Fase 3: F2006B

Modelación numérica de sistemas estocásticos

"Reto: Simulación y optimización de un modelo de difusión de epidemia"

Equipo: Dr. J Poncho

Profesores: Servando López Aguayo y Antonio Ortiz Ambriz

Arif Morán Velázquez - A01234442

Alberto Anaya Velasco - A01252512

Ishan Joel Don Wickramage Madawala Guzmán - A01704771

Borja Martínez Ramírez - A01234311

Mayra Stefany Gómez Triana - A01625609

Tecnológico de Monterrey

Monterrey, Nuevo León

08 de septiembre de 2023

La sorpresiva compañía RetoTec© , siempre a la vanguardia científica y tecnológica, ha regresado y ahora desea realizar un proceso de vacunación y manejo de una serie de epidemias. Dichos brotes de enfermedades se han presentado en 4 regiones del mundo, en donde no se tiene acceso a ninguna otra información extra, más que a los datos o modelos que se comparten en este documento y que, por lo tanto, servirán como una primera base para lograr decidir contra qué tipo de enfermedad se está tratando en cada caso. Por todo lo anterior, se ha lanzado una interesante competencia entre ocho equipos finalistas de clase mundial: **Dr. J Poncho**, Los profetas del Caos, Servandioses y Potenciales retrasados.

Los cuatro casos en donde RetoTec© solicita que se haga una predicción del tipo de epidemia correspondiente son:

- **I Datos de una epidemia contaminados con ruido.**
- **II Caso a) de modelación de una epidemia con agentes.**
- **III Caso b) de modelación de una epidemia con agentes.**
- **IV Datos de una epidemia con modelación estocástica.**

Se busca predecir a qué enfermedad corresponde cada caso (dando un estimado en porcentaje de la seguridad de predicción), en base a que se tiene, para estas cuatro regiones, antecedentes de una distribución R_0 dada por:

R_0	Enfermedad
1 - 2	Gripe porcina
2 - 5	Covid-19
5 - 7	Paperas
7 - 12	Varicela
12 - 18	Sarampión

Arquitectura red neuronal
La arquitectura de nuestra red neuronal se define de la siguiente manera:
1. Capa de entrada: la capa de entrada se define implícitamente por la variable 'x', que contiene los valores de tiempo.
2. Capa oculta: Se especifica una capa oculta con 10 neuronas mediante la variable 'hiddenLayerSize'. Esta capa es la que permite que la red aprenda patrones y relaciones en los datos.
3. Capa de salida: la capa de salida se crea automáticamente para coincidir con la dimensión de 'y_data', que contiene los datos originales.
La red se entrena durante 5 épocas utilizando los datos de entrada 'x' y los datos objetivo 'y_data'. Además, los datos se dividen en conjuntos de entrenamiento, validación

y prueba utilizando la función 'dividerand' con una relación de división de 70% para entrenamiento, 15% para validación y 15% para prueba.
En resumen, la arquitectura de la red neuronal es una red feedforward con una capa oculta de 10 neuronas, una capa de entrada implícita y una capa de salida implícita que coincide con la dimensión de los datos objetivo.

1. **Datos de una epidemia contaminados con ruido.**

   En esta primera predicción será necesario el archivo *reto1* para posteriormente hacer la predicción correspondiente de R_0 En este caso, se te pide, además de los parámetros optimizados, que muestres la configuración de red neuronal que se utilizó, y además, que se comente el porqué se prefirió dicha arquitectura de red neuronal. Recuerda, la idea es convencer a RetoTec© que tu equipo es el mejor.

   Como primer paso, suavizamos los datos originales de esta epidemia, usando una red neuronal. Como resultado de este proceso conseguimos la figura 1.

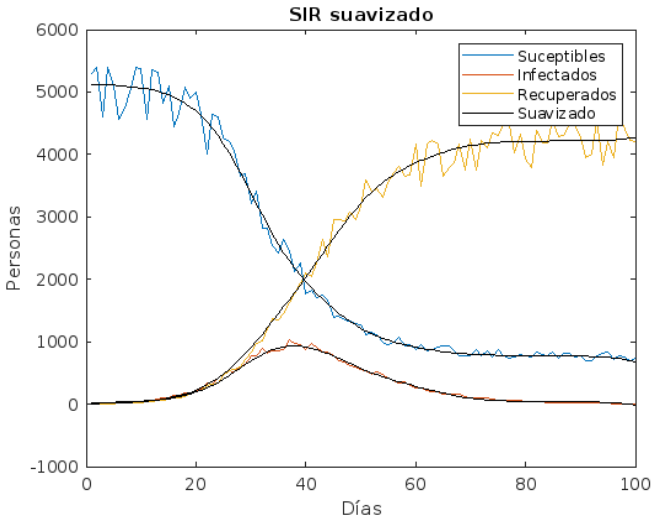

Figura 1. Datos suavizados, problema 1.

A continuación, definimos la función objetivo que nos ayudará a determinar cual es valor de R_0 adecuado:

$$\left(\dot{S} + \frac{\beta SI}{N}\right)^2 + (\dot{I} + \gamma I)^2 + (\dot{R} - \gamma I)^2 = 0 \tag{1}$$

Al graficar esta función objetivo para distintos todos los valores de β y γ conseguimos las siguientes dos representaciones visuales del pozo que se forma.

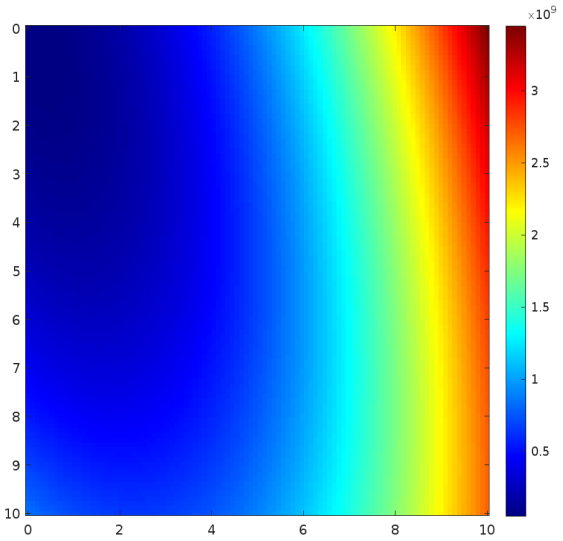

Figura 2. Representación bidimensional de la función objetivo.

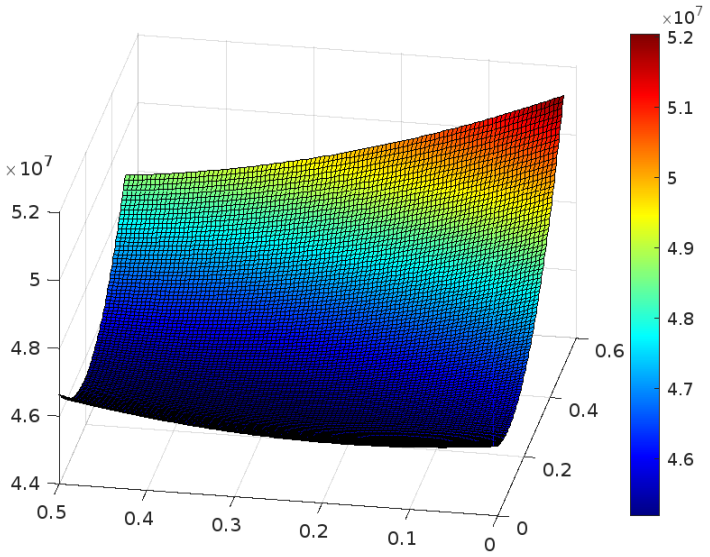

Figura 3. Representación tridimensional de la función objetivo.

Y tras el proceso de optimización de esta función objetivo, con el algoritmo genético llegamos a los siguientes resultados de β, γ, y R_0 :

$$\beta \approx 0.3203$$

$$\gamma \approx 0.1486$$

$$R_0 = \frac{\beta}{\gamma} \approx 2.15545\pm \tag{2}$$

Ahora comparamos la figura 1, con un SIR determinístico con $\beta = 0.3218$, y $\gamma = 0.1509$.

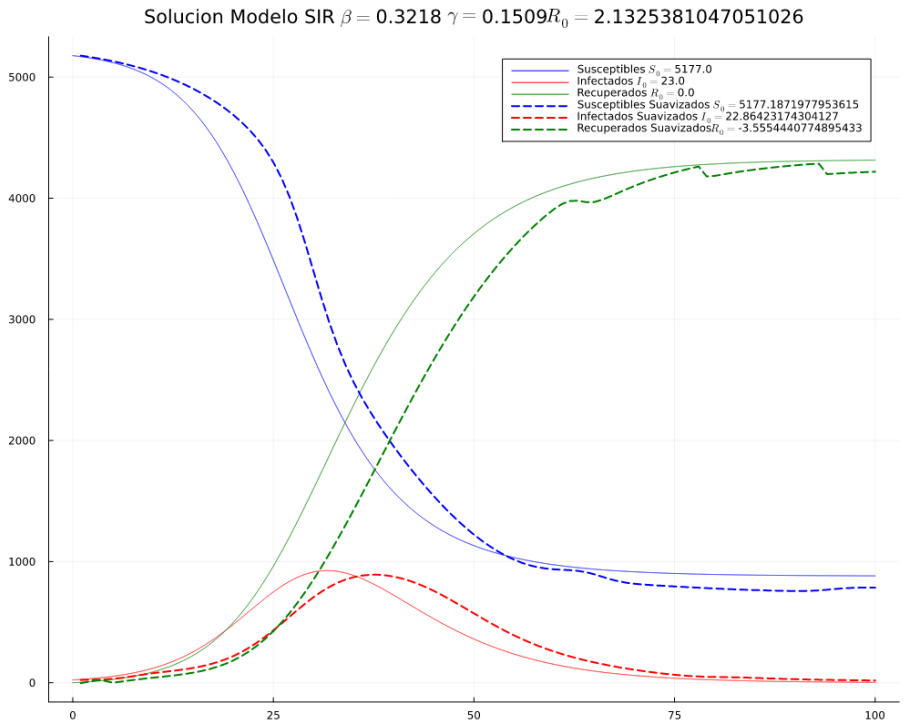

Figura 4. Comparación dado un β y γ

Estos resultados nos hacen suponer que, en este caso, la epidemia en cuestión es de Covid-19, con un 84% de certidumbre, por lo que también podría llegar a ser una epidemia de gripe porcina.

1. **Caso a) de modelación de una epidemia con agente**

   En este caso, se sabe que los datos reales corresponden en gran medida con una simulación hecha en Netlogo basada en el modelo epiDEM Basic. Se te pide adecuar dicho modelo para mostrar las curvas SIR, considerando además que se usará como valores de Netlogo lo siguiente: [initial people 500, infection chance 2, recovery chance 10, average recovery time 50]. Se te pide especialmente en este caso, que reportes qué cambios se hicieron al modelo original de Netlogo, y además cómo se pasaron los datos de dicha simulación de agentes a Matlab. Finalmente, ¿es congruente el R obtenido por la simulación de Netlogo con el valor obtenido por el modelo SIR ya sintonizado? ¿O se trata de otra definición de R?

   Al observar los resultados de la simulación podemos notar que el R_0 de Netlogo está definido de una forma distinta. Comúnmente se define como $\frac{\beta}{\gamma}$, pero al calcular R_0 de esta forma, llegamos a un valor de 2.14, mientras que el promedio de las 10 corridas de Netlogo da como resultado 2.584799.

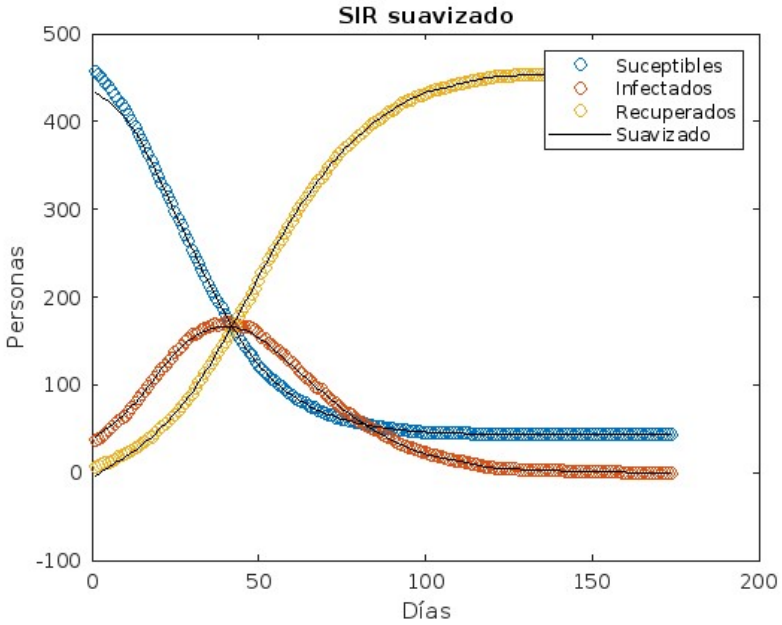

Figura 5. Datos suavizados, problema 2.

A partir de estos datos, asumimos que la epidemia es de Covid-19, y tenemos una certeza del 82% de esto. Aún así no podemos descartar la posibilidad de que esta epidemia sea en realidad de gripe porcina.

1. **Caso b) de modelación de una epidemia con agentes.**

   En este tercer caso, se sabe nuevamente que los datos reales corresponden en gran medida a una simulación hecha en Netlogo basada en el modelo edpiDEM Basic. Se te pide nuevamente adecuar dicho modelo, para mostrar las curvas SIR, considerando además que se usará como valores de Netlogo lo siguiente: [initial people 500, infection chance 2, recovery chance 90, average recovery time 50]. Para este caso, especialmente, se te solicita que reportes, además, un promedio de 10 ejecuciones del proceso de Netlogo, para conocer que tanta desviación estándar existe en los datos mostrados en cada simulación. ¿Es congruente el valor de R_0 obtenido por la simulación de Netlogo, con el predicho por el modelo SIR ya sintonizado? ¿O se trata de otra definición de R_0?

   El R_0 promedio de las 10 corridas con los parámetros dados para este ejemplo fue de 1.4636747, por otra parte el R_0 calculado a partir del modelo SIR fue de 1.632, una diferencia de poco más del 10%. Esto parece indicar diferentes definiciones de R_0. Para comprobar si este es el caso, decidimos calcular manualmente el valor de R_0 para varios tiempos cualquiera, según su definición $R_0 = \beta\gamma$, y encontramos valores de 3, 0, 0.5, 5, y 0.49998, cuando la Netlogo indica 1.011478, 0.930822, 1.079855, 1.03141, y 1.544813 respectivamente. Este último descubrimiento nos confirma que Netlogo calcula R_0 por algún método alternativo al convencional.

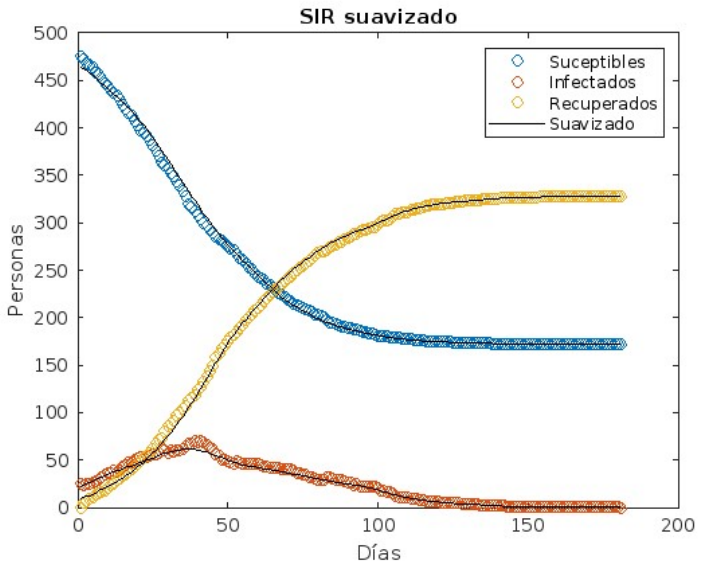

Figura 6. Datos suavizados, problema 3.

Esto nos deja entrever que la epidemia probablemente es de gripe porcina, y tenemos una certidumbre calculada del 82% de que en efecto sea el caso, pero también podría llegar a ser, una epidemia de Covid-19.

1. **Datos de una epidemia con modelación estocástica.**

   En este último caso, se tienen datos que no están equiespaciados, ya que provienen de un modelo estocástico del tipo Doob-Gillespie. Se te pide descargar el archivo reto4.mat. En este último escenario, además de las predicciones, se te pide que antes de realizar un pre-procesamiento con redes neuronales, realices un “moving average" de los datos previamente facilitados. Reporta los resultados obtenidos con tus algoritmos genéticos, para el caso de usar y no usar la técnica de “moving average" para los datos.

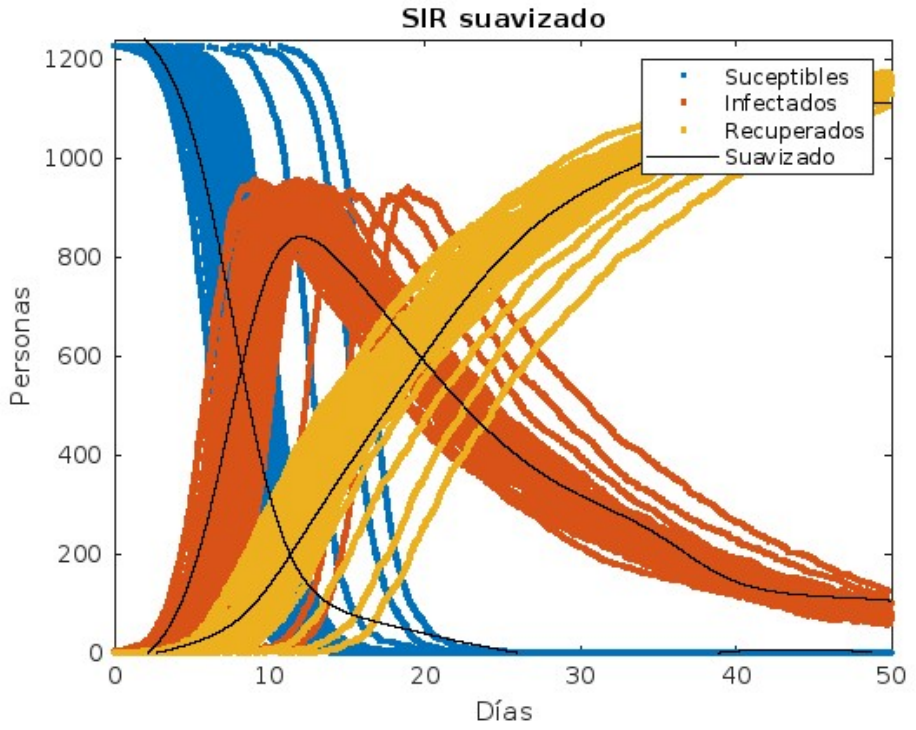

Figura 7. Datos suavizados, problema 4.

$$\beta \approx 0.5785$$
$$\gamma \approx 0.0597$$

$$R_0 = \frac{\beta}{\gamma} \approx 9.690117 \pm \tag{3}$$

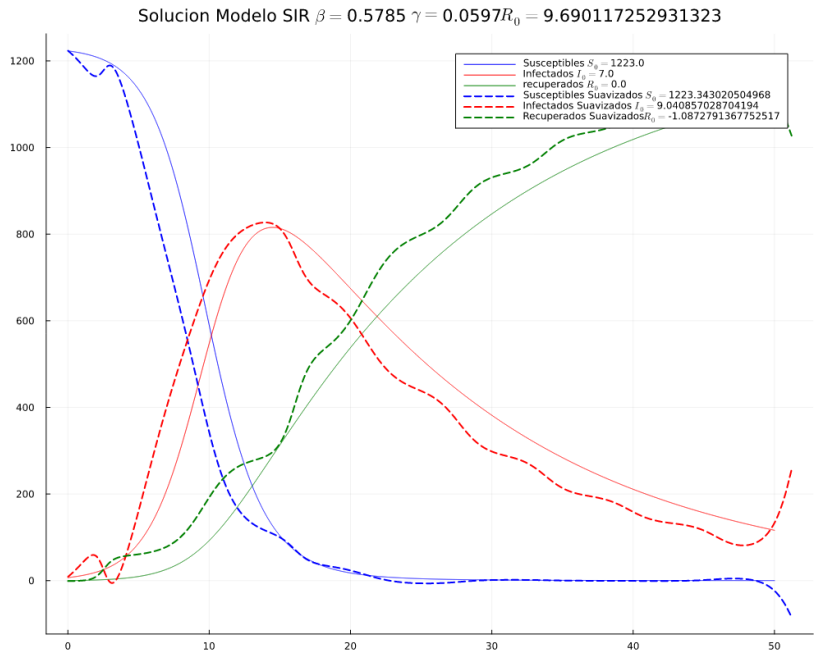

Figura 8. Comparación modelo SIR determinístico dado un β y γ

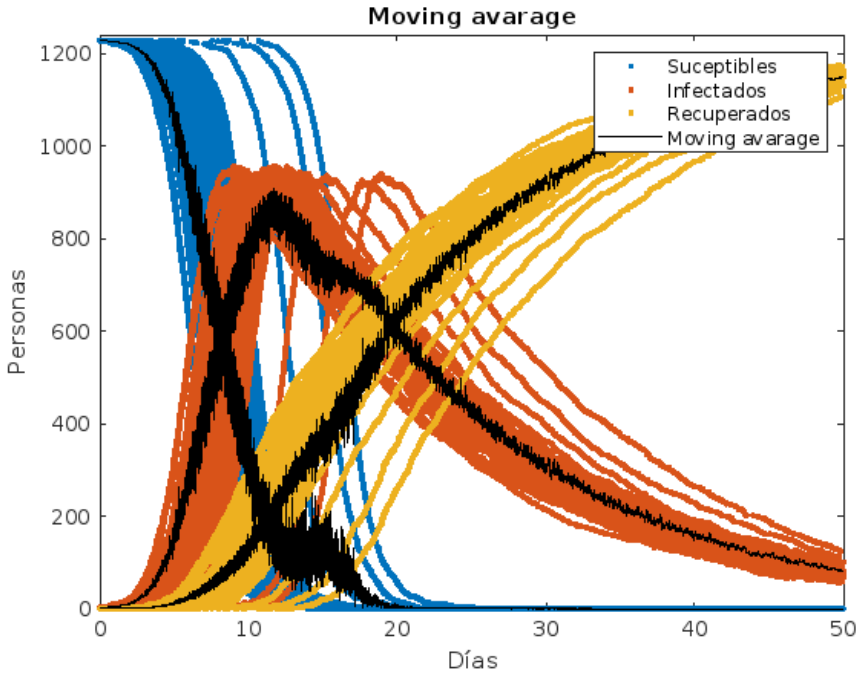

Figura 9. Media móvil con ventana de 50.

Esta información nos lleva a suponer que esta epidemia es de varicela, para la cual tenemos una certidumbre del 83%. Aun así, existe una pequeña posibilidad de que la epidemia sea de sarampión, o incluso paperas.

1. **Cálculo de incertidumbre**

   Para calcular la incertidumbre y darnos una idea de la certeza con la que podemos asegurar que el R_0 calculado es el real asociado a la enfermedad, podemos realizar una simplificada aproximación de los errores estimados que se van acumulando en cada parte del proceso.

   Primeramente, se considera que en los datos existe un error en la estimación de infectados y recuperados debido a los errores de precisión de las pruebas para cada enfermedad. Tomando un estimado, se tiene que puede existir alrededor de un 11% de error en cuanto al diagnóstico de la enfermedad, considerando un alto grado de error para la gripe porcina y la Covid-19, y mucho más bajos para enfermedades con síntomas palpables o visibles como las paperas, varicela y sarampión (CDC, 2016).

   A continuación, procedemos a resumir el proceso de Bootstrap utilizado para calcular la incertidumbre:

   Análisis de Bootstrap

   Utilizamos el método Bootstrap para evaluar la incertidumbre en nuestras estimaciones de R_0. Este método implica la generación repetida de múltiples muestras

de datos a partir de nuestro conjunto de datos original. Cada muestra se obtiene seleccionando aleatoriamente observaciones con reemplazo, lo que simula la variabilidad inherente en los datos de casos reales. Luego, aplicamos el mismo proceso de ajuste del modelo SIR a cada muestra generada, lo que nos permite obtener una distribución de estimaciones de R_0 basadas en diferentes conjuntos de datos de muestra. La variabilidad en estas estimaciones Bootstrap refleja la incertidumbre en nuestros resultados y nos permite calcular un intervalo de confianza alrededor de nuestra estimación puntual de R_0.

El intervalo de confianza obtenido a través del análisis de Bootstrap nos proporciona una medida de la incertidumbre en nuestras estimaciones de R_0. Al combinar esta información con la consideración de errores de precisión en la detección de la enfermedad, podemos evaluar con mayor confianza la robustez de nuestros resultados y proporcionar una estimación más precisa de la certeza asociada a nuestro valor calculado de R_0.

1. **Reflexiones individuales.**

Arif Moran:
A través de este curso, tuvimos la oportunidad de explorar una extensión del modelo SIR determinístico, como lo fue estudiar este modelo estocásticamente. Asimismo trabajamos en un acercamiento a un caso laboral real, esto a través de los puntos más importantes en el análisis de datos, el suavizado a través de métodos más sofisticados como las redes neuronales como mejora del promedio móvil. Considero que el modelo con el que trabajamos no solo nos permitió trabajar en nuestras habilidades de resolución de problemas, sino también de interpretación. Una de las cosas fundamentales a lo largo fue el entendimiento de algoritmos estocásticos del reto, además se trabajó en la interpretación de un fenómeno real, como lo fue identificar las enfermedades que existían a partir de R_0 . Algo importante a recalcar es que una componente fundamental es la interpretación de esta constante. Si bien en este caso se tomó como $\frac{\beta}{\gamma}$, en casos que buscan apegarse con mayor certeza a la realidad se emplea el uso de una definición diferente y no constante.

Alberto Anaya Velasco
Este reto nos permitió analizar y evaluar la importancia y utilidad de los modelos estocásticos para simular procesos complejos que resultan muy difíciles o incluso imposibles de modelar de manera determinista. Todos los fenómenos que suceden a nuestro alrededor parecerían contar con un grado de aleatoriedad, por lo que es de gran utilidad saber, además de lo predecible, saber identificar y aprender a manejar los elementos aleatorios de los procesos.
En esta última entrega comenzamos a ver la utilidad de redes neuronales para generar pruebas de regresión de los modelos SIR contaminados con ruido, estocásticos y simulados con agentes. Además, también nos basamos en algoritmos genéticos, de nuevo con el uso de redes neuronales para generar una predicción de las enfermedades a las que pertenecían los conjuntos de datos. De esta forma podemos finalmente ver cómo los algoritmos de aprendizaje y redes neuronales nos permiten, aunque no necesariamente entender mejor, poder predecir de manera más precisa fenómenos estocásticos y reducir su incertidumbre.
Ishan Joel Don Wickramage Madawala

Durante el transcurso de este curso, hemos tenido la oportunidad de adentrarnos en el intrigante ámbito de la modelización de fenómenos, adoptando un enfoque estocástico que difiere del tradicional modelo determinista SIR. En el marco de esta experiencia, exploramos en profundidad cómo los algoritmos de redes neuronales pueden emerger como herramientas poderosas para el suavizado de datos y la mejora de la precisión de las predicciones, en contraposición a los métodos más simples como los promedios móviles. Desde una perspectiva personal, considero que este desafío no solo ha enriquecido mis habilidades en la resolución de problemas, sino que también ha profundizado mi capacidad para interpretar fenómenos del mundo real. La constante R0 se convirtió en un elemento de análisis que aprendimos a descifrar, y descubrimos que su interpretación puede variar significativamente según el contexto y la definición empleada. Esta experiencia me ha instado a abrazar la complejidad inherente a los procesos y a buscar soluciones innovadoras en el ámbito del análisis de datos.

Borja Martinez
En esta última fase del reto se nos presentó un problema bastante único en su estructura, esto llevó a que el trabajo en equipo fuera aún más importante que en otros proyectos, pero también a que surgiera de una forma poco habitual, pues en vez de dividir las actividades por hacer según los puntos faltantes, dividimos los puntos faltantes en actividades, y al repetirse muchas de ellas, decidimos que estas actividades serían las divisas, para posteriormente ser replicadas en los demás puntos. Esto nos llevó a un trabajo extremadamente colaborativo, y transversal.

Mayra Stefany Gómez
Este reto ha representado un interesante viaje hacia la comprensión de los modelos estocásticos y su relevancia en la simulación de fenómenos complejos. A medida que explorábamos las profundidades de la aleatoriedad en nuestro entorno, hemos desarrollado una apreciación más profunda de la importancia de no solo comprender lo predecible, sino también de abordar la incertidumbre inherente a los procesos. Durante esta última fase, la incorporación de las redes neuronales en la generación de pruebas de regresión para los modelos SIR contaminados con ruido resultó ser una revelación. Además, el uso de algoritmos genéticos, junto con las redes neuronales, para predecir enfermedades a partir de datos complejos, ilustra de manera concluyente cómo estas herramientas pueden contribuir a la reducción de la incertidumbre en la predicción de fenómenos estocásticos. Este reto ha enfatizado la importancia de adaptarse y emplear enfoques innovadores para abordar problemas únicos y complejos.

Drive códigos:
`https://drive.google.com/drive/folders/1eSJkZyD7V7Kv3b5HthtB21KPkfJ_RVGz?usp=sharing`

Referencias
CDC. (2016). *Rapid influenza diagnostic tests.* `https://www.cdc.gov/flu/professionals/diagnosis/clinician_guidance_ridt.htm`