UNIVERSIDAD CARLOS III

INTELIGENCIA ARTIFICIAL EN LAS ORGANIZACIONES

2023-2024 GRADO EN INGENIERÍA INFORMÁTICA

APLICACIÓN DE RNA

GRUPO 81

Diego Caballero García-Alcaide NIA:100451177

100451177@alumnos.uc3m.es

Diego Calvo Engelmo NIA:100451091

100451091@alumnos.uc3m.es

Álvaro Bernal Torregrosa NIA:100451179

100451179@alumnos.uc3m.es

Raúl Ágreda García NIA:100451269

100451269@alumnos.uc3m.es

ÍNDICE

PARTE I: REGRESIÓN

1. Preprocesamiento de datos

A continuación vamos a explicar los pasos llevados a cabo durante el preproceso de los datos para prepararlos para entrenar los modelos. Los pasos que hemos llevado a cabo son:

- Filtrado de atributos
- Agrupamiento de datos por países
- Cambio de contagios acumulados a contagios diarios
- Corrección de datos negativos
- Análisis de características
- Selección de países y días

Filtro de atributos

Como primer paso del preproceso, hemos analizado los atributos de los datos recibidos. Tras esto, hemos visto que entre ellos, se encontraban los atributos de latitud y longitud. Analizando estos valores, y dado que disponemos de las columnas "país" y "provincia", consideramos que estos dos atributos son irrelevantes.

Esto lo fundamentamos en el hecho de que el crecimiento de los casos de covid va a estar más relacionado con el país a analizar más que por la posición geográfica de dicho país, y en concreto, de su capital. Esto es así ya que depende principalmente de la cantidad de gente de un país, sus políticas y costumbres, lo que determinará cómo evolucionarán los casos. Por ejemplo, si analizamos una región del sur de Francia, seguramente su posición geográfica sea más cercana a la capital de Andorra, pero la evolución de los casos se asemeje más a la evolución francesa. Es por esto que hemos decidido eliminar dicho atributo.

Por otro lado, dado que cuando recibamos unos datos a predecir, lo que necesitamos será saber cuántos días previos al nuevo dia disponemos para predecir los datos, hemos decidido reetiquetar las columnas correspondientes a los días, comenzando desde el penúltimo día como el día -1, hasta el primer día de los datos, en concreto el día -1142. A su vez, el último día de los datos se corresponderá con la "clase" a predecir por el modelo, que en este caso la necesitamos para entrenarlo. De este modo, el modelo sabrá entender en base a -x días previos como predecir los siguientes días.

Agrupamiento de datos por países

En segundo lugar, hemos analizado los atributos de país y provincia; dándonos cuenta que para cada país, podríamos encontrarnos múltiples provincias o regiones asociadas a este. Analizando los valores de dichas regiones, llegamos a la conclusión de que en la mayoría de los casos, se trataban de regiones marginales, islas, o eventos que estaban clasificados como pertenecientes a cierto país.

Sin embargo, dada la naturaleza de los datos a estudiar, la evolución de los contagios, que dichas regiones se clasifiquen como de un país no quiere decir que su evolución vaya a ser similar, ya que esto vendrá determinado más por la distancia geográfica que por la distribución política de regiones. Es por esto, que hemos tomado la decisión de eliminar aquellas provincias que no pertenezcan a la misma región que el país principal, y fusionando en una sola fila aquellas provincias colindantes. En concreto, los cambios que hemos realizado son los siguientes:

- **Canadá:** Al igual que algunos países, Canadá estaba dividida por regiones y había tres filas que se referían a situaciones especiales que se dieron por la pandemia. Estas tres son "Diamond Princess", "Grand Princess" y "Repatriated Travellers". Los dos primeros se refieren a cruceros que tuvieron una gran incidencia durante los primeros meses de la pandemia, y la última son de datos de viajeros repatriados a sus países. Creemos que estas 3 filas de datos son irrelevantes por lo que hemos decidido borrarlas y las demás las juntamos en 1 fila, haciendo la suma de los contagios para cada uno de los días.
- **China**: China estaba dividida por todas las regiones que conforman China continental, decidimos incluir todas incluida la fila “Unknown”.
- **Dinamarca, Francia, Países Bajos, Nueva Zelanda y Reino Unido:** Estos países están divididos por sus territorios en ultramar y el territorio continental. Debido a que los territorios de ultramar tienen datos muy pequeños que son insignificantes para nuestro modelo, hemos decidido mantener solo el continental.
- **Filas especiales:** Además de países, el archivo contenía datos especiales de las olimpiadas de verano e invierno y del crucero “MS Zaandam”. Como el período de estos eventos es muy reducido y los casos son insignificantes, hemos decidido eliminarlos.

Cambio de contagios acumulados a contagios diarios

Hemos decidido usar los contagios diarios en lugar de los acumulados ya que, además de que el objetivo es que el modelo prediga el número de contagios del día siguiente, creemos que puede aportar varias ventajas:

- Este cambio hace que los datos puedan ser interpretados mejor.
- La red neuronal puede ser más sensible a los cambios lo que nos puede dar una mejor precisión.
- Durante las primeras semanas de la pandemia los casos crecían exponencialmente lo que podía hacer que el modelo se fijará más en estos cambios que en otros más sutiles.

Corrección de datos negativos

Tras haber transformado los datos en el paso anterior, nos hemos dado cuenta de que en ciertos días, el incremento de contagios es negativo. Esto puede deberse entre otras cosas al error humano, si días previos se registraron más casos de los correctos. Por tanto, debemos considerar estos datos como errores y tomar una decisión para solucionarlo.

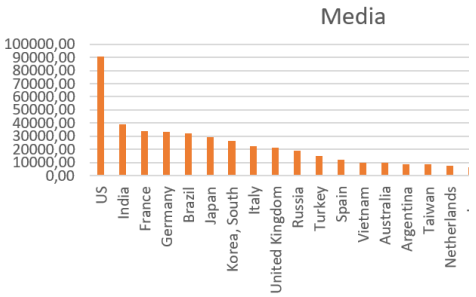
Una de las opciones que nos planteamos es propagar este decremento de casos hacia delante o hacia atrás. Es decir, establecer ese día como 0 casos, e ir restando a días posteriores o anteriores el decremento de casos correspondiente al día erróneo. Sin embargo, esta solución nos supone un problema, debido a que en múltiples casos, la propagación del decremento supondría arrastrar dicho valor negativo muchos días. Esto eliminaría los casos de estos días y por tanto también se vería perjudicado el modelo, ya que perderíamos la información de evolución de estos días.

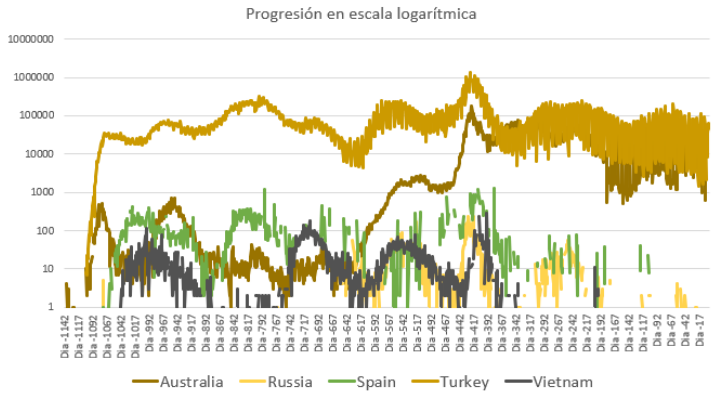
Por tanto, para poder preservar dicha información, y dado que lo que nos interesa es más la evolución de datos diarios que el cómputo global de casos totales, hemos tomado la decisión de establecer esos valores negativos a 0. De este modo, solucionamos el error y no afectamos a la información de evolución de los casos que aportan los días adyacentes al día erróneo.

Análisis de características

Como se indica en el enunciado de la práctica, tras haber realizado el análisis y la limpieza de los datos y atributos hemos decidido realizar el análisis de sus características. Tomando en primer lugar la media y la varianza.

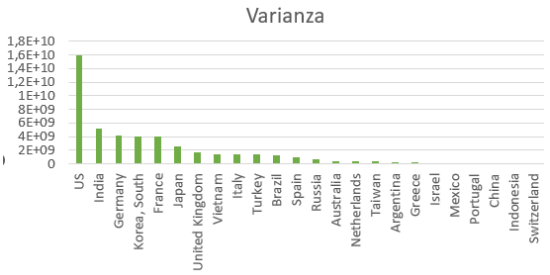
Tras haber realizado la media del conjunto de países, hemos buscado aquellos países que tenían una media similar a España, ya que es sobre este país sobre el que vamos a generar el modelo de predicción. Cuando hemos hecho esto, hemos visto que los países que tienen una media más cercana a España son Rusia, Turquía, Vietnam y Australia. Sin embargo, cuando comparamos los datos de estos cuatro países con España, vemos como no se corresponden con el comportamiento de España. Esto lo podemos apreciar en la gráfica "Progresión en escala logarítmica"; en la cual convertimos los datos usando escala logarítmica para eliminar las diferencias en cuanto a cantidad total de contagios (que puede variar entre países) y observar simplemente la evolución en cuanto a crecimiento o decrecimiento de los datos diarios. Por tanto, consideramos esta medida poco representativa.

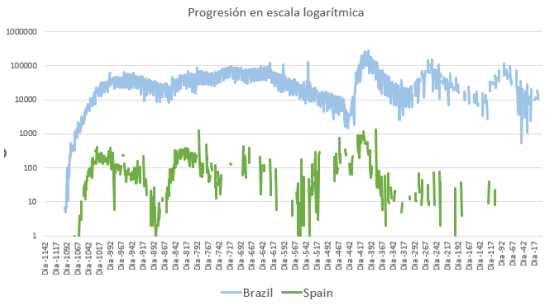




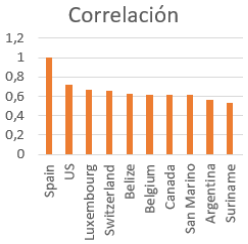
En segundo lugar, analizamos la varianza en la evolución de los casos, observamos que los países más similares a España son Turquía, Brasil, Rusia y Australia. Ya sabemos que Turquía,

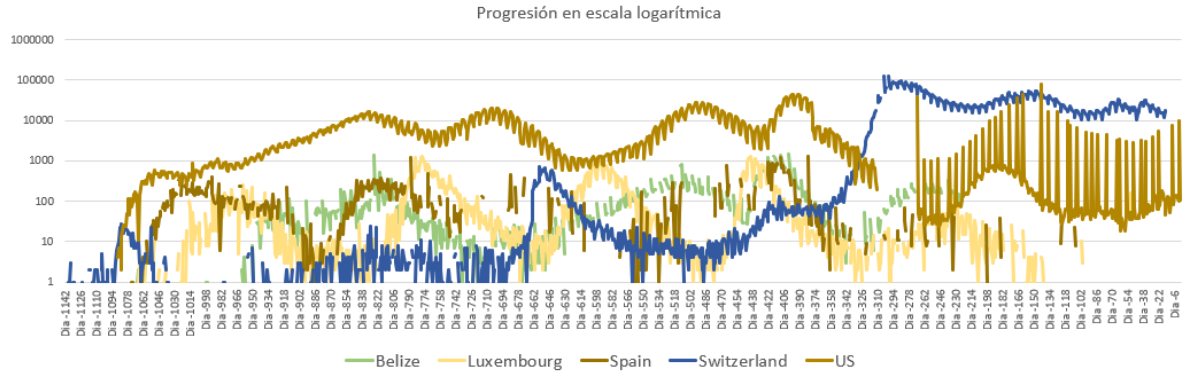
Rusia y Australia no comparten similitudes con España, y si analizamos Brasil, podemos ver que ocurre de nuevo lo mismo.





En tercer lugar, hemos decidido analizar el coeficiente de correlación de Pearson entre los países, ya que este indicador cuantifica la relación lineal entre dos conjuntos de datos (países); y por tanto si dos países tienen un coeficiente cercano a 1, implica que son similares en términos de evolución. Tras obtener dicho coeficiente, vemos que los cuatro países con mayor correlación con España son US, Luxemburgo, Suiza y Belice. Si analizamos como en casos anteriores la progresión en escala logarítmica, vemos como los valores de las distintas series están más cerca, y sobre todo en la primera mitad podemos ver países con alta similitud a la evolución de España como Luxemburgo o Belice.





Por tanto, consideraremos este indicador como el principal a la hora de determinar qué subconjunto de países vamos a utilizar para entrenar nuestro modelo.

Selección de países

Como hemos comentado en el apartado anterior, nos hemos basado en el coeficiente de correlación de Pearson entre España y el resto de países para determinar qué subconjunto seleccionar. En base a esto, hemos tomado la decisión de tomar el subconjunto de los 30 mejores países en cuanto a este indicador. Consideramos que dicha cantidad es suficiente para poder entrenar el modelo y disponer de instancias suficientes; y eliminando de este modo al otro conjunto de países que no están tan relacionados con el comportamiento de España.

Si por el contrario hubiésemos escogido todo el subconjunto; esto podría provocar que el modelo aprendiese patrones que no se corresponden con el comportamiento de España, y por tanto este modelo sería de peor calidad. Por ello, los países que hemos seleccionado para entrenar el modelo son:

Andorra	France	Norway
Argentina	Hungary	Portugal
Belgium	Iceland	San Marino
Belize	Ireland	Seychelles
Bolivia	Israel	Spain
Bosnia and Herzegovina	Italy	Suriname
Bulgaria	Kuwait	Switzerland
Canada	Lebanon	US
Denmark	Luxembourg	United Kingdom
Dominican Republic	Montenegro	Uruguay

Por tanto, finalizamos el preprocesamiento de datos con estos 30 países, que utilizaremos como entrada para poder entrenar nuestro modelo de predicción.

---

2. Entrenamiento del modelo

En primer lugar, realizamos numerosas pruebas para la selección de atributos. Tras probar diferentes combinaciones en el rango de días seleccionados para entrenar al modelo sacamos las siguientes conclusiones:

- Seleccionar un rango muy amplio de días provoca que el modelo aprenda patrones que son irrelevantes para la predicción de los contagios del día siguiente. Es decir, entrenar al modelo con días de hace 1 o 2 años no influye positivamente en la predicción de contagios que habrá mañana, sino al revés, debido a la lejanía de tiempo. En especial teniendo en cuenta que la evolución de contagios en los países seleccionados es tan cambiante, como hemos visto en las gráficas del apartado anterior.
- Los rangos de días donde más parecidas son las evoluciones de los países y donde más datos significativos hay, son las que obtienen un mejor rendimiento del modelo.

Por tanto, visualizando las gráficas de contagios de los países seleccionados y la gráfica ya mencionada de correlación en escala logarítmica, hemos realizado pruebas con los periodos que más se adaptan a estas condiciones.

Finalmente, obtuvimos que el rango de días con el que se obtenía un mejor rendimiento de los modelos era del **día -851 al día -751**, siendo el día -751 la clase o etiqueta.

Cabe destacar que hemos elegido como clase el día -751 porque analizando los datos, hemos visualizamos que el tercer día posterior al -750 tenía valor 0. Por tanto, como nuestro objetivo es predecir los 3 días siguientes, cogimos un día antes para poder comparar mejor la predicción de ese tercer día cuando utilicemos el modelo final elegido.

En resumen, nuestra selección de atributos para los modelos consiste en:

- 100 atributos de entrada (99 días anteriores al día de predicción y el atributo 'Países')
- 1 atributo de salida o clase (Día siguiente a los días utilizados como valores de entrada).

Una vez seleccionados los atributos probamos distintas arquitecturas para dar con el modelo que mejor se ajustase a nuestros datos de entrenamiento. Para ello generamos los siguientes modelos:

Modelo	Epochs	Nº Capas ocultas	Tamaño capas ocultas	Función de activación
DeepLearning-01	10	2	50, 50	Rectifier
DeepLearning-02	100	2	50, 50	Rectifier
DeepLearning-03	1	2	50, 50	Rectifier
DeepLearning-04	20	4	50, 50, 50, 50	Rectifier
DeepLearning-05	20	2	80, 50	Rectifier
DeepLearning-06	20	5	80, 60, 40, 20, 10	Rectifier
DeepLearning-07	20	1	50	Rectifier
DeepLearning-08	10	2	50, 50	Tanh
DeepLearning-09	20	4	50, 50, 50, 50	Tanh
DeepLearning-10	20	5	80, 60, 40, 20, 10	Tanh

Por último, mencionar que todas las pruebas han sido realizadas con validación cruzada de **3 folds** o particiones puesto que al haber seleccionado 30 países, cada fold está formado por 10 países, resultando en 2 particiones para entrenar y 1 para test, es decir, se utilizan 20 países para entrenar los modelos y 10 países para evaluarlo (proporción 1/3 test, 2/3 train). Con esto conseguimos que los resultados sean lo más realista posible en comparación con el rendimiento que tendrá el modelo final.

Cabe destacar que hemos utilizado una local **random seed** para todos los procesos de validación cruzada para evitar obtener resultados muy distintos en cada ejecución del proceso. Con esta semilla lo que se consigue es que las particiones o folds generados para la validación

cruzada no se vuelvan a generar aleatoriamente en cada ejecución, sino que sean siempre los mismos, y por ende, que los resultados del rendimiento de los modelos sean muy parecidos cada vez que se ejecuta el proceso 'modelos.rpm'.

---

3. Análisis de resultados

En este apartado vamos a recoger los resultados obtenidos tras el entrenamiento de los modelos anteriores. Para ello, hemos utilizado como medidas de calidad la correlación y el RMSE; de modo que cuanto mayor índice de correlación y menor RMSE obtenga un modelo, mejor será la calidad de este. Por otro lado, dado que hemos utilizado validación cruzada, nos aseguramos que los datos de test no se comparten con los datos de entrenamiento y no se produce '*data leakage*' ni overfitting en el modelo.

A continuación, representamos en la siguiente tabla los resultados obtenidos:

Modelo	Correlación	RMSE
DeepLearning-01	0.869 +/- 0.078	6269.766 +/- 6752.072
DeepLearning-02	0.922 +/- 0.053	4546.469 +/- 2404.519
DeepLearning-03	0.615 +/- 0.533	6529.364 +/- 2183.470
DeepLearning-04	0.971 +/- 0.013	3133.206 +/- 1791.915
DeepLearning-05	0.878 +/- 0.141	3527.377 +/- 2146.881
DeepLearning-06	0.972 +/- 0.011	4181.412 +/- 3186.275
DeepLearning-07	0.775 +/- 0.304	6105.130 +/- 6307.189
DeepLearning-08	0.862 +/- 0.181	8345.932 +/- 7306.849
DeepLearning-09	0.879 +/- 0.077	7682.153 +/- 7268.834
DeepLearning-10	0.888 +/- 0.103	9382.847 +/- 6065.197

Tras analizar los modelos, podemos sacar las siguientes conclusiones. En primer lugar, vemos cómo aumentar el número de epochs (ciclos que se reentrena el modelo con el conjunto de datos) por encima de 20 no supone mucho beneficio; de modo que manteniendo este valor nos aseguramos que no se produzca overfitting en el modelo y reducimos el tiempo de entrenamiento. No obstante, poner un valor muy bajo como en el caso del modelo 'DeepLearning-03' tampoco es adecuado, ya que el modelo no tiene opción de reajustar los pesos y por tanto el resultado obtenido no es óptimo.

En segundo lugar, si analizamos los resultados tras modificar el número de capas ocultas y el tamaño de dichas capas, vemos como los resultados no varían significativamente. Sí es cierto que los resultados de los modelos con 4 o 5 capas mejoran con respecto a los de únicamente 2 capas, por lo que podemos concluir que hacer la red más profunda, es decir, añadir más capas, puede beneficiar al rendimiento de los modelos, siempre y cuando se tenga en cuenta el no añadir demasiadas, con tal de evitar el overfitting. Sin embargo, al no trabajar con un gran conjunto de datos (solo 30 países/instancias), los modelos no se ven tan afectados por el efecto de estos parámetros. No obstante, cabe destacar que casos extremos como 'DeepLearning-07', donde hemos establecido el número de capas a 1, si que vemos un deterioro considerable en comparación al resto de modelos.

En tercer lugar, si observamos la función de activación de los distintos modelos, podemos ver como 'Rectifier' funciona mejor que 'Tanh', aportando mejores resultados en todos los casos. Esto puede deberse a que 'Tanh' funciona generalmente mejor para casos en los que las salidas de la red estén centradas alrededor del cero; por lo que nuestros datos no se ajustan a estas condiciones.

Por todo ello, en base a lo explicado anteriormente, hemos obtenido que los mejores modelos en cuanto a la correlación y el RMSE son 'DeepLearning-04' y 'DeepLearning-06'. Para determinar qué modelo elegir, hemos optado por escoger aquel modelo que obtiene mejores valores en conjunto; y por tanto, dado que el modelo 04 obtiene bastante menos RMSE que el 06, ha sido el modelo que hemos seleccionado para quedarnos como modelo final. Este modelo lo hemos exportado y lo usaremos en el siguiente apartado para predecir los datos de los dos países correspondientes.

4. Predicción de valores futuros

Tras haber seleccionado el modelo 'DeepLearning-04' en el apartado anterior, vamos a utilizarlo para predecir el valor de tres días siguientes en España, y tomaremos como segundo país Reino Unido. Hemos seleccionado este país de manera arbitraria de entre los países seleccionados para entrenar el modelo, los cuales deberían seguir un comportamiento similar al de España. A continuación se muestra una tabla con los resultados obtenidos.

País	Tipo de valor	Día 1 (-750)	Día 2 (-749)	Día 3 (-748)
España	Predicción	8480	13480	3346
	Valor real	10829	14515	11435
Reino Unido	Predicción	11727	13920	11171
	Valor real	127171	12057	12027

Como podemos ver, si nos centramos en España en primer lugar, podemos observar como los dos primeros días se predicen bastante bien, sobre todo el segundo, pero vemos una grán diferencia en el tercer día entre el valor predicho y el real. Esto nos ha extrañado en un primer momento, pero tras analizar los resultados, creemos que esto se debe a la naturaleza de los datos.

España es un país que en los fines de semana tiene 0 casos de covid registrados, ya que esos días no se reportaban datos. Esto afecta negativamente al modelo, ya que en vez de entender una progresión constante de los casos, cada cierto periodo (5 días) se produce una bajada drástica de los mismos. Por ello, consideramos que el modelo al intentar predecir el tercer día, consideró que se encontraba dentro de un periodo de descenso de casos, y por tanto predice un valor menor que el real.

No obstante, si vemos los datos obtenidos con Reino Unido, podemos ver cómo los valores obtenidos se ajustan más a los valores reales. Esto es indicativo de que el modelo, aunque no es perfecto, tiene un comportamiento aceptable y que no solo es capaz de predecir valores del país objetivo, España, sino que también puede predecir los de países con comportamientos ligeramente diferentes. También, cabe destacar que, probablemente los datos de Reino Unido se ajustan mejor a los valores reales, debido a que hay una mayor cantidad de datos de calidad con respecto a los datos con los que se ha entrenado España.

5. Conclusiones

Tras realizar esta primera parte de la práctica, hemos llegado a las siguientes conclusiones:

- **Los datos de entrenamiento son muy importantes**. Hemos visto durante el desarrollo de la práctica que los datos que se van a utilizar para entrenar un modelo de red de neuronas tienen gran relevancia para el proceso; ya que son el punto de partida de todo el proceso y que, sin datos de calidad, es difícil obtener un modelo que sea capaz de predecir de forma adecuada valores desconocidos.
- **La importancia del preprocesado**. En datasets muy grandes, como los que podemos obtener en la actualidad gracias a la expansión de Internet y las nuevas tecnologías, los tiempos de procesamiento de algoritmos de aprendizaje automático pueden ser muy elevados. Además, no siempre más datos implican mejores resultados, ya que estos datos pueden contener información que no queremos que el modelo aprenda. En este caso, no nos es útil que el modelo aprenda cómo se comportan ciertos países que nada tienen que ver con cómo se comporta la evolución del Covid en España. Es por esto que la parte de análisis y procesado de datos adquiere un papel fundamental en el entrenamiento de este y otros tipos de modelos.

PARTE II: SERIES TEMPORALES

1. Preparación de los datos

Para comenzar con el preproceso de los datos, obtuvimos el dataset de datos de las vacunas de cada país; del cual nos quedamos únicamente con las columnas de 'date', 'location', y 'new_vaccinations'. Este último atributo contiene los datos de las vacunas aplicadas diariamente, justo lo que necesitamos para entrenar nuestro modelo.

Tras esto, observamos que había *missing values* en algunas celdas, quedando el número de vacunas aplicadas como vacío, por lo que rellenamos esos valores con 0 para evitar posibles errores futuros. Con los datos ya correctos, extrajimos de cada país el conjunto de datos pertenecientes a cada uno, los cuales trasladamos a otro fichero para poder trabajar con ellos de forma individual junto con los datos de los casos diarios.

Al hacer esto, y juntar los datos de los casos con las vacunas, vimos cómo las fechas de ambos conjuntos de datos no cuadran. Por ello, tuvimos que eliminar los datos anteriores al 22/01/20 y posteriores al 09/03/23; quedándonos así con el subconjunto de fechas que tenían datos de ambos atributos.

En último lugar, dado que el conjunto de datos es bastante grande y observamos que había muchas fechas en la parte inicial y final que carecían de vacunas. Por ello, para reducir el tamaño del conjunto de datos y quedarnos con las fechas que tienen información relevante para el entrenamiento del modelo, hemos seleccionados los siguientes periodos para ambos países:

- Inicio → 15/03/2021
- Fin → 30/11/2021

Por ello, utilizaremos los días 1, 2 y 3 de diciembre para predecir y evaluar el modelo. La selección de este periodo se ha hecho en base a dos los siguientes objetivos:

- Obtener un rango de días en el que el modelo pueda observar tanto una disminución como un aumento de los casos.
- Un periodo significativo en la administración de vacunas.

---

2. Entrenamiento del modelo

A continuación, vamos a realizar el entrenamiento de los 10 modelos para cada uno de los países. Para ello, planteamos los mismos modelos con las diferentes combinaciones de parámetros, y escogeremos para cada país la configuración de modelo que mejor se adapte a cada uno.

Los diez modelos generados son los siguientes:

Modelo	Tamaño de ventana	Training cycles	Learning Rate	Momentum
NeuralNet-01	10	200	0.01	0.9
NerualNet-02	10	200	0.01	0.5
NeuralNet-03	10	200	0.01	0.01
NeuralNet-04	20	200	0.01	0.9
NeuralNet-05	20	100	0.01	0.9
NeuralNet-06	20	500	0.01	0.9
NeuralNet-07	20	200	0.005	0.9
NeuralNet-08	30	200	0.01	0.9
NeuralNet-09	30	200	0.02	0.9
NeuralNet-10	30	200	0.001	0.9

Estos modelos han partido de la modificación del tamaño de ventana, y hemos ido probando para cada tamaño distintas variaciones de parámetros. Estos modelos serán utilizados en el siguiente paso para predecir tanto los datos de España como los datos de Reino Unido.

---

3. Análisis de resultados

Tras haber ejecutado los modelos anteriores con España, hemos obtenido los siguientes resultados:

Modelo	Correlación	RMSE
NeuralNet-01	0.670 +/- 0.396	4197.832 +/- 4386.686
NerualNet-02	0.603 +/- 0.413	4623.136 +/- 4291.353
NeuralNet-03	0.596 +/- 0.413	4692.446 +/- 4397.464
NeuralNet-04	0.709 +/- 0.394	3486.785 +/- 3985.442
NeuralNet-05	0.801 +/- 0.222	4111.896 +/- 4235.224
NeuralNet-06	0.798 +/- 0.223	4112.609 +/- 4264.897

NeuralNet-07	0.801 +/- 0.222	4078.241 +/- 4171.349
NeuralNet-08	0.761 +/- 0.232	5096.500 +/- 5636.561
NeuralNet-09	0.764 +/- 0.229	5270.299 +/- 6241.283
NeuralNet-10	0.816 +/- 0.208	4711.667 +/- 4625.695

Como podemos ver, obtenemos a simple vista unos valores algo peores a los obtenidos en la Parte I de la práctica. Esto, en primera instancia puede deberse a que en este caso, utilizamos el modelo Neural Net en vez del modelo de DeepLearning. Además, en la parte I, utilizamos los datos de múltiples países, los cuales aportan cierta información sobre el comportamiento de la pandemia que puede ser significativa para el modelo. Ante estos resultados, viendo que las correlaciones eran similares pero que, en concreto, el modelo 'NeuralNet-04' tenía un RMSE menor al resto; hemos considerado este modelo como el mejor para predecir los datos de España.

Tras esto, hemos entrenado los mismos modelos con Reino Unido, obteniendo los siguientes valores:

Modelo	Correlación	RMSE
NeuralNet-01	0.522 +/- 0.406	3429.787 +/- 2910.338
NerualNet-02	0.514 +/- 0.419	3274.277 +/- 2633.921
NeuralNet-03	0.521 +/- 0.408	3274.964 +/- 2608.649
NeuralNet-04	0.547 +/- 0.399	3441.909 +/- 3357.464
NeuralNet-05	0.490 +/- 0.338	4835.476 +/- 4753.683
NeuralNet-06	0.482 +/- 0.340	4898.141 +/- 4747.166
NeuralNet-07	0.486 +/- 0.338	4848.441 +/- 4724.942
NeuralNet-08	0.430 +/- 0.345	6119.041 +/- 5408.764
NeuralNet-09	0.384 +/- 0.297	6887.410 +/- 5529.317
NeuralNet-10	0.478 +/- 0.336	5601.804 +/- 5058.889

Tras obtener estos resultados, vemos que hemos obtenido una correlación bastante baja, en comparación con el caso de España. No obstante, si nos fijamos en el RMSE obtenido por los modelos, vemos como los valores no se alejan tanto a los obtenidos con el país anterior. Por tanto, al igual que en caso anterior, hemos seleccionado el modelo con menor RMSE; en concreto 'NeuralNet-03', ya que tiene RMSE similar al modelo 02, pero menor varianza del RMSE.

4. Predicciones de valores futuros

Tras haber determinado los modelos anteriores para ambos países, hemos usado los modelos seleccionados para predecir los valores de los tres días siguientes. Cabe destacar que, dado que los modelos solo predicen los casos futuros, hemos tenido que rellenar a mano los datos de las vacunas de cada día que se iba prediciendo, para que el modelo pudiese usarlos como información para predecir los días posteriores. Tras haber ejecutado este proceso, hemos obtenido las siguientes predicciones:

País	Tipo de valor	Día 1	Día 2	Día 3
España	Predicción	7578	9661	11461
	Valor real	10536	14500	13738
Reino Unido	Predicción	41740	42856	43570
	Valor real	47235	53067	50573

Si analizamos en primer lugar los valores de España, podemos ver cómo a pesar de que los dos primeros días tienen una predicción peor que en el modelo propuesto en la primera parte, vemos como el tercer día se acerca más al valor real. Creemos que esto es debido a la naturaleza del modelo de serie temporal; que a diferencia del modelo anterior, no se ha visto tan afectado por los valores 0 en los casos de covid de los fin de semana.

Por otro lado, si analizamos lo obtenido con Reino Unido, vemos como aunque la diferencia entre la predicción y el valor real sea mayor que la de España, en proporción el error producido es algo menor, ya que en los datos seleccionados, UK tiene mayor magnitud de casos COVID. Por otro lado, si nos centramos solo en la evolución de los días, vemos como los valores se mantienen constantes y cerca de lo esperado. Esto es debido a que, a diferencia de España, Reino Unido reportó un registro de datos más constante, por lo que el modelo no se ve contaminado por los valores nulos y por tanto aporta resultados más consistentes.

5. Conclusiones

Tras haber obtenido los resultados anteriores, hemos podido extraer las siguientes conclusiones:

- **Relevancia de las vacunas diarias**. Tras haber obtenido los resultados de la parte II, y comparándolos con los obtenidos en la Parte I, consideramos que tener un registro de las vacunas no tiene un impacto realmente significativo a corto plazo. Dado el contexto del problema, como los contagios se producen antes que el reporte de los mismos, conocer que en un día se han vacunado x personas no será determinante para saber los casos que se reportará mañana, ya que estos sujetos lo habrán cogido previamente. Por ello, y dado que el modelo de series temporales se basa en el análisis de franjas de días previos, no muy grandes, para predecir los días siguientes; consideramos que no se ven

los resultados muy afectados por los valores de las vacunaciones; y por tanto, la calidad de los resultados no es significativamente mejor que los obtenidos en la parte I.

- **Calidad de los datos**: al igual que en la parte I, consideramos que la calidad de los datos de entrenamiento es de vital importancia para el desarrollo de un modelo de aprendizaje automático; debido a que la existencia de valores nulos o defectuosos pueden influir en el correcto aprendizaje del modelo y perjudicar la calidad del mismo.

INVESTIGACIÓN: Casos similares y noticias relacionadas con la práctica.

Para esta práctica hemos investigado sobre modelos que se han aplicado en la vida real antes, durante y después de la pandemia del COVID-19. Estos incluyen tanto aplicaciones de modelos de inteligencia artificial, como modelos matemáticos.

1. Mapa de riesgo de propagación de COVID-19 por contagio comunitario en España:

Este artículo describe un modelo matemático para predecir la propagación de la epidemia de COVID-19 en España. consiste en una versión adaptada de los modelos epidemiológicos en tiempo discreto y se ajusta específicamente a la dinámica de transmisión del virus SARS-CoV-2 en la población española, con el fin de estimar la tasa de riesgo en cada municipio en España, considerando tres factores principales:

1. Dinámica de transmisión: divide la población en diferentes grupos, como susceptibles, expuestos, asintomáticos, infectados, hospitalizados y recuperados. Basados en probabilidades derivadas de estudios epidemiológicos del COVID-19.
2. Movilidad: se registran los viajes laborales entre municipios y dentro de ellos. Esto ayuda a comprender cómo se propaga la infección en diferentes áreas y permite simular el impacto de restricciones de movilidad.
3. Demografía: la población se divide en tres grupos según la edad: jóvenes , adultos y mayores, ya que afecta de manera diferente a estos grupos.

Este modelo también cuenta con limitaciones como que no puede predecir la importación de casos internacionales y depende de parámetros epidemiológicos en tiempo real, los cuáles en la primera fase de la epidemia en España, donde la mayoría de casos eran importados, enfrentó dificultades debido a la falta de datos en tiempo real. Sin embargo, también tiene otras ventajas ya que cuenta con parámetros para actualizar los estudios epidemiológicos; la influencia del periodo asintomático; permite estimar un mapa de riesgos de nuevos casos, anticipándose a la propagación del virus por individuos asintomáticos; y las restricciones de movilidad masiva (cuarentena) pueden ser fácilmente introducidas, permitiendo obtener nuevos valores de riesgo bajo esas medidas. (Arenas & Gómez-Gardeñes, n.d.)

2. Inteligencia artificial para analizar las medidas adoptadas durante la pandemia

Cada país decidió gestionar las medidas restrictivas de manera individual dependiendo del número de casos que había en cada región. Estas medidas coartaban la libertad individual priorizando la salud global de los ciudadanos, pero esto hizo que hubiera muchas críticas sobre si estas medidas eran exageradas. La universidad de Oxford decidió recoger estas restricciones que estaban activas en 236 países o regiones del mundo, esta gran cantidad de datos ha llevado a que se pueda investigar si las restricciones ayudaban al control del virus y se llegó a la conclusión de que independientemente de las medidas adoptadas en las regiones o países si estas eran sostenidas en el tiempo daba lugar a que los casos siguieran el mismo ritmo de crecimiento y decrecimiento.

Es por ello que un grupo de investigación valenciano decidió crear un modelo predictivo y prescriptivo para el análisis de los casos y las posibles medidas a adoptar. Para la realización de este modelo se dio prioridad aquellas medidas que tenían un impacto significativo en el número de casos. De esta forma se podían detectar qué medidas eran más eficaces en cada momento específico. (Oliver & Conejero, 2022)

BIBLIOGRAFÍA

Arenas, A., & Gómez-Gardeñes, J. (n.d.). *Mapa de riesgo de propagación de COVID-19 por contagio comunitario en España*. #COVID-19RISK. https://covid-19-risk.github.io/map/spain/es/

Oliver, N., & Conejero, J. A. (2022, January 19). *Venciendo a la pandemia con inteligencia artificial*. The Conversation. Retrieved October 5, 2023, from https://theconversation.com/venciendo-a-la-pandemia-con-inteligencia-artificial-174732