
**UNIVERSIDAD DE ANTIOQUIA**
**FACULTAD DE INGENIERÍA**
**INTELIGENCIA ARTIFICIAL**

# Proyecto inteligencia artificial

Informe final

**Integrantes**
Manuela Arteaga Arango
Rosa Puerta Campo

**Documentos**
1.152.710.365
1.002.389.631

**Docente**
Raúl Ramos Pollan

**Medellín-Antioquia**
**2023**

**Proyecto analitica de datos**

**Abstract**

La pandemia del COVID-19 ha representado una grave amenaza global, desafiando los sistemas de salud con su alta tasa de contagio y demanda que sobrepasa las capacidades de las UCI. A pesar de los esfuerzos de control, la disminución en el autocuidado y la persistencia de nuevos contagios han aumentado la necesidad de buscar soluciones efectivas. En este informe, se expone un proyecto que tiene como objetivo anticipar el desarrollo de la pandemia a través de un modelo predictivo fundamentado en datos históricos. Se centra particularmente en la predicción de dos aspectos cruciales: los nuevos casos y las nuevas defunciones. Se detalla el proceso de identificación y visualización del conjunto de datos, así como la metodología utilizada para comprender y preparar los datos para su análisis.

**Introducción**

La pandemia del COVID-19, declarada por la OMS como una emergencia sanitaria mundial, ha impuesto desafíos sin precedentes a los sistemas de salud en todo el mundo. El SARS-CoV-2, causante de esta pandemia, ha ejercido una presión abrumadora sobre las infraestructuras médicas, evidenciando la necesidad urgente de estrategias efectivas para controlar su propagación. A pesar de los avances en la producción de vacunas, la distribución equitativa y el seguimiento continuo de la evolución de la pandemia siguen siendo aspectos críticos en la lucha contra este virus altamente contagioso.

Con el objetivo de prever la trayectoria futura de la pandemia, este estudio se centra en el desarrollo de un modelo predictivo sólido que pueda estimar la evolución de la pandemia en diferentes países. Este enfoque puede ser muy útil para la planificación de recursos de salud, y la preparación para posibles cambios en la propagación del virus.

**Metodología**

Para el desarrollo de este proyecto se definieron diferentes etapas para garantizar el buen análisis. Las cuales se pueden definir como

1. Identificación y visualización del conjunto de datos.
2. Exploración descriptiva del dataset.
3. Iteraciones de desarrollo.
   3.1 Preprocesado de datos.
   3.2 Modelos supervisados.
   3.3 Modelos no supervisados.
   3.4 Resultados, métricas y curvas de aprendizaje.
4. Retos y consideraciones de despliegue.
5. Conclusiones.

**Desarrollo y análisis**

1. **Identificación y visualización de dataset**

Se generó un código para automatizar el proceso de descarga y preparación de los datos sobre COVID-19 que vamos a utilizar en este estudio. En este código se realizan una serie de acciones para descargar, descomprimir y visualizar un conjunto de datos relacionados desde Kaggle. En primer lugar, se configura el archivo kaggle.json para poder acceder a Kaggle y descargar el conjunto de datos. Luego, se instalan algunas bibliotecas útiles, como numpy, matplotlib y pandas. A continuación, se crea un directorio .kaggle en el directorio de inicio y se copia el archivo kaggle.json allí, asegurando que tenga los permisos adecuados. Luego, se descarga un conjunto de datos llamado covid19-dataset desde Kaggle y se descomprime. Finalmente, se lee el archivo CSV resultante llamado 'owid-covid-data.csv' en un DataFrame llamado dataset y se muestra su contenido, el cual contiene 166326 filas y 67 columnas.

2. **Exploración descriptiva del dataset**

Nuestro dataset contiene información sobre los casos de covid 19 recopilados alrededor de algunos países del mundo. Algunas de las columnas que nos llamaron la atención fueron las siguientes:

- location: Contiene el país a donde está asociado el registro
- total_cases: Contiene el conteo de casos positivos totales hasta el momento
- new_cases: Contiene la información de nuevos casos de Covid reportados.
- total_deaths: Contiene la información de las muertes causadas por el covid hasta ese día.
- new_deaths: Contiene la información de las nuevas muertes reportadas en esa fecha.
- icu_patients: Contiene el conteo de pacientes en UCI hasta ese día.
- hosp_patients: Contiene el conteo de pacientes que han sido hospitalizados hasta el momento.
- people_vaccinated: Contiene el conteo de personas que han sido vacunadas hasta el momento.
- new_vaccinations: Contiene el número de nuevos vacunados ese día.
- date: Contiene la fecha del registro.

Sin embargo, new_cases y new_deaths son las columnas de especial interés en nuestro estudio.

En la segunda parte, el objetivo principal fue comprender cómo es la distribución de la información y cómo se comportan los datos.

Para ello se muestran las columnas disponibles en el conjunto de datos y se identifican los países involucrados en los registros. Luego, se cuenta el número de registros por país para determinar cuál país tiene la mayor cantidad de datos y se seleccionan los países que tiene la mayor cantidad de datos registrados en latino américa para su estudio.

Además, para entender mejor cómo se comportan los datos y cuánta información contiene, se llevaron a cabo mediciones estadísticas media, mediana, desviación estándar) para las columnas relacionadas con el total de muertes, total de casos y nuevos casos. Se generan histogramas para visualizar la distribución de estos datos.

Estas pruebas proporcionaron información valiosa sobre la naturaleza de los datos, como si siguen una distribución específica, si existen correlaciones significativas entre variables, entre otros aspectos. Este análisis inicialmente reveló que Argentina y México tenían el mayor número de registros, lo que sugiere que estos dos países podrían ser candidatos adecuados para un análisis más profundo en el proyecto.

## 3. Iteraciones de desarrollo

Dado que nuestro proyecto aborda dos variables objetivo, específicamente new_cases y new_deaths, hemos optado por llevar a cabo un procesamiento de datos individual y separado para cada caso. La decisión de adoptar este enfoque se basa en la aspiración de obtener resultados óptimos al implementar modelos y evaluar las métricas pertinentes. La confianza reside en que realizar un procesamiento de datos centrado en la variable objetivo para cada una de las metas proporcionará una limpieza de datos que se ajusta a la naturaleza de lo que se pretende pronosticar en cada escenario, lo cual, a su vez, se traducirá en un rendimiento mejorado de los modelos.

### 3.1 Preprocesado de datos (new_cases como variable objetivo)

La primera fase del procesado se enfocó en la identificación de datos faltantes. Para abordar esta cuestión, se creó un código que calculó la cantidad de valores ausentes en cada columna del conjunto de datos. Durante este proceso, se identificaron columnas con una cantidad significativa de datos nulos, llegando incluso a registrar hasta 160,630 valores faltantes. La magnitud de estos valores ausentes en un conjunto de datos que consta de 166,326 filas resalta la relevancia de resolver esta problemática de datos incompletos.

Prosiguiendo con el análisis, se exploró la distribución de los datos correspondientes a la variable objetivo seleccionada y se constató que esta distribución no sigue una forma gaussiana. Asimismo, se examinaron los tipos de datos representados en cada columna del conjunto de datos. Se profundizó en el análisis de las columnas numéricas, evaluando estadísticas clave como la media, la desviación estándar y otros parámetros relevantes.
Además, se exploró la correlación entre las distintas columnas para obtener una comprensión más clara de las relaciones entre las variables. Este análisis buscaba esclarecer la influencia de estas relaciones en el conjunto de datos y en las predicciones futuras.

Después de completar la fase inicial de reconocimiento de datos, avanzamos hacia la etapa de limpieza y preparación de los datos. En este proceso, se aplicó una transformación logarítmica a la columna de new_cases y se procedió a la codificación de las columnas no numéricas para facilitar su utilización en la implementación posterior de modelos. Se abordaron las columnas con datos faltantes mediante la aplicación de diversas estrategias:

- Eliminación de columnas con más del 50% de datos faltantes.
- Exclusión de filas en las que la columna objetivo (new_cases) presentara valores nulos.
- Remoción de NaN en las columnas categóricas, sustituyéndolos por la moda.
- Eliminación de columnas cuya correlación con "new_cases" fuera inferior a 0.1.

- Relleno de los valores faltantes en las columnas numéricas mediante la media.

  Estas medidas fueron implementadas con el objetivo de garantizar la calidad y coherencia de los datos para la construcción eficaz de modelos predictivos.

### 3.1.2 Modelo 1 supervisado Random Forest Regressor (new_cases como variable objetivo)

En la evaluación del rendimiento de nuestro modelo predictivo supervisado, empleamos diversas métricas, entre las que se incluyen MAE, MSE, RMSE, R2 Score, MAE Ratio y RMSE Ratio. El modelo supervisado seleccionado para este análisis fue el Random Forest Regressor y se encontraron los mejores hiperparametros. Este algoritmo de aprendizaje automático se basa en la construcción de múltiples árboles de decisión durante el entrenamiento y la fusión de sus predicciones para obtener una estimación más precisa y robusta. Al implementar el Random Forest Regressor, buscamos aprovechar su capacidad para manejar conjuntos de datos complejos y proporcionar resultados precisos en la predicción de variables numéricas, como es el caso de nuestro proyecto.

### 3.1.3 Modelo 2 no supervisado PCA + Random forest (new_cases como variable objetivo)

En la evaluación del rendimiento de nuestro modelo, que combina un algoritmo no supervisado con un algoritmo predictivo, nos centramos en la métrica del mejor RMSE, buscando minimizar su valor. Esta elección se fundamenta en la importancia de obtener predicciones precisas y fiables en nuestro contexto. Además, durante este proceso, encontramos los mejores hiperparametros para lograr un rendimiento optimizado. En particular, utilizamos un modelo que integra técnicas no supervisadas y predictivas, buscando aprovechar las fortalezas de ambos enfoques para obtener resultados más robustos y precisos.

### 3.1.4 Resultados, métricas y curvas de aprendizaje.

**Para Modelo 1 (Random Forest Regressor, Target: new_cases)**

- Los resultados obtenidos de las métricas se muestran en la siguiente tabla:

| Modelo | MAE | MSE | RMSE | R2 Score | MAE Ratio | RMSE Ratio |
|---|---|---|---|---|---|---|
| Random Forest Regressor | 5758.6728 | 1.1390e+09 | 33749.1965 | 0.8474 | 0.4773 | 2.7972 |

```
RMSE Test:  31315.47500 (± 2236.94892020 )
RMSE Train:  27998.81903 (± 1123.09450975 )
```

Se observa que el valor del R2 Score no es considerablemente alto, y, por otro lado, el valor del RMSE no muestra una reducción, lo cual es lo contrario a lo que se espera.

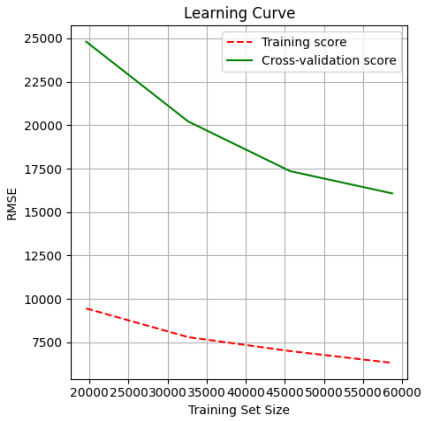
- Mejores hiperparametros:

| max_depth | min_samples_leaf | n_estimators | random_state |
|---|---|---|---|
| 17 | 1 | 60 | 23 |

```
RMSE del Random Forest en entrenamiento:  5376.41572
RMSE del Random Forest seleccionado:  239827233.32266
```

Un RMSE alto en el conjunto de evaluación indica que el modelo no está prediciendo bien los resultados en datos que no ha visto durante el entrenamiento.
El valor extremadamente alto en este caso (239827233.32266) sugiere que el modelo podría estar enfrentando problemas significativos al generalizar a datos de prueba. Es posible que haya un problema de overfitting.

- Curva de aprendizaje:




En la curva de aprendizaje, se observa que el rendimiento del RMSE en el conjunto de entrenamiento mejora a medida que aumenta el tamaño de los datos. Esta tendencia también se refleja en la curva correspondiente al conjunto de prueba o validación. Este comportamiento es alentador, indicando que el modelo presenta menor error en las predicciones a medida que se incrementa el tamaño de los datos. Esta mejora es evidente hasta el dato 60,000, como se muestra en la gráfica. Se espera que al alcanzar el dato 163,133, que representa la totalidad de las filas, el modelo exhiba un RMSE considerablemente reducido.

Sin embargo, es importante señalar que existe la posibilidad de overfitting, y se sugiere considerar modelos menos complejos en futuras predicciones.

**Para Modelo 2 (PCA + Random forest, Target: new_cases)**

- Los resultados obtenidos de las métricas se muestran en la siguiente tabla:

| Modelo | RMSE 1 elemento | RMSE 3 elementos | RMSE 5 elementos | RMSE 7 elementos | RMSE 9 elementos |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| PCA + Random forest | 18302.5301 | 19018.1254 | 19441.6721 | 12472.0221 | 14717.5879 |

- Mejores hiperparametros:

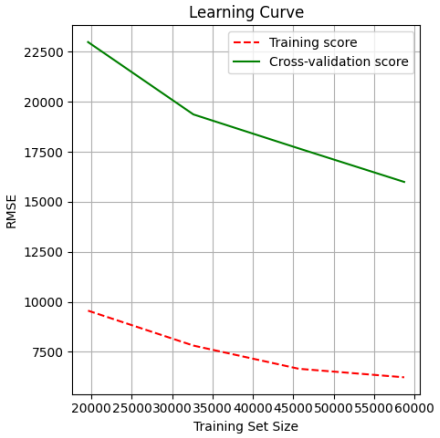
  PCA

  ```
  Mejor RMSE:  12472.02219   ; obtenido con  7 componentes
  ```

  Mejor Random Forest para PCA obtenido

| max_depth | n_estimators | random_state |
|---|---|---|
| 19 | 60 | 23 |

- Curva de aprendizaje:




En la evolución de la curva de aprendizaje, se nota que a medida que se amplía el conjunto de entrenamiento, el desempeño del RMSE mejora. Este patrón también se evidencia en la curva correspondiente al conjunto de prueba o validación. Esta dinámica es prometedora, ya que sugiere que el modelo disminuye su margen de error en las predicciones a medida que se incrementa la cantidad de datos. Este progreso es notable hasta el punto de 60,000 datos, según se aprecia en el gráfico. Se anticipa que al llegar al dato 163,133, que representa la totalidad de las filas, el modelo mostrará un RMSE significativamente reducido.

No obstante, se destaca la importancia de tener en cuenta la posibilidad de overfitting, y se recomienda considerar modelos menos complejos en futuras predicciones.

### 3.2. Preprocesado de datos (new_deaths como variable objetivo)

Durante la elaboración de los modelos se llegó a dos tipos de preprocesados distintos, según las necesidades, es decir para la clasificación o la regresión, por lo cual es necesario procesar la información de forma distinta.
El preprocesado para regresión consistió en completar los campos que presentaban información NaN, para eso se decide eliminar las filas que no tuvieran el dato que se esperaba predecir ("new_deaths"), para posteriormente completar los datos faltantes teniendo en cuenta el promedio.

Posteriormente se establecen gráficos de dispersión y una matriz de correlación que permite establecer una relación entre atributos con el fin de establecer cuáles características podrían generar mejores resultados en el modelo de regresión. Es de resaltar que se crea un atributo sintético con el fin de establecer una relación entre algunos atributos de entrada y la salida de interés.

Se tuvieron en cuenta diferentes factores como:

**I. Establecer el sistema de clasificación:**

Se estableció cómo sería el sistema de clasificación, el cual constaría de 3 niveles que son:

- Riesgo elevado (3): Donde se indica que el estado de la pandemia es altamente peligroso.
- Riesgo moderado (2): Donde indica que el estado de la pandemia es peligroso pero que no representa un riesgo elevado, sin embargo es pertinente realizar monitoreos y prepararse para posibles escaladas en el riesgo.
- Riesgo bajo (1): Donde indica que el estado de la pandemia no es sustancialmente peligroso y se cuentan con las herramientas para combatirlo eficientemente.

**II. Establecer cómo se puede llenar el sistema de clasificación:**

El siguiente paso consistió en definir la base del sistema de clasificación global y su proceso de llenado para generar el coeficiente de clasificación. Se propuso establecer la variable de clasificación general en función de variables de clasificación parciales. Las variables parciales, junto con sus definiciones matemáticas y métricas de evaluación, son las siguientes:

**A. Porcentaje de nuevos contagios:**

Esta variable cuantifica la proporción de nuevos contagios en relación con la población total del país. Matemáticamente se define como:

$$Métrica\ porcentual = \frac{\#\ nuevos\ casos}{\#\ Población} * 100 * factor\ de\ escalado$$

Los nuevos casos corresponden al valor de la columna 'new_cases' y la población al valor de 'population'. Se utilizó un factor de escalado de 1000 para amplificar la métrica y facilitar la clasificación.

Luego de obtener el valor porcentual de la métrica el siguiente paso fue entonces clasificarla en 3 niveles, de forma similar a la métrica global de clasificación, sin embargo aquí se usaron umbrales para establecer estos límites, por ello entonces la clasificación fue la siguiente:

- ➔ Riesgo elevado (3): Si el valor de la métrica porcentual es mayor al 20%.
- ➔ Riesgo moderado (2): Si el valor de la métrica porcentual es mayor al 5% pero menor o igual al 20%.
- ➔ Riesgo bajo (1): Si el valor de la métrica porcentual es menor o igual al 5%

**A. Porcentaje de mortalidad:**

Esta variable busca cuantificar la letalidad de la enfermedad utilizando datos de nuevas muertes y nuevos casos. Matemáticamente se define como:

$$Métrica\ porcentual = \frac{\#\ nuevas\ muertes}{\#\ Nuevos\ casos} * 100$$

Las nuevas muertes corresponden al valor de la columna 'new_deaths' y los nuevos casos al valor de 'new_cases'.

Esta métrica también se clasificó de acuerdo al nivel de amenaza de la siguiente forma:

- ➔ Riesgo elevado (3): Si el valor de la métrica porcentual es mayor al 10%.
- ➔ Riesgo moderado (2): Si el valor de la métrica porcentual es mayor al 3% pero menor o igual al 10%.
- ➔ Riesgo bajo (1): Si el valor de la métrica porcentual es menor o igual al 3%

### 3.2.1 Modelos supervisados con new_deaths como variable objetivo

#### 3.2.1.1 Primer modelo: Regresión lineal (new_deaths)

La regresión lineal, empleada tanto en Machine Learning como en estadística, ofrece una aproximación para modelar la relación entre una variable escalar dependiente “y” y una o más variables explicativas, denotadas como “X” [2]. En este contexto, la variable dependiente es la cantidad de nuevas muertes en un pais, en este primer caso fue tomado Argentina ("new_deaths"). Las variables explicativas consideradas son: "total_cases", "new_cases", "total_deaths", "people_vaccinated_per_hundred", "people_fully_vaccinated_per_hundred", "total_vaccinations", "people_vaccinated", "people_fully_vaccinated", "new_cases_per_million", y un atributo sintético definido como "death_rate".

Es crucial destacar que para construir el modelo, se divide el conjunto de datos en dos segmentos: entrenamiento y prueba. Inicialmente, se realiza una separación utilizando el 80% de los datos para el entrenamiento y el 20% restante para la validación o prueba. Posteriormente, se implementa la función bootstrap para dividir los datos de manera adicional. Esta técnica de remuestreo implica extraer muestras repetidas del conjunto de datos original con reemplazo, asignando el 30% de las muestras a los datos de prueba.

#### 3.2.1.2 Segundo modelo: Máquina de soporte vectorial (new_deaths)

Las máquinas de soporte vectorial (SVM) son una serie de algoritmos de aprendizaje supervisado, enfocados usualmente con problemas que involucran clasificación y regresión. Se alimenta de datos etiquetados en clases, con el objetivo de construir un modelo capaz de predecir la clase de una nueva muestra. Una SVM presenta los datos como puntos en el espacio, separando las clases por medio de hiperplanos que se definen mediante vectores (de aquí el nombre), con los cuales buscan la correspondencia del valor con un grupo o no [5].

#### 3.2.1.3 Resultados del primer modelo: Regresión lineal (new_deaths)

Mediante este modelo se obtuvieron las siguientes dos resultados importantes:

1. Métricas obtenidas al dividir el set de datos sin utilizar la función bootstrap:

| Error medio absoluto | Suma residual de los cuadrados (MSE) | R2- score | Promedio de validación |
|---|---|---|---|
| 439.35 | 627055.12 | -0.22 | -80.9822 |

2. Métricas obtenidas al dividir el set de datos al implementar la función bootstrap:

| Error medio absoluto | Suma residual de los cuadrados (MSE) | R2- score |
|---|---|---|
| 86.1925 | 12642.2865 | 0.187 |

De acuerdo a estos resultados se puede observar que la implementación de la función bootstrap mejoró significativamente las métricas de rendimiento del modelo. La reducción en el error medio absoluto y la suma residual de los cuadrados indica una mejora en la precisión y en la capacidad del modelo para ajustarse a los datos. Sin embargo, el coeficiente R2-score sigue siendo bajo en ambos casos, lo que sugiere que el modelo puede necesitar mejoras adicionales para explicar mejor la variabilidad en los datos observados. Sin embargo el valor de R2-score con la función bootstrap es de valor mayor y positivo.

Además de analizar estos valores es necesario conocer la estabilidad del modelo, lo cual se puede observar mediante una curva de aprendizaje como se muestra en la imagen 1.

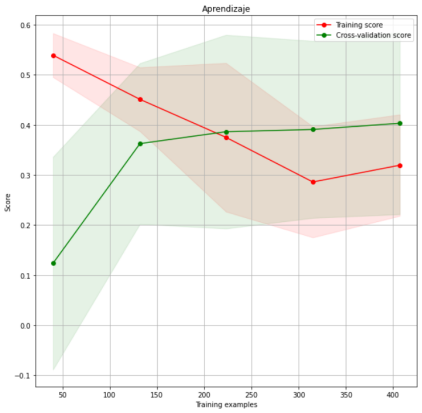



**Imagen 1. Curva de aprendizaje del primer modelo**

Se puede evidenciar sesgo, ya que los modelos no tienden a presentar una convergencia adecuada, es decir los datos no presentan una mejora en los datos de validación del modelo utilizado y los datos de entrenamiento tienden a disminuir.

#### 3.2.1.4 Resultados del segundo modelo: máquina de soporte vectorial

Se generó una iteración, donde el algoritmo iría evaluando el valor de C, de 1.0 a 5.0 con saltos de 0.1, y variando el tipo de gamma como auto o cómo scale. El dataset resultante es el siguiente:

| | C | Gamma | Precision train | Recall train | F1-Score train | Precision test | Recall test | F1-Score test |
|---|---|---|---|---|---|---|---|---|
| **0** | 1.0 | auto | 0.998081 | 0.998081 | 0.998081 | 0.658438 | 0.454062 | 0.383975 |
| **1** | 1.0 | scale | 0.643716 | 0.376579 | 0.258463 | 0.603321 | 0.358560 | 0.237200 |
| **2** | 1.1 | auto | 0.997142 | 0.997103 | 0.997121 | 0.665455 | 0.458629 | 0.391377 |
| **3** | 1.1 | scale | 0.643716 | 0.376579 | 0.258463 | 0.603321 | 0.358560 | 0.237200 |
| **4** | 1.2 | auto | 0.997142 | 0.997103 | 0.997121 | 0.665455 | 0.458629 | 0.391377 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **77** | 4.8 | scale | 0.638677 | 0.380292 | 0.276592 | 0.597715 | 0.356419 | 0.253479 |
| **78** | 4.9 | auto | 0.999064 | 0.999022 | 0.999042 | 0.665455 | 0.458629 | 0.391377 |
| **79** | 4.9 | scale | 0.638677 | 0.380292 | 0.276592 | 0.597715 | 0.356419 | 0.253479 |
| **80** | 5.0 | auto | 0.999064 | 0.999022 | 0.999042 | 0.665455 | 0.458629 | 0.391377 |
| **81** | 5.0 | scale | 0.638677 | 0.380292 | 0.276592 | 0.597715 | 0.356419 | 0.253479 |

82 rows × 8 columns

**Imagen 2. Dataset de configuraciones del modelo máquina de soporte vectorial.**

Además se extrajo el mejor modelo de forma sintética, arrojando entonces el siguiente resultado:

| | C | Gamma | Precision train | Recall train | F1-Score train | Precision test | Recall test | F1-Score test |
|---|---|---|---|---|---|---|---|---|
| **10** | 1.5 | auto | 0.999064 | 0.999022 | 0.999042 | 0.665455 | 0.458629 | 0.391377 |

**Imagen 3. Mejor configuración de parámetros del modelo máquina de soporte vectorial.**

Significa entonces que el mejor modelo es aquel con un valor de C=1.5 y un Gamma calculado automáticamente. Las matrices de confusión son las siguientes:

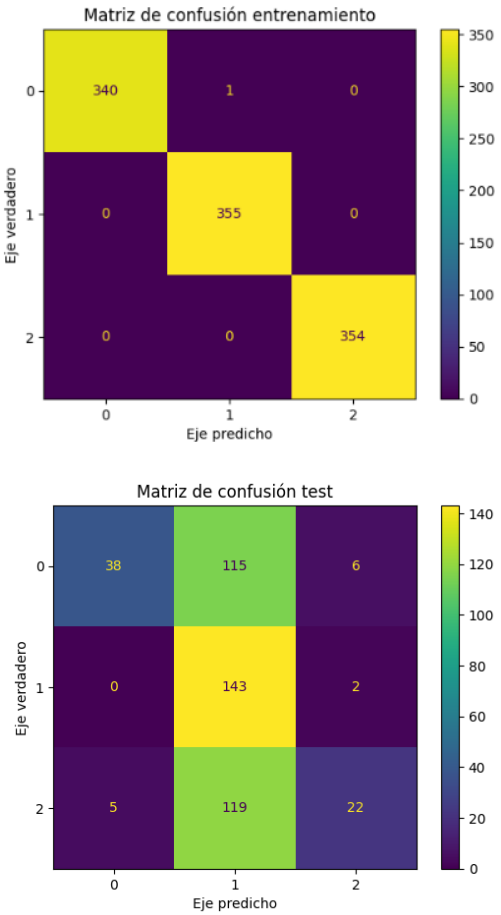



**Imagen 4 y 5. Matrices de confusión modelo máquina de soporte vectorial.**

Como se observa, parece que el modelo aprende de memoria los datos pues la matriz de confusión de entrenamiento es perfecta, sin embargo, al aplicar los datos de test, se evidencia que el algoritmo realmente no está siendo capaz de clasificar correctamente los datos, y que se encuentra sesgado, pues la gran parte de sus clasificaciones las hace como un riesgo intermedio (reflejado en la matriz con el valor 1 en los ejes). Esto puede evidenciar que la diferencia entre los distintos grupos no es tan fácilmente discriminable por un algoritmo de este tipo. La curva de aprendizaje que se obtiene es la siguiente:

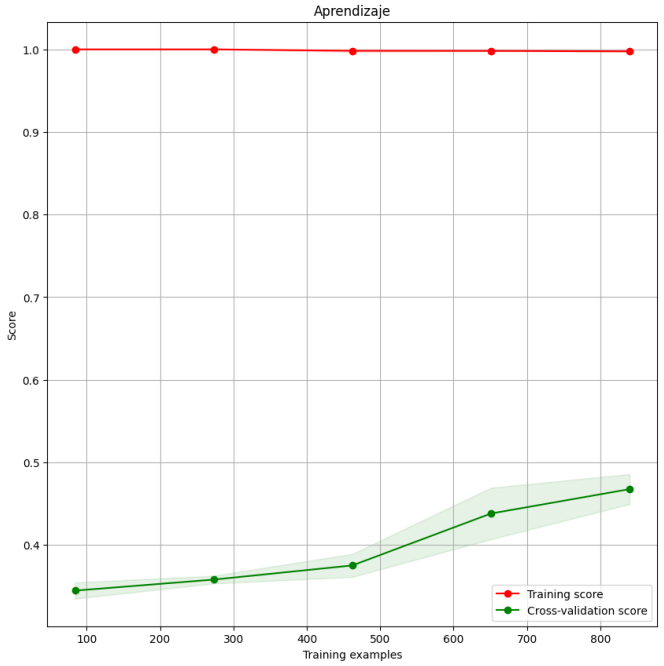



**Imagen 6. curva de aprendizaje del modelo máquina de soporte vectorial.**

La curva de aprendizaje para este modelo nos indica un sesgo, por lo que el modelo hizo suposiciones sobre los datos de entrenamiento. Esto conduce a una simplificación excesiva del modelo y puede provocar un error elevado tanto en los conjuntos de entrenamiento como de prueba. Sin embargo, esto también hace que el modelo sea más rápido de aprender y fácil de entender.

**Retos y consideraciones de despliegue**

- Se encontró que no todos los problemas son fácilmente clasificables, pues los datos se encuentran mezclados y lo que podría suponer una tarea medianamente fácil de clasificación para un humano, puede convertirse en una tarea compleja incluso para una máquina.
- Para desplegar un algoritmo de este tipo es necesario validar pertinentemente los modelos, de forma que los resultados que arrojen sean consecuentes con los reales, pues de esto supone entonces la fiabilidad y confianza del sistema en cuestión.
- Abordar la eficiencia computacional y el manejo de grandes conjuntos de datos para garantizar un rendimiento óptimo del modelo, especialmente en situaciones de escalabilidad.
- Investigar la inclusión de características temporales más avanzadas, como patrones estacionales o eventos específicos, para mejorar la capacidad predictiva del modelo en el contexto de la evolución temporal de la pandemia.

**Conclusiones**

Durante el desarrollo de nuestro modelo predictivo para la evolución de la pandemia del COVID-19, hemos subrayado la importancia crítica de un riguroso preprocesamiento de datos. La selección de variables objetivo, 'new_cases' y 'new_deaths' en nuestro caso, ha demandado enfoques distintos para optimizar la precisión de las predicciones.

Aunque los modelos tienen limitaciones, hemos identificado áreas para mejoras futuras. La incorporación de datos externos y la exploración de enfoques multivariados podrían enriquecer significativamente el modelo. Es esencial comprender que la construcción de modelos predictivos es un proceso iterativo que requiere ajustes continuos y actualizaciones. A pesar de estas limitaciones, el modelo puede proporcionar información valiosa para respaldar la toma de decisiones informada. La comunicación transparente de los resultados y la consideración de las incertidumbres son fundamentales para una correcta interpretación y aplicación en el contexto de la evolución de la pandemia.

**Referencias**

[1] COVID-19 dataset. (s.f.). Kaggle: Your Machine Learning and Data Science Community. Tomado de: https://www.kaggle.com/datasets/georgesaavedra/covid19-dataset

[2] ¿Cómo sé si mi modelo de predicción es realmente bueno?. Tomado de: https://datos.gob.es/es/blog/como-se-si-mi-modelo-de-prediccion-es-realmente-bueno

[3] Métricas De Evaluación De Modelos En El Aprendizaje Automático. Tomado de: https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-elaprendizaje-automatico

[4] Interpretabilidad de los modelos de Machine Learning. Tomado de: https://quanam.com/interpretabilidad-de-los-modelos-de-machine-learning-primera-parte/

[5] Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021). Tomado de: https://ourworldindata.org/covid-vaccinations