Universidad Carlos III de Madrid - Grado en Ingeniería Informática

INTELIGENCIA ARTIFICIAL EN LAS ORGANIZACIONES

Práctica 1 - Redes de Neuronas Artificiales

Curso 2021/2022

Jorge Rodríguez Fraile, Grupo 83
Franco Exequiel Schüler Allub, Grupo 83
Xu Chen, Grupo 83
Enrique Ángel Arrabal Ruiz, Grupo 83

Índice

Índice de tablas

Índice de ilustraciones

Introducción

El objetivo principal de esta práctica es la utilización de Redes de Neuronas Artificiales para la predicción de casos confirmados de contagios de COVID-19. Para el uso de este tipo de arquitecturas hemos utilizado la herramienta Weka. Para ello, hemos precisado de realizar numerosas experimentaciones de manera que pudiésemos elegir los 6 mejores modelos obtenidos, tal y como indica el enunciado.

En primer lugar, expondremos el contexto (o el estado del arte) de predicciones de este tipo de datos que ya se hayan hecho con anterioridad. A continuación, procederemos a documentar todos y cada uno de los experimentos realizados, así como los 6 mejores modelos obtenidos para, exponer los resultados finales para los 2 mejores modelos capaces de predecir 15 días de contagios confirmados y muertes, para finalmente, explicar las conclusiones. Los países que se han seleccionado han sido son: España y Colombia.

En este enlace podrá acceder al Excel en el que podrá encontrar los distintos modelos que hemos generado a lo largo de la práctica, y mediante este otro enlace se puede acceder a la descarga de los diferentes ficheros con los que se ha desarrollado este proyecto.

Contexto

En esta sección, incluiremos cualquier información conocida por nosotros, casos similares, artículos científicos y otros documentos relacionados con nuestra temática. En concreto, hemos encontrado dos artículos relacionados con la predicción de casos y muertes causados por COVID. Para cada uno de ellos hemos incluido un resumen general y los resultados obtenidos por los autores.

Estimación y predicción de casos de COVID-19 en metrópolis brasileñas[3]

Haciendo uso de Google Académico[5], hemos encontrado un artículo muy interesante sobre estimación y predicción de casos confirmados de COVID-19. El objetivo de este estudio realizado en 2020 fue la estimación de la tasa de transmisión del virus, así como el número de muertes causadas por el mismo. Los métodos utilizados consistieron en la creación de modelos matemáticos destinados a predecir este tipo de datos en las capitales brasileñas más importantes. Los resultados obtenidos fueron equiparados con los resultados actuales referentes a COVID-19 y las conclusiones fueron que las medidas tomadas por los gobiernos no fueron para nada suficientes.

Como es sabido por todos, el COVID-19 es un virus que causa problemas en el sistema respiratorio. La pandemia comenzó a mediados del 2020 y, debido a su elevado ratio de contagio, se extendió exponencialmente a todo el mundo. Ya hemos comentado que en

Brasil las medidas establecidas no fueron suficientes, y tampoco lo fue en el resto del mundo, Los datos en Brasil en concreto fueron bastante preocupantes y, por lo tanto, la motivación de este artículo en particular es más que notable. Para entender las características de esta enfermedad en el pueblo brasileño, se aplicaron modelos matemáticos basados en ecuaciones diferenciales y centrados en tres grupos distintos de individuos: susceptibles, infectados y recuperados.

Paralelamente, se pudo predecir la tasa de reproducción de la enfermedad, que indica el número de personas sanas que puede infectar una persona contagiada. También se investigó el número de casos diarios, el día con más infectados y el número de muertes.

Como resultados principales, se puede afirmar que se consiguieron predecir la mayor parte de los casos. En el sur del país se observó que los casos obtenidos fueron considerablemente similares con los casos reales. En la zona norte tan solo se investigó una ciudad y el gran aumento en los casos fue también bastante análogo al modelo obtenido.

Prediction of Covid19 with the use of Random Forests Algorithm and Artificial Neural Networks[4]

Haciendo uso también de Google Académico[5], expondremos un paper realizado en 2021 sobre predicción de Covid19 con la utilización de algoritmos de machine learning así como redes de neuronas artificiales. En concreto, se expondrá la implementación de un modelo que trate de analizar y predecir la propagación de este virus mediante técnicas de Inteligencia Artificial en lenguaje Python.

En primer lugar, se hizo un análisis exploratorio de los datos, obteniendo medias, modas, varianzas, desviaciones típicas, etc. para tratar de extraer algún tipo de información adicional de la gran cantidad de datos de los que se disponían. A continuación, se utilizaron algoritmos de clasificación y de regresión (Random Forest), así como un modelo de red neuronal que tratara de generalizar lo más posible los contagios generados por COVID.

Parte 1

Procedimiento

Obtención de los datos

El primer paso en cualquier tipo de predicción es la obtención de la información necesaria para la misma. En nuestro caso, hemos obtenido los datos de ***The Humanitarian Data Exchange*** [1]**,** una entidad que recopila datos relativos a enfermedades o epidemias. En

concreto, los datos son proporcionados por la Universidad Johns Hopkins[2] haciendo uso de diversas fuentes.

Para nuestro caso práctico, nos interesan los datos relativos a los casos confirmados de contagio de COVID-19, clasificados por países y regiones. Sin embargo, estos datos no se pueden exportar directamente a Weka, ya que poseen una serie de características que deben ser corregidas.

Por ejemplo, muchos países aparecían con comas, guiones, apóstrofes y asteriscos, esto provocaba que Weka no comprendiera el tipo de datos y, por lo tanto, no permitía visualizar y abrir el archivo.

Modelos

En esta sección se expondrán todos los modelos considerados para predecir día a día los contagios confirmados de COVID-19. Emplearemos el perceptrón multicapa, una red de neuronas capaz de hacer regresión sobre datos no lineales, basándose en datos previos de ejemplo que en este caso son los descritos en la sección anterior. Para generar estos modelos se considerarán los siguientes parámetros:

- **Capas ocultas:** Número de capas ocultas que empleará nuestra red de neuronas artificial, estas son las encargadas de realizar el procesamiento por lo que afectará en gran medida a los resultados.
- **Factor de aprendizaje:** Indicará cómo de rápido irán cambiados los pesos asociados a las neuronas de la red.
- **Tiempo de entrenamiento:** Número de ciclos de entrenamiento con los datos de entrenamiento que se realizarán.

A continuación, mostraremos todos los modelos que hemos considerado para este caso. Estos seguirán unos criterios de identificación y tendrán unas características propias, las cuales darán más importancia a los porcentajes de errores de cada modelo.

La codificación que se ha elegido para nombrar los modelos sigue el siguiente esquema XX-YY-ZZ-WW, tal que:

- **XX:** El número de capas de la red neuronal.
- **YY:** El factor de aprendizaje, que determina en qué medida aprenderá la red.
- **ZZ:** Número de ciclos de entrenamiento de la red.
- **WW:** Días considerados para predecir.

Modelo	Cross-validation RAE	Cross-validation RRSE	Percentage split RAE	Percentage split RRSE
5-0,01-500-610	31,27%	45,04%	16,63%	17,89%
5-0,01-500-610	29,39%	42,05%	16,10%	20,12%
5-0,05-500-610	29,81%	43,80%	17,40%	18,94%
5-0,1-1000-610	**29,61%**	**41,67%**	**14,50%**	**18,42%**
5-0,1-500-610	29,15%	40,82%	16,19%	20,91%
5-0,3-1000-610	30,71%	44,09%	15,52%	19,04%
5-0,3-500-610	36,57%	41,53%	15,92%	19,79%
7-0,15-500-610	34,56%	48,56%	24,25%	22,64%
8-0,01-1000-610	29,88%	48,23%	25,80%	23,01%
8-0,01-500-610	31,87%	48,79%	23,10%	21,43%
8-0,05-1000-610	29,67%	49,10%	29,60%	25,23%
8-0,05-500-610	31,52%	50,67%	16,93%	18,24%
8-0,1-1000-610	29,80%	47,61%	16,58%	20,96%
8-0,1-500-610	31,04%	45,60%	18,88%	19,06%
8-0,3-2000-610	32,00%	48,78%	19,32%	20,60%
8-0,3-500-610	34,29%	48,78%	19,32%	20,60%
10-0,05-500-610	29,16%	45,46%	17,92%	21,71%
10-0,1-1000-610	30,19%	47,72%	19,13%	23,37%
10-0,1-500-610	28,61%	44,43%	14,43%	19,63%
10-0,3-500-610	40,59%	49,98%	15,04%	18,43%
5-0,01-1000-90	19,46%	19,30%	26,55%	45,82%
5-0,01-1000-90	29,14%	43,78%	17,23%	18,09%
5-0,05-500-90	19,34%	19,27%	21,26%	27,49%
5-0,1-1000-90	**18,02%**	**17,79%**	**29,02%**	**43,93%**
5-0,1-500-90	**19,28%**	**19,41%**	**22,48%**	**28,73%**
10-0,05-500-90	**17,87%**	**16,82%**	**28,71%**	**33,63%**
10-0,1-500-90	17,83%	17,71%	29,74%	42,27%

Tabla 1. La tabla muestra las tasas de error generadas por cada modelo

- **RAE**: Relative Absolute Error.
- **RRSE**: Root Relative Squared Error.

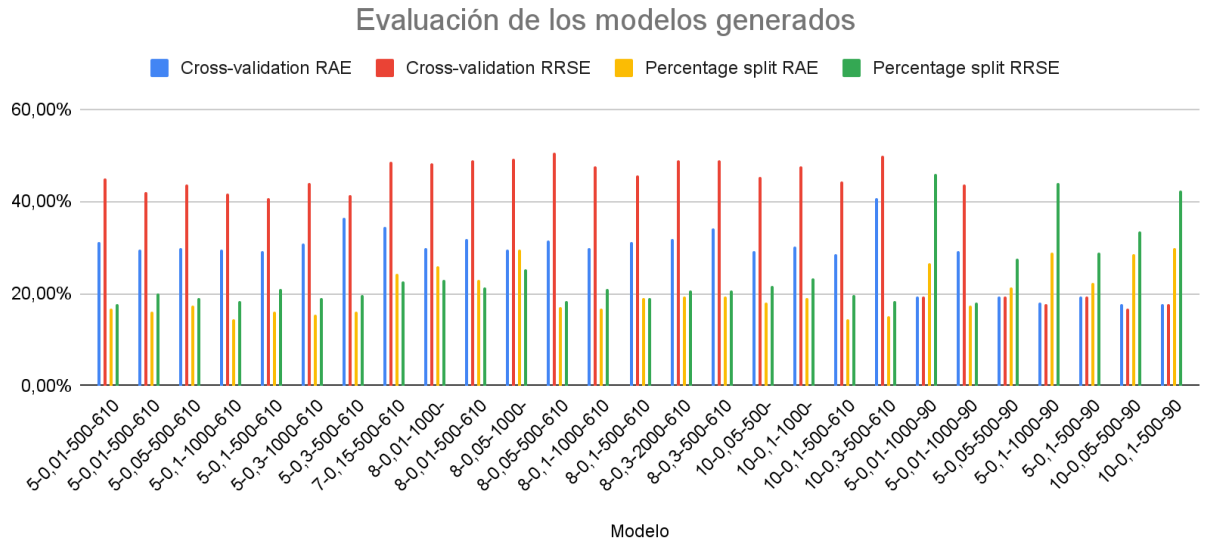

Ilustración 1. El gráfico muestra la comparativa de cada modelo

Estos son los resultados obtenidos de todos los modelos que hemos considerado, se ha proporcionado la representación tabular para poder ver más detalle los resultados, pero en la representación gráfica se puede ver mucho mejor.

Como se puede ver, los mejores resultados no solamente están asociados al número de capas ocultas de la red, dado que los modelos de 7 a 10 capas son peores en general que los de 5, aun dándoles más ciclos por si necesitaban más entrenamiento. Es por esto por lo que en los siguientes modelos que generamos (los de 90 días) nos centramos solo en 5 y 10 capas.

En cuanto al factor de aprendizaje, vimos que el valor por defecto de 0,3 era demasiado alto, por lo que probamos a reducirlo a 0,1, 0,05 y 0,01. Esta reducción como se puede ver mejoran los errores, sin embargo pudimos ver que reducir el factor a partir de un punto no produce mejoras y nos quedamos con el factor 0,1 como uno de los mejores para este problema.

Tras experimentar con los 610 días y ver que los modelos que se generaban no eran tan buenos como se esperaba, se redujo a 90 días. De esta manera solo considera para el entrenamiento los últimos 3 meses de contagios.

Con esta reducción del número de y con el conocimiento adquirido de los modelos de 610 días se generaron otros 7 nuevos modelos, que esta vez sí que daban errores asumibles para este problema.

En las siguientes 6 subsecciones se exponen los resultados de los 6 mejores modelos:

Modelo 1

Parámetros

- 5 capas
- Tasa de aprendizaje de 0,1
- 1000 ciclos
- 90 días

5-0,1-1000-90

Predicción

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
Real	4.950.356	4.951.251	4.951.726	4.948.513	4.950.253	4.951.675
5-0,1-1000-90	4.961.082	4.985.635	5.002.465	4.970.545	4.985.635	5.000.552

Error cometido

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
5-0,1-1000-90	0,22%	0,69%	1,02%	0,45%	0,71%	0,99%

Como podemos observar, se obtienen resultados bastante prometedores, con porcentajes de error menores a la unidad. Esto puede parecer contraproducente con los parámetros elegidos, ya que tenemos muy pocas capas y una tasa de aprendizaje bastante alta (generalmente, se usan tasas cercanas a las centésimas). Sin embargo, esto se ha visto contrarrestado con el gran número de ciclos (1000) que ha provocado que el modelo aprenda lo suficiente como para realizar predicciones considerablemente buenas.

Modelo 2

Parámetros

- 5 capas
- Tasa de aprendizaje de 0,1
- 500 ciclos
- 90 días

Predicción

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
Real	4.950.356	4.951.251	4.951.726	4.948.513	4.950.253	4.951.675
5-0,1-500-90	5.114.900	5.114.900	5.114.900	5.231.749	5.231.749	5.231.749

Error cometido

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
5-0,1-500-90	3,32%	3,31%	3,30%	5,72%	5,69%	5,66%

Para este modelo en concreto hemos obtenido resultados que, si bien no son malos ni mucho menos, son considerablemente peores que los obtenidos para el Modelo 1. Como ya comentamos en aquel modelo, un algoritmo ejecutado con pocas capas, una tasa de aprendizaje relativamente grande y pocos ciclos puede traducirse en peores resultados. Esto es exactamente lo que ha pasado con este modelo.

Modelo 3

Parámetros

- 10 capas
- Tasa de aprendizaje de 0,05
- 500 ciclos
- 90 días

Predicción

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
Real	4.950.356	4.951.251	4.951.726	4.948.513	4.950.253	4.951.675
10-0,05-500-90	5.150.136	5.168.532	5.187.656	5.302.957	5.318.890	5.333.319

Error cometido

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
10-0,05-500-90	4,04%	4,39%	4,76%	7,16%	7,45%	7,71%

En este caso tenemos más capas que en el modelo anterior y menor tasa de aprendizaje, pero los mismos ciclos. A pesar de esto vemos que los resultados no mejoran y es que muchas veces más capas no implica mejores resultados. Lo mismo ocurre con la tasa de aprendizaje, un algoritmo que trate de afinar más la solución no tiene por qué ser mejor.

Modelo 4

Parámetros

- 5 capas
- Tasa de aprendizaje de 0,1
- 1000 ciclos
- 610 días

Predicción

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
Real	4.950.356	4.951.251	4.951.726	4.948.513	4.950.253	4.951.675
5-0,1-1000-610	4.911.532	4.897.915	4.900.592	4.892.974	4.912.907	4.932.853

Error cometido

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
5-0,1-1000-610	0,78%	1,08%	1,03%	1,12%	0,75%	0,38%

Para este modelo tenemos los mismos parámetros que el Modelo 1, pero en este caso hemos entrenado con todos los días disponibles. Los errores cometidos son ligeramente mayores que en el modelo ya mencionado; esto puede deberse a la gran cantidad de datos que se han utilizado para entrenar. Puesto que queremos predecir los nuevos contagios en los próximos días a la actualidad, los datos referentes a hace más de un año no son para nada relevantes para el problema en cuestión.

Modelo 5

Parámetros

- 5 capas
- Tasa de aprendizaje de 0,05
- 500 ciclos
- 90 días

Predicción

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
Real	4.950.356	4.951.251	4.951.726	4.948.513	4.950.253	4.951.675
5-0,05-500-90	5.117.304	5.136.210	5.154.941	4.799.335	4.808.288	4.818.005

Error cometido

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
5-0,05-500-90	3,37%	3,74%	4,10%	3,01%	2,87%	2,70%

Consecuentemente, teniendo una combinación relativamente mala de parámetros podemos intuir que los resultados no serán muy buenos. Esto se debe a que hemos seleccionado pocas capas y pocos ciclos, a pesar de una tasa de aprendizaje baja.

Modelo 6

Parámetros

- 10 capas
- Tasa de aprendizaje de 0,1
- 500 ciclos
- 90 días

Predicción

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
Real	4.950.356	4.951.251	4.951.726	4.948.513	4.950.253	4.951.675
10-0,1-500-90	4.936.611	4.953.580	4.971.541	5.334.052	5.351.554	5.369.355

Error cometido

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
10-0,1-500-90	0,28%	0,05%	0,40%	7,79%	8,11%	8,44%

Para este supuesto, tenemos un número alto de capas y de tasa de entrenamiento y, sin embargo, hemos empleado 500 ciclos de entrenamiento para los últimos 90 días de datos. Como podemos observar, para España se obtienen grandes resultados (cercano al 0% de error) pero para Colombia los resultados empeoran. Esto puede deberse a que el modelo se ha ajustado muy bien a los datos de España, pero para los de Colombia los hiperplanos no han conseguido ajustarse a la función.

Resultados

Como esperábamos las predicciones al principio son más precisas y tienen menor error, pero cuanto más predecimos sobre lo predicho más falla, aunque no aumenta en gran medida para los 3 días considerados.

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
Real	4.950.356	4.951.251	4.951.726	4.948.513	4.950.253	4.951.675
5-0,1-1000-90	**4.961.082**	**4.985.635**	**5.002.465**	**4.970.545**	**4.985.635**	**5.000.552**
5-0,1-500-90	5.114.900	5.114.900	5.114.900	5.231.749	5.231.749	5.231.749
10-0,05-500-90	5.150.136	5.168.532	5.187.656	5.302.957	5.318.890	5.333.319
5-0,1-1000-610	4.911.532	4.897.915	4.900.592	4.892.974	4.912.907	4.932.853
5-0,05-500-90	5.117.304	5.136.210	5.154.941	4.799.335	4.808.288	4.818.005
10-0,1-500-90	4.936.611	4.953.580	4.971.541	5.334.052	5.351.554	5.369.355

Tabla 2. Resultados de las predicciones obtenidas por cada modelo

	Día 1 ESP	Día 2 ESP	Día 3 ESP	Día 1 COL	Día 2 COL	Día 3 COL
5-0,1-1000-90	**0,22%**	**0,69%**	**1,02%**	**0,45%**	**0,71%**	**0,99%**
5-0,1-500-90	3,32%	3,31%	3,30%	5,72%	5,69%	5,66%
10-0,05-500-90	4,04%	4,39%	4,76%	7,16%	7,45%	7,71%
5-0,1-1000-610	0,78%	1,08%	1,03%	1,12%	0,75%	0,38%
5-0,05-500-90	3,37%	3,74%	4,10%	3,01%	2,87%	2,70%
10-0,1-500-90	0,28%	0,05%	0,40%	7,79%	8,11%	8,44%

Tabla 3. Errores cometidos por cada modelo

En la tabla anterior se pueden ver los valores predichos por los 6 modelos y sus correspondientes errores de predicción.

Para facilitar la visualización de la progresión de los errores en los 3 días predichos se ha representado mediante dos gráficas separadas, una para España y otra para Colombia.

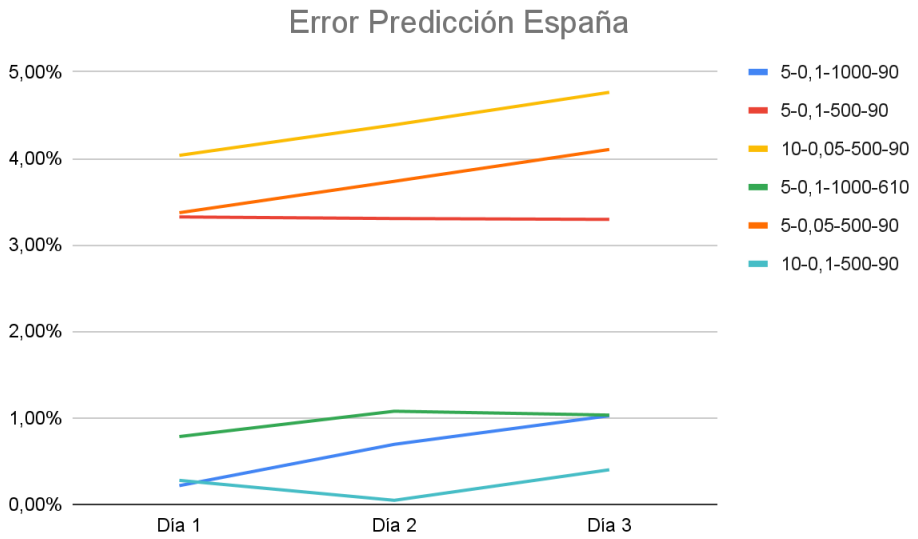

Ilustración 2. El gráfico muestra una comparativa entre los errores cometidos por las predicciones de cada modelo elegido para el caso de España

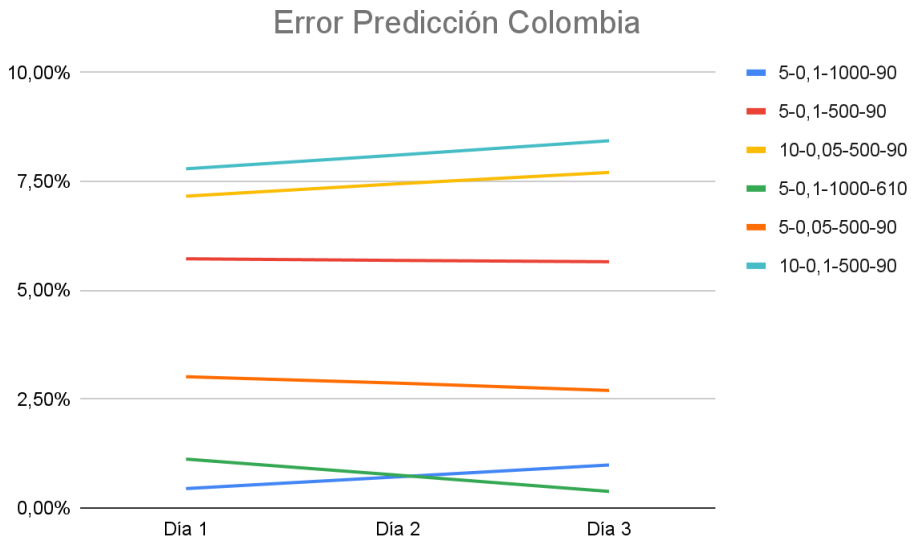

Ilustración 3. El gráfico muestra una comparativa entre los errores cometidos por las predicciones de cada modelo elegido para el caso de Colombia.

Podemos apreciar que el mejor modelo es el de **5 capas ocultas**, **factor de aprendizaje 0,1**, entrenado en **1000 ciclos** con los **últimos 90 días**. Los errores para este modelo son prácticamente menores al 1 %, por lo que podemos decir que es un modelo que realiza su función de una manera eficaz.

Parte 2

En este apartado se van a desarrollar distintos modelos para la predicción de casos confirmados y fallecidos en España y Colombia. Esta predicción será para 15 días.

Para esta parte, los datos serán presentados en forma de series temporales. Esto quiere decir que para generar el modelo, se utilizarán los k valores anteriores para obtener los siguientes valores. Cada valor que se utiliza será una entrada para la red de neuronas artificiales.

Lo primero que haremos será generar el conjunto de entrenamiento que utilizará la red de neuronas, que serán los casos confirmados y los fallecidos totales en ese día, sin ser acumulados. Cómo hemos mencionado, estos son los valores que se utilizarán como entradas a la red.

Una vez tengamos los conjuntos de datos, pasamos al proceso de entrenamiento, donde se ha entrenado un perceptrón multicapa y hemos obtenido los modelos que se explicarán en el siguiente apartado.

España

Para realizar la predicción de casos confirmados y fallecidos en España de 15 días, hemos realizado varias pruebas con distinto número de capas ocultas, ciclos, razones de aprendizaje, así como valores de lag length. Con esto, queríamos estudiar los distintos efectos que generaban estos parámetros en una tarea de predicción de 15 instancias.

En la siguiente gráfica se pueden ver todos los modelos que hemos probado con su respectiva media entre el error de predicción de confirmados y muertes para los 15 días:

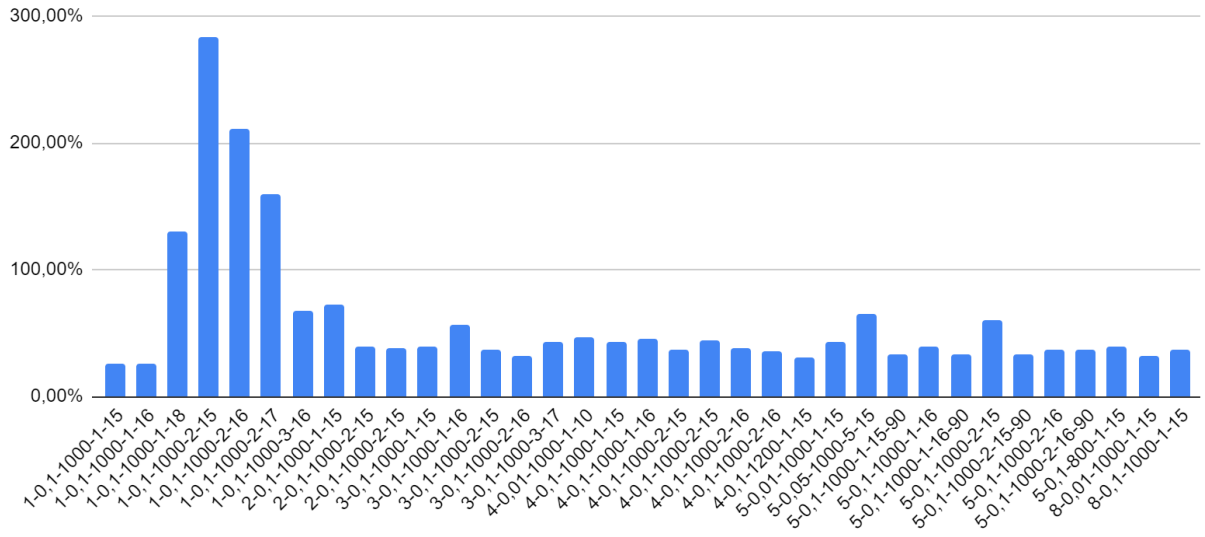

Ilustración 4. Media de errores cometidos por todos los modelos generados para el caso de España.

Tras un procedimiento orientado a prueba y error, hemos obtenido finalmente un modelo cuyo valor de los errores resultan ser los más bajos en relación con los demás modelos. Los valores obtenidos para este modelo están reflejados a continuación.

	9/9	9/10	9/11	9/12	9/13	9/14	9/15	9/16	9/17	9/18	9/19	9/20	9/21	9/22	9/23
Real Confirmados	4.763	4.440	0	0	7.804	3.261	3.723	4.075	3.222	0	0	5.988	2.450	2.840	3.031
Real Muertos	71	72	0	0	103	155	90	101	44	0	0	118	82	102	100
4-.1-1000-1-15	3.998	3.295	-404	150	6.483	3.521	3.787	3.165	1.495	274	1.214	5.172	2.438	3.603	2.712
	95	82	-74	-91	53	84	71	85	73	-74	-85	47	77	74	77

Tabla 4. Resultados obtenidos para cada modelo en el caso de España.

Para obtener este modelo, hemos utilizado los siguientes parámetros:

- 4 capas ocultas
- Tasa de aprendizaje 0,1
- 1000 ciclos
- Mínimo lag 1
- Máximo lag 15

Como podemos observar en la tabla, los fines de semana no se proporcionaron datos en España, y por ello, tanto los datos para confirmados como para muertos tienen un valor de 0 en esas fechas. Ese suceso hace que las predicciones empeoren ligeramente, puesto que son sucesos que alteraciones negativas y bruscas en la regularidad de los datos. Estos se pueden ver los días 11 y 12 de septiembre, así como el 18 y 19 del mismo mes.
Para el caso de los casos confirmados, podemos ver que nuestro modelo predice de una manera más que correcta los valores para esos días. Para el resto de los días, el modelo obtenido se ajusta también muy bien a los valores obtenidos por los datos reales, generando dos gráficas casi idénticas y obteniendo un error del **17,81 %**.

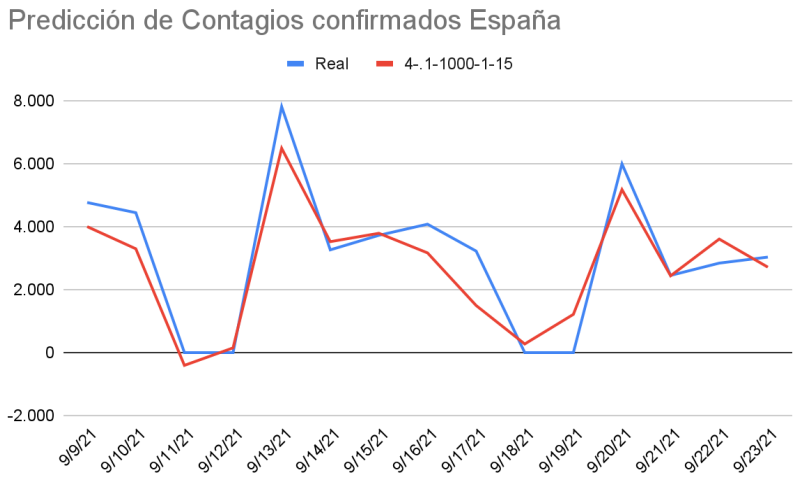

Ilustración 5. Comparativa entre los valores de contagios predichos por el mejor modelo obtenido y los valores reales para el caso de España

No ocurre lo mismo en el caso de los fallecidos. Aquí, a pesar de que nuestro modelo predice suficientemente bien los fallecidos para los próximos 15 días, tiene un porcentaje de error del **32,87 %**, lo que hace que el error medio para este modelo suba desde un que habíamos obtenido con los casos confirmados a un **25,34 %** total. Este resultado es más que aceptable y por ello hemos decidido quedarnos con este modelo.

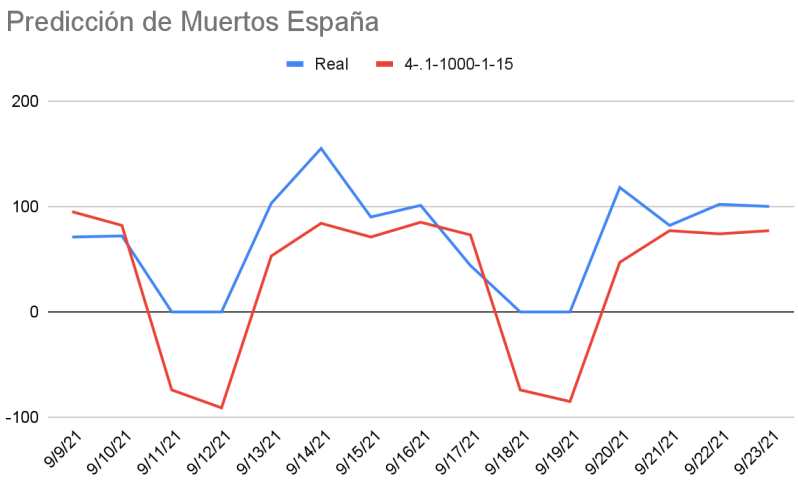

Ilustración 6. Comparativa entre los valores de fallecidos predichos por el mejor modelo obtenido y los valores reales en el caso de España.

Colombia

En cuanto a la predicción de casos confirmados y muertes en Colombia, se han realizado varias pruebas, de la misma manera que con España, obteniendo finalmente el mejor modelo de entre todos ellos. Los modelos que se han considerado al igual que en la parte anterior se muestran en la siguiente gráfica:

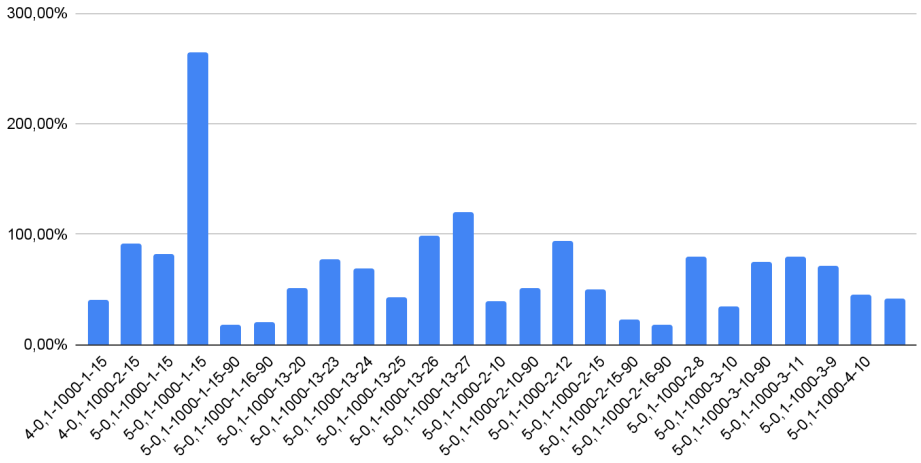

Ilustración 7. Comparativa entre la media de los errores cometidos entre todos los modelos generados para el caso de Colombia

Se puede ver que el modelo que de media da mejor resultados en las predicciones es el que utiliza los siguientes parámetros:

- 5 capas ocultas
- Tasa de aprendizaje 0,1
- 1000 ciclos
- Mínimo lag 1
- Máximo lag 15
- 90 días

Las predicciones realizadas por este modelo, junto a los datos reales se pueden ver a continuación:

	9/9	9/10	9/11	9/12	9/13	9/14	9/15	9/16	9/17	9/18	9/19	9/20	9/21	9/22	9/23
Real Confirmados	1.803	1.772	1.806	1.671	1.314	1.435	1.570	1.484	1.544	1.655	1.813	1.185	1.373	1.581	1.608
Real Muertos	53	49	63	55	40	26	40	29	44	34	35	29	38	44	26
5-0,1-1000-1-15-90	1.823	1.849	1.373	1.273	1.181	1.124	1.177	1.184	1.103	1.233	1.475	1.518	1.740	1.527	1.273
	64	49	49	44	35	38	38	42	39	32	30	29	31	31	26

Tabla 5. Resultados obtenidos con el mejor modelo seleccionado para el caso de Colombia.

El porcentaje de error del número de confirmados y el número de muertes resulta ser bastante bajo, sin embargo, si echamos un vistazo a las gráficas que se encuentran debajo, podemos observar que en ciertos días la diferencia entre ellos aumenta anormalmente debido a una predicción poco precisa o acertada.

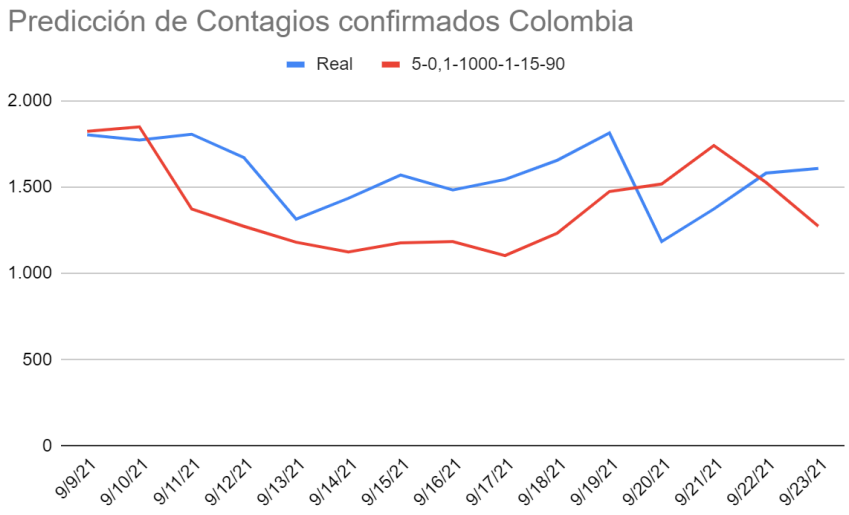

Ilustración 8. Comparativa entre los valores de contagios predichos y los valores reales para el caso de Colombia

En la gráfica de contagios confirmados en Colombia, hay días en los que la brecha entre datos reales y predicción no resulta muy prometedora. Sin embargo, dado que el porcentaje de error que conlleva este modelo para esta gráfica resulta ser de tan solo **18,8 %**, lo cual es bastante atractivo comparado con los demás modelos en los cuales obteníamos unos porcentajes muy elevados, por lo que prescindimos de ellos.

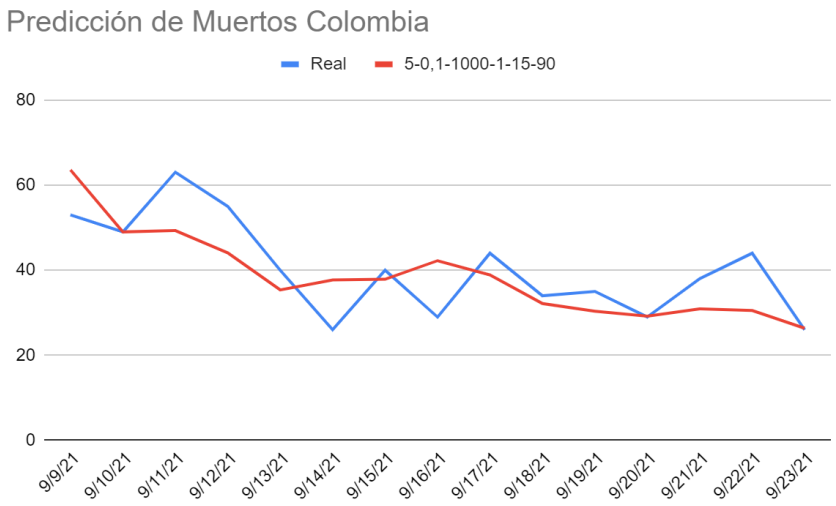

Ilustración 9. Comparativa entre los valores de fallecidos predichos y los valores reales para el caso de Colombia.

En el caso de la gráfica de muertes en Colombia, el recorrido de las dos rectas son casi las mismas, exceptuando los días en los que se producen los picos máximos y mínimos de la predicción, mientras que para los datos reales estos picos resultan ser más suaves. Para esta gráfica, el porcentaje de error obtenido es de tan solo **16.72 %**, esto es bastante positivo.

El porcentaje medio obtenido en este modelo ha sido de **17,76 %**, siendo este el más bajo que hemos encontrado entre todos los modelos realizados, lo cual resulta bastante decente, dado que en algunos modelos hemos obtenido unos porcentajes medio poco decentes y un tanto alarmantes.

Conclusiones

En definitiva, en esta práctica nos hemos podido hacer idea, aunque sea de manera reducida, lo que supone encontrar un buen modelo que se adapte a los datos de los que se dispone. Si bien las redes neuronales funcionan muy bien en la actualidad, hemos podido ver de primera mano lo difícil que es encontrar un modelo que generalizara mejor o que encontrara las predicciones óptimas. En total, hemos generado más de 60 modelos aproximadamente, lo que puede dar una idea de la búsqueda exhaustiva de modelos cada vez mejores.

Nos ha resultado bastante atractivo el ser capaces de obtener modelos que predigan el número de contagios confirmados y de muertes de un país en concreto y ver los errores que comete dicho modelo tras realizar una comparación con los datos reales.

Finalmente, a pesar de que todos habíamos estado familiarizados con Weka, la utilización de otros plugins o características externas nos ha enriquecido considerablemente. Lo cierto es que las series temporales eran relativamente desconocidas por la mayoría de nosotros, por lo que al principio nos costó un tiempo entender qué era lo que estábamos haciendo y qué era lo que queríamos conseguir. Pero al final conseguimos comprender la finalidad de estas tareas.

Referencias

[1] The Humanitarian Data Exchange [en línea] Acceso: 29/09/2021. https://data.humdata.org/

[2] Universidad Johns Hopkins [en línea] https://www.jhu.edu/

[3] Sousa, G. J. B., Garces, T. S., Cestari, V. R. F., Moreira, T. M. M., Florêncio, R. S., & Pereira, M. L. D. (2020). Estimación y predicción de casos de COVID-19 en metrópolis brasileñas. Revista Latino-Americana de Enfermagem, 28.

de Moraes Batista, A. F., Miraglia, J. L., Donato, T. H. R., & Chiavegatto Filho, A. D. P. (2020). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. *medRxiv*.

[5] Google Académico [en línea] Acceso: 07/10/2021. https://scholar.google.es/schhp?hl=es