos en la previsón futura de nuevos casos de infección confirmados según el experimento de Rustam y otros (2020).

En el presente proyecto se desea plantear un enfoque distinto, utilizando la transformación de una curva epidemiológica completa diferente para el aprendizaje del modelo. Para este caso, se propone el estudio de un proyecto en que se utilizan técnicas de machine learning para la previsión de curvas de contagios a través de curvas de epidemias similares.

2.3. Conclusiones sobre el estado del arte

El conjunto de proyectos que componen el estado del arte permite obtener las siguientes conclusiones:

En primer lugar, puede observarse que los proyectos anteriormente presentados son consistentes en la fuente de la que se obtienen los datos de contagios de COVID-19. El repositorio de la Johns Hopkins University es un set de datos accesible, completo, y consistente, que se mantiene actualizado con periodicidad. Por este motivo, se propone el uso de dicha fuente para extraer los datos históricos de contagios de COVID-19.

Si embargo, existe una gran carencia en los proyectos del apartado anterior. En las propuestas anteriores se hace uso de los propios datos reales de la enfermedad para realizar la previsión. En el caso de utilizar datos de un estadio temprano de la enfermedad, los resultados mostrarían un sistema inestable con tendencia creciente, por lo que la previsión únicamente sería válida para periodos pequeños posteriores a los datos reales. En otras palabras, no se podría hacer una previsión acertada de las etapas de recesión del virus, ni de su estimación temporal.

Para mejorar la precisión de previsión con respecto a la curva de contagios esperada, se propone la utilización de una curva completa de una enfermedad infecciosa real con características similares. El modelo podrá ofrecer una previsión de las fases terminales de la pandemia.

Finalmente, se decide iniciar la exploración de las técnicas de machine learning propuestas para los proyectos anteriores, habiendo demostrado buenos resultados en dichas investigaciones. Sin embargo, se deberá valorar la utilización de algoritmos alternativos que resulten efectivos en la interpretación y el modelado de curvas temporales.

3. Descripción general de la contribución del TFM

Este apartado contiene una descripción clara y concisa de la propuesta, indicando los objetivos, la metodología de trabajo y el planteamiento general para sus realización. Se complementa con la descripción de los resultados esperados considerando el alcance y las potenciales limitaciones, y finalmente se incluye un presupuesto y la planificación temporal del proyecto

3.1. Objetivos

El objetivo del presente trabajo es evaluar la eficacia de un modelo de previsión de las curvas de contagios de COVID-19, utilizando como referencia las curvas completas de pandemias anteriores. Para alcanzar el objetivo principal se plantean los siguientes objetivos secundarios:

- Construcción de un modelo basado en una curva de contagios completa y real.
- Evaluación del eficacia del procedimiento aplicado a los datos de diferentes países europeos (UK, DE, FR, IT, y ES).
- Valoración del modelo completo frente a investigaciones anteriores.
- Creación de un panel de visualización (Dashboard) que incluya los resultados obtenidos y métricas clave.

3.2. Metodología del trabajo

Para el proyecto se propone una metodología agile. La metodología agile realiza un desarrollo rápido para habilitar el uso del producto en el menor tiempo posible, y posteriormente la realización de mejoras e iteraciones en función del feedback recibido para el producto en distribución. La sección Trabajos Futuros, describe mejoras que no serán incorporadas en el producto inicial, pero que junto con el feedback de los usuario que consuman el producto, se podrán valorar en nuevas iteraciones.

Para el desarrollo del primer producto funcional, se plantea el siguiente proceso, que está alineado con los objetivos del apartado anterior.

- **Paso 1:** Búsqueda de fuentes de datos de contagios de la pandemia de COVID-19 para UK, DE, FR, IT, y ES (EU5) así como un análisis y acondicionamiento preliminar de los datos.
- **Paso 2:** Localización del punto de madurez de una curva de COVID-19 con respecto a una curva completa. Se deberá encontrar un método de procesamiento de datos capaz de equiparar ambas curvas, para lograr localizar dicho punto.
- **Paso 3:** Entrenar el modelo con las curvas base, mediante algoritmos de Machine Learning. Diversos algoritmos deberán ser estudiados para encontrar aquel que realice el mejor ajuste para el caso. Para el desarrollo software, será utilizado un país modelo.
- **Paso 4:** Evaluar el comportamiento de la curva para cada país. La definición del mejor modelo vendrá determinada por los resultados obtenidos para cada país, en función de los algoritmos y datos utilizados.
- **Paso 5:** Complementar la evaluación con una comparativa con los antecedentes ubicados en el contexto del proyecto, para validar la propuesta.
- **Paso 6:** Desarrollo del dashboard de visualización de resultados.

Adicionalmente, la recolección de material con fines educativos y didácticos, y la composición de documentación en forma de memoria técnica, se realizará a lo largo de todo el proceso de manera paralela a los pasos anteriores.

3.3. Descripción general de los componentes de la propuesta

Los componentes que forman el conjunto de la descripción de la propuesta están definidos en el este apartado. Otorgan una visión global del proyecto.

3.3.1. Planteamiento de la propuesta y resultados esperados

Para lograr los objetivos anterior e implementar la metodología, se realizará un desarrollo en código del procesamiento de los datos. El lenguaje de programación utilizado será Python, ya que se dispone de gran cantidad de librerías aplicables al análisis de ellas. De hecho, la implementación de los algoritmos de machine learning se simplifica en gran medida con la

utilización de este lenguaje de programación, limitando el desarrollo a apenas la configuración del algoritmo.

Se desarrollará mediante código algoritmos y funciones que realicen las transformaciones y el procesado de las curvas de contagios necesarios para ser utilizados en las técnicas de machine learning. Posteriormente, se utilizará el mismo código para la experimentación con dichos algoritmos hasta disponer de diversos modelos funcionales que evaluar. El planteamiento propone realizar todo este desarrollo software con un único país hasta lograr un funcionamiento correcto, y posteriormente efectuar pruebas con el resto de EU5 para depurar errores y asegurar el funcionamiento en todos los casos.

Se deberá incluir una etapa de evaluación una vez finalizado la etapa de desarrollo, para poder validar el procedimiento frente los antecedentes descritos, pues deberá obtenerse una mejora o una ventaja frente a los mismos. En estos resultados se esperan predicciones capaz de estimar a largo plazo, y que sigan la forma típica de una curva de contagios en un caso epidemiológico.

Como etapa final, se construirá un panel de visualización que deberá incluir información relevante sobre la epidemia y la estimación de diversos modelos. Se espera obtener vistas claras y sencillas pero completas de la información de mayor interés.

3.3.2. Alcance y limitaciones

El alcance de la propuesta es el siguiente:

- Se podrá obtener un modelo que complete la curva de evolución del COVID-19 de 5 países europeos.
- Se pretende obtener el punto de madurez de la pandemia en tiempo real en cada país (EU5), utilizando datos actualizados.
- Se podrá evaluar el procedimiento frente a los que se han descrito en el estado del arte.
- Se podrán observar los resultados en un dashboard que permite una visualización cómoda y clara.

Las principales limitaciones esperadas que presenta la propuesta se describen a continuación:

- El modelo presenta una menor precisión local, al no utilizar datos de COVID-19 para el entrenamiento del modelo, ni evaluar los patrones locales de los datos reales de COVID-19.
- El modelo, para la previsión del COVID-19, tiene una obsolescencia muy temprana, dado el nivel de madurez actual de la curva de COVID-19.
- El procedimiento tan solo puede aplicarse para curvas de COVID-19 que hayan pasado su punto de inflexión, de manera que se pueda estimar la altura que llegará a alcanzar la curva acumulada total.

3.3.3. Listado de participantes

Para la realización de la propuesta será necesario un número mínimo de participantes:

- El estudiante a cargo de la propuesta del TFM. Encargado de llevar a cabo la investigación para completar la propuesta, y de realizar el desarrollo para obtener los modelos presentados como el resultado del proyecto.
- El director de la propuesta de TFM. Supervisa el avance del proyecto, asegurando que cumple con los requisitos mínimos de calidad y cualidad innovativa.

3.3.4. Tecnologías implicadas

Este apartado va en línea con la Descripción general del contexto del proyecto, ya que se ha utilizado para introducir una de las tecnologías implicadas más relevantes: el machine learning, y por extensión, el procesamiento de datos. Sin embargo, en el contexto de la Industria 4.0, existe otra tecnología con un gran impacto en la propuesta que son las técnicas de visualización y el business intelligence.

3.3.5. Arquitectura, componentes e integración de tecnologías

Desde una visión global de la propuesta, para el desarrollo del presente proyecto con respecto a la arquitectura se utilizará un ordenador personal con conexión a internet. Por otro lado, el ordenador estará equipado del software siguiente:

- Anaconda Navigator: se utilizará para desarrollar el programa que albergará el procesamiento de los datos y construcción de los modelos. El lenguaje utilizado para esta aplicación es Python, en el entorno de Jupyter Notebook.
- Microsoft Excel: permite una visualización rápida del formato y contenido de los datos, de manera que su procesamiento mediante el lenguaje Python, sea más rápido y sencillo.
- Microsoft Word: se utilizará para elaborar la documentación fundamental del proyecto, así como la memoria académica del presente trabajo.
- Adobe Acrobat Reader: permite la lectura y análisis del estado del arte, y de la documentación adicional para el desarrollo del proyecto.
- Tableau Desktop: software para la visualización interactiva de datos orientado hacia el business intelligence. Se utilizará para el desarrollo del dashboard de visualización y análisis.

Desde el punto de vista de los datos, se puede definir una arquitectura diferente, y es el flujo de almacenamiento y transformación que sufren, descrito de manera gráfica en la Figura 2.

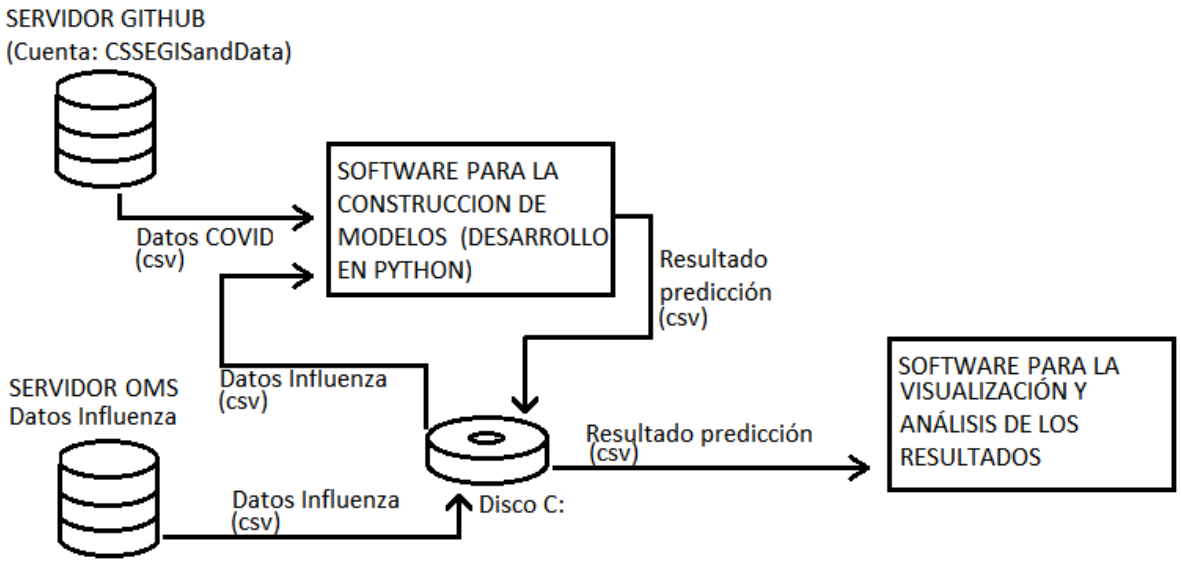

Figura 2. Flujo de almacenamiento y transformación de los datos para la propuesta. (Elaboración propia)

Los datos de COVID-19 son automáticamente descargados de GitHub en cada ejecución del código de construcción de modelos, de manera que se trabajará siempre con la versión más actualizada. Por otro lado, los datos de curvas base no necesitan una descarga automática,

debido a que son datos completos y finitos, por lo que estarán ya almacenados en la ubicación del código. Estos datos alimentan el código de Python y se utilizarán para la construcción de los modelos, y generación de resultados, que son nuevamente almacenados en el disco. Finalmente, el software de visualización, Tableau Desktop, lee estos resultados y los incorpora a una proyecto que permite visualizarlos y analizarlos de manera sencilla.

3.3.6. Presupuesto y retorno esperado de la inversión

Este apartado describe el presupuesto y el retorno de la inversión esperados para el proyecto. Con respecto a los costes del proyecto, éstos se pueden dividir en 5 categorías: recursos humanos, hardware, software, bibliografía y datos, y suministros.

3.3.6.1. Costes de recursos humanos

Se trata de los costes asociados al personal encargado del desarrollo del proyecto. Se cuantifica de manera exclusiva para el proyecto un total de 260h a lo largo de 4 meses, para un Analista titulado al que se le atribuye una retribución de 15€/h. Teniendo en cuenta estos datos se contabiliza el coste por recursos humanos en la Tabla 2.

Recursos humanos			
Concepto	Nº horas	Coste/hora	Coste imputable
Analista titulado	260 h	15 €/h	3.900 €
Coste Total			3.900 €

Tabla 2. Desglose de los costes totales asociados a los recursos humanos. (Elaboración propia)

3.3.6.2. Coste de hardware

La fuente de costes del hardware incluye únicamente un ordenador portátil, en este caso un ASUS N551JK para el que asume un tiempo de depreciación de 5 años. El coste imputable asociado al hardware se puede consultar en la Tabla 3.

Hardware					
Concepto	Cantidad	Precio	Dedicación	Tiempo de depreciación	Coste imputable
Ordenador portátil ASUS N551JK	1	1.000 €	4 meses	60 meses	67 €
Coste Total					67 €

Tabla 3. Desglose de los costes totales asociados al hardware. (Elaboración propia)

3.3.6.3. Costes de software

En este caso se hace referencia al software descrito en el aparatado Arquitectura, componentes e integración de tecnologías. Se utilizan dos programas de distribución gratuita (Anaconda Navigator y Acrobat Reader DC versión gratuita), y una licencia de Microsoft Office obtenida a través de una asociación con la UNIR. Por otro lado, el software de Tableau Desktop tiene un coste anual de 840 USD ($), utilizado únicamente durante el último mes de desarrollo. El coste total asociado al software aparece descrito en la Tabla 4.

Software					
Concepto	Cantidad	Precio	Dedicación	Tiempo de depreciación	Coste imputable
Licencia Microsoft Office	1	0 € *	4 meses	36 meses	0 €
Licencia Anaconda Navigator	1	0 €	4 meses	24 meses	0 €
Acrobat Reader DC básico	1	0 €	4 meses	36 meses	0 €
Tableau Desktop	1	$ 840	1 meses	12 meses	59 €
Coste Total					0 €
					** Licencia gratuita por asociación a la UNIR*

Tabla 4. Desglose de los costes totales asociados al software. (Elaboración propia)

3.3.6.4. Costes de bibliografía y datos

Este componente hace referencia al coste incurrido por el acceso a la documentación necesaria para la realización del proyecto, así como por el acceso y utilización de las fuentes de datos. Toda la documentación, así como los datos obtenidos son gratuitos y de libre distribución y uso, por lo que el coste nulo para esta categoría se refleja en la Tabla 5.

Bibliografía y datos	
Concepto	Coste imputable
Bibliografía	0 €
Set de datos Influenza Estacional 18/19 (OMS)	0 €
Set de datos Influenza A H1N1 (OMS)	0 €
Set de datos de COVID-19 (John Jopkins University)	0 €
Coste Total	0 €

Tabla 5. Desglose de los costes totales asociados a bibliografía y datos. (Elaboración propia)

3.3.6.5. Costes de suministros

En la Tabla 6 se muestra el coste imputable desglosado asociado a los suministros que se han necesitado para este proyecto. Dado que los suministros se han utilizados para usos externos a este proyecto, tan solo se ha incluido el coste asociado al periodo de uso exclusivo para el desarrollo del trabajo. Los dos suministros principales son Internet y electricidad; cualquier otro suministro se considera poco significativo frente a éstos, y se ha despreciado en el cálculo.

Suministros				
Concepto	Consumo	Nº horas	Coste/unidad	Coste imputable
Electricidad	0,11 kW	260 h	0,1284 €/kWh	4 €
Conexión a internet POST Luxembourg	4 meses	65 h/mes	51,99 €/mes	19 €
Coste Total				22 €

Tabla 6. Desglose de los costes totales asociados a los suministros. (Elaboración propia)

3.3.6.6. Coste total del proyecto

Agregando todas las fuente de coste, se obtiene un coste total para el proyecto de 4.092€, reflejado en la Tabla 7.

Coste Total	
Concepto	Coste imputable
Recursos humanos	3.900 €
Hardware	67 €
Software	59 €
Bibliografía y datos	0 €
Suministros	67 €
Coste Total	4.092 €

Tabla 7. Desglose de los costes totales asociados al proyecto. (Elaboración propia)

3.3.6.7. Ingresos esperados y retorno a la inversión

Dado el bajo presupuesto asignado a este proyecto, cuya mayor fuente de gasto puede atribuirse a los recursos humanos, no es necesario que se perciban ingresos importantes para retornar la inversión. Cabe destacar que la pandemia de COVID-19 ha supuesto un auténtico reto para la sociedad y que existen múltiples iniciativas para paliar sus efectos ofrecidas de manera voluntaria y sin esperar remuneración.

Sin embargo, para cubrir los gastos de desarrollo, se han realizado gran número de inversiones en investigaciones relacionadas con la epidemia. Al tratarse de un trabajo fin de máster, se va a considerar los ingresos percibidos por una beca de investigación sencilla. Se asumen unos ingresos mensuales de 900 euros, con una inversión inicial de 1000 euros, como se describe en la Tabla 8.

Ingresos			
Concepto	Ingreso mensual	Meses	Coste imputable
Analista titulado	900 €/mes	4 meses	3.600 €
Inversión inicial			1.000€
Coste Total			4.600 €

Tabla 8. Ingresos percibidos para el proyecto. (Elaboración propia)

En este caso, los gastos son cubiertos por los ingresos, incluyendo la inversión inicial, por lo que el retorno a la inversión sería inmediato.

3.3.7. Planificación general

La previsión para el desarrollo de las fases del proyecto viene definida en el diagrama de la Figura 3. Las fases descritas coinciden en gran medida con los pasos planteados en la metodología de la propuesta. Se ha decido utilizar una planificación quincenal, dado que se espera finalizar el primer producto funcional en un periodo de 4 meses.

		2021						
ID	Nombre	mar-2	abril-1	abril-2	mayo-1	mayo-2	junio-1	junio-2
Aplicación de técnicas Machine Learning para la previsión de las curvas de contagios de COVID en Europa								
T0	Documentación y recolección del estado del arte							
T0.1	Documentación básica							
T0.1.1	Concepción de la idea y búsqueda de antecedentes							
T0.1.2	Recolección de material							
T0.2	Recolección, documentación y estudio de técnicas y resultado de experimentación académica							
T1	Recolección y procesado básico de datos de contagios							
T1.1	Estudios de las posibilidades y selección de curvas base							
T1.2	Procesado, visualización y análisis de las curvas base, y selección final							
T2	Gran procesamiento de los datos y experimentación con algoritmos de machine learning							
T2.1	Procesado de los datos							
T2.1.1	Generación de funciones de transformación de fechas							
T2.1.2	Generación de funciones de transformación logística							
T2.1.3	Generación de alogritmos de localización del punto de madurez							
T2.1.4	Integración del código para transformaciones de los datos funcionales para EU5							
T2.2	Implementación práctica del código de proceso							
T2.2.1	Desarrollo con un único país							
T2.2.2	Experimentación con EU5							
T2.3	Pruebas de funcionamiento del conjunto integrado							
T2.3.1	Implementación y experimentación de algoritmos de machine learning con un único país							
T2.3.2	Experimentación con el conjunto de EU5, y con todos los algoritmos seleccionados y curvas base							
T3	Obteción de resultados y evaluación							
T3.1	Utilización del algoritmo para todos los países y modelos y extracción de resultados							
T3.2	Evaluación del los resultados y de la metodología							
T3.2.1	Evaluación de los resultados de la predicción de los diferentes modelos generados							
T3.2.2	Evaluación global de la metodología frente al estado del arte							
T4	Generación de un panel de visualización (Dashboard)							
T4.1	Creación y experimentación de la estructura de datos y alogoritmo para el almacenamiento							
T4.2	Creación de las visualización y diseño gráfico							
T5	Documentación del progreso y generación de la memoria del trabajo							

Figura 3. Cronograma de la planificación del proyecto. (Elaboración propia)

4. Desarrollo de la propuesta para el modelo de previsión de las curvas de COVID-19

Esta capítulo plantea el detalle del desarrollo de la propuesta para la construcción de modelos de previsión de casos acumulados de COVID-19. La descripción de la contribución se divide en cinco partes claras y diferenciadas, que siguen de manera aproximada los pasos descritos para el procedimiento planteado:

- Análisis preliminar de los datos: se corresponde con el paso 1 del procedimiento, y tiene como objetivo presentar las fuentes de datos utilizadas, y sacar información básica de ellas.
- Previsión temporal preliminar y determinación del punto de madurez de las curvas de COVID-19: se corresponde con el paso 2 del procedimiento, y realiza las transformaciones más importante y aplicación algoritmos de preparación en los datos.
- Algoritmos de Machine Learning para la transformación de la curva de referencia en los datos conocidos de COVID-19: correspondiente con el paso 3, describe como se realiza la aplicación práctica de las técnicas de machine learning elegidas, y se describen las decisión generales aplicadas a estas técnicas.
- Resultados y análisis de los modelos obtenidos: equivalente a los pasos 4 y 5 del procedimiento, presenta los resultados obtenidos con cada uno de los modelos desarrollados y realiza una valoración del modelo más adecuado para cada país, y una valoración global del método.
- Dashboard para visualización y análisis de resultados: se refiere el paso 6 del procedimiento, y describe los elementos del dahsboard de visualización y la información que proporciona al usuario.

4.1. Análisis preliminar de los datos

Un análisis preliminar constituye una fase fundamental al iniciar un proyecto basado en datos. Esta fase permite conocer su estructura y obtener las primeras visualizaciones, de manera que se puedan sacar las primeras informaciones y conclusiones.

4.1.1. Recolección y análisis de los datos de COVID-19

La fuente de los datos de contagios de COVID-19 será el "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" que se actualiza en su perfil de Github (Dong, Du H, & Gardner). En el apartado Conclusiones sobre el estado del arte, se propuso la utilización de esta fuente, ya que es la más utilizada en todas las investigaciones científicas consultadas. Adicionalmente, la motivación para la utilización de dicha fuente de datos es la siguiente:

- Los datos son coherentes, completos, y están estructurados.
- La recolección y publicación se realiza por una institución académica que los provee de autoridad.
- La actualización de los datos tiene una periodicidad diaria.
- Tiene un acceso sencillo y gratuito.
- Cumple con los requisitos para ser utilizados en la propuesta, disponiendo de datos de todos los países para los que se realizará el estudio.

El repositorio contiene los datos de contagios globales acumulados en el archivo en formato csv "time_series_covid19_confirmed_global.csv", estructurado con las siguientes columnas:

- Province/State (Provincia/Estado): Contiene el estado o provincia en países que disponen de territorios lejanos la ubicación del territorio principal. Para el territorio principal, este campo está vacío.
- Country/Region (País/Región): País al que se refieren los datos. En el repositorio hay información de 195 países.
- Lat: Latitud de la ubicación del país, o provincia/estado, si corresponde.
- Long: Longitud de la ubicación del país, o provincia/estado, si corresponde.
- Las columnas restantes contienen los datos para cada una de las columnas anteriores, encabezadas la fecha que se corresponde con el dato. La serie temporal comienza el 22 de enero de 2020, y finaliza en la actualidad. En este caso, la mayor parte del análisis se ha realizado con los datos que fechan hasta el 21 de junio de 2021.

Se utiliza la programación del proyecto para realizar una fase inicial de análisis y transformación de la fuente de datos. En primer lugar, se ha eliminado la información que no se corresponda con los países de estudio y del territorio principal. Posteriormente, se ha

transformado la información de manera que exista una columna con cada una de las fechas del set, y una columna por país con la incidencia acumulada de casos por fecha.

Una representación de los datos recolectados a fecha de mayo de 2021, bajo este nuevo formato, puede observarse en la Figura 4.

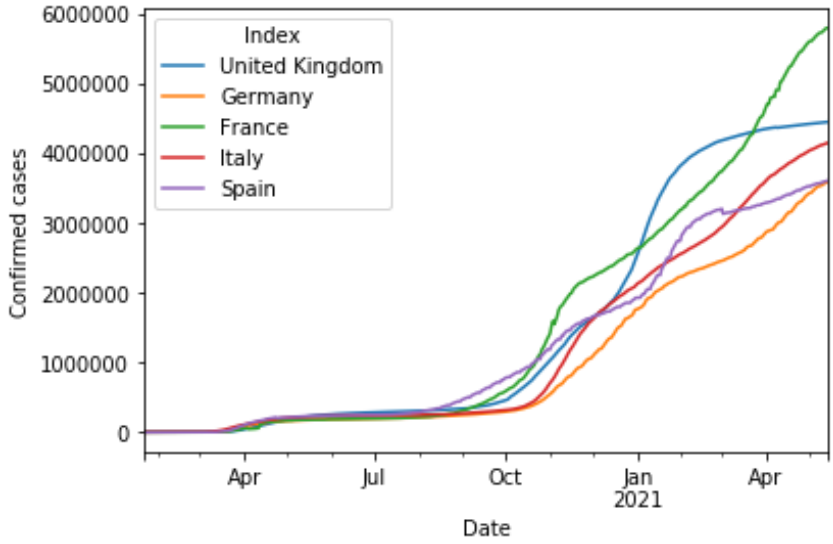

Figura 4. Visualización preliminar de los datos de casos acumulados de COVID-19 a fecha de 15 de mayo de 2021. (Elaboración propia)

De esta visualización inicial se puede apreciar que los datos de prácticamente todos los países siguen una tendencia cíclica, correspondiente con las olas de mayor incidencia. También se podría anticipar que las curvas de UK y ES parecen tener un estado de madurez mayor, y que previsiblemente la tendencia de contagios sea más baja que en otro países, especialmente FR. Adicionalmente, se puede destacar que al tener UK una curva casi completa, nos avanza que la forma final para la curva de contagios sigue una función sigmoide, en estado incompleto para la mayor parte de países.

4.1.2. Recolección y análisis de los datos de influenza estacional

Una de las comparaciones que más se ha realizado desde el inicio de la pandemia, es la del COVID-19 con la gripe estacional. La influenza estacional, o gripe estacional, también es provocada por un virus y, desde el punto de vista del cuadro sintomático, presenta síntomas en cierta medida similares, como la tos, el malestar general, o la fiebre. Sin embargo, el interés de la comparativa se fundamenta en que el virus que provoca la gripe es bien conocido por la

comunidad científica y permite realizar previsiones acerca del evolución de la epidemia de COVID-19.

En este caso, se propone el uso de una curva de contagios conocida de la gripe estacional para prevenir la futura incidencia del COVID-19. Existen varios motivos por los que esta enfermedad se considera apropiada:

1. Tiene una alta contagiosidad con respecto a otras enfermedades, también por vía aérea a través de expulsión de virus durante la respiración o el habla.
2. Tiene un impacto a nivel global.

Sin embargo, la primera limitación con respecto a su uso es que, frente a la gripe común, el COVID-19 no se considera una enfermedad estacional que remite de manera natural con el cambio de clima y temperatura. La segunda limitación observada, es que el COVID-19 presenta un patrón en olas que no se espera encontrar en los datos de gripe. Adicionalmente, el origen de estas olas se puede ver influido por causas naturales, como mutaciones y nuevas variantes del virus, o la acción humana, relajación o endurecimiento de las medidas de contención, sin poder separar fácilmente cada componente.

Se procede a realizar un rápido análisis de los datos de gripe, para identificar qué año puede ser el más adecuado para ser utilizado en el modelo. Los datos se obtienen de la página web de la Organización Mundial de la Salud (WHO Global Influenza Surveillance and Response System (GISRS), 2021), pudiendo acceder a los datos de incidencia semanales de múltiples países desde el año 1995. En este caso la frecuencia de los datos es semanal y, por tanto, se decide extraer los datos de los países europeos EU5 posteriores a 2010, y anteriores a 2020. Se considera que los datos anteriores a 2010 no son comparables pues se deberá utilizar datos de un periodo lo más cercanos posible al actual, para poder capturar un estilo de vida, y nivel de globalización similar al actual. Por otro lado, se descartan los datos del año 2020 y 2021 ya que, debido al surgimiento de la epidemia de COVID, la gripe estacional generó datos muy bajos y controlados por las medidas de contención.

El formato de los datos obtenidos es el siguiente:

- Country (País), Who Region (Región OMS), Influenza transmission zone (Región de transmisión de la gripe): localizan geográficamente los datos.

- Year (Año), Week (Semana), Start date (Fecha de inicio), End date (Fecha de finalización): localizan temporalmente los datos.
- Número de espécimes: número de espécimes recibidos y número de espécimes procesados
- Número de virus de gripe A detectados por subtipo: A (H1), A (H1N1) pdm09, A (H3), A (H5), A (sin subtupo), A (Total).
- Número de virus de gripe B detectados por subtipo: B (linaje Yamagata), B (linaje Victoria), B (linaje no determinado), B (Total).
- Número total de positivos del virus de influenza estacional.
- Número total de negativos del virus de influenza estacional.
- Tipo de brote.

Al realizar un representación simple de los datos del número total de positivos de influenza entre 2010 y 2020 por país, se obtiene el gráfico de la Figura 5.

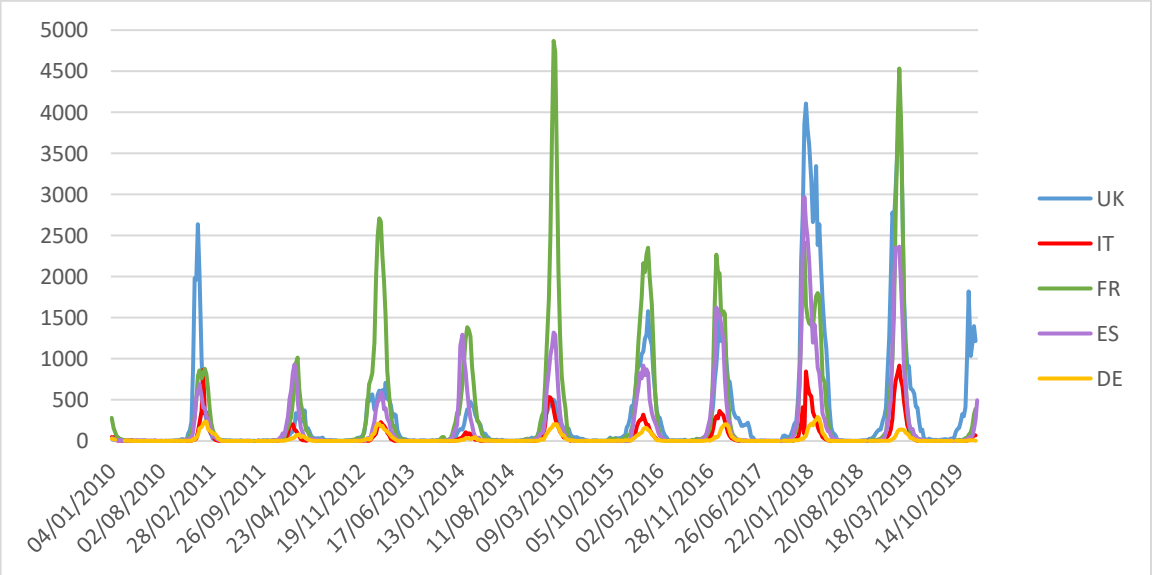

Figura 5. Número total de positivos de influenza estacional semanales para EU5 desde el año 2010 al año 2019 según los datos de la OMS. (Elaboración propia)

Destacan los brotes del final de año de 2014, 2017, y 2018, y en especial estos dos últimos, ya que tuvieron mayor impacto de manera generalizada en todos los países. Para asegurar la elección del pico de contagios, se incluye el gráfico de la incidencia acumulada para los picos de 2014 a 2019, asumiendo que un ciclo acaba y empieza el 1 de julio.

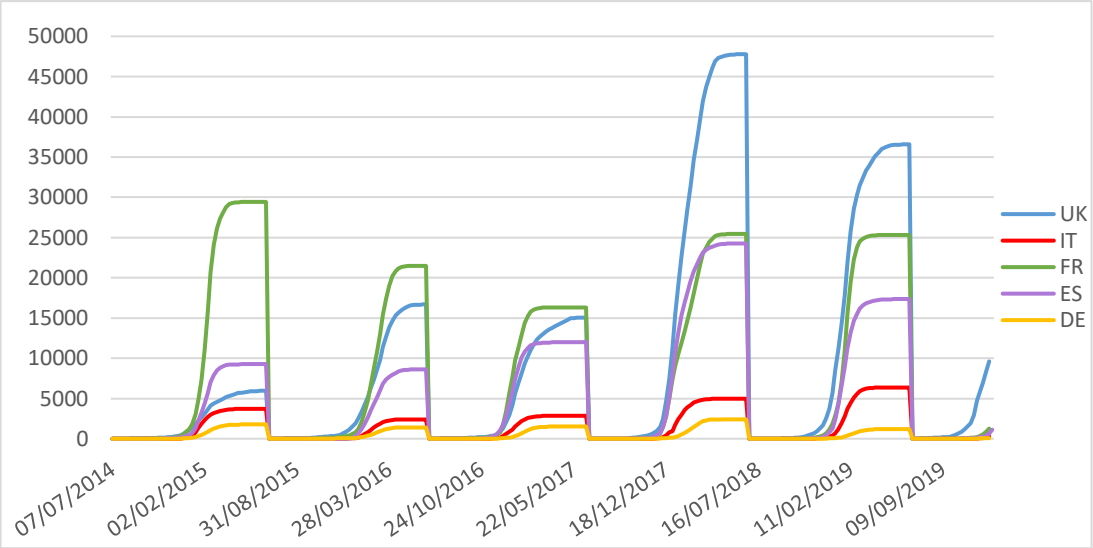

Figura 6. Número total de positivos de influenza estacional acumulados por ciclo anual para EU5. (Elaboración propia)

Observando la Figura 6, se observa que los datos de finales 2018 presentan una representación más parecida a la que muestran los datos de COVID-19 de la Figura 4, siendo además los datos del segundo año con un número de contagios acumulados de la década pasada. También se considera un punto a favor, que la muestra seleccionada es la más reciente de todas las que se ha recolectado, de manera que se puedan maximizar la similitud de variables ocultas relacionadas con las costumbres contemporáneas de cada país

Para la utilización de los datos en el modelo, se almacenan en un archivo csv la fecha de inicio del periodo semanal, y una columna por país que contenga la incidencia acumulada, desde el 1 julio de 2018 hasta el 1 de julio de 2019.

4.1.3. Recolección y análisis de los datos de la influenza H1N1 de 2009: Gripe A

El subtipo H1N1 es el provocante de grandes pandemias mundiales, la más destacable, la pandemia de influenza de 1918, para la cual estimaciones recientes sugieren que provocó entre 50 y 100 millones de muertos en todo el mundo (Spreeuwenberg, Kroneman, & Paget, 2018). Esta epidemia se expandió rápidamente durante los años 1918 y 1920, con tres grandes olas de mortalidad, provocando sintomatología que principalmente generaban fiebre y

afectaban a las vías respiratorias. Sin embargo, aunque ésta se considere la mayor epidemia del siglo pasado, presenta diversos problemas a la hora de ser utilizada para el modelo:

1. No se ha encontrado una fuente de datos clara que recoja los datos de contagios debidos al subtipo de influenza H1N1 durante el periodo de 1918 a 1920.
2. Existen estimaciones acerca de la curva de mortalidad causada por el virus de influenza de 1918, como la que realiza Ansart y otros (2009). Sin embargo, aunque los estudios son públicos, no se ha localizado una fuente de datos que refleje la línea temporal de los resultados obtenidos en dichos estudios. Adicionalmente, de utilizar estos datos se requeriría asumir que la curva de contagios es similar en forma a la curva de mortalidad.
3. Estas estimaciones tienen una variabilidad muy alta dependiendo del estudio que se observe, por lo que de tomar una fuente no se podrían considerar datos de gran exactitud.
4. Los patrones estimados para las muertes por la influenza de 1918 muestran 3 olas mucho más diferenciadas que las que se están observando par los casos de COVID-19. Este patrón por olas podría alterar la previsión final provocando un resultado erróneo.

Sin embargo, el subtipo H1N1 provocó otra pandemia global mucho más reciente: la pandemia de Gripe A. La pandemia de gripe A se expandió durante la segunda parte del año 2009 y principios de 2010 a nivel global. Esta pandemia no tuvo un impacto comparable con la de la COVID-19, ya que tuvo una mortalidad baja (entre 150 y 575 mil víctimas) cuando las estimaciones apuntan a que el contagio se produjo entre el 11 y el 21% de la población mundial. (Dawood, y otros, 2012)

La recolección de datos de esta pandemia se ha realizado nuevamente en el portal de datos de la OMS, filtrando para este caso el brote de 2019, y únicamente aquellos casos identificados con el subtipo de influenza H1N1pdm09. La representación de dichos datos puede consultarse en la Figura 7 y la Figura 8.

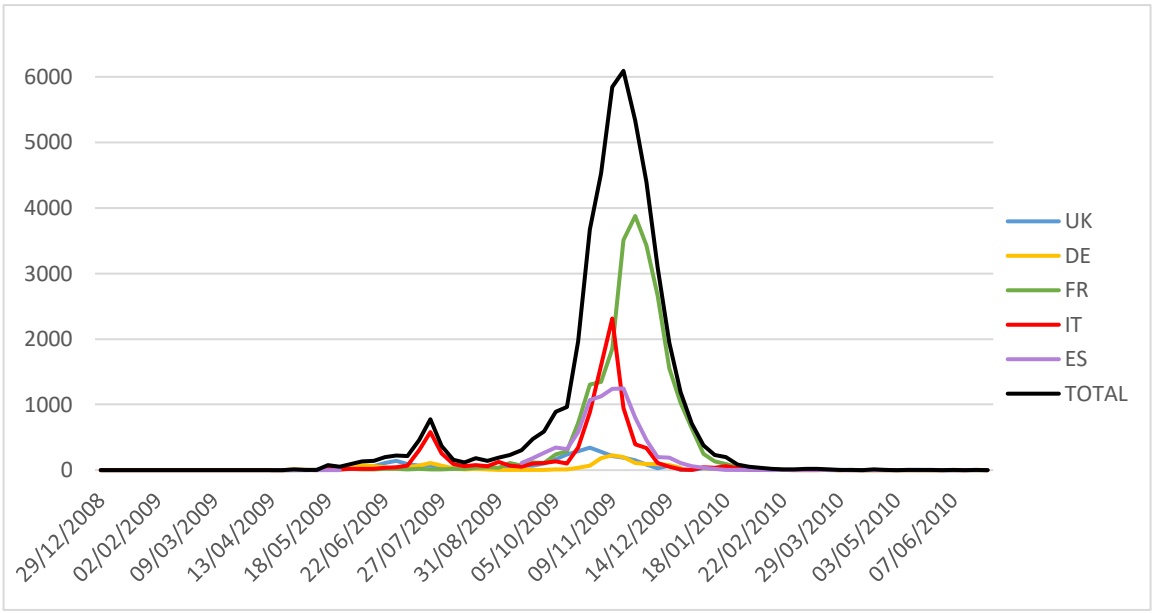

Figura 7. Número total de positivos de influenza A (H1N1pdm09) semanales por ciclo anual para EU5. (Elaboración propia)

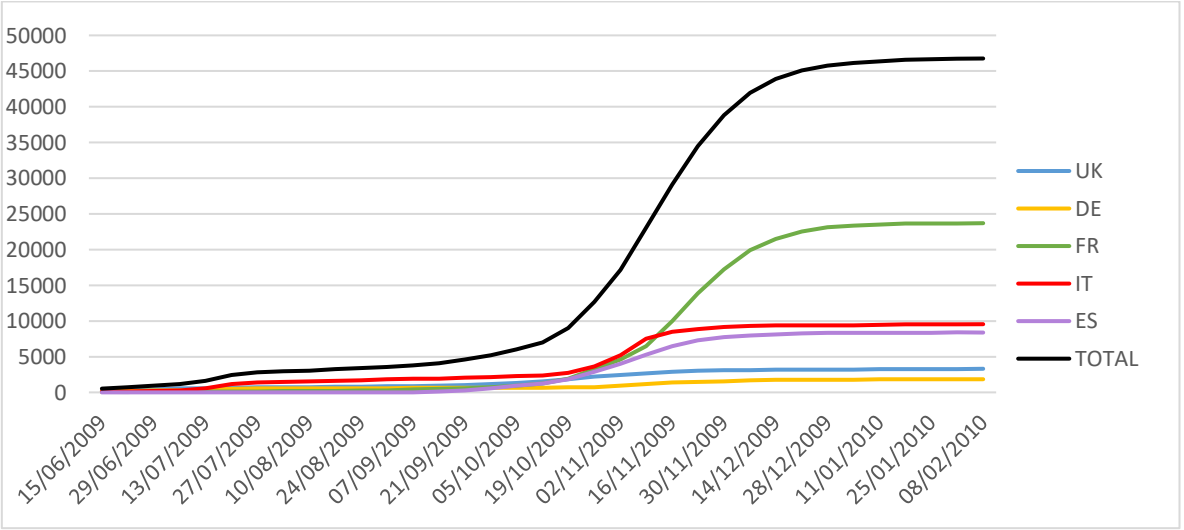

Figura 8. Número total de positivos de influenza A (H1N1pdm09) acumulados por ciclo anual para EU5. (Elaboración propia)

Al igual que se ha realizado para los datos de influenza estacional, se han almacenado en un csv los datos de la incidencia acumulada por país y se ha agregado una columna adicional que contenga los datos de la fecha inicial de la semana a la que se refieren. Adicionalmente, se ha decidido eliminar los datos anteriores al 27 de julio de 2009, ya que el escalón que puede

observarse al inicio de la gráfica de la Figura 8 no tiene gran impacto en la estimación de la parte final de la gráfica, pero puede perjudicar el análisis y alterar los resultados en ciertas transformaciones de los datos.

Se destaca en la Figura 8 la pequeña magnitud de las curvas de Reino Unido y Alemania, no llegando a superar los 5000 casos. En estos casos las perturbaciones aleatorias representan un componente mucho mayor, considerando la magnitud de la muestra, y se prevén dificultades al utilizar estar curvas como base para el aprendizaje utilizando el procedimiento que se propone en el presente trabajo.

4.1.4. Transformación de los datos tipo fecha en "Microsoft Excel date serial numbers"

Uno de los problemas fundamentales de la utilización de técnicas de análisis de datos, y algoritmos matemáticos en series temporales, es la interpretación del eje temporal. En muchos casos, los datos introducidos con formato de fecha no son compatibles con muchas de las funciones existentes y que desean utilizarse para la construcción del modelo. Por este motivo, se desea utilizar una transformación del dato de fecha a un formato número, que crezca a un ritmo unitario.

Se concluyó que la equivalencia que realiza el software de Microsoft Excel para la interpretación de los datos de fechas podría ser adecuado para el caso. Microsoft Excel utiliza el número 1 para el 1 de enero de 1900, y asigna en sentido creciente el siguiente número en la serie, a la siguiente fecha. De esta manera al 2 de enero de 1900 le corresponde el número 2, y al 1 de mayo de 2021 le correspondería el número 44317. A este número de ahora en adelante se le denominará Excel date serial number (EDSN).

Por tanto, se ha desarrollado una pareja de funciones capaz de realizar las transformaciones: excel_date convierte una fecha en EDSN utilizando la ecuación *(1)*, y inv_excel_date convierte un ESDN en una fecha con la ecuación *(2).*

$$excel\ date\ serial\ number = (fecha - (1989/12/30))_{días} \qquad (1)$$

$$fecha = (1989/12/30) + (excel\ date\ serial\ number)_{días} \qquad (2)$$

Se utiliza esta conversión para incluir en cada conjunto de datos una columna adicional una columna adicional con el ESDN equivalente a los datos de la columna que contiene la línea temporal. De esta manera, dependiendo del caso, podría utilizarse un formato otro.

4.2. Previsión temporal preliminar y determinación del punto de madurez de las curvas de COVID-19

Uno de los elementos clave del procedimiento para el desarrollo del modelo, es la localización de punto de madurez de las curvas de COVID-19 con respecto a su equivalente en las curvas de referencia utilizadas, en este caso, la curva de casos de influenza A y de influenza estacional para el brote de 2018-19. Tras el análisis preliminar del apartado anterior, se observó que todas la curvas de infección acumulada tenían una forma que podría relacionarse con la de una curva logística con un periodo de crecimiento exponencial, seguida de un periodo de decrecimiento exponencial. Con esta observación, se decide utilizar la transformación de las diferentes curvas de contagios, tanto de COVID-19 como de otros brotes de referencia, en una curva logística de ecuación conocida, para realizar una equivalencia comparable entre ambas curvas y así obtener el denominado punto de madurez de la curva de COVID-19 sobre la curva de referencia.

Se define la función logística con la ecuación *(3)*, de manera que estará definida con la determinación del valor de los parámetros c, a y b, y considerando t la variable independiente temporal.

$$eq\_logistica = \frac{c}{1 + a \cdot e^{-b \cdot t}} \qquad (3)$$

Estos parámetros a, b y c se calculan utilizando la función curve_fit de scipy. Esta función permite utiliza métodos de reducción de mínimos cuadrados no lineares para encontrar un ajuste para una función f, a partir de un conjunto de datos. En este caso, la función f es la ecuación (3) y el conjunto de datos será en cada caso los datos de contagios acumulados, frente al ESDN. Cabe destacar que se ha ajustado el ESDN de manera que en todos los casos la serie empiece en 0, ya que al utilizar el ESDN directamente, la variabilidad en los datos temporales resultaba insignificante frente a su módulo, y por tanto la función no lograba generar una equivalencia con éxito.

Si cogiéramos como ejemplo, el caso de España (ES), y aplicamos el algoritmo, las ecuaciones de las tres curvas de contagios serían las siguientes:

$$eq\_logistica_{COVID-19,ES} = \frac{3957342{,}634}{1 + 291{,}884 \cdot e^{-0{,}0169494 \cdot t}}$$

$$eq\_logistica_{SeasonalFlu18,ES} = \frac{17248.685}{1 + 68674803{,}039 \cdot e^{-0{,}0833047 \cdot t}}$$

$$eq\_logistica_{H1N1pdm09,ES} = \frac{8397.025}{1 + 4239{,}203 \cdot e^{-0{,}0848251 \cdot t}}$$

Este ajuste puede observarse en la Figura 9 en la que se puede comparar la progresión de los casos reales en la línea de puntos azul, frente al modelo de aproximación a una curva logística de la línea roja. En el caso de los datos de COVID-19, la función ha relacionado la curva incompleta de casos, con la primera parte de la curva logística, permitiendo así estimar de manera preliminar el final de la curva.

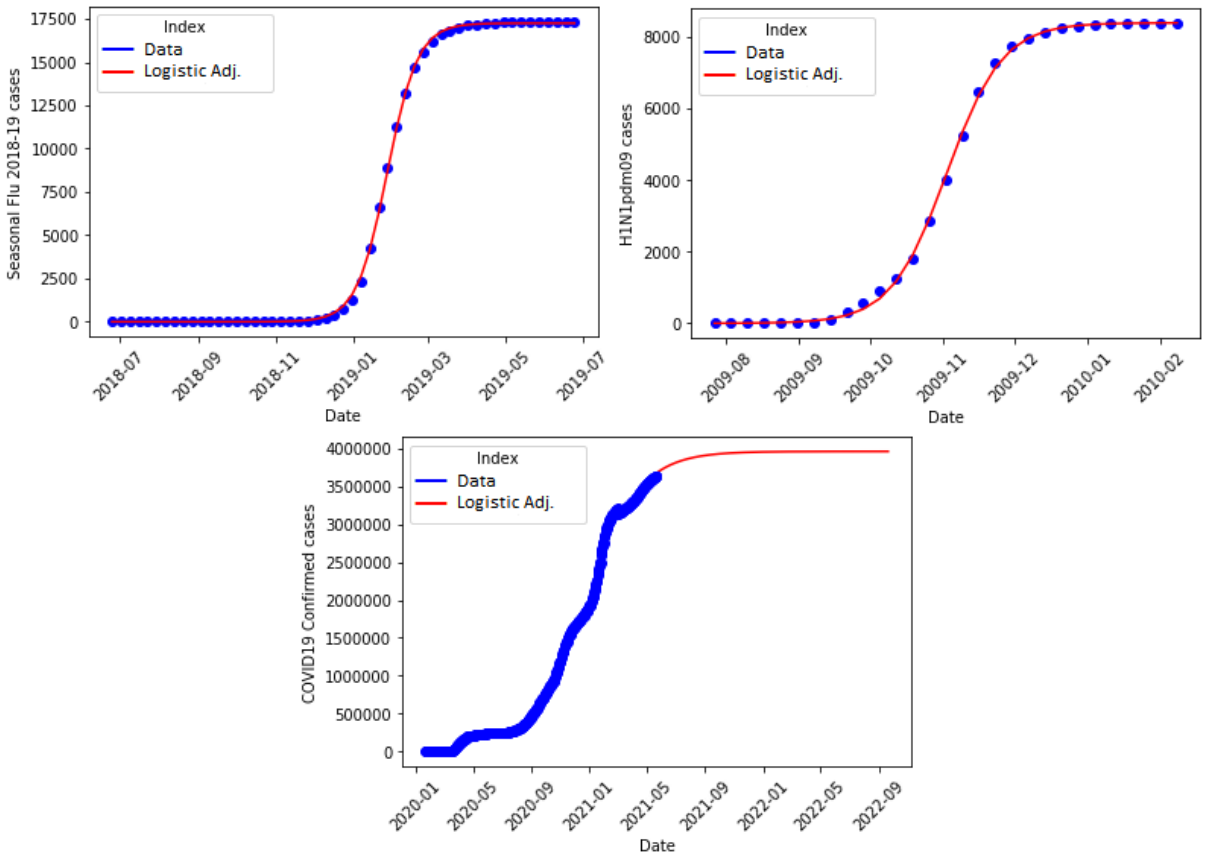

Figura 9. Comparativa de las curvas reales de contagios de ES a la aproximación por medio de una función logística. (Elaboración propia)

De esta manera se puede proporcionar la misma estimación para el resto de los países europeos (EU5), cuyos resultados son los mostrados en la Figura 10.

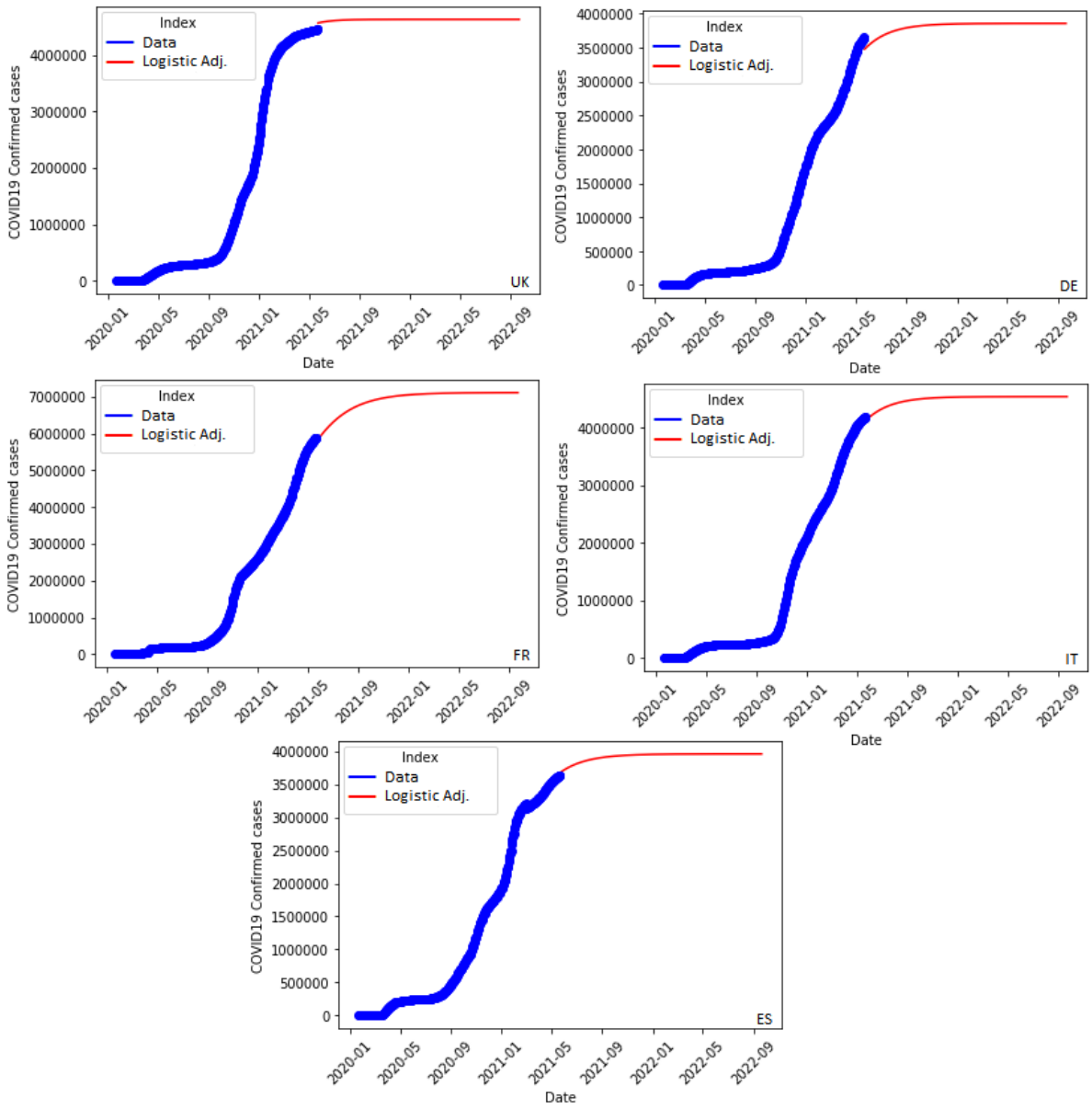

Figura 10. Previsión preliminar de los casos de COVID19 para EU5 basada en la aproximación a una curva logística. (Elaboración propia)

Sin embargo, aunque esta representación constituye una predicción preliminar, existen diversos motivos por los que el desarrollo de un modelo más elaborado es necesario. El motivo principal, es que la presente previsión está basada en la suposición de que las curvas de COVID-19 presentan una forma sigmoide perfecta, asumiendo también la antisimetría entre la parte de crecimiento y decrecimiento exponencial de una curva logística. Sería mucho más

adecuado continuar con el procedimiento planteado inicialmente para el proyecto, y utilizar la evolución de la forma de una curva de contagios real para realizar la previsión.

4.2.1. Ajuste manual de las curvas influenza A de Reino Unido, Alemania e Italia.

Una vez, evaluado el procedimiento para todos los países, se comprueba que el método para las transformación no es compatible con las curvas de Influenza A H1N1 para Alemania y Reino Unido. Ya se predijo en el apartado Recolección y análisis de los datos de la influenza H1N1 de 2009: Gripe A que, debido a los datos bajos de las muestras, es posible que no se pudieran utilizar en la metodología. En los datos de Italia encuentra el mismo problema, no porque los datos sean muy bajos, si no por el escalón en los contagios en el mes de julio de 2009. Como se observa en la Figura 11, el resultado el ajuste de la curva a la función logística no se aproxima a la curva real de los datos, por lo que no se recomienda su utilización con el proceso propuesto en este documento. De utilizarse, se podría situar erróneamente el punto de madurez de la curva, y así obtener un resultado final con una tendencia errónea.

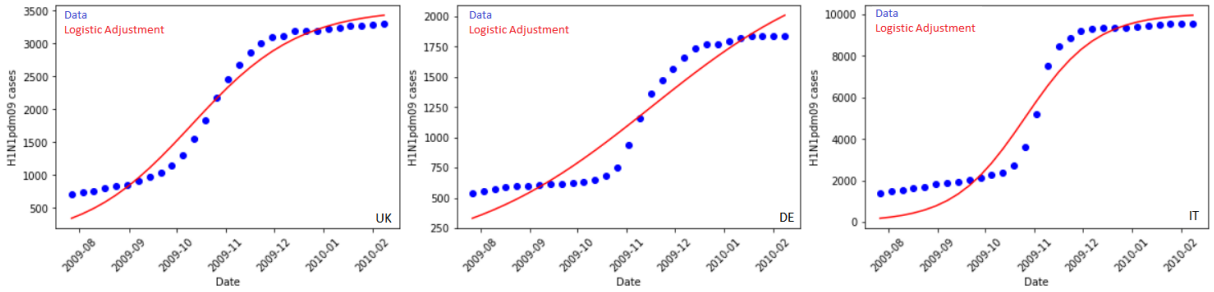

Figura 11. Ajuste logístico automático obtenido para los datos de H1N1 de UK, DE, y IT. (Elaboración propia)

No obstante, se ha ajustado manualmente las curvas de H1N1 de estos 3 países, intentando priorizar un ajuste más acertado para el final de la curva. De esta manera, en lugar de utilizar el ajuste automático de la curva, los parámetros de estas curvas serán fijos y determinados de manera iterativa, hasta lograr la mejor aproximación. En la X se muestra el resultado de este ajuste manual.

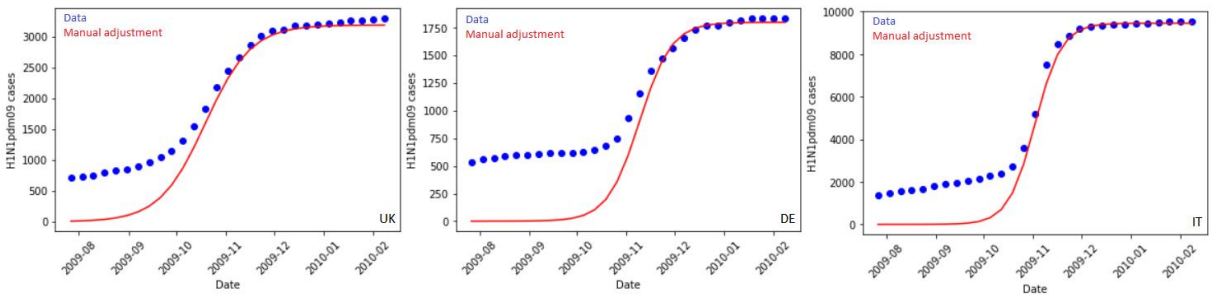

Figura 12. Ajuste logístico manual para los datos de H1N1 de UK, DE, y IT. (Elaboración propia)

Los parámetros resultantes de todas las curvas base ajustadas pueden consultarse en la Tabla 9 y la Tabla 10. Los parámetros de la curva de COVID-19 son dependientes de su avance.

	UK	DE	FR	IT	ES
a	280481	16532857	1886232227	70104273	68674699
b	0,058	0,071	0,097	0,081	0,083
c	36207,681	1189,739	25359,494	6357,16	17248,684

Tabla 9. Parámetros del ajuste logístico de las curvas de influenza estacional de 2018/19. (Elaboración propia)

	UK*	DE*	FR	IT*	ES
a	357,809	36315,502	17756,738	128027,453	4239,208
b	0,07	0,1	0,084	0,12	0,085
c	3192	1800	23866,007	9452	8397,025
					*Ajuste manual

Tabla 10. Parámetros del ajuste logístico de las curvas de Influenza A H1N1. (Elaboración propia)

4.2.2. Localización del punto de madurez de las curvas y ajuste de la escala temporal

La representación de las curvas mediante el modelo logístico nos permite obtener ecuaciones definidas. Estas ecuaciones serán utilizadas en esta sección para localizar el punto de madurez de las curvas. El modelo logístico que matemáticamente se representa con la ecuación *(3)*, tiene una representación gráfica que viene reflejado en la Figura 13, donde se pueden apreciar los ejes de las asíntotas y el punto de inflexión central. La Figura 13 proviene de una la

explicación que Espina Marconi realiza del modelo logístico en la Serie de estudios económicos (1984), que ha sido alterada para coincidir su nomenclatura con la de la ecuación *(3)*.

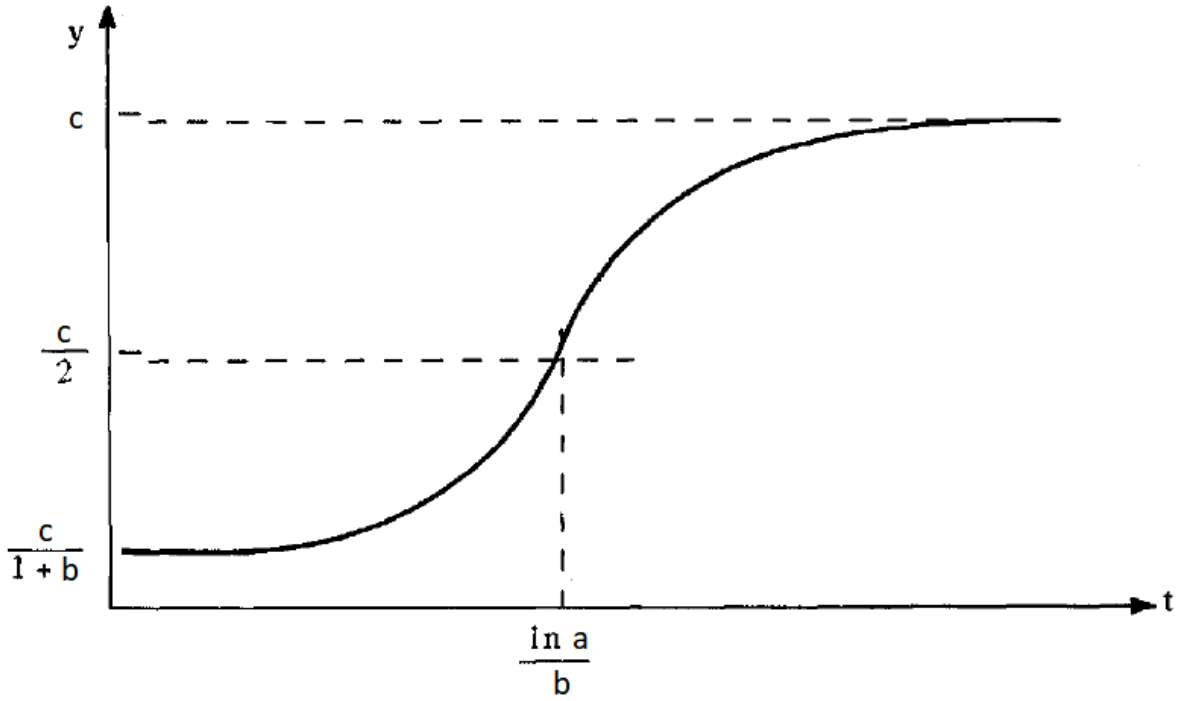

Figura 13. Representación gráfica del modelo logístico. **(Espina Marconi, 1984)** *(Imagen alterada)*

Se parte de que para la curva 1 y la curva 2 tienen las siguiente ecuaciones:

$$y_1(t) = \frac{c_1}{1 + a_1 \cdot e^{-b_1 \cdot t}} \; ; \; y_2(t) = \frac{c_2}{1 + a_2 \cdot e^{-b_2 \cdot t}}$$

Tpm se corresponde con la fecha del último dato de COVID, y permitiría obtener así el valor de casos para esa fecha según el modelo logístico.

$$y_{1,tpm} = \frac{c_1}{1 + a_1 \cdot e^{-b_1 \cdot tpm1}}$$

Equiparando ambos modelos logísticos según la Figura 13 se obtiene la siguiente igualdad:

$$\frac{c_1 - y_{1,tpm}}{c_1 - \frac{c_1}{2}} = \frac{c_2 - y_{2,tpm}}{c_2 - \frac{c_2}{2}}$$

$$c_2 - (c_1 - y_{1,tpm)} \cdot \frac{c_2}{c_1} = y_{2,tpm}$$

Finalmente, con la ecuación *(3)* se obtiene la fecha del punto de madurez de la curva 1 sobre la curva 2.

$$t_{pm2} = \frac{\ln\left(\left(\frac{c_2}{y_2(t)} - 1\right)\frac{1}{a_2}\right)}{-b_2} \qquad (4)$$

Al implementar este razona en el software y aplicarla a la curva de España, resulta en que el día 2021-05-23, sería el equivalente 2009-12-02 para los datos de la curva de influenza H1N1pdm09, y equivalente al 2019-02-27 en la curva de gripe estacional de 2018. Si se observa nuevamente la Figura 9, este resultado gráficamente demostraría la equivalencia entre los datos contagios acumulados de COVID-19, y los datos recogidos para el gripe A hasta el 2 de diciembre de 2009, o los datos de gripe estacional hasta el 17 de febrero de 2019.

De manera análoga se va a acotar el punto inicial equivalente en los datos de la curva de referencia, utilizando el primer dato que registrado para la curva de COVID-19 que data del 22 de enero de 2021. Utilizando la misma metodología para el caso de España, se concluye que este punto es equivalente a la fecha de 2009-08-27 para los datos de influenza H1N1, y a la de 2018-11-20 para los datos de gripe estacional.

Se decide utilizar la metodología para aplicar una transformación a las curvas de referencia y lograr obtener así una curva que coincida en la escala temporal con los datos de COVID-19. Esta transformación permitirá utilizar para todos los datos de contagio la misma escala temporal y así reducir el problema en una dimensión, de manera que los algoritmos que se utilicen para encontrar la transformación de una curva en otra solo deban tener en cuenta los datos de contagios.

Los paso a seguir para obtener el set de datos transformado son los siguientes: para cada punto temporal en la escala de datos de COVID-19 se obtiene la fecha equivalente en las curvas de referencia. Este dato nos permite asignar valores de la curva de referencia a la escala temporal de la curva de COVID-19. Sin embargo, las fechas obtenidas pueden encontrarse dentro de un intervalo, por lo que se ha decido aplicar una interpolación lineal entre los valores en los extremos del intervalo, para localizar el valor del punto intermedio. Al realizar esta última etapa se incurre en una nueva aproximación que conlleva un error, sin embargo, globalmente considera que de esta manera se genera un error más pequeño que usar el valor de la ecuación logística equivalente.

Se puede observar en la Figura 14 la evaluación el método descrito anteriormente aplicada al país modelo. Se comprueba que la transformación se ha efectuado correctamente con un resultado prácticamente idéntico al original

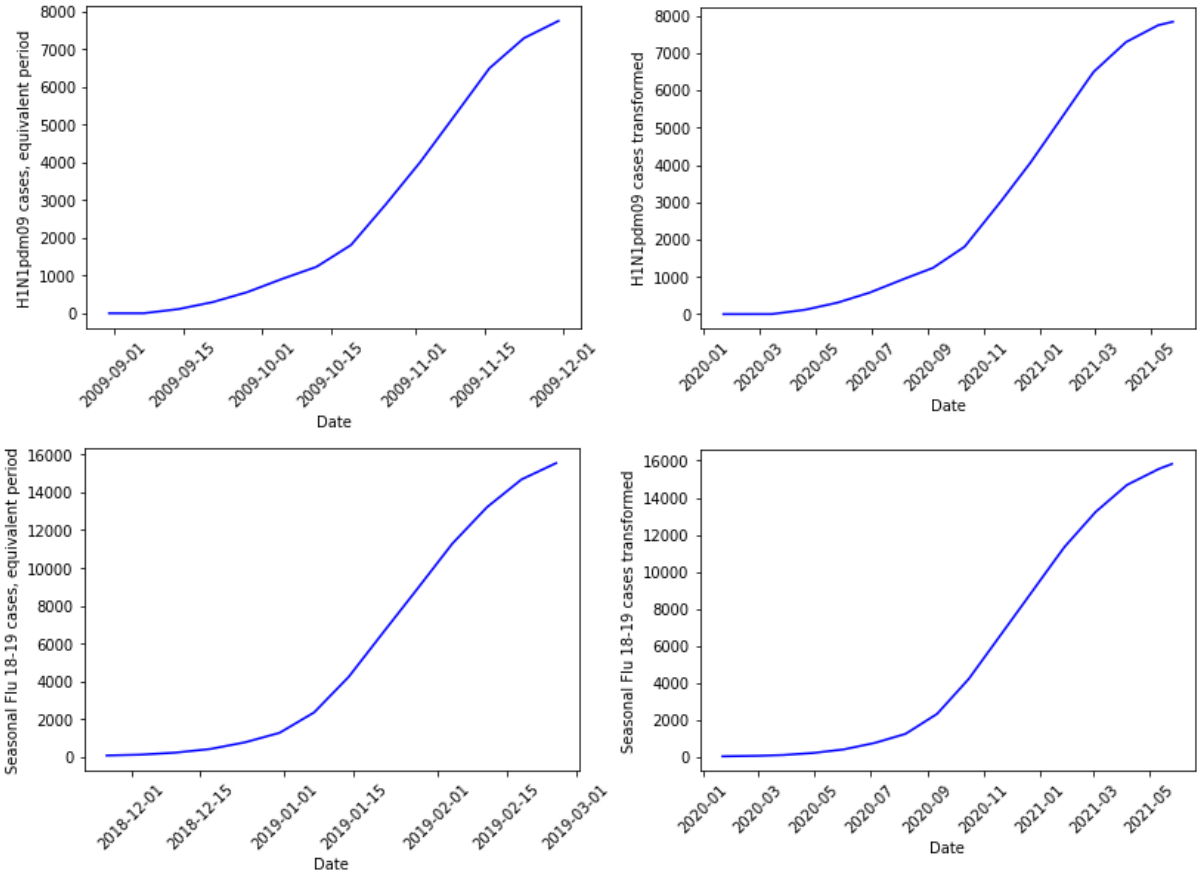

Figura 14. Comparativa de las curvas de referencia tras la transformación de su escala temporal y las curvas originales en el periodo equivalente. (Elaboración propia)

Finalmente se realiza una etapa final en el procesado de las muestras: el escalado. La magnitud de los datos de COVID-19 es mucho mayor que la magnitud de cualquiera de las curvas utilizadas como base, por lo que un escalado evita que el algoritmo tenga que estimar el salto de magnitud entre la entrada y la salida del modelo, con su consecuente error. Por lo general, el escalado es considerado un ejercicio de buenas prácticas en el aprendizaje de series con algoritmos de regresión. En este caso, se ha realizado un escalado sencillo, definido en la ecuación *(5)*.

$$coef_{escalado} = \frac{X_{max,COVID} - X_{min,COVID}}{X_{max,base} - X_{min,base}} \qquad (5)$$

El coeficiente se multiplicará por los datos de entrenamiento antes de ser utilizados por el algoritmo, pero también a los datos que alimentarán al modelo construido para obtener la predicción.

4.3. Algoritmos de Machine Learning para la transformación de la curva de referencia en los datos conocidos de COVID-19

Tras la preparación descrita en los apartados anteriores, se puede iniciar el estudio de predicciones utilizando algoritmos de Machine Learning. En este caso, se desea utilizar algoritmos de la rama de la regresión ya que, dentro del aprendizaje supervisado, estos algoritmos permiten predecir una salida a partir de una entrada determinada. Para este caso, la entrada sería la curva base, y la salida la curva de COVID-19. Se probarán en este estudio aquellos algoritmos que mejor comportamiento muestran en la documentación referenciada en capítulo Contexto y estado del arte: el algoritmo Long-Short-Term-Memory (LSTM), Linear Regression (LR), Least Absolute Shrinkage and Selection Operator (LASSO). Se decide utilizar estos algoritmos por dos motivos principales: en primer lugar, han probado tener buen resultado en la predicción de curvas epidemiológicas demostrándose en los resultados de dichos experimentos; en segundo lugar, la comparativa con estos experimentos para la validación del procedimiento es más justa si se utilizan algoritmos similares. No obstante, en el apartado 5.1 Trabajos Futuros se propone la aplicación de algoritmos alternativos utilizando el mismo procedimiento para evaluar si se pueden obtener mejores predicciones.

El objetivo en este capítulo es aplicar cada uno de estos algoritmos a una combinación de datos COVID-H1N1 y COVID-Influenza Estacional, para encontrar el modelo que mejor refleja la transformación entre los datos de una curva base y la curva del COVID-19. También se evaluará la calidad de la predicción durante un periodo de prueba de 10 días. En ambos casos, se utilizará la raíz del error cuadrático medio (RMSE) como indicador de calidad del modelo, tras observar que esta métrica es altamente utilizada en los estudios recolectados para el estado del arte. Adicionalmente, el RMSE es una métrica que permite cuantificar el error sin cancelaciones entre desviaciones positivas y negativas, por lo que resulta fiable para la comparación de predicciones y ajustes.

El procedimiento seguido para el entrenamiento y predicción aplicado a todos los algoritmos y curvas base es el siguiente:

1. Declaración y configuración del modelo, que deberá ser uno diferente para cada algoritmo y curva base de datos de influenza estacional o influenza H1N1.
2. Procesamiento de los datos hasta el punto de madurez de COVID-19 y curva base para su utilización.
3. Entrenamiento del modelo y sus parámetros internos utilizando como entrada los datos de la curva base anteriores al punto de madurez e introduciendo como salida esperada los datos de la curva de COVID-19.
4. Procesamiento de los datos posteriores al punto de madurez de la curva base.
5. Predicción de los datos futuros de COVID-19 alimentando el modelo entrenado con los datos de la curva base posteriores al punto de madurez.
6. Obtener la métrica de análisis RSME del modelo final ajustado comparando la salida del modelo con los datos reales de COVID-19

Se describirá a continuación los detalles de la aplicación de este proceso a cada uno de los algoritmos utilizados.

4.3.1. Ajuste con el algoritmo Long-Short-Term Memory (LSTM)

Este algoritmo es el que mejores resultados mostró en el experimento de Tian, Luthra, & Zhang (2020). Para aplicar el siguiente algoritmo, se utiliza la librería Keras de Python, y en particular que se utilizara el algoritmo LSTM Sequential. Todos los modelos LSTM propuestos para el ejercicio contienen una única variable, pues es uno de los primeros parámetros a configurar en el algoritmo LSTM: el número de variables (number of features), en este caso 1. Por otro lado, tradicionalmente este tipo de algoritmos son utilizados en ejercicios de regresión en los que se utilizan los datos pasado para predecir el futuro, que es precisamente el modo de funcionamiento de los modelos descritos en el estado del arte. En estos casos, la variable de entrada al modelo se replica en conjuntos desfasados, de manera que el modelo aprenda sobre la evolución de la salida para cada uno de los saltos en la variable de entrada o time steps. Para este caso, en el que se quiere transformar la curva base en la curva de COVID, el número de time steps es 1 para usar el algoritmo. El tercer parámetro es el número de

muestras que nuevamente, es solo 1. La entrada deberá ser tridimensional, aunque solo contenga información una de las dimensiones.

EN la prueba inicial se utiliza la siguiente configuración para un LSTM sencillo o Vanilla. Este modelo utiliza apenas una capa oculta y una capa de salida para proporcionar una predicción. El ajuste del modelo se realiza utilizando el error cuadrático medio.

Se decide probar si el resultado mejora utilizando una variación del algoritmo LSTM. El algoritmo LSTM bidireccional permite que el modelo aprenda la secuencia hacia delante (forward) y hacia atrás (backwards), y concatene ambas representaciones. Con respecto a la configuración, se utilizará la misma que se propuso para el modelo LSTM básico.

4.3.2. Ajuste con el algoritmo Linear Regression (LR)

Los mejores resultados para el experimento de Rustam, y otros (2020) se obtuvieron con un modelo de regresión lineal. Para el procesamiento de los datos de COVID-19 con el algoritmo LR se utilizará la librería de Python sklearn, importando los métodos de LinearRegression. La configuración de este algoritmo es la más sencilla, pues apenas consta de parámetros; tan solo se deberá introducir los datos de entrada de la curva base y de la curva de COVID-19 para el periodo conocido. El modelo se encargará de encontrar la relación entre ambas curvas asumiendo que se sigue una ecuación lineal.

4.3.3. Ajuste con el algoritmo Least Absolute Shrinkage and Selection Operator (LASSO)

Es otro de los algoritmos que mostró un resultado favorable en el análisis de Rustam, y otros (2020). Para la implementación de este algoritmo se utiliza nuevamente la librería sklearn, y en particular los métodos de LassoCV. En este método, su parámtro principal es α (alpha). Este parámetro determina el balance entre la minimización de la suma de cuadrados de los residuos y la minimización de la suma de los coeficientes de la regresión. El funcionamiento consistiría en excluir variables para mejorar la precisión de la estimación, y más variables serán excluidas con un valor α más alto. Se decide utilizar la versión LassoCV en lugar de Lasso, para que el algoritmo encuentre el mejor ajuste de α. La configuración adicional es muy sencilla: tal y como se realiza en el resto de los algoritmos, se introducen los datos conocidos de COVID-19 junto con los datos de la curva base para encontrar la transformación.

4.4. Resultados y análisis de los modelos obtenidos

Este apartad refleja los resultados de la experimentación entre las diversas fuentes de datos, algoritmos, y países a los que serán aplicados. En total, para cada país existe un total de 8 combinaciones, 2 curvas base como entrada aplicadas a 8 algoritmos de Machine Learning diferentes. El objetivo principal de esta tarea es determinar qué modelo funciona mejor para cada país y sacar conclusiones de la exactitud obtenida. Tras el entrenamiento de cada uno de los modelos con los datos anteriores al punto de madurez, se han introducido los datos de curva base posteriores al punto de madurez para predecir el comportamiento futuro del COVID-19. Dado que las curvas de COVID-19 de UK y de ES tienen una gran madurez, acercándose a una estabilidad en los contagios acumulados, se ha decido utilizar los datos hasta el 23 de abril de 2021, de manera que se pueda probar el procedimiento con una curva de la epidemia más incompleta. Para el resto de los países se ha usado el conjunto de datos de COVID-19 completos, con fechas de 21 de junio de 2021.

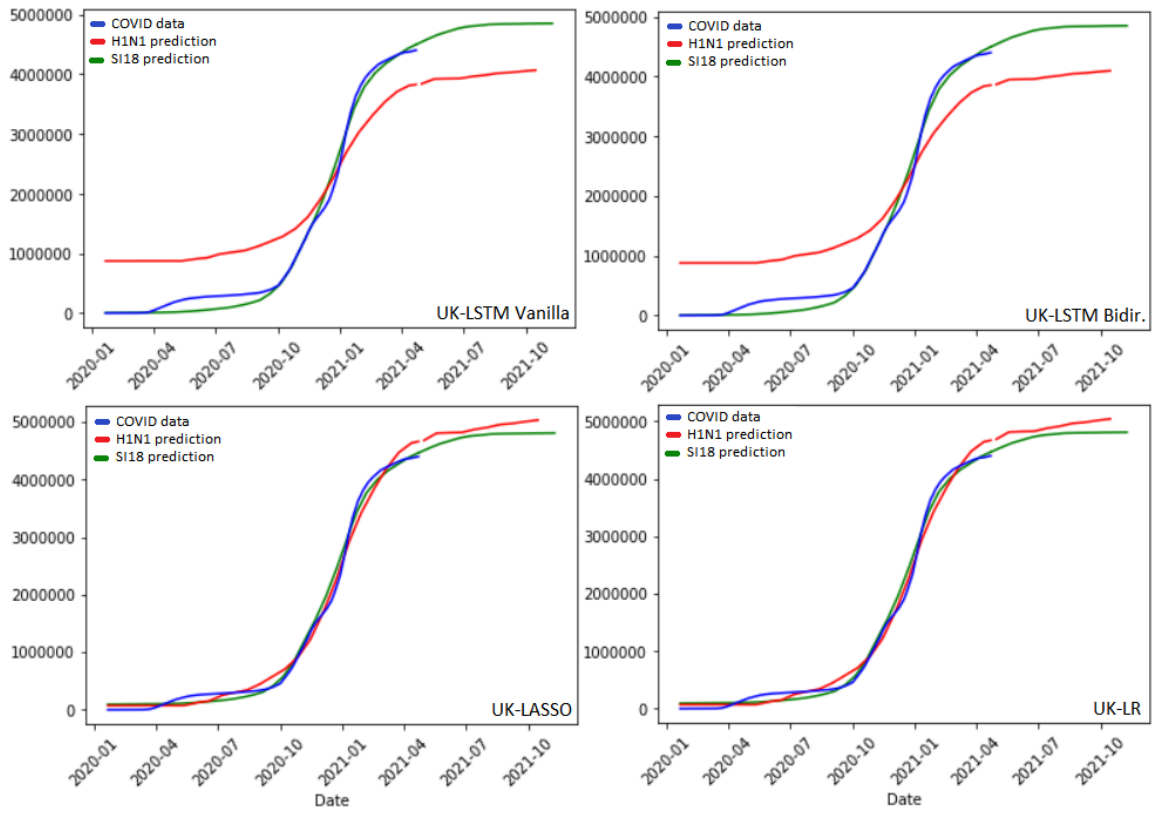

Figura 15. Predicción de la incidencia acumulada de COVID-19 para UK a partir del 23 de abril de 2021 para cada modelo estudiado. (Elaboración propia)

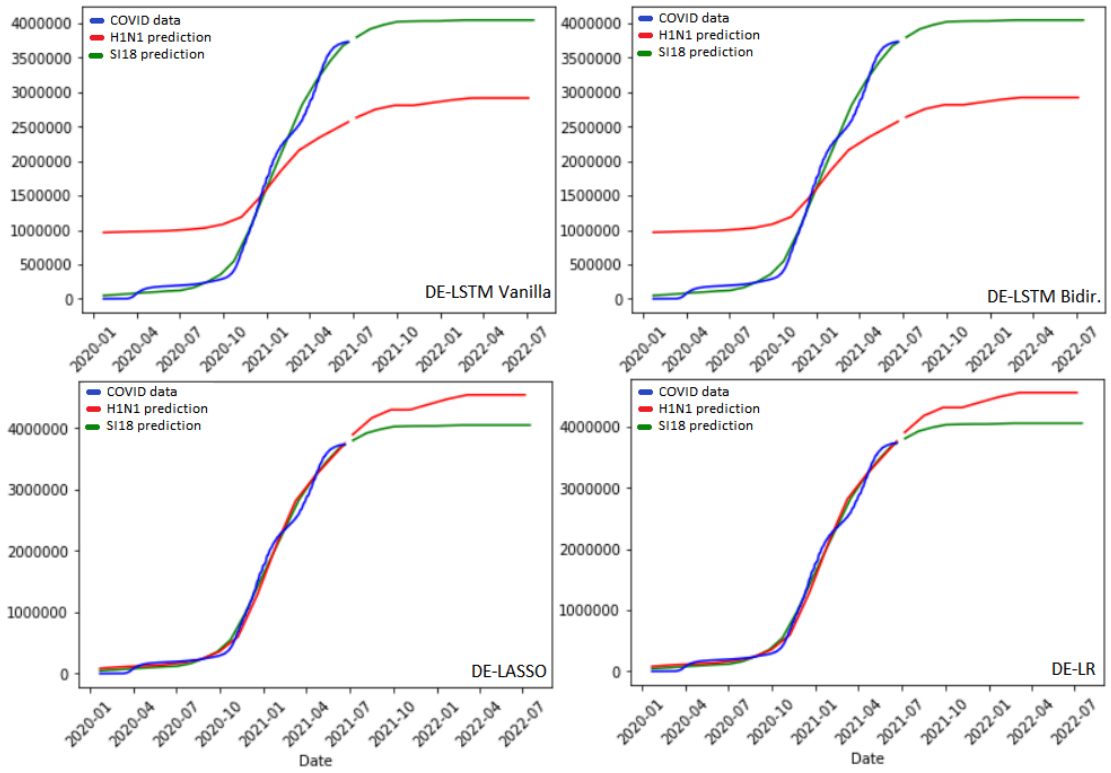

Figura 16. Predicción de la incidencia acumulada de COVID-19 para DE a partir del 21 de junio de 2021 para cada modelo estudiado. (Elaboración propia)

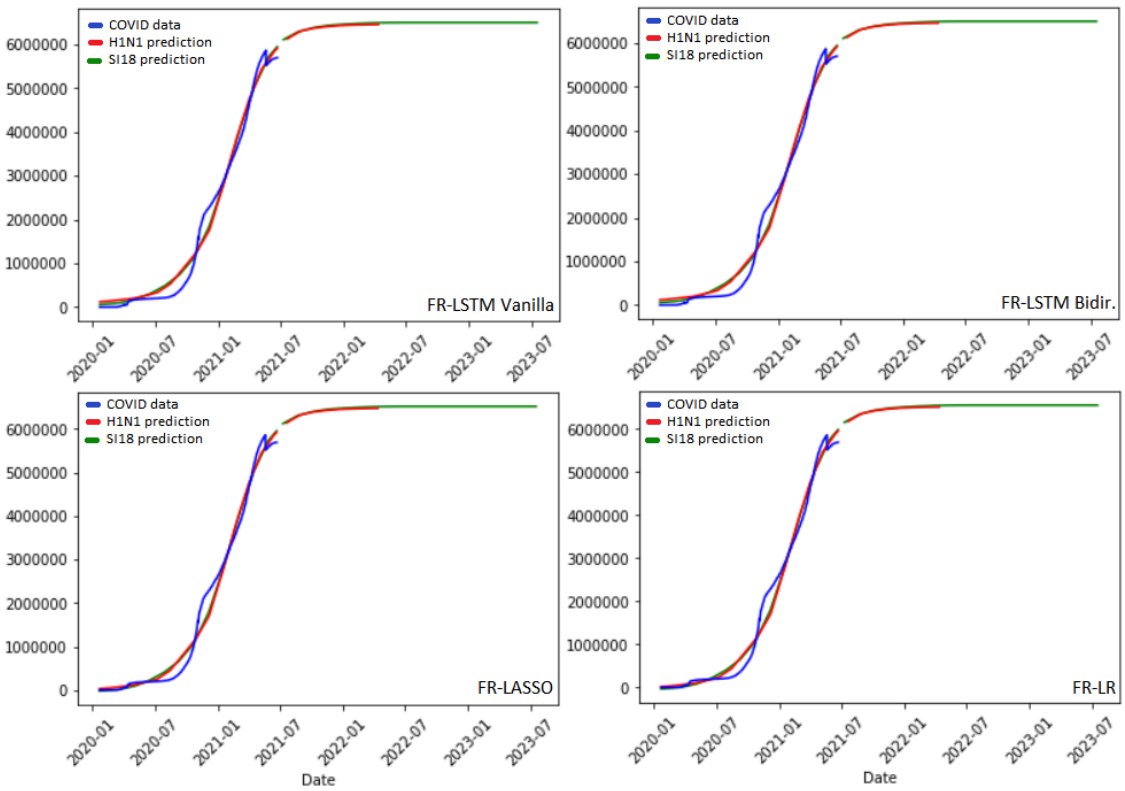

Figura 17. Predicción de la incidencia acumulada de COVID-19 para FR a partir del 21 de junio de 2021 para cada modelo estudiado. (Elaboración propia)

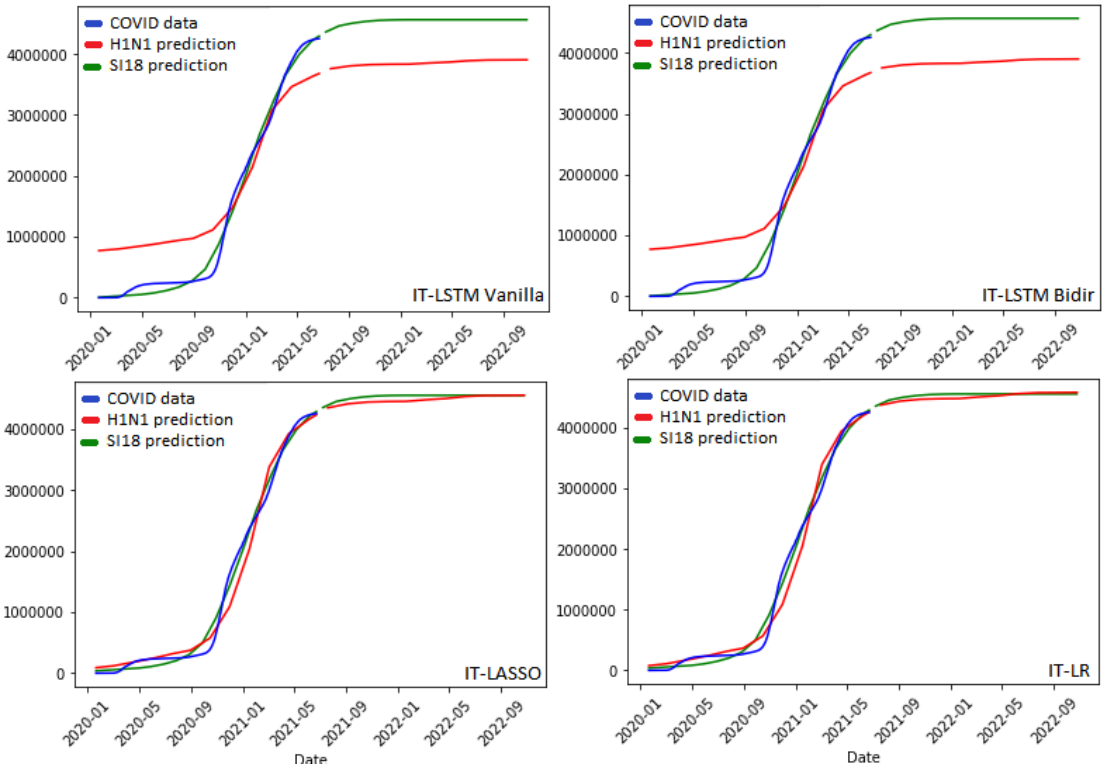

Figura 18. Predicción de la incidencia acumulada de COVID-19 para IT a partir del 21 de junio de 2021 para cada modelo estudiado. (Elaboración propia)

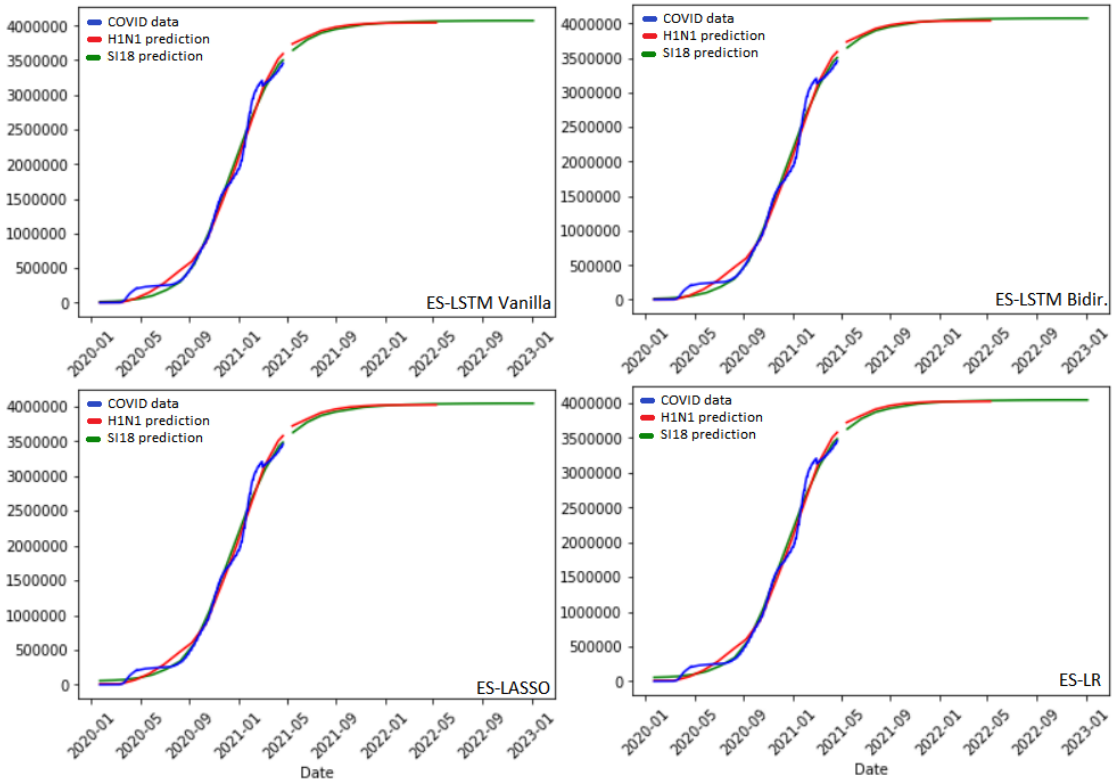

Figura 19. Predicción de la incidencia acumulada de COVID-19 para ES a partir del 23 de abril de 2021 para cada modelo estudiado. (Elaboración propia)

4.4.1. Análisis de la exactitud de la predicción

Como se indica en el apartado anterior, la métrica utilizada en estas pruebas es el RMSE. En este caso se compara el error entre los datos originales de COVID-19 que se han reservado para la evaluación y los datos predichos obtenidos en la salida de cada modelo para este mismo periodo. Para realizar este análisis, se ha decidido utilizar el set desde el inicio de los datos (22 de enero de 2020) hasta 10 días antes del final de la muestra (10 días antes del 21 de junio de 2021), dejando esos 10 días restantes para evaluar la precisión de la predicción del modelo. En el caso de UK y ES cuyas curvas se han tomado hasta el 22 de abril de 2021, el entrenamiento realizará con los datos hasta el 22 de abril, y se evaluará el periodo entre el 22 de abril y el 21 de junio. Ya que, como se indica más adelante, el análisis no es comparable entre diferentes países y tampoco es el objetivo para este caso. Los resultados obtenidos para el RMSE de la predicción de los modelo descritos anteriormente se pueden consultar en la Tabla 11.

		ML Algorithm			
Country	**Curve**	**LSTM Vanilla**	**LSTM Bidirectional**	**LASSO**	**LR**
UK	H1N1pdm09	675.496	652.454	209.359	217.826
	SI18	320.053	320.727	279.509	282.855
DE	H1N1pdm09	1.018.953	1.022.270	399.376	439.531
	SI18	297.000	296.559	275.109	298.603
FR	H1N1pdm09	578.314	602.038	589.305	663.538
	SI18	879.071	900.042	888.530	957.258
IT	H1N1pdm09	468.120	488.513	118.962	200.902
	SI18	337.593	330.836	316.110	319.314
ES	H1N1pdm09	203.199	209.012	202.394	204.883
	SI18	377.160	387.557	343.767	346.523
		Train set: (UK, ES): 2020-01-22 - 2021-04-22; (DE, FR, IT): 2020-01-22 - 2021-06-11			
		Validation set: (UK, ES): 2021-04-23 - 2021-06-21; (DE, FR, IT): 2021-06-12 - 2021-06-21			
		Best result per country/curve in green, worst result in red			

Tabla 11. RSME de la predicción de cada modelo en el periodo de evaluación. (Elaboración propia)

Estos datos, pueden representarse gráficamente en la Figura 20 para facilitar el análisis.

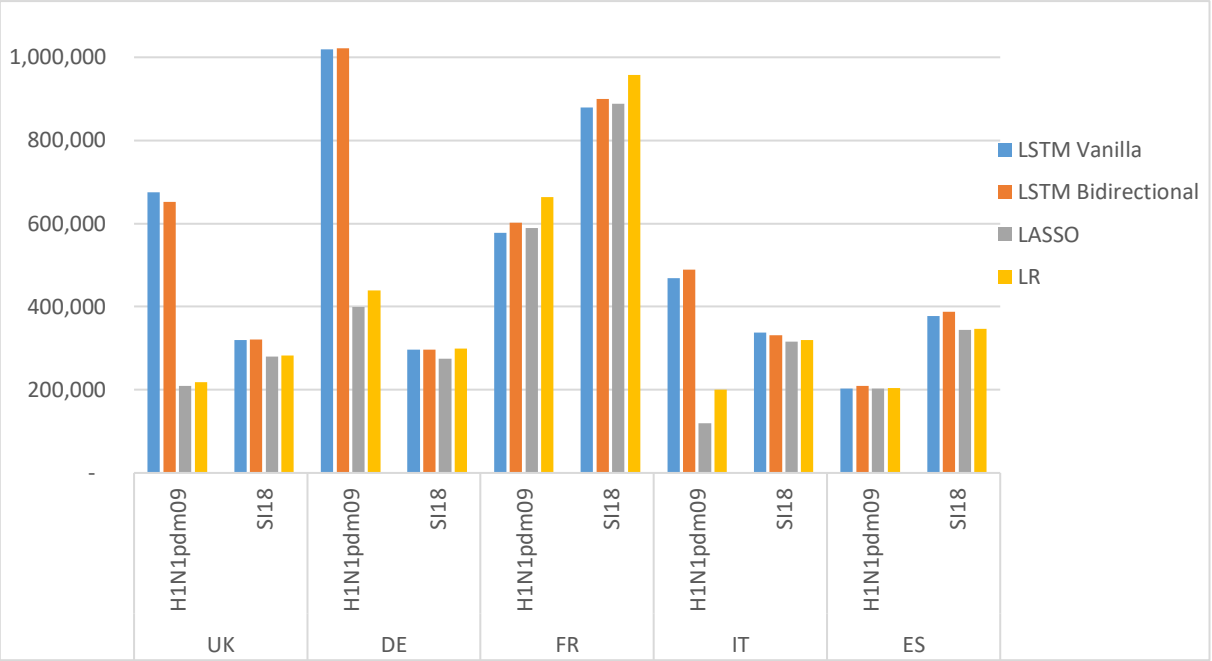

Figura 20. RSME de la predicción de cada modelo en el periodo de evaluación. (Elaboración propia)

4.4.2. Análisis del ajuste del modelo

Este procedimiento busca obtener una predicción a largo plazo, bajo la suposición de que el comportamiento de las curvas de COVID-19 se asemeja al de pandemias anteriores. En este caso el ajuste se ha evaluado para el periodo completo: del 22 de abril de 2020, al 21 de junio para DE, FR e IT, y al 23 de abril para UK y ES. Frente al análisis anterior que permite evaluar a precisión a corto plazo, esta prueba permite asegurar la fiabilidad del resultado a largo plazo, asegurando la efectividad de la transformación las curvas base y la curva de COVID-19 al construir el modelo. Para ello se han introducido los datos de la curva base al modelo para el periodo completo, y se ha calculado el error entre la curva predicha y la de los datos de COVID-19 reales. Los resultados del RMSE para este análisis vienen reflejado en la Tabla 12.

Country	Curve	ML Algorithm			
		LSTM Vanilla	*LSTM Bidirectional*	*LASSO*	*LR*
UK	H1N1pdm09	679.102	679.337	137.509	137.445
	SI18	134.956	134.951	114.317	114.306
DE	H1N1pdm09	775.935	775.871	112.837	112.701
	SI18	91.702	91.729	91.704	91.656
FR	H1N1pdm09	245.020	245.033	233.184	232.613
	SI18	227.799	227.783	217.548	217.015
IT	H1N1pdm09	95.513	95.597	97.236	95.022
	SI18	102.365	102.361	98.477	98.468
ES	H1N1pdm09	106.222	106.205	105.974	105.968
	SI18	105.643	105.626	100.963	100.957
					Best result per country/curve in green, worst result in red

Tabla 12. RSME del ajuste de cada modelo en el periodo completo. (Elaboración propia)

Nuevamente, se incluye una representación gráfica de los datos, para facilitar su comprensión, en la Figura 21.

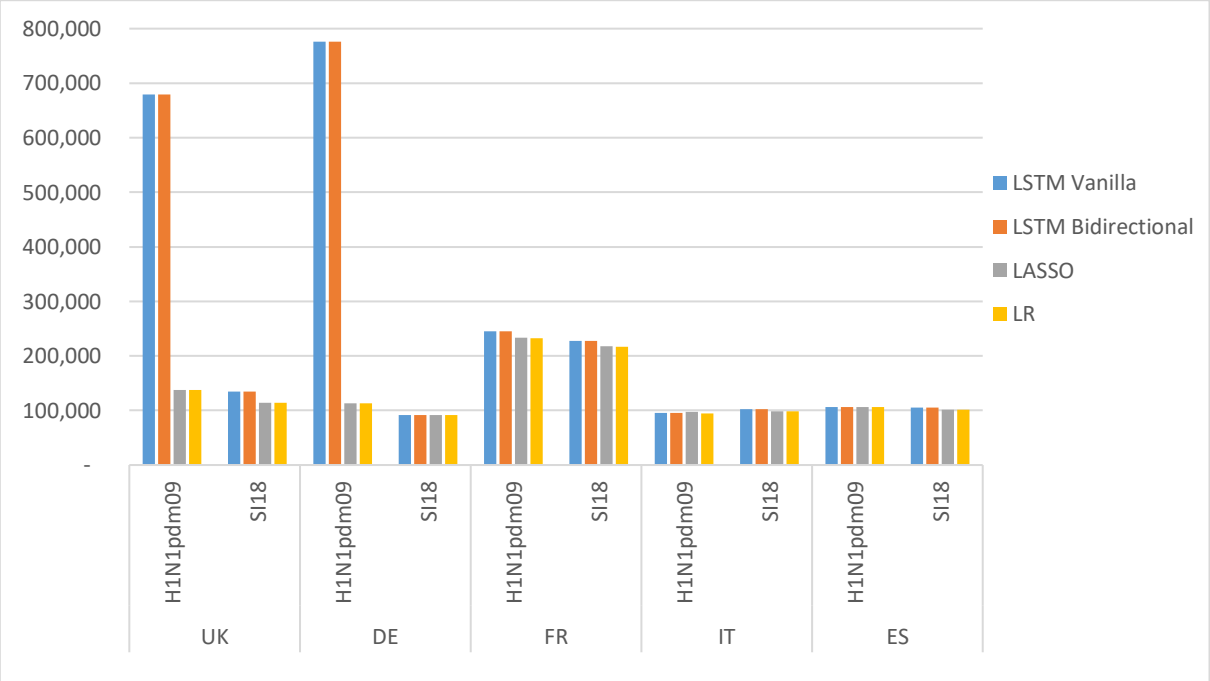

Figura 21. RSME del ajuste de cada modelo en el periodo completo. (Elaboración propia)

Se observa un ajuste muy malo para los modelos que utilizan los algoritmos LSTM en los datos Influenza A (H1N1) en UK y DE. Una segunda vista que permite ajustar el eje a la magnitud del resto de los datos se puede observar en la Figura 22.

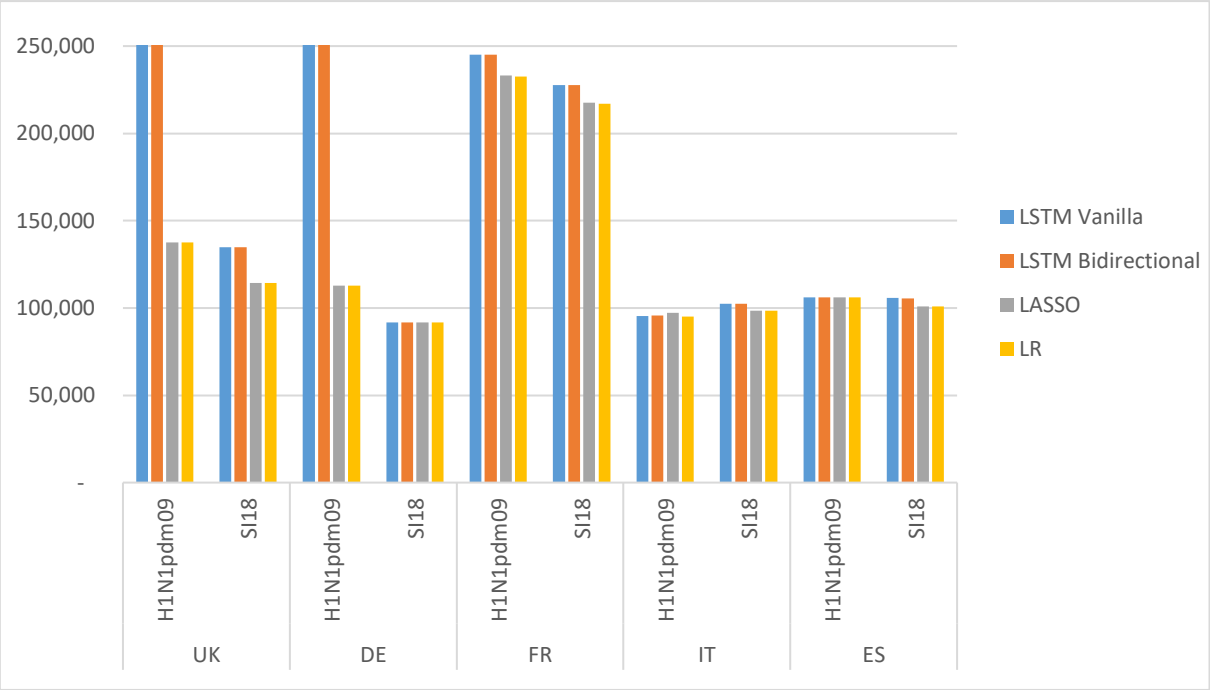

Figura 22. Vista del eje ajustado para el RMSE de la predicción de cada modelo tras la etapa de experimentación. (Elaboración propia)

En este análisis como en el anterior, cabe destacar que la métrica de RMSE no es comparable entre diferentes países, pues el error será mayor en países que tengan valores de contagios acumulados más altos. En caso de querer comparar el grado de similitud del modelo y la realidad, se recomienda obtener una métrica de error porcentual o utilizar otro tipo de métricas de ajuste como R^2. Sin embargo, para este caso no se estima necesario, ya que se desea obtener el mejor ajuste para cada país, y por tanto realizar las comparaciones entre modelos aplicados a los datos de cada país.

4.4.3. Conclusiones globales

En un análisis global de los datos, al analizar la Tabla 11 y la Figura 20 se observa que la predicción a corto plazo tiene un comportamiento más auténtico para los modelos que se han desarrollado con LASSO. La excepción se encuentra en los modelos de FR donde las mejores predicciones se realizan con los modelos LSTM Vanilla, y los modelos LASSO quedan en segundo lugar. Desde el punto de vista de las curvas base utilizadas, las curvas que utilizan los datos de la Influenza A de 2009 (H1N1pdm09) suelen cometer errores menores en la predicción a corto plazo, con respecto a las curvas de influenza estacional del año 18/19 (SI18).

Existen tres excepciones importantes, que se corresponden con los modelos LSTM de curva H1N1 pdm09 en UK, DE, IT. Se trata precisamente de los países en los que se ha tenido que realizar un ajuste logístico manual.

Desde el punto de vista del ajuste, en la Tabla 12 y la Figura 21 se observa un ajuste ligeramente mejor en los modelos que utilizan los datos de influenza estacional, siendo especialmente notable para UK y DE. Generalmente, el ajuste es mejor para aquellos modelos que usan curvas base con más datos: en la Figura 8 se observa que UK y DE tuvieron una curva de contagios mucho más pequeña que el resto de los países durante la epidemia de H1N1 de 2009, y por otro lado la incidencia acumulada de la curva de influenza estacional de 2018/19 es mayor que la incidencia acumulada de la curva de influenza A de 2009, comparando con la Figura 6.

Desde el punto de vista de los algoritmos usados para entrenar el modelo, existen diferencias menos notables, a excepción del uso de algoritmos LSTM en las curvas de H1N1 de UK y DE. Sin embargo, analizando el etiquetado del mejor y peor resultado de RSME por país y curva base en la Tabla 12, se observa que el algoritmo Linear Regression (LR) tiene el mejor resultado en todos los casos. Por otro lado, el peor resultado se observa para los modelos que usan LSTM, y en especial el LSTM básico o Vanilla. Aunque la diferencia entre modelos es mínima en los datos de IT y ES con la curva de influenza A, y los de UK con la curva de influenza estacional.

4.4.4. Recomendación por país del modelo a utilizar

De acuerdo con los resultados obtenidos se elaboran recomendaciones de uso de cada modelo. Se busca una concordancia en los resultados de ambos análisis, no obstante, en caso de conflicto entre directrices claramente contradictorias, se priorizará un modelo con un ajuste adecuado, pues a largo plazo el error en el ajuste puede acumular mayor error en la precisión de la predicción que el observado en el corto plazo. Con estas consideraciones, las recomendaciones para cada país son las siguientes:

- Reino Unido (UK): En este caso se deberían descartar definitivamente el resultado de los modelos LSTM-H1N1. La recomendación es clara hacia el uso de los modelos LASSO y LR. La curva de H1N1 obtiene resultados de más precisión, sin embargo, los modelos

que han utilizado la curva de SI18 se han ajustado mejor a los datos de COVID-19. Finalmente, se recomienda el uso de LASSO-SI18 para UK, considerando como opciones también aceptables LASSO-H1N1, y los modelos con LR.

- Alemania (DE): Nuevamente se debe evitar el uso de los modelos LSTM-H1N1. Los modelos recomendados son aquellos que usan la curva base SI18, con un ajuste sutilmente mejor para LR-SI18, y una predicción algo mejor en LASSO-SI18. Cabe destacar que los resultados del modelo con H1N1 para LASSO y LR muestran una tendencia más a la alza que los que usan SI18, y que, observando el error de predicción, el error es menor en esta curva. Por estos motivos, se recomienda el uso de LASSO-SI18, aunque también sería una buena opción LR-SI18, o el resto de los modelos que utilizan la curva de influenza estacional.
- Francia (FR): Se observa un ajuste mejor en los resultados de SI18, no obstante, en la predicción es muy evidente que la curva H1N1 genera modelos mejores, por lo que se priorizará esta última conclusión. Por otro lado, mientras que el ajuste apunta al uso de LASSO o LR, la predicción es mejor en el modelo LSTM-Vanilla seguida de LASSO. A la vista de estos resultados, la mejor opción es la curva LASSO-H1N1.
- Italia (IT): En este caso, no existen diferencias sustanciales en el ajuste de los modelos, por lo que la elección estará marcada principalmente por los resultados de la predicción. Al observarse mejores predicciones con la curva H1N1 y mejores aún con LASSO, ésta será la recomendación para Italia, considerando también LR-H1N1 una buena opción.
- España (ES): Nuevamente, se trata de un país cuyos modelos tienen un ajuste de calidad similar. Tal como ocurre con Italia, la curva de H1N1 genera modelos con mejores predicciones, y entre ellos LASSO se encuentra a la cabeza, aunque las diferencias con el resto de los modelos basados en la curva de Influenza A son muy pequeñas. Finalmente, se recomienda el modelo LASSO-H1N1, pero también se considera cualquier modelo construido con H1N1 una buena opción.

4.4.5. Comparativa respecto a los modelos propuestos en el estado del arte

Para esta comparativa se hace referencia a los estudios de Tian, Luthra, y Zhang (2020) y de Rustam, y otros (2020). En primer lugar, es necesario comparar el contexto y el objetivo del

análisis de estos estudios frente a los del presente procedimiento. Como ya se indicó en el apartado Conclusiones sobre el estado del arte, la principal ventaja que se observa en este procedimiento es que se logra obtener una previsión de muy a largo plazo, llegando a incluir predicciones para el año 2022 para todos los países excepto UK que llega hasta el final del año 2021. En el otro lado, Tian, Luthra, y Zhang solo hace una previsión para el periodo de validación, únicamente 5 días partiendo del 8 de abril de 2020, sin poder vislumbrar la asíntota de los casos acumulados que se observaría al final de la pandemia. En el caso de Rustam, y otros, la previsión se hace para los próximos 10 días partiendo del 17 de marzo de 2020. Aunque, como se indicó en el apartado de Alcance y limitaciones, el procedimiento que se describe en este documente no podría aplicarse en el contexto de estas investigaciones.

Desde el punto de vista de la predicción, Tian, Luthra, y Zhang analizan el modelo con 5 días de evaluación, mientras que en el presente trabajo y en el texto de Rustam, y otros se analiza una predicción de 10 días. Si bien una diferencia de 5 a 10 días tiene bajo impacto en el valor del RMSE, sí que lo tiene la magnitud de los datos que se están evaluando. Para ello, se decide obtener el porcentaje de error con respecto a la media de los datos en evaluación para los tres análisis. En el caso de los datos de Tian, Luthra, y Zhang, se deberá revertir la normalización basada en los casos por millón de la población del país. Dichos resultados se muestran en la Tabla 13, la Tabla 14, y la Tabla 15.

Country	Curve	Data average in validation period	ML Algorithm (%RMSE)			
			LSTM Vanilla	LSTM Bidirectional	LASSO	LR
UK	H1N1pdm09	4.481.434	15,07%	14,56%	4,67%	4,86%
	SI18		7,14%	7,16%	6,24%	6,31%
DE	H1N1pdm09	3.726.846	27,34%	27,43%	10,72%	11,79%
	SI18		7,97%	7,96%	7,38%	8,01%
FR	H1N1pdm09	5.684.918	10,17%	10,59%	10,37%	11,67%
	SI18		15,46%	15,83%	15,63%	16,84%
IT	H1N1pdm09	4.249.478	11,02%	11,50%	2,80%	4,73%
	SI18		7,94%	7,79%	7,44%	7,51%
ES	H1N1pdm09	3.633.505	5,59%	5,75%	5,57%	5,64%
	SI18		10,38%	10,67%	9,46%	9,54%
Validation set: (UK, ES): 2021-04-23 - 2021-06-21; (DE, FR, IT): 2021-06-12 - 2021-06-21						
Best result per country/curve in green						

Tabla 13. % RMSE de la media de casos considerados para las predicciones de los modelos. (Elaboración propia)

Country	ML Algorithm (% RMSE)		
	HMM	Hierarchical Bayes	LSTM RNN
Germany	21,53%	5,59%	5,67%
Italy	17,16%	8,89%	10,16%
US	20,22%	10,87%	1,63%
Taiwan	12,13%	111,38%	6,70%
Japan	13,28%	23,04%	17,74%
South Korea	14,24%	6,75%	2,01%
Validation set: 2020/4/9 - 2020/4/14			
Best result per country in green			

Tabla 14. % RMSE de la media de casos considerados para las predicciones de los modelos de Tian, Luthra, & Zhang **(2020)**. *(Elaboración propia)*

ML Algorithm	%RMSE per cumulated cases
LR	7,05%
LASSO	2,81%
SVM	13,94%
ES	3,09%
Validation set: 2020/03/18 - 2020/03/27. Best result per country in green	

Tabla 15. % RMSE de la media de casos considerados para las predicciones de los modelos de Rustam, y otros **(2020)**. *(Elaboración propia)*

Si se comparan los resultados de la predicción de la propuesta con los de los ejemplos mostrados en el estado del arte, se demuestra que los resultados más favorable de cada uno de estos dos experimentos tienen porcentajes más bajos de manera generalizada que en la propuesta del presente proyecto. Sin embargo, es importante destacar que la diferencia es muy pequeña, manteniendo un orden de magnitud muy cercano, y que, en el único caso comparable por país, Italia, el porcentaje de error es menor utilizando el nuevo procedimiento.

Teniendo en cuenta que el porcentaje de error obtenido con esta nueva metodología está en la magnitud del de los estudios anteriores, y considerando además las ventajas analizadas al inicio de la comparativa con respecto a las predicciones a largo plazo, queda validada el presente procedimiento como un método de predicción a largo plazo para datos de contagios acumulados en epidemias.

4.5. Dashboard para la visualización y el análisis de los resultados de predicción

La visualización de los resultados es una parte importante en los proyectos orientados a predicciones. Es importante lograr vistas sencillas, accesibles y limpias de manera que una persona que no conozca el software de construcción de las predicciones sea capaz de leer e interpretar los resultados rápido y de manera autosuficiente. Por este motivo, se propone un dashboard de visualización de los resultados en el contexto del Business Intelligence de la Industria 4.0.

4.5.1. Arquitectura de datos para los resultados

Una vez obtenidos los resultados en el código, deberán ser almacenados de manera que el software de visualización, Tableau, pueda leerlos. La arquitectura de los datos se construye sobre tres tablas distintas.

Model: contiene la conexión entre el model_id, y los datos de la curva base y algoritmo utilizados para ese modelo. Además, incluye la información de recomendación de cada modelo. Es una tabla estática, que no se ve afectada por la salida del código de construcción de modelos, y tiene la siguiente estructura:

- Model_id (key): de números enteros con rango de 0 a 39, es único para cada tipo de modelo, asociándose a un país, una curva base, y una algoritmo de machine learning.
- Country: puede ser “UK”, “DE”, “FR”, “IT”, “ES”. Identifica al país con el que se corresponde con el model_id
- Curve: puede ser “H1N1pdm09” o “SI18”. Identifica a la curva base que se corresponde con el model_id
- Algorithm: puede ser “LSTM Vanilla”, “LSTM Bidirectional”, “LASSO”, “LR”. Identifica al algoritmo de machine learning que se corresponde con el model_id
- Recommended: de números enteros con rango 0 a 1. Identifica aquel modelo que se recomienda por país con un 1, y mantiene el 0 para el resto de los modelos.

Se incluye esta tabla en el Anexo A para aclarar que model_id corresponde con cada modelo, y como se ha marcado la recomendación.

MetaData: contiene información general de cada resultado generado para un modelo determinado. Está compuesta de las siguiente columnas:

- Output_id: identifica el resultado. Los resultados generados juntos para un país tendrán el mismo output_id, aunque contengan varios modelos.
- Country: país al que corresponde el conjunto de resultados.
- Model: se corresponde con el model id de los resultados obtenidos.
- RMSE: El valor del RMSE de la predicción para un periodo de 10 días. Evalúa la calidad del modelo.
- Active: de cara a la visualización, el valor por defecto es 1, pero de querer evitar la visualización de un resultado, se podría cambiar a cero.

La clave de esta tabla sería la combinación entre output_id y model.

Data: Contiene el detalle de los resultados generados. Tiene la siguiente estructura:

- Output_id: se trata del output_id del resultado al que se refiere la información.
- Country: se trata del país al que se refiere el resultado
- Model: Contiene el model_id del modelo al que se refiere el resultado.
- Date: Contiene la fecha de la que datan los contagios acumulados.
- COVID_cases: contiene los datos de contagios ya sean históricos, o predicciones.
- Type: puede ser “Actual” o “Prediction” en función de si contiene información de datos históricos, o datos de la predicción.

La clave de esta tabla sería la combinación entre output_id, model, y date.

Una representación gráfica de las relaciones entre las tres tablas puede observarse en la Figura 23.

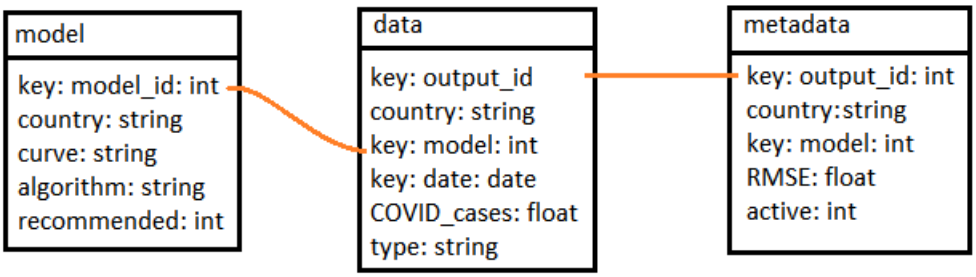

Figura 23. Arquitectura de las tablas de datos de los resultados. (Elaboración propia)

Estos datos son almacenados en tres archivos de extensión csv distintos con nombres "Model.csv", "MetaData.csv", y "Data.csv", dónde solo Metada y Data son actualizados tras cada ejecución del código de construcción de modelos.

4.5.2. Dashboard de visualización de resultados

Se hace uso de la herramienta de visualización Tableau para desarrollar un dashboard que permita analizar los resultados de manera sencilla e intuitiva. La solución ofrece 4 vistas distintas que presentan la información de una manera diferente. Por un lado, se necesitan vistas que ofrezca datos numéricos o de texto de los resultados. Por otro lado, también se necesitan vistas gráficas de la progresión de los contagios. Desde otra perspectiva se desea tener vistas fijas en el modelo recomendado por país, y vistas que permitan analizar uno a uno cada uno de los modelos propuestos. Con esto, se han generado las siguientes 4 vistas:

- **Overview: Best Model:** tabla numérica del último resultado activo de los modelos recomendados.
- **Overview: Model Deep Dive:** tabla numérica del último resultado activo de cada uno de los modelos probados.
- **Graph: Best Model:** vista gráfica del último resultado activo de los modelos recomendados.
- **Graph: Model Deep Dive:** vista gráfica del último resultado activo de cada uno de los modelos probados.

4.5.2.1. Overview

Las vistas "Overview: Best Model" y "Overview: Model Deep Dive" se unen en una sola página llamada "Overview". En ella se proporciona la siguiente información:

- **Maximum COVID cases:** número máximo de casos registrados para la curva predicha. Se identifica con el máximo número de casos que se espera para la epidemia en el país.
- **Last Date Actuals:** fecha del último dato histórico.
- **Last Date Predictions:** fecha del último dato generado en la predicción.

- **Model prediction RMSE:** valor del RMSE del modelo en el periodo de evaluación (10 después del último dato histórico).
- **Epidemic predicted end date:** esta fecha se ha calculado como la primera fecha en alcanzar un 97% del valor más alto de la serie. Es orientativa y se basa en el aplanamiento total de la curva en su tramo final.

Estas métricas se proporcionan para las dos vistas. La diferencia principal reside en que, en la vista "Best Model", estas métricas son mostradas por país para el modelo recomendado para ese país. Las métricas en "Model Deep Dive" puede aparecer para todos los resultados generados. En la Figura 24 se muestra la página completa "Overview" que incluye los datos del "Best Model" y el "Model Deep Dive".

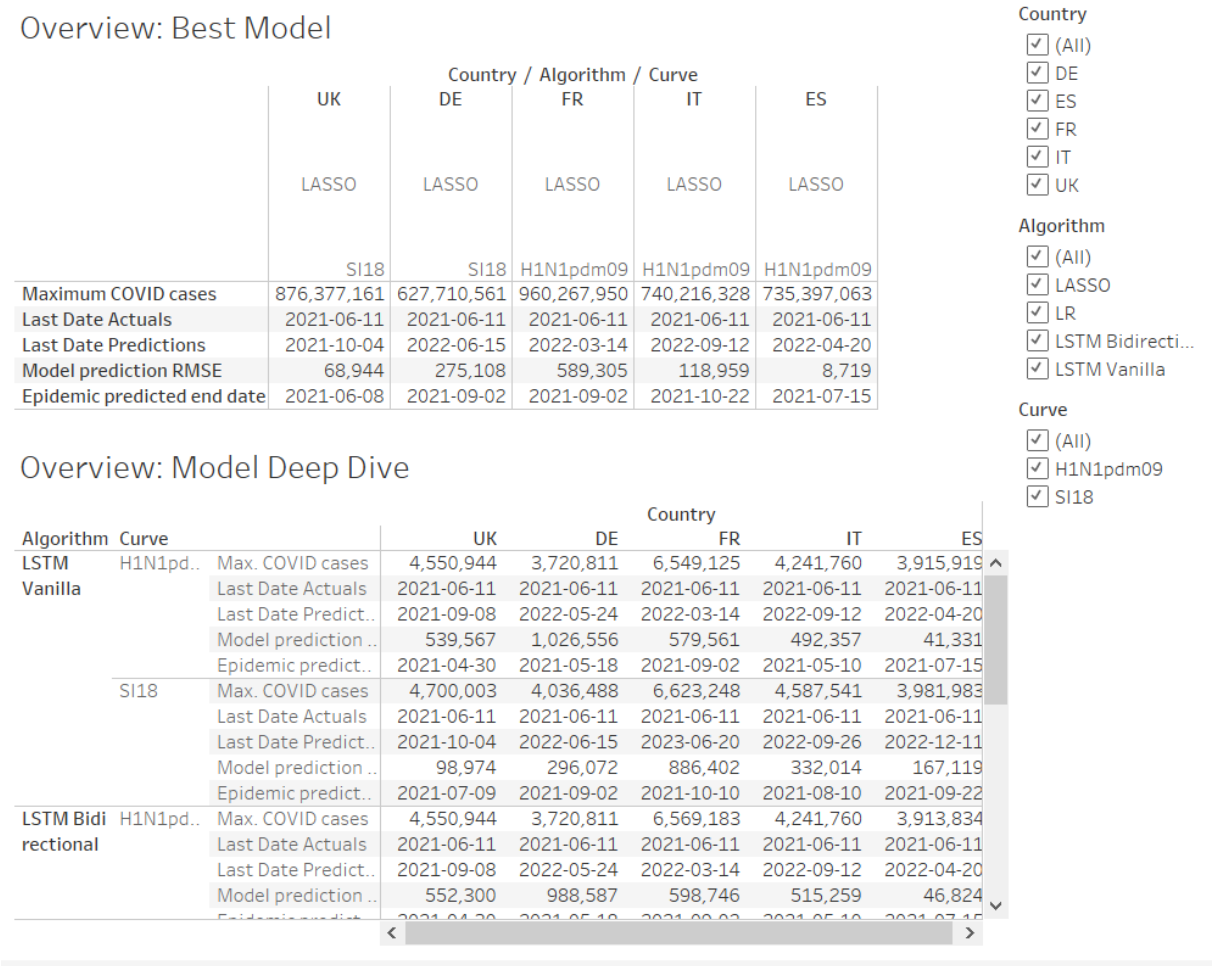

Figura 24. Página Overview del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia)

La página tiene además una serie de filtros que permiten reducir las tablas a la información que se desea aislar.

- **Country:** permite seleccionar los países para los que se muestra la información.
- **Algorithm:** permite seleccionar los resultados de los modelos que se han construido a partir de un algoritmo determinado.
- **Curve:** permite seleccionar los resultados de los modelos que se han construido a partir de una curva determinada.

Un detalle importante, es que mientras que el filtro Country afectará a ambas tablas simultáneamente, los filtros Algorithm y Curve tan sólo afectarán a la tabla "Model Deep Dive". Esto es porque "Best Model" siempre ha de mostrar la información del modelo recomendado. En la Figura 25 puede verse un ejemplo en el que se ha filtrado una serie de países, afectado a ambas tablas, y que también se han filtrado resultados de modelos construidos con LR y LSTM Bidirectional y la curva SI18, afectando únicamente a la tabla "Model Deep Dive".

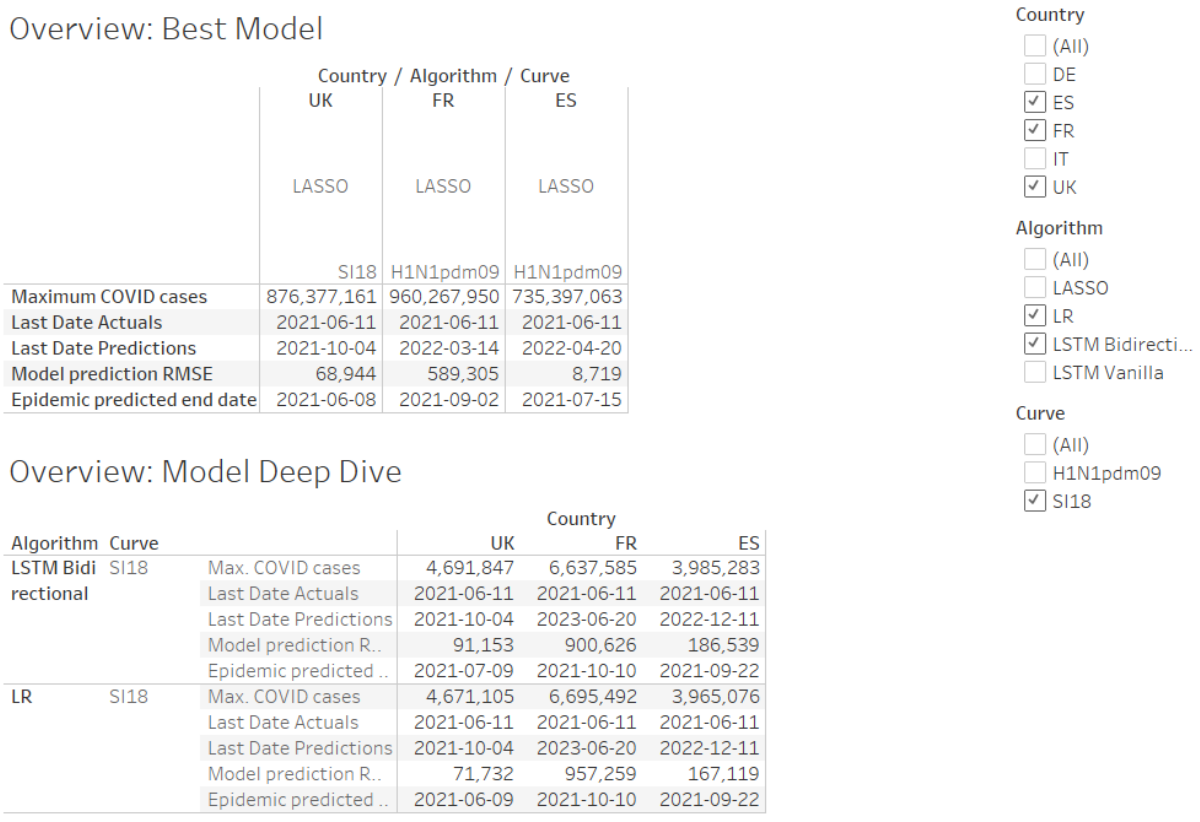

Figura 25. Demostración del funcionamiento de los filtros de la página Overview del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia)

Como información adicional que se proporciona, al posicionar el cursor sobre la tabla se obtiene la curva base y el algoritmo usados en el modelo, y el output id de los resultados.

4.5.2.2. Graph: Best Model

Las vistas gráficas se han ubicado en páginas diferentes para permitir una mejor visualización de los datos. En primer lugar, tenemos la gráfica de la predicción de los modelos recomendados "Graph: Best Model". Esta gráfica indica en colores diferentes los datos históricos (azul) y la predicción (naranja) para los resultados del modelo recomendado de cada país. La Figura 26 muestra la página "Graph: Best Model" completa con la representación de las curvas de los cinco países.

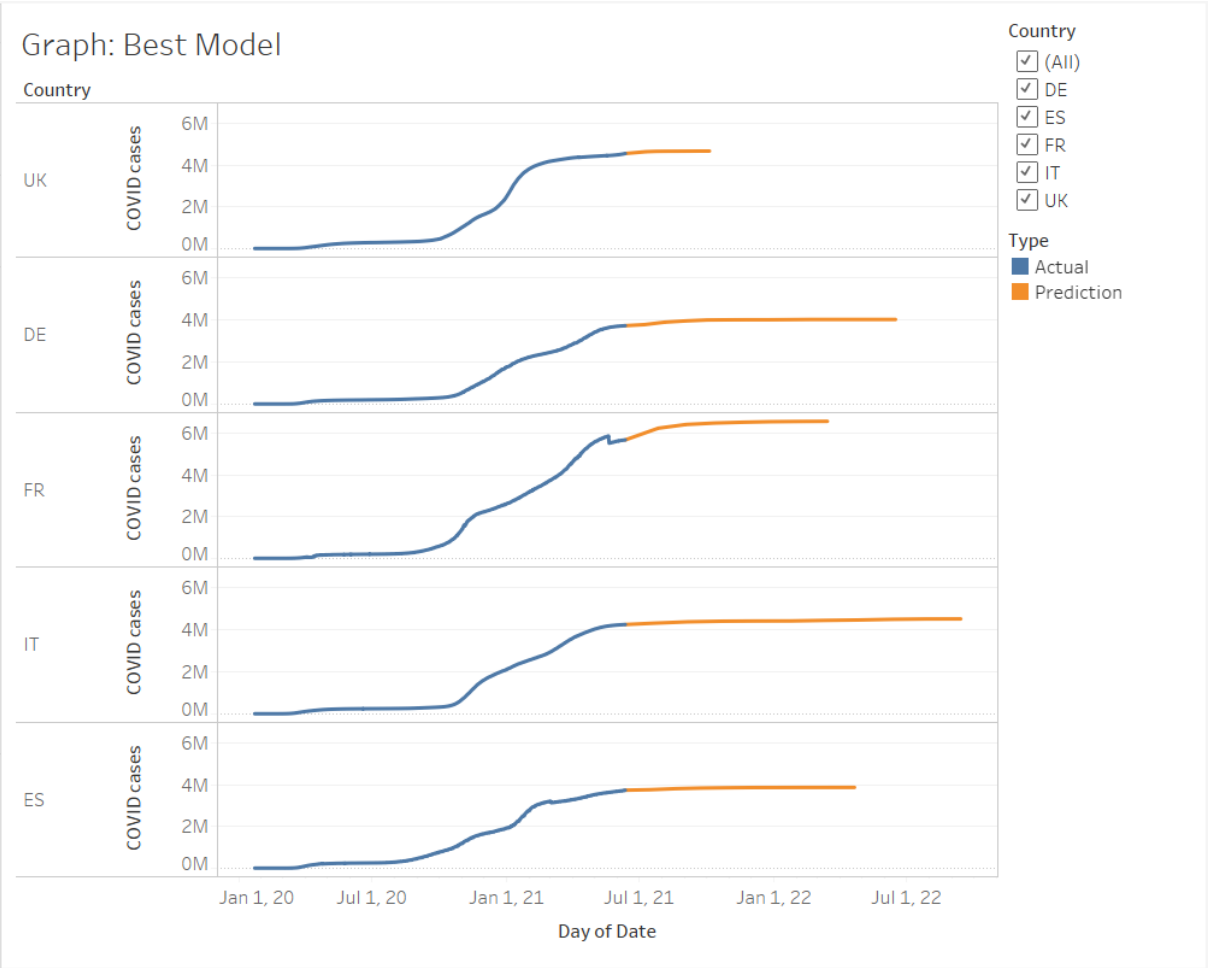

Figura 26. Página Graph: Best Model del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia)

Esta vista también incluye un filtro que permite aislar las curvas de países determinados. Cuanto menor sea el número de países filtrados, más grande será la visualización y mejor se podrá realizar el análisis. Adicionalmente, si se posa el cursor sobre la curva, no sólo se obtiene el valor concreto de contagios y la fecha de cada punto, también se puede obtener el tipo de

dato (histórico o predicción), el algoritmo y la curva base usados en el modelo y el output id del resultado.

4.5.2.3. Graph: Model Deep Dive

Existe una versión más extensa de la vista anterior, que contiene información de todos los modelos probados: "Graph Model Deep Dive". Esta vista proporciona la gráfica de contagios para el último resultado activo de cada modelo probado. Como se observa en la Figura 27, la composición de la vista es muy similar a la del "Graph: Best model", a excepción de los filtros de la derecha que permiten filtrar el último resultado de cada uno de los modelos disponibles, basándose en la curva base y el algoritmo usado para su construcción.

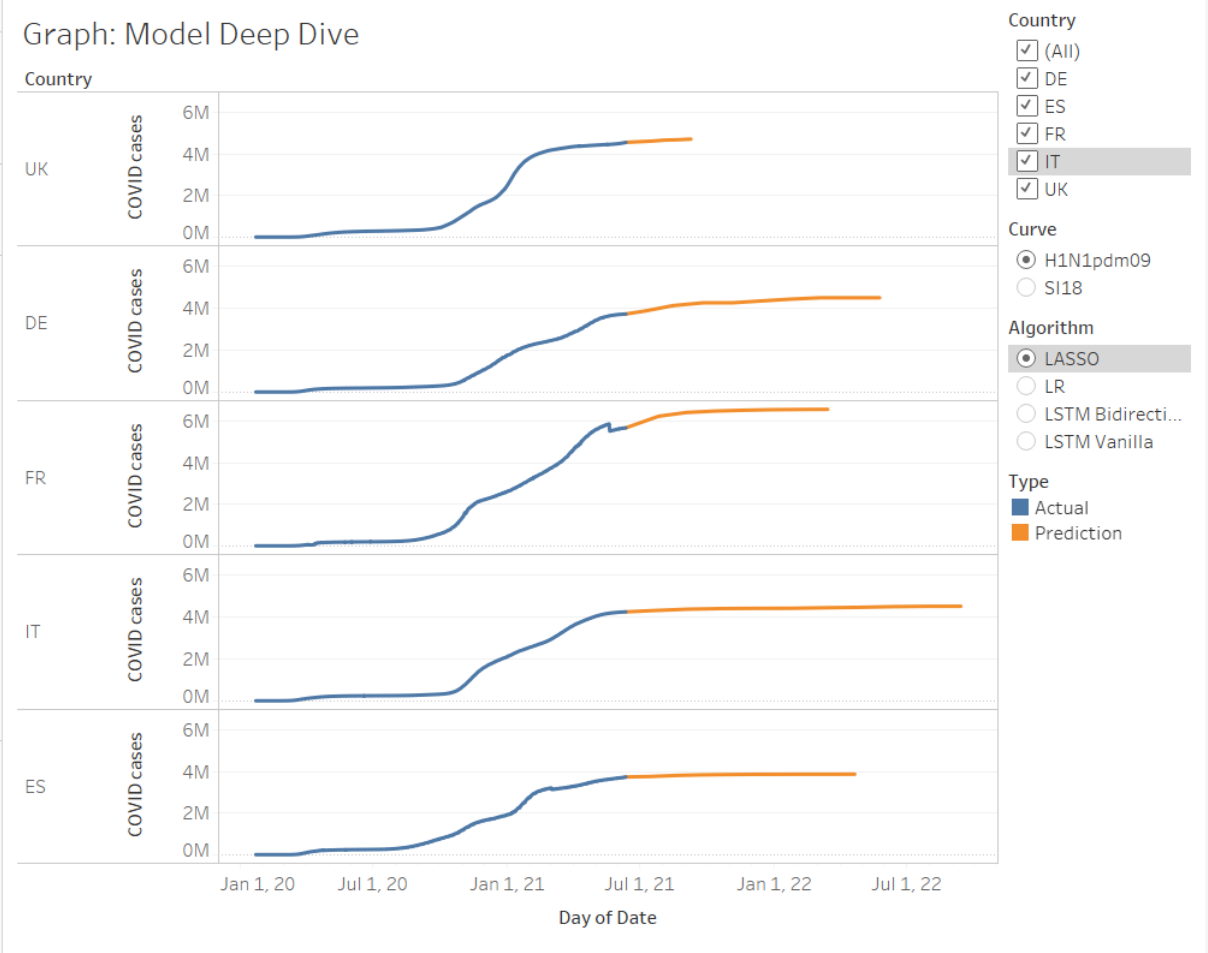

Figura 27. Página Graph: Model Deep Dive del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia)

Los filtros de la curva base y el algoritmo son de selección única, permitiendo filtrar un único modelo por país simultáneamente, de manera que la vista resulte sencilla de realizar. El filtro aplicado al país mantiene la capacidad de seleccionar varias opciones simultáneamente y funciona exactamente igual que el de la vista "Graph: Best Model". De hecho, resulta interesante filtrar un número menor de países para tener una vista más grande y detallada de la gráfica, como se muestra en la Figura 28 en la que se ha filtrado el resultado de Alemania (DE) de la previsión usando el modelo LR-SI18.

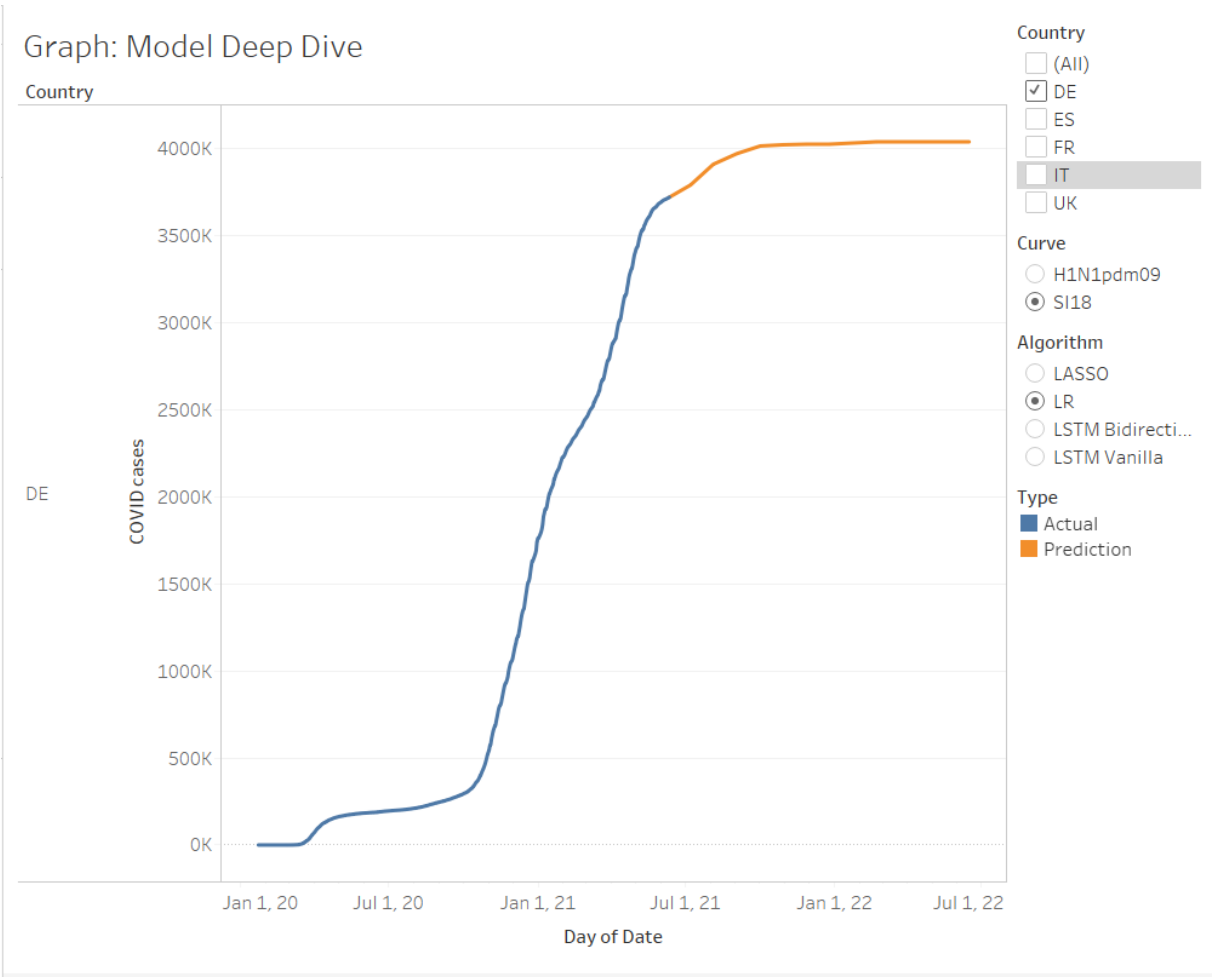

Figura 28. Ejemplo de filtrado en la página Graph: Model Deep Dive del Dashboard de la previsión de la epidemia de COVID-19. (Elaboración propia)

5. Conclusiones y trabajos futuros

Tal y como se describe en el apartado Motivación, existe una necesidad para la realización de predicciones de la actual pandemia de COVID-19. Se requiere el estudio de nuevos procedimientos que puedan aportar información adicional para anticipar la situación epidemiológica, y tras el análisis del Contexto y estado del arte, se concluye que hay una falta de estudios que proporcionen:

1. predicciones a largo plazo.
2. predicciones que logren generar una curva de contagios en la línea de lo observado en situaciones epidemiológicas similares.

En este contexto, se ha planteado un proyecto que propone un procedimiento para solventar los problemas anteriores, y al mismo tiempo, complementar los resultados de estudios actuales con una fuente de información adicional. Finalmente, la contribución principal del proyecto viene definida por:

1. La obtención de un modelo capaz de realizar predicciones a largo plazo.
2. La obtención de una estimación del fin de la pandemia para cinco países europeos.
3. Incorporación de datos de epidemias anteriores para aumentar la fiabilidad del resultado a largo plazo.

A la vista de los resultados descritos a lo largo del documento, que se han obtenido a través del procedimiento descrito en Metodología del trabajo, se pueden enumerar los siguientes hitos. Los hitos están en la línea de los objetivos definidos al inicio del proyecto (apartado Objetivos).

- **Análisis de fuentes de datos para su uso como curva base.** En primer lugar, se ha realizado una selección de las curvas que presentaban mayor potencial para la construcción de modelos de predicción. Las curvas de la epidemia de influenza A de 2009 y la influenza estacional del año 18/19 fueron posteriormente evaluadas mediante su incorporación en diversos modelos, hasta finalmente lograr obtener una recomendación para cada país.
- **Uso de diversos algoritmos de machine learning para la generación de resultados.** el capítulo 4 describe el procesamiento completo de los datos para que éstos puedan ser

utilizados en diversos algoritmos de machine learning. Se han probado dos variantes del algoritmo LSTM, pero también el algoritmo LASSO y LR, todos ellos con probados resultados al utilizarse con series temporales.

- **Integración software del código para la construcción de modelos.** Se ha desarrollado el código mediante el lenguaje de Python, de manera que en una única secuencia se puedan descargar los datos más actualizados de COVID-19 y procesarlos para obtener una predicción. Se describe el desarrollo software a lo largo de todo el capítulo 4, a excepción del último apartado.
- **Construcción de diversos modelos de predicción mediante el procedimiento planteado para cinco países distintos:** Mediante el procesamiento de los datos de todas las curvas de contagios, y el uso de diversos algoritmos de machine learning, se han construido hasta 8 modelos para cada uno de los cinco países que forman parte de este proyecto.
- **Evaluación de la eficacia del procedimiento aplicado a los datos de diferentes países europeos:** En el apartado 4.4. Resultados y análisis de los modelos obtenidos se realiza una valoración de todos los modelos desde el punto de vista del ajuste del modelo, y de la precisión en la predicción a corto plazo. De esta manera, se ha realizado una recomendación del modelo a utilizar para cada país de acuerdo de acuerdo con el comportamiento observado.
- **Valoración del modelo completo frente a investigaciones anteriores.** Puede consultarse la valoración en el apartado 4.4.5 Comparativa respecto a los modelos propuestos en el estado del arte, en el que se valida el procedimiento basándose en las ventajas exclusivas que estas predicciones aportan y los resultados obtenidos con respecto a la métrica de valoración (RMSE).
- **Creación de un panel de visualización (dashboard) que incluya los resultados obtenidos y métricas clave:** descrito en profundidad en el aparatado 4.5, contiene cuatro vistas diferente que proporcionan tanto información gráfica como numérica. La información se presenta de manera clara, y de interpretación sencilla, indicando claramente en vistas diferenciados los resultados de los modelos recomendados.

Teniendo en cuenta el cumplimiento de los punto anteriores, finalmente se concluye que el presente trabajo ha cumplido con su objetivo principal: la evaluación de la eficacia de modelos

para previsión de las curvas de contagios de COVID-19 obtenidos a partir de las curvas completas de epidemias anteriores, y la validación del procedimiento planteado para ello.

5.1. Trabajos Futuros

Tras la valoración del trabajo efectuado en las conclusiones anteriores, se pueden enumerar las diferentes líneas de desarrollo que se abren en este punto. De esta manera, se plantea la siguiente lista de trabajos futuros que pueden complementar o mejorar la propuesta, ampliando su alcance.

- **Ensayo con un mayor número de algoritmos de machine learning**. En el proyecto actual se han probado ciertos algoritmos sencillos, sin embargo, se puede intentar alcanzar mejores resultados con algoritmos más complejos o analizando un conjunto mayor de algoritmos.
- **Incorporar factores externos o implementar aprendizaje de varias variables.** En este momento el modelo aprende de un única variable que es la curva de contagios. Sin embargo, se podría utilizar algoritmos que analizaran otras variables, como los días festivos, la tasa de vacunación, o el grado de severidad de las restricciones, para incorporarlos a la previsión, y mejorar su precisión a corto plazo.
- **Implementar y evaluar el resultado metodología en epidemias diferentes, pero de características similares a la COVID-19.** Con la epidemia de COVID-19 bajo control y vislumbrando el final de la pandemia, se valora la posibilidad de utilizar el procedimiento propuesto para aplicarlo a otros casos epidemiológicos. Deberán ser de características similares a las enfermedad de las curvas base, o no pondrán ser utilizadas, y nuevas curvas base evaluarse para estos casos.

Referencias bibliográficas

Allenby, G. M., Rossi, P. E., & McCulloch, R. E. (2005). Hierarchical Bayes Models: A Practitioners Guide. *SSRN Electronic Journal*. doi:10.2139/ssrn.655541

Ansart, S., Pelat, C., Boelle, P.-Y., Carrat, F., Flahault, A., & Valleron, A.-J. (junio de 2009). Mortality burden of the 1918-1919 influenza pandemic in Europe. *Influenza and other respiratory viruses, 3*, 99-106. doi:10.1111/j.1750-2659.2009.00080.x

Baños, J., Brotons, C., & Farré, M. (1998). *Glosario de Investifación Clínica y Epidemiológica* (Vol. 23). Barcelona: Monografías Dr. Antonio Estévez.

Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P.-Y., . . . Widdowson, M.-A. (septiembre de 2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet Infectious Diseases, 12*(9), P687-695. doi:10.1016/S1473-3099(12)70121-4

Dong, E., Du H, & Gardner, L. (s.f.). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infectious Diseases, 20*(5), 533-534. doi:10.1016/S1473-3099(20)30120-1

El Mundo Gráficos. (31 de marzo de 2021). Mapa del coronavirus: expansión en cifras del Covid-19 en el mundo. *El Mundo*. Recuperado el 3 de abril de 2021, de https://www.elmundo.es/ciencia-y-salud/salud/2020/03/02/5e5cd4ebfc6c83632e8b4644.html

Espina Marconi, L. (1984). El modelo logístico. En *Serie de estudios económicos.* Santiago de Chile: Departamento de informaciones estadísticas y publicaciones del Banco Central de Chile.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural computation, 9*, 1735-80. doi:10.1162/neco.1997.9.8.1735

Instituto Nacional de Estadística. (s.f.). *Estimación del número de defunciones semanales (EDeS) durante el brote de covid-19. Tablas de resultados.* Recuperado el 4 de abril de 2020, de Instituto Nacional de Estadística: https://www.ine.es/experimental/defunciones/experimental_defunciones.htm#tablas_resultados

Jackson, J., Weiss, M., Schwarzenberg, A., & Nelson, R. (2020). *Global Economic Effects of COVID-19.* Congressional Research Service.

Jurafsky, D., & Martin, J. H. (2020). Hidden Markov Models. En *Speech and Language Processing* (pág. Appendix A).

Lee, S. Y., Lei, B., & Mallick, B. (2020). Estimation of COVID-19 spread curves integrating global data and borrowing information. *PLoS ONE*. doi:10.1371/journal.pone.0236860

Najafimehr, H., Mohamed Ali, K., Safari, S., Yousefifard, M., & Hosseini, M. (16 de abril de 2020). Estimation of basic reproduction number for COVID-19 and the reasons for its differences. *International Journal of Clinical Practice, 74*(8). doi:10.1111/ijcp.13518

Nyce, C. (2007). *Predective Analystics White Paper.* Malvern: American Institute for Chartered Property Casualty Underwriters (AICPCU).

Russell, S., & Norvig, P. (2010). *Artificial Intelligence A Modern Approach* (3rd ed.). New Jersey: Pearson Education, Inc.

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B.-W., Aslam, W., & Choi, G. S. (2020). COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access, 8*, 101489-101499. doi:10.1109/ACCESS.2020.2997311

Spreeuwenberg, P., Kroneman, M., & Paget, J. (2018). Reassessing the Global Mortality Burden of the 1918 Influenza Pandemic. *American journal of epidemiology*(187), 2561-2567. doi:10.1093/aje/kwy191

Tian, Y., Luthra, I., & Zhang, X. (julio de 2020). Forecasting COVID-19 cases using Machine Learning models. doi:10.1101/2020.07.02.20145474

WHO Global Influenza Surveillance and Response System (GISRS). (15 de mayo de 2021). Laboratory confirmed data from the Global Influenza Surveillance and Response System (GISRS). Recuperado el 15 de mayo de 2021, de https://apps.who.int/flumart/Default?ReportNo=12

World Health Organization. (12 de octubre de 2020). *Coronavirus disease (COVID-19) | Q&A*. Recuperado el 3 de abril de 2020, de World Health Organization: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19

Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z., & Chen, D. (2020). Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. *International Journal of Infectious Diseases*, 874-875. doi:10.1016/j.ijid.2020.02.033

Anexo A. Tabla estática 'model'

model_id	country	curve	algorithm	recommended
0	UK	H1N1pdm09	LSTM Vanilla	0
1	UK	H1N1pdm09	LSTM Bidirectional	0
2	UK	H1N1pdm09	LASSO	0
3	UK	H1N1pdm09	LR	0
4	UK	SI18	LSTM Vanilla	0
5	UK	SI18	LSTM Bidirectional	0
6	UK	SI18	LASSO	1
7	UK	SI18	LR	0
8	DE	H1N1pdm09	LSTM Vanilla	0
9	DE	H1N1pdm09	LSTM Bidirectional	0
10	DE	H1N1pdm09	LASSO	0
11	DE	H1N1pdm09	LR	0
12	DE	SI18	LSTM Vanilla	0
13	DE	SI18	LSTM Bidirectional	0
14	DE	SI18	LASSO	1
15	DE	SI18	LR	0
16	FR	H1N1pdm09	LSTM Vanilla	0
17	FR	H1N1pdm09	LSTM Bidirectional	0
18	FR	H1N1pdm09	LASSO	1
19	FR	H1N1pdm09	LR	0
20	FR	SI18	LSTM Vanilla	0
21	FR	SI18	LSTM Bidirectional	0
22	FR	SI18	LASSO	0
23	FR	SI18	LR	0
24	IT	H1N1pdm09	LSTM Vanilla	0
25	IT	H1N1pdm09	LSTM Bidirectional	0
26	IT	H1N1pdm09	LASSO	1
27	IT	H1N1pdm09	LR	0
28	IT	SI18	LSTM Vanilla	0
29	IT	SI18	LSTM Bidirectional	0
30	IT	SI18	LASSO	0
31	IT	SI18	LR	0
32	ES	H1N1pdm09	LSTM Vanilla	0
33	ES	H1N1pdm09	LSTM Bidirectional	0
34	ES	H1N1pdm09	LASSO	1
35	ES	H1N1pdm09	LR	0
36	ES	SI18	LSTM Vanilla	0
37	ES	SI18	LSTM Bidirectional	0
38	ES	SI18	LASSO	0
39	ES	SI18	LR	0