Universidad Carlos III de Madrid

Inteligencia Artificial en las Organizaciones

Práctica 1: Aplicación de RNA

Beatriz Sonsoles Encinas Muñoz
Lucas Gallego Bravo
Alejandro García Berrocal
Pablo Martín Muñoz

Índice

Objetivos

La pandemia causada por el COVID-19 ha tenido un gran impacto en nuestra sociedad y es un suceso que pasará a la historia como una de las peores pandemias globales, con más de 695 millones de casos confirmados y alrededor de 7 millones de muertes. Sin embargo, las medidas tomadas para frenar los efectos de la pandemia y la carrera contrarreloj para encontrar una cura, mejoraron la situación en gran medida.

A esta lucha contrarreloj contribuyó el análisis de datos realizado a través del Aprendizaje Automático. De esta manera, se pudieron identificar a los grupos de personas con mayor riesgo de contagio, además de mejorar la tasa de aciertos del diagnóstico de los pacientes, prever la propagación, rastrear el origen del virus, etc.

Esta práctica tiene como objetivo ser capaces de predecir la propagación del virus, para lo cual se hará uso de Redes de Neuronas Artificiales (RNA) que serán entrenadas con datos sobre los casos confirmados, proporcionados por The Humanitarian Data Exchange y recolectados por la Universidad Johns Hopkins, y sobre la evolución de la vacunación mundial, localizados en la web de Our World In Data. Para facilitar el análisis de estos datos, la herramienta RapidMiner tendrá un papel importante.

La práctica consta de dos partes diferenciadas, regresión y series temporales:

- La primera parte consiste en resolver un problema de regresión para obtener un modelo, el cual se determinará probando distintas arquitecturas, variando los meta-parámetros (número de entradas, número de nodos en las capas ocultas, número de capas ocultas…) y analizando los valores de parámetros como el error cuadrático medio, el error absoluto, el error cuadrático relativo o la correlación. Posteriormente, el modelo seleccionado será usado para predecir tres días consecutivos en España y en otro país a nuestra elección.

- La segunda parte debe resolver este problema mediante series temporales, que son una sucesión de datos medidos temporalmente y ordenados cronológicamente. Para entrenar la red neuronal, se deberá crear un conjunto de atributos conocido como Windowing. Los dos mejores modelos (uno para cada uno de los países seleccionados) serán determinados tras probar diferentes arquitecturas y comparar el desempeño y los resultados de los mismos. Posteriormente, serán usados para predecir los tres siguientes valores de las dos series temporales.

Procesamiento de los Datos

El procesamiento de los datos es una parte fundamental previa al análisis de los mismos. Para ello, como se ha indicado en el enunciado al principio de la práctica se ha cambiado las fechas por número de días de tal manera que cada fecha ahora sigue el formato de 'dia-#'. La siguiente fase de procesamiento de datos se lleva a cabo en un fichero python al que se ha llamado 'analisis_limpieza.ipynb', es decir un notebook de python.

Al analizar los datos se encuentran atributos que no se consideran relevantes para las predicciones de los días por el un modelo, esos atributos son la longitud, latitud y la provincia o estado.

El fichero resultante deja el país o región y los días. Se consideró que lo mejor para los datos era agrupar las provincias o estados de un país en una sola entrada del país de tal manera que no se tienen en cuenta las provincias sino el país en completo. Dicho esto se encuentran distintas entradas de un mismo país que contiene datos distintos a los otros puesto que pertenecían a distintas provincias, en este caso se considera que esas entradas son duplicados y para no perder la información de esas zonas las sumamos en uno dando lugar a una entrada de un país con todos los datos del dicho. Como otros procesamientos adicionales se han eliminado entradas que corresponden a las Olimpiadas de de los años 2020 y 2021 y se han convertido en un uno distintas zonas del Congo que aparecían en los datos como 'Congo (#)'.

Este notebook además de realizar el procesamiento de los datos generan los ficheros del tipo csv para la prueba de modelos y predicciones, para los modelos se usará el fichero 'datos_ia_sum.csv' para entrenar con todos los paises o en un sólo grupo limitado de países de Europa (España, Portugal, Francia, Italia, Alemania, Grecia y Bélgica) llamado 'datos_paises.csv'. Para las predicciones se utilizará el fichero 'data_spain_italy.csv'.

Es importante resaltar que cuando se obtengan estos ficheros hay que aplicar otro procesamiento en Excel. Este proceso implica restar los valores de un día con el día anterior dado que estos datos son los valores acumulados desde el inicio y no lo ocurrido en dicho día. Una vez finalizado este proceso ya se podría utilizar para entrenar modelos y predecir los valores de los siguiente días. Todo ese procesamiento de datos es correspondiente a la parte uno de esta práctica.

En la parte dos seguimos el mismo camino pero únicamente desde Excel, es decir, escogeremos los países que gustemos para entrenar y generar modelos y predecir con estos. Utilizando las columnas de casos confirmados y vacunas de estos y

dejando los datos de igual manera que como los de la parte uno para su uso. Finalmente, los ficheros usados para esta última parte están limitados a los días que hemos seleccionado para entrenar los modelos y con los que predecir además de encontrar todos los ficheros originales y necesarios para predicciones en el fichero zip, pero no se encuentran los ficheros generados por el notebook de python.

I. Regresión (RNA)

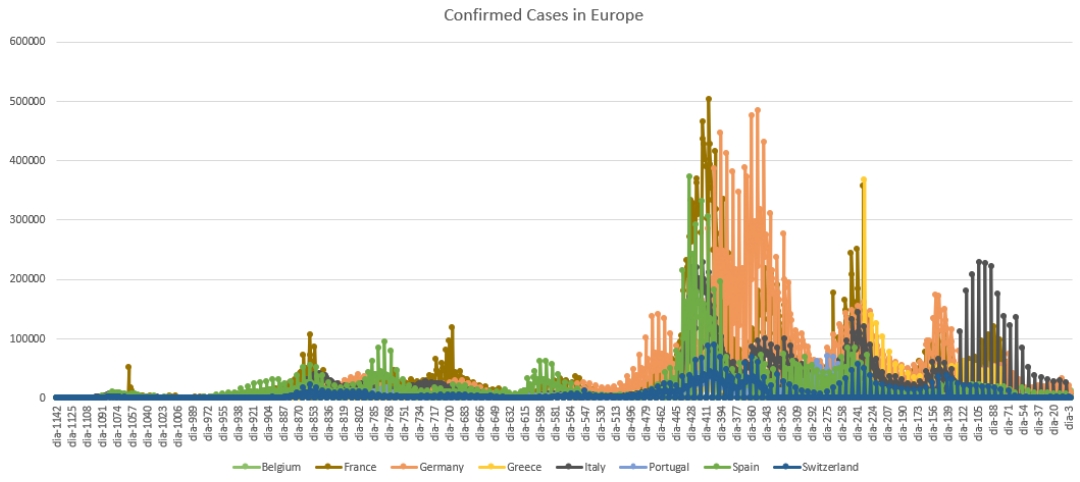
En esta primera parte de la práctica, se debe resolver el problema de regresión planteado usando Deep Learning, una RNA multicapa que usa tanto feed-forward como back-propagation para ajustar los pesos de las conexiones entre las neuronas para generar el mejor modelo. Los datos con los que se entrenará al modelo proceden del fichero csv llamado "time_series_covid19_confirmed_global", localizado en la web de The Humanitarian Data Exchange.

Entrenamiento

El proceso de entrenamiento de la red neuronal consiste en introducir el conjunto de datos previamente procesado en la arquitectura seleccionada para obtener la predicción deseada. Para conseguir resultados óptimos, se procederá a experimentar con diferentes arquitecturas y meta-parámetros hasta que se de con la mejor arquitectura.

Los datos de entrada seleccionados para realizar la predicción con el modelo han sido elegidos según la cercanía de ciertos países a España, ya que tendrán datos más similares a los de España que un país que no sea europeo.

A partir de la siguiente gráfica, que muestra los casos confirmados de COVID-19 en Bélgica, Francia, Alemania, Grecia, Italia, Portugal, España y Suiza, elegimos un periodo.



Como se puede observar en la gráfica, un rango de días interesante parece ser desde el día-601 hasta el día-541. Para poder probar con

diferentes datos de entrada, decidimos tomar cuatro intervalos de días diferentes: desde el día-561 hasta el 541, desde el 566 hasta el 541, desde el 571 hasta el 541, y desde el 601 hasta el 541, es decir, 20, 25, 30 y 60 días antes de la “Clase” (día-540), que es el atributo elegido como clase predecida.

Sin embargo, para que la red tenga una mayor variedad de datos con la que entrenar al modelo usamos los datos de todos los países, previamente procesados.

Tras elegir los datos de entrada, probamos diferentes arquitecturas. La siguiente tabla muestra las diferentes configuraciones de 11 modelos probados, indicando el nombre del modelo, los días que han sido elegidos como atributos de la red neuronal, y el número y tamaño de las capas ocultas, así como el Root Mean Squared Error (RMSE) y Mean Absolute Error (MAE) obtenidos después de aplicar *Cross-Validation*:

Nombre	Días Elegidos	Epochs	Capas Ocultas (número x tamaño)	RMSE	MAE	Correlación
Modelo 1	541-571	10	4x20	2831.858	1060.431	0.969
Modelo 2	541-561	10	2x20 2x15	2632.029	990.118	0.913
Modelo 3	541-566	10	4x20	2958.27	1116.06	0.912
Modelo 4	541-571	20	4x50	3216.755	1159.304	0.976
Modelo 5	541-571	20	4x60	3712.368	1260.216	0.974
Modelo 6	541-571	20	5x50	3691.262	1255.111	0.94
Modelo 7	541-601	20	4x60	3686.801	1197.148	<0.95
Modelo 8	541-571	20	4x60	3028.987	1084.545	0.936
Modelo 9	541-571	30	4x60	3395.559	1257.999	0.964
Modelo 10	541-601	30	4x50	3024.307	1140.703	0.924
Modelo 11	541-601	30	4x52	3284.682	1244.429	0.898

Como se ha explicado anteriormente, los días elegidos comprenden cuatro rangos distintos, uno de 20 días, otro de 25, otro de 30 y otro de 60 días antes del día a predecir. Con estos cuatro conjuntos de datos, probamos

a modificar también el número de epochs y el tamaño y número de capas ocultas.

Las epochs representan el número de veces que los datos de entrenamiento son pasados por la red neuronal, y su valor por defecto en RapidMiner son 20 epochs. Realizar más epochs implica entrenar la red neuronal más veces con todos los datos, por lo que puede significar la reducción de errores en el modelo. No obstante, demasiadas epochs pueden causar el *overfitting* o sobreajuste del modelo (haciendo así que su rendimiento sea peor con datos no vistos), al igual que muy pocas epochs pueden resultar en *underfitting*. Por tanto, para nuestros modelos probamos tanto con las 20 epochs predeterminadas como con 10 y 30 epochs.

En cuanto al número de capas ocultas de la red neuronal, a mayor número de capas, mayor es el aprendizaje del modelo y mejor es el rendimiento cuando el conjunto de datos es extenso. No obstante, también aumenta el riesgo de *overfitting* y el coste computacional al ser la red más profunda. Ya que contamos con una gran cantidad de datos, decidimos probar con 4 y 5 capas ocultas, en vez de las 2 capas ocultas que se crean de manera predeterminada. La intención es que, de esta manera, la red pueda tener un mayor aprendizaje y producir resultados con una buena precisión.

Finalmente, otro meta-parámetro con el que experimentamos es el tamaño de las capas ocultas. Al igual que con el número de capas ocultas, a mayor tamaño de las capas, mejor es el aprendizaje y mayor es la probabilidad de *overfitting*. Además, un mayor tamaño significa mejorar la capacidad de generalización, por lo que funciona mejor con datos que no haya analizado previamente. Teniendo todo esto en cuenta, decidimos probar con un rango de valores para el tamaño que varían desde 15 neuronas hasta 65, pasando por 20 (el valor predeterminado), 50 y 52 neuronas.

Nos basaremos tanto en los valores del Root Mean Squared Error como en los del Mean Absolute Error y la correlación para evaluar los meta-parámetros elegidos.

El RMSE es la métrica más usada para problemas de regresión y se rige por la siguiente fórmula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (valor\ aproximado_i - valor\ actual_i)^2}{N}}$$

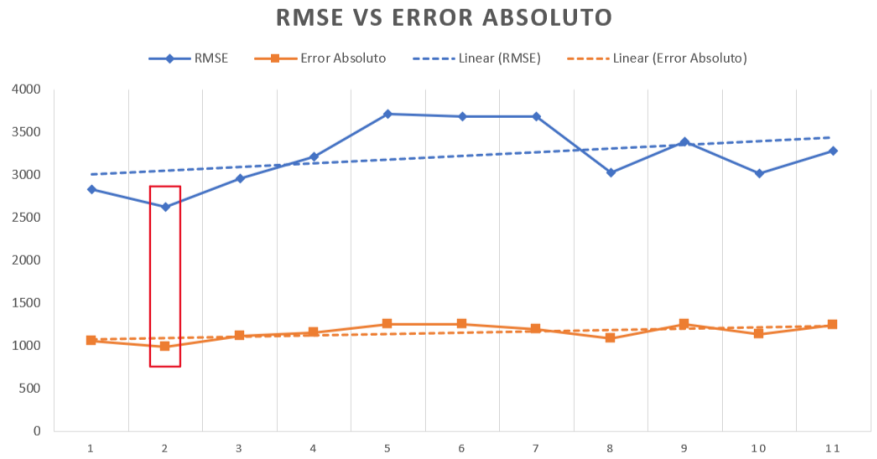
El MAE es una métrica más simple que el RMSE pero también es ampliamente usada en regresión. La fórmula para calcular este error es:

$$MAE = \sum_{i=1}^{N} \left| valor\ aproximado_i - valor\ actual_i \right|$$

El coeficiente de correlación mide la fuerza de la relación entre la "Clase" y los valores predichos. Por lo general, valores del coeficiente de correlación superiores a 0.95 pueden resultar en multicolinealidad, lo cual hace que el modelo sea inestable, así que los modelos que tengan una correlación por encima de este valor no serán tomados en cuenta.

Por tanto, si primero ordenamos los modelos de menor a mayor RMSE, observamos que los mejores modelos usan tanto 20 como 30 días como atributo seleccionado y la mayor parte realiza 10 epochs. En cuanto al número y tamaño de capas ocultas, los mejores modelos tienen 4 capas con un tamaño de 20 y 15 neuronas. Si estudiamos el MAE de los modelos, los mejores tienen, de nuevo, tanto un rango de 20 como de 30 días, usan 10 epochs, y contienen 4 capas ocultas, con alrededor de 20 neuronas de tamaño.

La descripción gráfica de los dos errores mencionados se muestra en la siguiente imagen, donde la recta de puntos representa la media de los respectivos errores:



Usando las características predominantes descubiertas, elegimos como mejor modelo el **Modelo 2**, marcado en rojo en la gráfica superior. Este modelo, como se puede observar en la tabla, usa como atributos los 20 días anteriores a la "Clase", realiza 10 epochs y contiene 4 capas ocultas, dos capas ocultas de 20 neuronas de tamaño, y otras dos con 15 neuronas. El siguiente mejor modelo es el Modelo 1, pero debido a su alto coeficiente de correlación, hemos decidido no usarlo para hacer las predicciones.

Problemas

Al calcular el incremento de casos confirmados en Excel, algunos de los valores resultantes eran negativos. Asumimos que esto era por casos en los que los contagiados se recuperaban del virus o fallecían.

Dado que estos valores negativos perjudicaban al rendimiento del modelo, decidimos convertirlos en ceros usando la función de Excel *IF(valor<0, 0, valor)*.

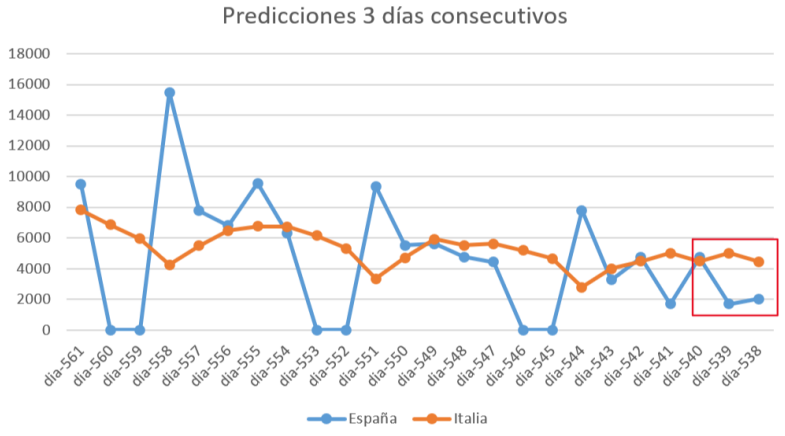
Resultados

Después de encontrar el mejor modelo para resolver el problema de regresión (**Modelo 2**), se puede usar el modelo para predecir 3 días consecutivos de la evolución del virus en España y en otro país elegido, que en nuestro caso es Italia. Por tanto, generamos un nuevo fichero de Excel con los datos del día-541 al día-561 y dejamos el último día ("Clase") vacío, ya que es el dato a predecir. Una vez obtenida la predicción, movimos los datos (incluyendo la predicción) una celda hacia la izquierda y volvimos a realizar la predicción.

Este proceso se repitió una última vez para así obtener la predicción de los 3 días consecutivos seleccionados. Los resultados de este proceso se muestran en la siguiente tabla:

País	Día-540	Día-539	Día-538
España	4745	1698	2020
Italia	4483	5024	4452

Gráficamente, podemos comprobar como las predicciones de los tres días consecutivos siguen una tendencia parecida a la de los días anteriores:



Dado que los valores negativos en los datos de España (en azul) fueron cambiados por ceros, tiene sentido que los datos del día-561 al día-541 tengan variaciones. No obstante, se puede apreciar en los 3 días predichos (en rojo) que el modelo lo ha tenido en cuenta y siguen la misma tendencia que los días anteriores.

En lo referente a Italia (naranja), los datos siguen un comportamiento mucho más regular, por lo que los 3 días predichos también continúan con esta conducta.

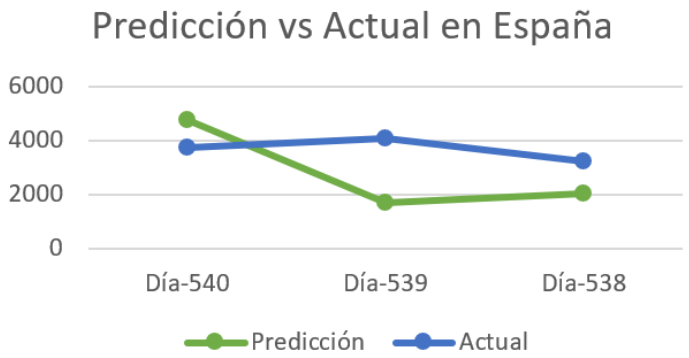
Comparación de los resultados

Para determinar si las predicciones son realmente correctas, las comparamos con sus valores reales, obtenidos del csv con el que entrenamos la red neuronal:

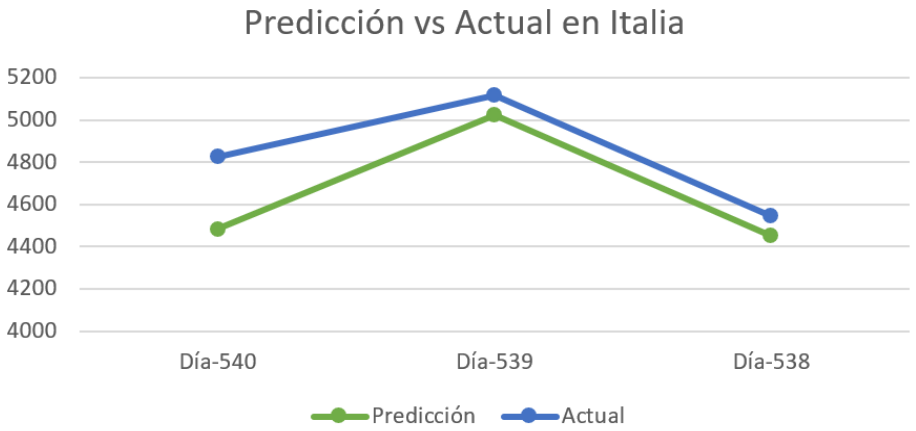
	Día-540		Día-539		Día-538	
País	**Predicción**	**Actual**	**Predicción**	**Actual**	**Predicción**	**Actual**
España	4745	3723	1698	4075	2020	3222
Italia	4483	4826	5024	5115	4452	4544

Si quisiéramos comparar los datos de las predicciones y los actuales de una manera más visual, podríamos usar las siguientes gráficas. Por simplicidad, se ha usado una gráfica para cada país, donde los 3 días predichos están remarcados en rojo.

Gráfica de España:



Gráfica de Italia:



Como podemos observar, en el caso de España, los valores aproximados y los reales son bastante similares, aunque el día-540 varíe ligeramente del valor esperado. Esto es debido a que, como se ha indicado anteriormente, hayan algunos ceros en los valores usados para la predicción, por lo que tiene sentido que las aproximaciones estén sujetas a cierto error.

No obstante, en el caso de Italia, que no tiene ningún cero en el rango elegido de 20 días, se puede apreciar una mejor correlación entre los valores aproximados y los reales en la gráfica superior.

En conclusión, podemos decir que el modelo encontrado es el más apropiado para la predicción de los 3 días consecutivos a nuestro periodo seleccionado, y que tanto las predicciones como los valores reales son similares, por lo que el problema de regresión se ha resuelto satisfactoriamente.

II. Series Temporales

En esta parte de la práctica, el problema de regresión previo se debe aplicar a dos conjuntos de datos diferentes. El primer conjunto de datos es el mismo que en la parte I de la práctica de casos confirmados, mientras que el segundo conjunto que son las vacunas aplicadas, se puede obtener de la web Our World In Data del fichero “owid-covid-data”. Este fichero de datos con dichos conjuntos será usado para entrenar la red y obtener el mejor modelo para cada uno de los países seleccionados.

Entrenamiento

Se han probado diferentes arquitecturas combinando parámetros del operador windowing y el sliding window validation, en éste último con diferentes parámetros del training window size como con el modelo de deep learning que se ha usado dentro de forma similar a la primera parte. Todas las combinaciones se encuentran detalladas en la siguiente tabla. El sliding window validation ha servido, de forma similar al cross-validation de la parte 1, a pesar de ser conceptos completamente distintos, para medir diferentes parámetros como el RMSE, MAE o correlación para decidir cuál de los modelos probados era el mejor para realizar las predicciones.

El intervalo de días es el mismo para cada uno de los modelos que va desde el 7/15/2021 al 9/13/2021

Nombre	Tamaño de Ventana	Training Window size	Epochs	Capas Ocultas (número x tamaño)	RMSE
Modelo 1	20	20	25	4x20	2245.508
Modelo 2	20	20	20	4x55	2666.022
Modelo 3	15	20	35	5x50	1781.408
Modelo 4	15	20	40	4x30	2293.217
Modelo 5	30	20	40	3x20,3x15	1504.942
Modelo 6	30	20	30	4x45	1496.899
Modelo 7	45	15	15	5x65	935.140
Modelo 8	15	14	40	5x20	3060.170

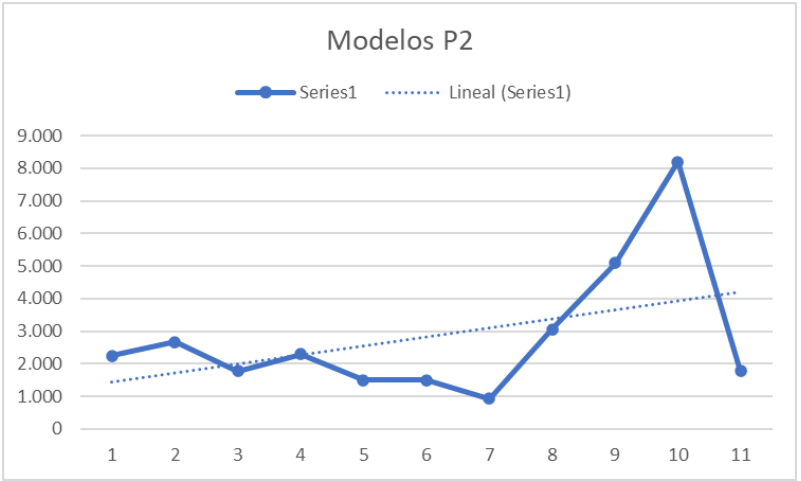
Modelo 9	55	5	20	4x30	5094.098
Modelo 10	55	5	20	4x55	8201.012
Modelo 11	40	20	30	2x10	1770.816

Se han probado diferentes tamaños de entrenamiento para las ventanas que indica la cantidad de días anteriores que utilizara el modelo para la predicción del siguiente. Observando la tabla, se observa que aquellos modelos que utilizaron un tamaño mucho menor de este parámetro tuvieron errores más altos como es en el caso de los modelos que con tamaños de 5 en ventanas de entrenamiento, en ambos tuvieron los errores más altos de todos los modelos. Sin embargo en los modelos que se han usado 20 o 15 no varió demasiado respectivamente.

En cambio con el tamaño de ventana no se noto diferencias significativas entre diferentes tamaños como 20, 15, 40 o 55. Si bien es verdad que los modelos de 55 tiene errores más altos, esto debe ser al parámetro de entrenamiento de ventana ya que el modelo 11o al 7, que pese a tener tamaños más altos como 45 o 40 los valores no se alejan como lo hacen los otros. Tanto los epochs como las capas ya se usaron de la misma forma como en la parte 1.

Para decidir que Modelo es el mejor y cual se usará para predecir los tres días se ha hecho uso del RMSE recogido por el sliding window.

La descripción gráfica del error mencionado se muestra en la siguiente imagen, donde la recta de puntos representa la media de los respectivos errores:



Tras analizar esto se ve que el modelo que obtuvo menos error fue el 7 por lo que será el modelo que usaremos, que si se observa la tabla anterior se ve que se usó un tamaño de ventana de 45 ,un entrenamiento de 15, 15 epochs y 5 capas ocultas de tamaño 65 cada una.

Problemas

Los días del primer fichero de la parte uno no se corresponden con el mismo número de días incluidos en el fichero para la parte dos. Mientras que el fichero de casos confirmados incluye desde el día 22/1/22 hasta el 9/3/23, el fichero con datos sobre la evolución de la vacunación, contiene información desde el día 3/1/20 hasta el 27/9/23. Por tanto, eliminamos los días que no se encontraban entre el rango más pequeño de la parte uno con el fin de evitar meses de valores igual a cero, lo que haría que el modelo no fuera tan preciso.

Al principio salían errores muy altos debido a que no se había limpiado bien el excel y había que cambiar el formato de ciertas columnas. Después de eso se probó con todas las fechas disponibles pero para mejores resultados acotamos el espacio temporal a dos meses, sin embargo los resultados no se diferenciaron mucho.

También se vio que ciertas combinaciones de los parámetros window size y training window size daban problemas. Uno dependía del otro de manera que no se podían alterar independientemente del otro. Si se variaba uno había que variar el otro para que funcionara.

Comparación de los Resultados

Para determinar si las predicciones son realmente correctas, las comparamos con sus valores reales, obtenidos del csv con el que entrenamos la red neuronal.

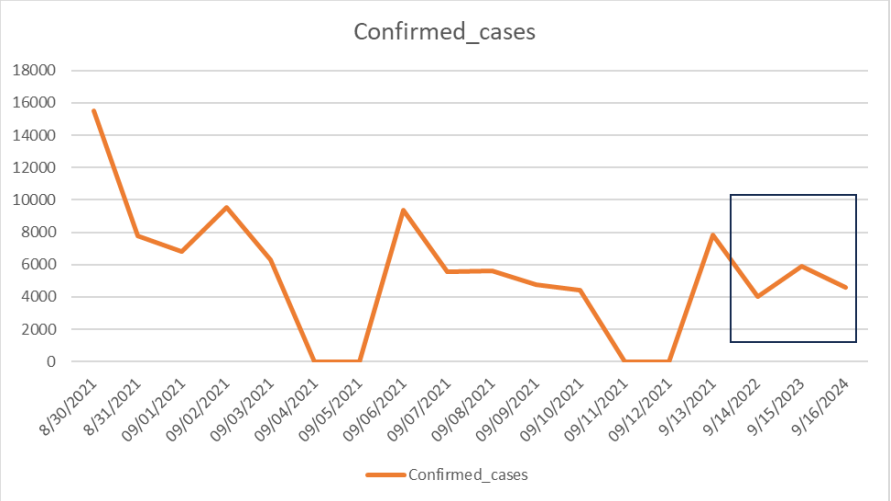
Para hacer las predicciones había que usar un csv igual el que se usaba para entrenar al modelo por lo tanto se guardaba además del modelo para usarlo luego un excel con el mismo número de atributos que espere el modelo.

Como el entrenamiento llegaba como se indicaba hasta el 9/13/2021, se han predecido los datos para el 9/14/2021, 9/15/2021 y 9/16/202.

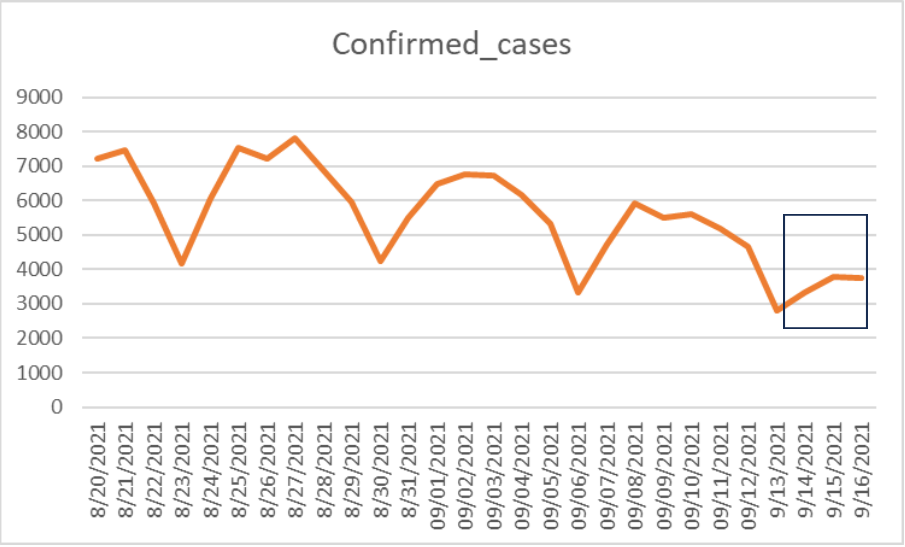
CASOS:

	9/14/2021		9/15/2021		9/16/2021	
País	**Predicción**	**Actual**	**Predicción**	**Actual**	**Predicción**	**Actual**
España	3999	3261	5905	3723	4580	4075
Italia	3346	4009	3791	4826	3753	5115

Casos España:



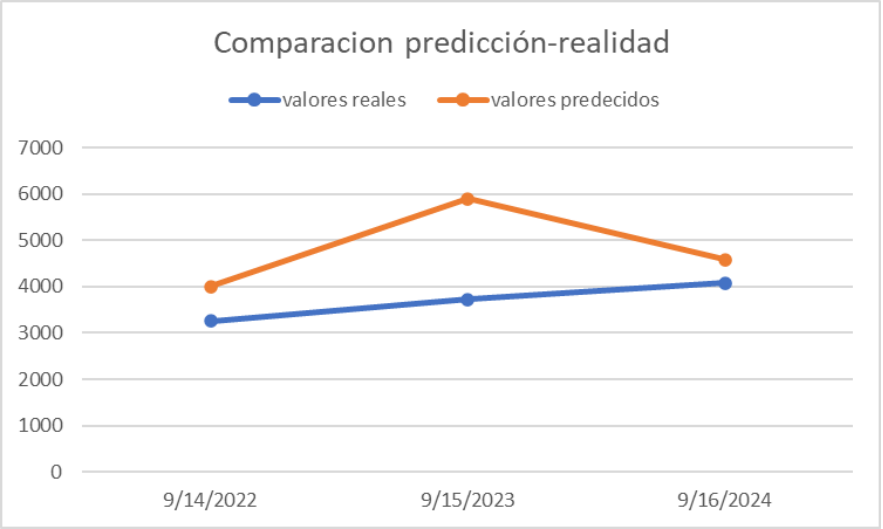
Casos italia:



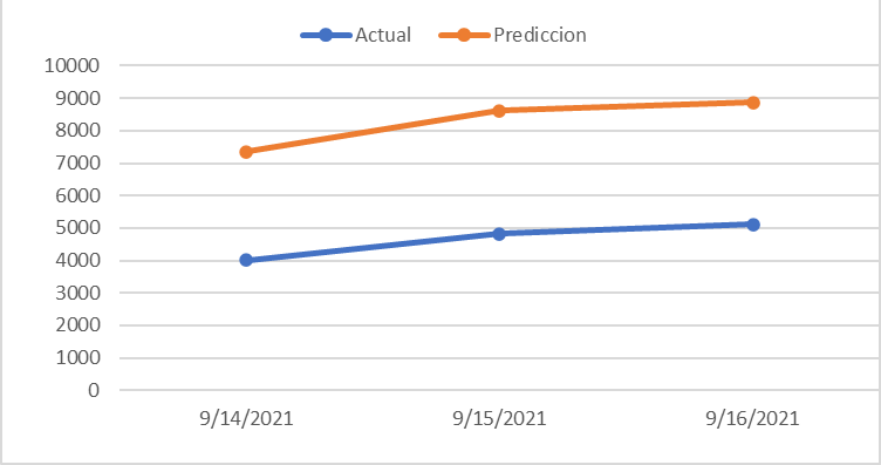
Si observamos las gráficas con los datos reales a los que se le ha añadido los últimos tres días predecidos se observa cómo siguen la tendencia que se estaba dando los días anteriores.

A continuación se muestra una gráfica con la comparación de ambos datos: predicciones y datos reales.

España:



Italia:

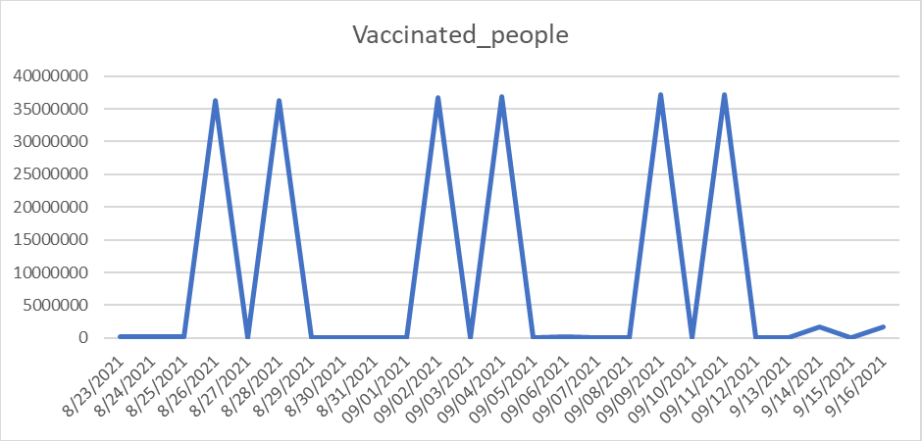


Se ve como los resultados han salido bastante cercanos a lo esperado, lo único que el segundo día en españa se alejó de más, viéndose tanto en la gráfica de la tendencia respecto a los anteriores días como en la comparación con los valores reales.

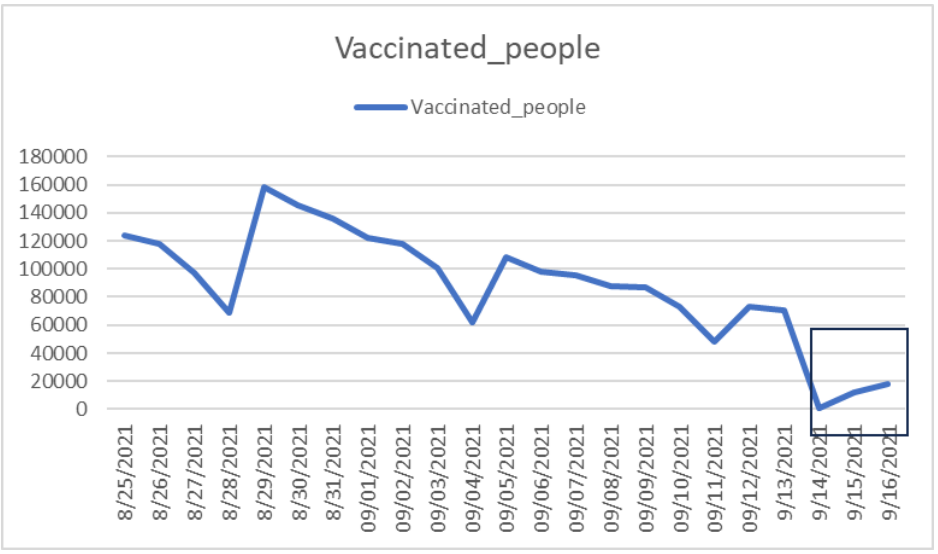
VACUNAS:

	9/14/2021		9/15/2021		9/16/202	
País	**Predicción**	**Actual**	**Predicción**	**Actual**	**Predicción**	**Actual**
España	1656437	38253	0	44303	1635550	37385758
Italia	607	68699	12397	62962	18088	75692

Vacunas España:



Vacunas Italia:



Observando ambas gráficas vemos que la predicción en España ha sido más acertada, aunque si bien los números no son iguales, se ve cómo sigue la tendencia de datos cercanos a 0 antes de volver a dispararse los datos, como pasa periódicamente ya que se ve una gráfica que se repite de forma constante.

Si bien para Italia los datos bajan de más puede ser debido a que las tendencias de Italia son muy distintas a las de España, a pesar de su similitud geográfica. Al igual que con España salen valores muy bajos

A continuación se muestra una gráfica con la comparación de ambos países con los valores de las predicciones y datos reales. Cabe destacar que en España los datos de los dos primeros días salen muy parecidos debido a que el tercer datos se aleja mucho de los valores anteriores.