PRÁCTICA 1
APLICACIÓN DE RNA

Inteligencia Artificial en las Organizaciones

Grado en Ingeniería Informática – Curso 2020/2021

Javier Cruz del Valle - 100383156

Gonzalo Fernández García – 100383212

Lucas González de Alba – 100383228

Índice

Introducción

Contexto

Las redes neuronales artificiales (RNA) pueden aplicarse contextos muy variopintos como reconocimiento de Imágenes, control de procesos, robótica, procesado del Lenguaje, diagnósticos médicos, análisis y procesado de datos o filtrado de ruido entre otros.

En nuestro caso se trata de análisis y procesado de datos mediante RNAs para predecir nuevos valores y estimar el comportamiento futuro de la serie temporal de contagios y fallecidos de la Covid19. Cabría preguntarse, ¿Se han planteado ya la utilización de RNAs en este contexto?

La respuesta es sí, este ejercicio de predicción es muy común y utilizado en diversas áreas por lo que muchos investigadores están intentando modelar el comportamiento de la Covid19 sobre pacientes y enfermos para reducir los vectores de contagio, aumentar las probabilidades de detección prematura. Para empezar, tenemos el caso de M. M. G. Lorenzo [1], que utiliza datos de expedientes médicos recogidos en Méjico para estimar mediante un perceptrón multicapa el grado de letalidad sobre enfermos por Covid19. De manera similar hemos encontrado otras publicaciones como la de Jairo Márquez Díaz [2], que presenta un sistema similar para la predicción de contagios o C. I. Orozco, E. Xamena, C. A. Martínez, y D. A. Rodríguez que presenta un modelo para el diagnóstico de enfermos a través de una radiografía pulmonar. En definitiva, cada vez son más los investigadores que proponen expandir el uso de RNAs a esta pandemia con el objetivo de aliviar la carga sanitaria y arrojar algo de luz sobre el problema.

Nuestro trabajo ha consistido en elaborar un modelo predictivo basado en aprendizaje automático mediante perceptrón multicapa

Planteamiento

El presente documento tiene la intención de realizar un estudio en el que se utilizan diferentes modelos y arquitecturas de redes neuronales para predecir la propagación de la enfermedad COVID-19.El estudio se va a realizar sobre un conjunto de datos que contiene información de 266 países, provincias y estados.

En la primera parte de la practica una vez hayamos probado diferentes modelos de RNA elegiremos el mejor y realizaremos la predicción para dos países concretos que son España y Singapur.

En la segunda parte se realizará la tarea de predicción mediante el uso de series temporales para determinar la evolución de los casos confirmados y la cantidad de fallecidos en los próximos días para España y Singapur.

Para la realización de dicha práctica hemos usado el entorno de análisis de datos 'Weka' que permite realizar el preprocesado de datos, modelos de redes neuronales y series temporales mediante la herramienta 'timeseriesForecasting'.

La metodología empleada para realización de la Parte I ha sido la siguiente:

La metodología empleada para realización de la Parte II ha sido la siguiente:

En este documento vamos a encontrar las siguientes secciones:

- Desarrollo del problema: en esta parte se explicará la obtención y procesado de datos, además de los procesos experimentados por los datos antes de ser usados para entrenar los modelos de redes neuronales.
- Parte I - Regresión: en esta sección del documento se expondrán diferentes modelos de RNA ,además estos se compararán para determinar cuál es el más adecuado a este problema. Una vez seleccionado el mejor modelo se realizará la predicción y se compararan los resultados de esta con la realidad.
- Parte II - Series temporales: en esta parte se expondrán diferentes modelos de RNA usados para la representación. También se realizará un estudio de los resultados obtenidos y una comparación de estos con la evolución real de la pandemia.
- Conclusión: sección del documento en la que se expondrán los problemas que hemos encontrado en la resolución de la práctica, además de una breve opinión personal sobre la misma.
- Bibliografía: última parte donde se listarán los enlaces utilizados en la redacción del contexto.

Desarrollo del problema

Los datos que se han empleado para el desarrollo de esta práctica han sido recopilados por "Center for Systems Science and Engineering" de la Universidad Johns Hopkins a partir de varias fuentes.

Estos datos han sido recopilados diariamente desde el día 22 de enero de 2020 para un total de 266 países, provincias y estados. Estos se encuentran disponibles en la página web: "data.humdata.org". Este fichero incluye un total de 266 instancias y 262 atributos.

Los atributos son:

- El nombre de la región o provincia
- El nombre del país
- La latitud
- La longitud
- El acumulado de los casos día a día desde el 22 de enero hasta el 6 de octubre (en total 258 días).

El fichero que contiene los datos inicialmente se trata de un archivo tipo .csv en el que los datos se encuentran separados por comas. En este fichero se han modificado varios caracteres que no son reconocidos por Weka y los nombres de los atributos correspondientes a las fechas.
Una vez realizadas estas modificaciones el archivo .csv se ha transformado en un archivo .arff que es tipo de fichero que utiliza Weka para los datos, en este proceso se añade una cabecera que incluye el nombre del fichero y los diferentes atributos que contiene con los distintos valores que pueden tomar cada uno de ellos.

El inicio de la cabecera del fichero de los datos finalmente queda así:

```
@relation Preprocesado.arff

@attribute Province/State {'Australian Capital Territory','New South Wales','Northern Territory',Queensland,'Sout
@attribute Country/Region {Afghanistan,Albania,Algeria,Andorra,Angola,'Antigua and Barbuda',Argentina,Armenia,Aus
@attribute Lat numeric
@attribute Long numeric
@attribute 'Dia -258' numeric
@attribute 'Dia -257' numeric
@attribute 'Dia -256' numeric
@attribute 'Dia -255' numeric
@attribute 'Dia -254' numeric
@attribute 'Dia -253' numeric
@attribute 'Dia -252' numeric
@attribute 'Dia -251' numeric
@attribute 'Dia -250' numeric
@attribute 'Dia -249' numeric
@attribute 'Dia -248' numeric
@attribute 'Dia -247' numeric
@attribute 'Dia -246' numeric
@attribute 'Dia -245' numeric
@attribute 'Dia -244' numeric
@attribute 'Dia -243' numeric
@attribute 'Dia -242' numeric
@attribute 'Dia -241' numeric
@attribute 'Dia -240' numeric
@attribute 'Dia -239' numeric
@attribute 'Dia -238' numeric
@attribute 'Dia -237' numeric
```

Para el desarrollo de las series temporales se han generado dos ficheros .arff nuevos que contiene la serie de casos confirmados y fallecidos acumulados día a día. Cada fichero corresponde a un país en este caso España y Singapur.

Al igual que los datos anteriores estos han sido recopilados por "Center for Systems Science and Engineering" de la Universidad Johns Hopkins a partir de varias fuentes y obtenidos de la página "data.humdata.org". En esta parte no solo se han obtenido los datos de casos acumulados también los de fallecidos.

El procesado de los datos ha sido similar a la parte anterior, la principal diferencia es que en esta ocasión se han eliminado los atributos de nombre del país, nombre de la provincia o estado, latitud y longitud. Esto supone que únicamente se han usado la serie de las fechas, la de los casos confirmados y la de fallecidos.

Finalmente, la cabecera del fichero de datos queda así:

```
@relation seriesSingapur.arff

@attribute Dia: {1/22/20,1/23/20,1/24/20,1/25/20,1/26/20,1/27/20,1/28/20,
@attribute Casos: numeric
@attribute Fallecidos: numeric

@data
1/22/20,0,0
1/23/20,1,0
1/24/20,3,0
1/25/20,3,0
1/26/20,4,0
1/27/20,5,0
1/28/20,7,0
1/29/20,7,0
1/30/20,10,0
1/31/20,13,0
2/1/20,16,0
2/2/20,18,0
2/3/20,18,0
2/4/20,24,0
2/5/20,28,0
2/6/20,28,0
2/7/20,30,0
2/8/20,33,0
2/9/20,40,0
2/10/20,45,0
2/11/20,47,0
2/12/20,50,0
2/13/20,58,0
2/14/20,67,0
2/15/20,72,0
2/16/20,75,0
```

Adjunto a este documento se pueden encontrar los ficheros de datos .arff utilizados para el desarrollo de la práctica.

Una vez que los datos han sido procesados para que se puedan utilizar en la herramienta de Weka, hemos iniciado un proceso mediante la aplicación de funciones para intentar mejorar los resultados que se obtiene de los modelos de RNA generados con estos. Las técnicas utilizadas han sido:

- Normalización: consiste en ajustar los valores de todos los atributos a un valor entre 0 y 1 siendo 0 el valor de la instancia mínima y 1 el de la máxima.
- NominalToBinary: convierte los atributos nominales en numéricos binarios, de manera que un atributo nominal con k valores se convierte en k atributos binarios.
- Randomize: mezcla aleatoriamente el orden de las instancias.

Finalmente debido a que los resultados obtenidos con los datos modificados de esta manera no eran mejores que los obtenidos con los datos normales, hemos tomado la decisión de utilizar los datos normales sin aplicarles ninguna técnica de las expuestas.

Parte I – Regresión (RNA – MLP)

Una vez cargados los datos correctamente procedemos a realizar varios modelos modificando enormemente los parámetros hasta encontrar un modelo aceptable e ir poco a modo modificando esos nuevos parámetros para ajustar la correlación lo más alta posible mientras los errores relativos se mantienen estables o decrecen.

Resultados del problema

Tras haber estado probando distintas combinaciones del modelo: número de épocas, número de capas ocultas, valores de neuronas para cada capa oculta, valor del ratio de aprendizaje, valor del momento, etc. Determinamos que el modelo que mejor se adapta a nuestro conjunto de datos es el siguiente:

```
Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.02 -M 0.1 -N 100 -V 0 -S 0 -E 100 -H 5
Relation:     Preprocesado.arff
Instances:    266
Attributes:   262
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

```
Time taken to build model: 0.92 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9641
Mean absolute error                  55446.7585
Root mean squared error             199109.1631
Relative absolute error                 26.8433 %
Root relative squared error             28.5776 %
Total Number of Instances              266
```

Con dicho modelo podemos observar que su coeficiente de correlación es bastante alto y sus errores relativos se encuentran dentro de un rango aceptable.

Gracias a la técnica de 'cross validation' hemos obtenido mejores resultados en las predicciones que veremos a continuación ya que con modelos entrenados mediante la separación de entrenamiento y test aunque los valores de correlación fueran mucho mayores y los errores ligeramente inferiores, sus predicciones eran totalmente erróneas respecto a los países utilizados (España y Singapur) en ello radica la razón por la que se apostó en obtener un buen modelo basado en 'cross validation' conocida como validación cruzada.

Análisis de los resultados

Pasamos a comentar los resultados predichos del número de contagios los tres días siguientes, empezaremos con España y continuaremos con Singapur. Finalmente, analizaremos detalladamente cada país.

Día -3

```
=== Predictions on test set ===

inst#,actual,predicted,error
1,?,775107.933,?
```

Día -2

```
=== Predictions on test set ===

inst#,actual,predicted,error
1,?,780630.274,?
```

Día -1

```
=== Predictions on test set ===

inst#,actual,predicted,error
1,?,800543.88,?
```

Día -3

```
=== Predictions on test set ===

inst#,actual,predicted,error
1,?,125130.403,?
```

Día -2

```
=== Predictions on test set ===

inst#,actual,predicted,error
1,?,126854.405,?
```

Día -1

```
=== Predictions on test set ===

inst#,actual,predicted,error
1,?,125914.38,?
```

Como podemos ver por los resultados, en España se predice que los próximos contagios irán aumentando alrededor de los 800K y en Singapur se estabilizarán sobre los 126K. Si comparamos dichos resultados con la realidad observamos que en España los datos se sitúan muy cercanos a los 800K, concretamente los datos reales son 789932, 789932, 813412; respectivamente. Esto quiere decir que no solo van en aumento, sino que se sitúan en torno a dichos valores. Con ellos podemos deducir que nuestro modelo adopta el comportamiento que se produce en la realidad y da un muy buen reflejo de los datos cuantitativamente.

Respecto a Singapur, si comparamos los resultados con los datos reales observamos que sus valores son muy dispares ya que sus valores deberían ser 57812, 57800, 57819, para los siguientes tres días respectivamente y nosotros casi los triplicamos. Si bien es cierto, el modelo sí replica el comportamiento (incluso cuando este es un tanto extraño si nos fijamos bien) aunque lo hace a una mayor escala.

Concluiremos el análisis de los resultados afirmando que nuestro modelo puede considerarse como aceptable en la predicción de los valores ya que obteniendo los errores medios en ambos países para los tres días obtenemos que en España es de 36996 y en Singapur de 204467, por lo que en media resultaría de un error de 120K que puede parecer mucho como valor absoluto, aunque no lo es tanto teniendo en cuenta que los datos alcanzan valores cercanos a 1M, lo cual supone un error relativo de un 12% en un plano mucho más general.

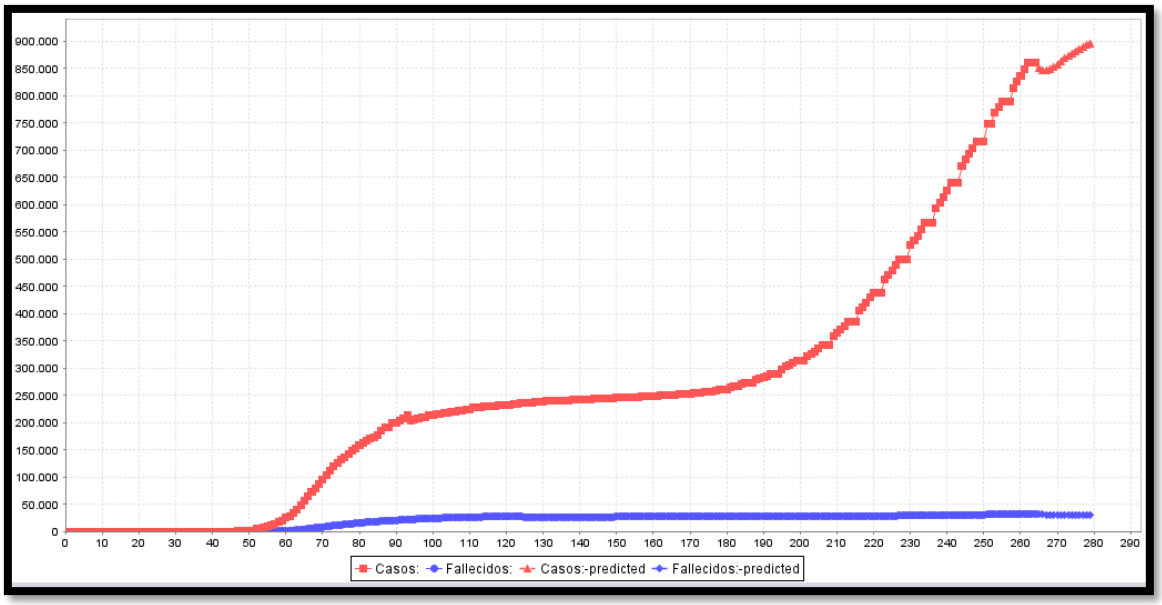
Parte II – Series temporales

Una vez se ha conseguido desarrollar un modelo predictivo fiable la siguiente tarea consistió en utilizar un modelo similar para estimar la serie temporal de contagios y fallecidos por Covid19 en España y Singapur. Para ello cargamos los datos procesados de nuestra serie temporal e iniciamos la experimentación.

Resultados del problema

Replicando el conocimiento que hemos adquirido en el apartado anterior hemos modelado nuestra red siguiendo una aproximación similar. Estos son los resultados obtenidos durante el testeo de los modelos anteriormente generados.

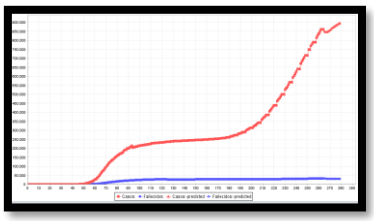
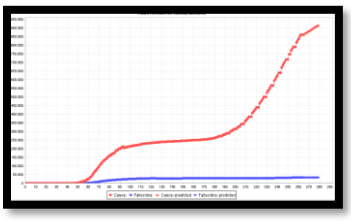
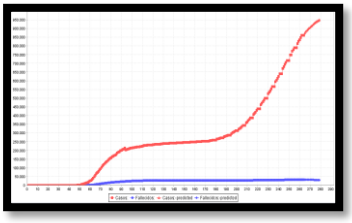
ESPAÑA



MultilayerPerceptron -L *0.02* -M 0.1 -N *100* -V 0 -S 0 -E 100 -H 5

Forecasting / lag: default / Periodicity: daily

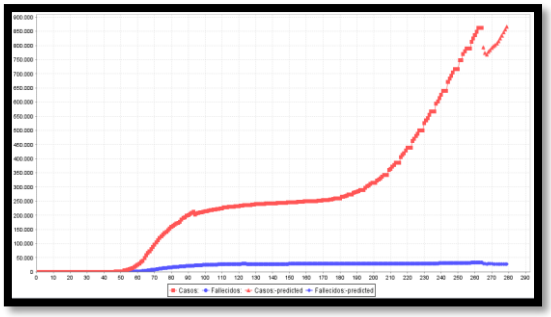
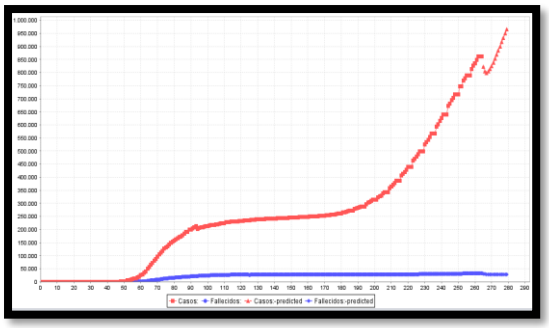
Aquí podemos observar cómo los modelos menos precisos generan predicciones que no se ajustan a la serie temporal, con el detalle de que requieren un ajuste de las variables marcadas como "lagged" para doblegar la curva.

Lag: 5-10 Lag: 10-20 Lag: 15-30

Después de estudiar esta predicción de serie temporal iniciamos la búsqueda de mejores valores para los modelos estudiados, encontrando algunos nefastos, otros tolerables, pero poco representativos, hasta alcanzar el modelo óptimo para la tendencia en España.

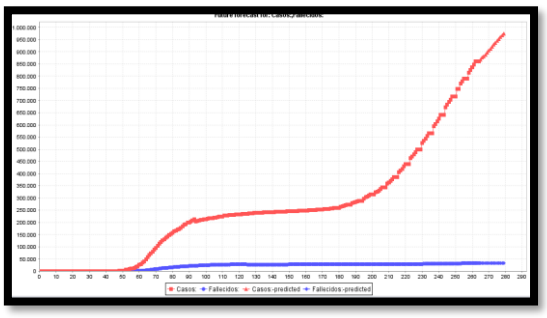
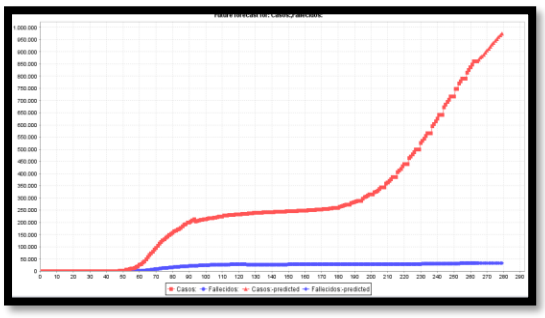
Ejemplos nefastos:

-L *0.01* -M 0.1 -N *250* -V 0 -S 0 -E 20 -H *175* -L *0.005* -M 0.1 -N *-500* V 0 -S 0 -E 50 -H *175*

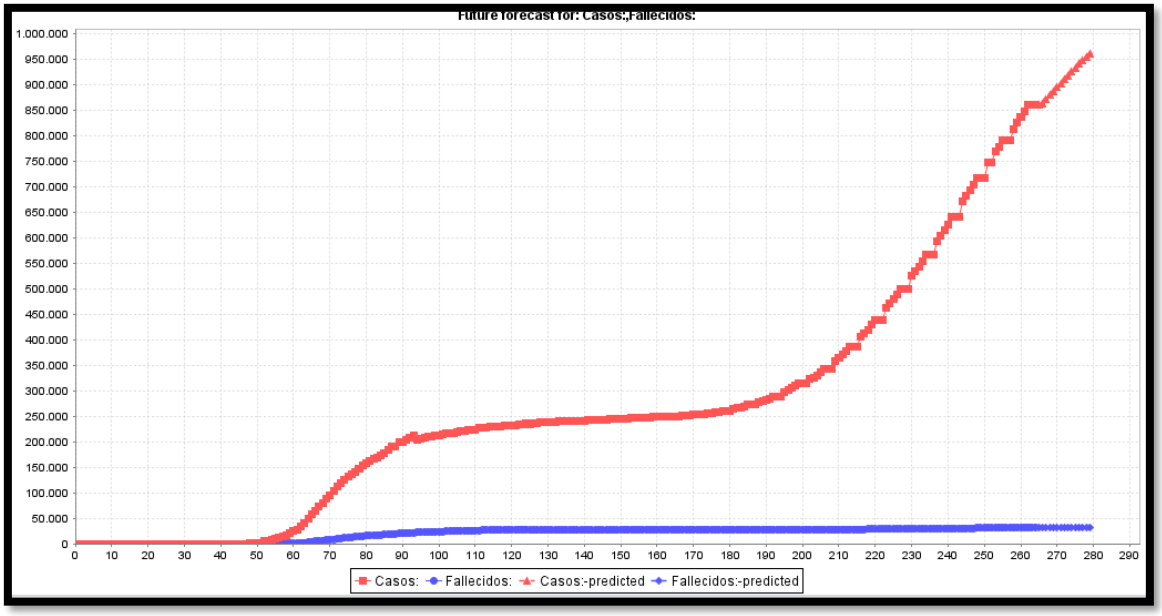
Forecasting / Lag: default / Periodicity: daily Forecasting / Lag: 1-3, Periodicity: daily

Ejemplos tolerables, pero poco representativos:

-L 0.*01* -M 0.1 -N *250* -V 0 -S 0 -E 1 -H *a* -L 0.*005* -M 0.1 -N *500* -V 0 -S 0 -E 50 -H *175*

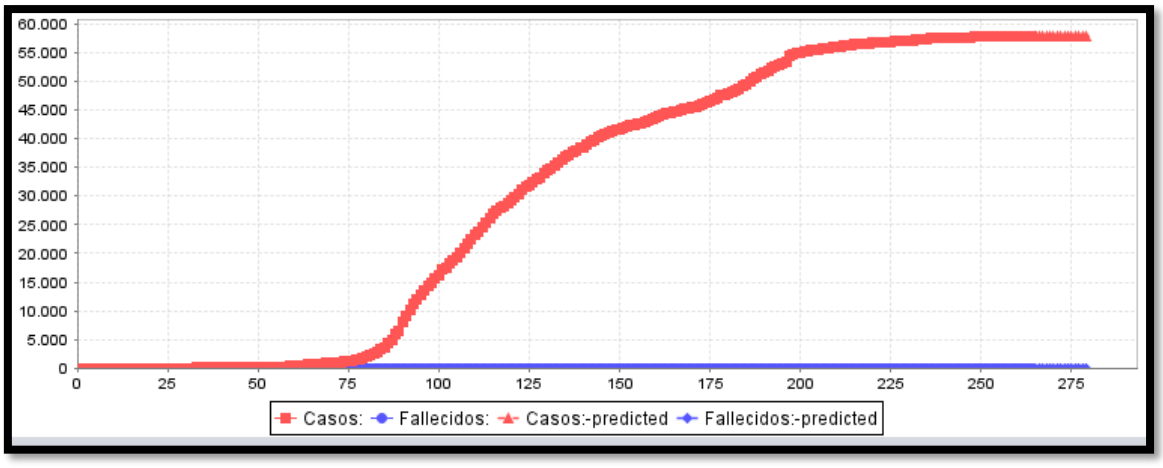
Forecasting / Lag: 3-8, periodicity: daily Forecasting / Lag: 3-8 / Periodicity: daily



MultilayerPerceptron -L *0.02* -M 0.1 -N *250* -V 0 -S 0 -E 20 -H *a*

Forecasting / Lag: default / Periodicity: daily

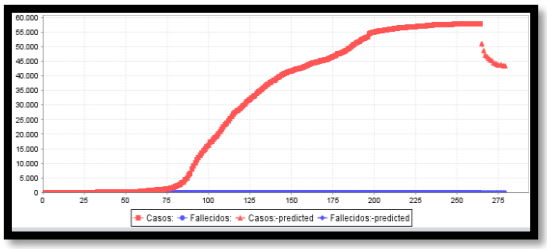
SINGAPUR

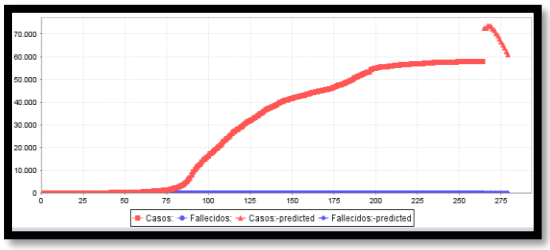


MultilayerPerceptron -L *0.05* -M 0.3 -N *560* -V 0 -S 0 -E 20 -H *20*

Forecasting / Lag: 8-18 / Periodicity: None

Ejemplos nefastos:





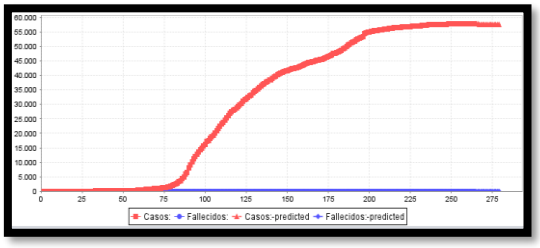
-L *0.01* -M 0.2 -N *200* -V 0 -S 0 -E 1 -H *175*

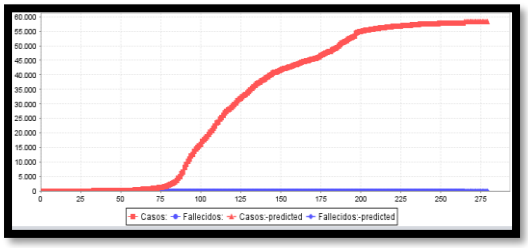
Forecasting / Lag: default / Periodicity: daily

-L *0.01* -M 0.2 -N *570* -V 0 -S 0 -E 20 -H *25*

Forecasting / Lag: 3-8 / Periodicity: none

Ejemplos más representativos:





-L *0.01* -M 0.3 -N *560* -V 0 -S 0 -E 20 -H *20*

Forecasting / Lag: 8-18 / Periodicity: none

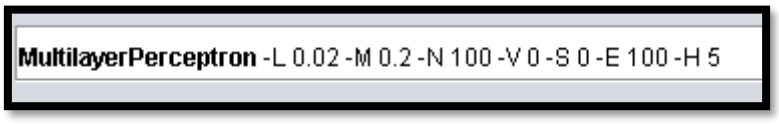
-L *0.01* -M 0.3 -N *560* -V 0 -S 0 -E 20 -H *20*

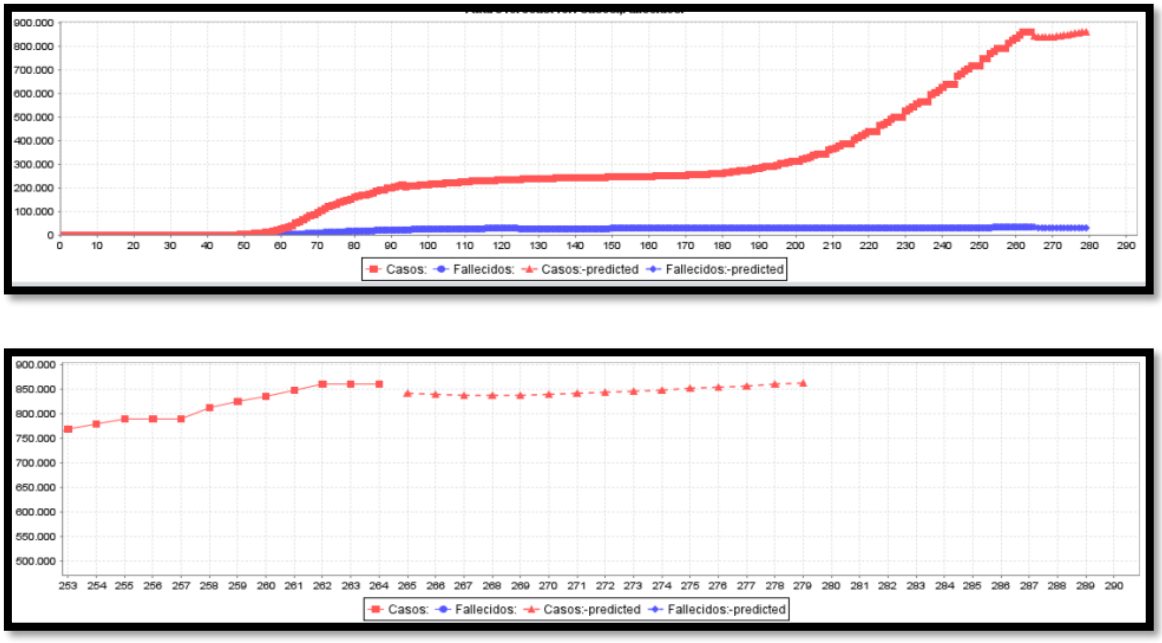
Forecasting / Lag: 8-11 / Periodicity: none

Análisis de los resultados

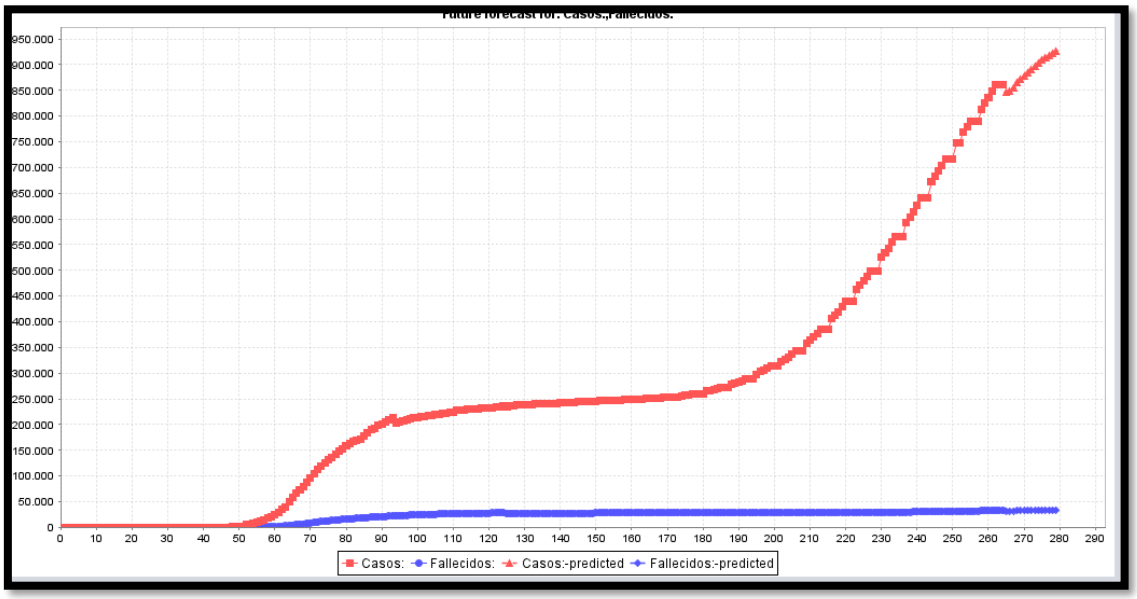
Seleccionamos el mejor de los modelos estudiados en la primera fase con el rango de "lag" por defecto.

ESPAÑA

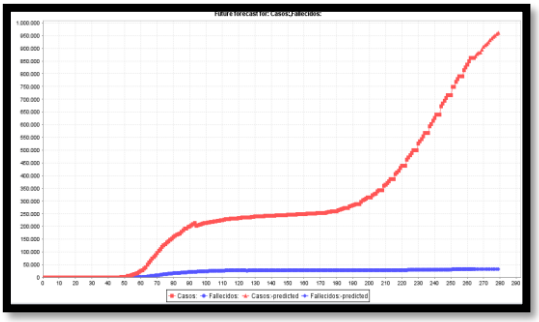




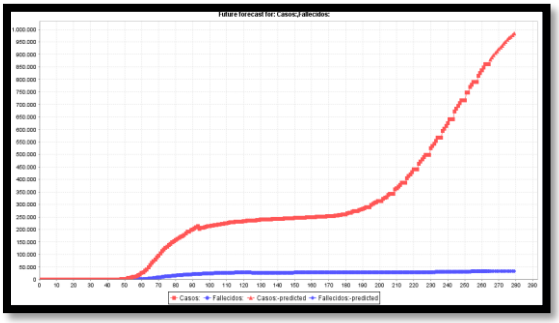
En este caso observamos cómo esta curva de contagios se adecua mejor que los modelos previos a la tendencia que presenta la serie temporal. Ajustando mejor los valores de mínimo y máximo lag conseguimos una gráfica que ajusta muy bien tanto casos confirmados como muertes por Covid19.



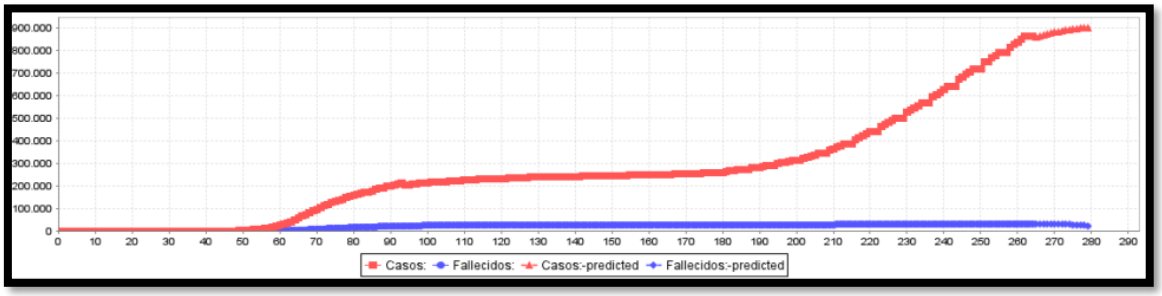
Lag: 1, 20



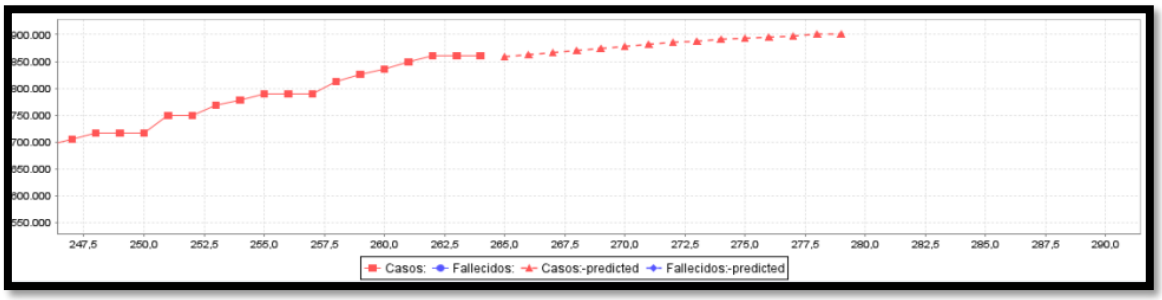
Lag: 4-5



Lag: 10-30



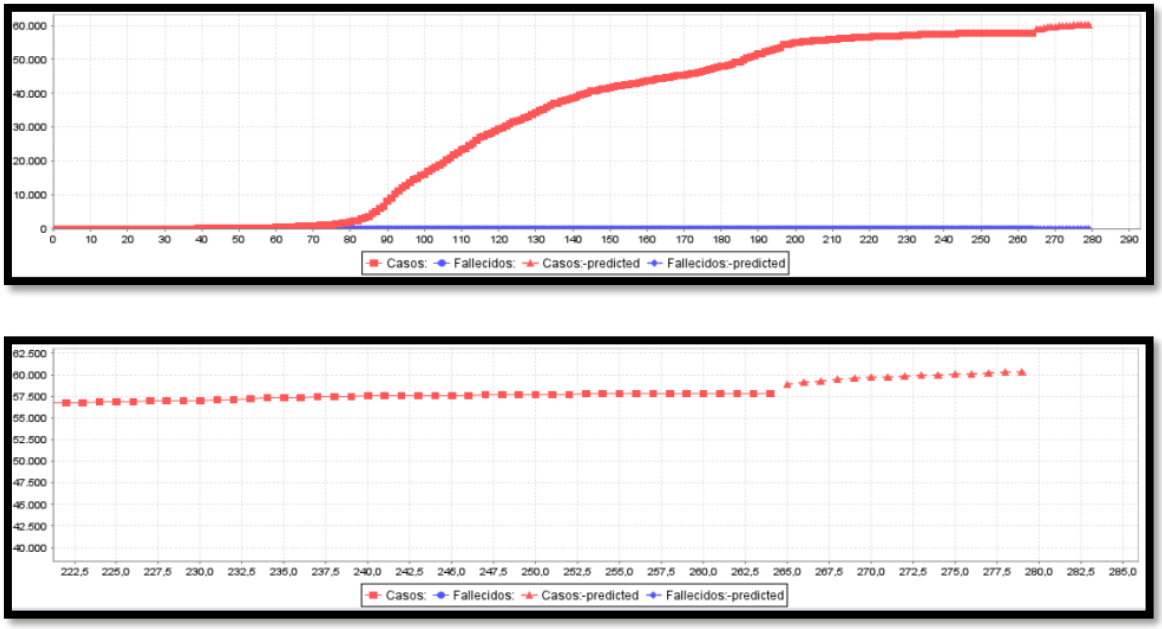
Lag 1-150



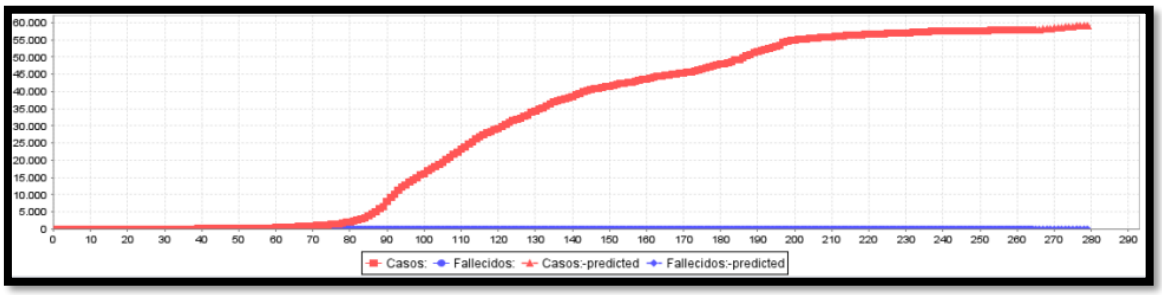
De nuevo seleccionamos el mejor de los modelos estudiados en la primera fase con el rango de "lag" por defecto.

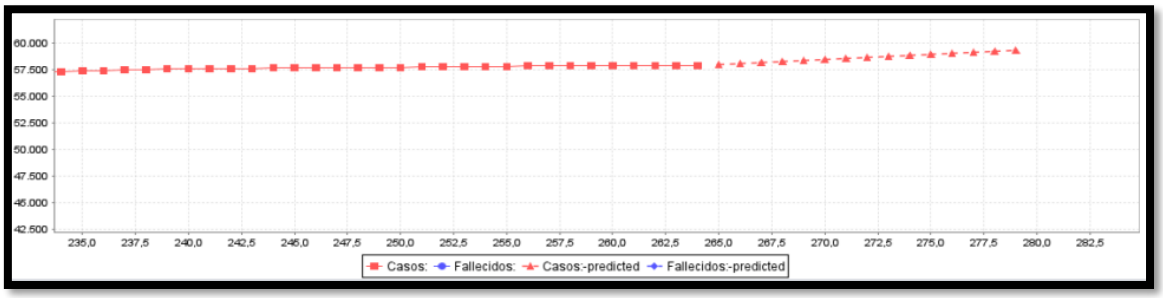
SINGAPUR

```
MultilayerPerceptron -L 0.02 -M 0.1 -N 100 -V 0 -S 0 -E 100 -H 5
```



A continuación, ajustamos la variable lag para conseguir una predicción que se ajuste mejor a la secuencia temporal de la serie en Singapur





Lag 1-150

Conclusión

Los mayores problemas encontrado a la hora de realizar la práctica se han encontrado en la utilización de los distintos parámetros que se pueden cambiar en el perceptrón multicapa en Weka lo que ha requerido investigación y muchas pruebas para aprender lo que hacía cada uno. Otro de los grandes problemas que hemos tenido es la cantidad de tiempo que requiere la realización de alguno de los modelos de RNA, especialmente aquellos que utilizan el método de 'cross validation'.

En términos generales, consideramos que esta práctica ha sido una experiencia enriquecedora en lo referido a comprender los conceptos y sobre todo el funcionamiento de las redes de neuronas artificiales, además nos ha permitido experimentar con diferentes modelos y arquitectura de RNA, así como con la herramienta de representación gráfica de Weka.

Pese a esto también creemos que la manera en la que se debe realizar la práctica mediante ensayo y error hasta obtener algún modelo aceptable es un poco tediosa y tal vez sería necesario incluir alguna pauta para hacer este proceso menos caótico.

Bibliografía

[1]

M. M. G. Lorenzo et al., "Adquisición de conocimiento sobre la letalidad de la COVID-19 mediante técnicas de inteligencia artificial," vol. 10, no. 3, p. 891, 2020.

[2]

J. E. M. Díaz, "Inteligencia Artificial y Big Data como soluciones frente al COVID-19," no. 50, pp. 315–331, 2020.

[3]

C. I. Orozco, E. Xamena, C. A. Martínez, and D. A. Rodríguez, "COVID–XR: A Web Management Platform for Coronavirus Detection on X-ray Chest Images," vol. 100, no. 1e, 2020.