Universidad de La Habana
Facultad de Matemática y Computación

Estimación de propagación de enfermedades respiratorias agudas y su impacto en los políclinicos de Cuba.

Autor:

José Antonio Concepción Alvarez
José Miguel Zayas Pérez
Dario Rodriguez Llosa
Alejandro Lamelas Delgado
Alejandro Ramírez Trueba
Ana Melissa Alonso Reina

Tutores:

Suilán
Elena

Informe del Proyecto de Machine Learning

Repositorio de Github

Introducción

Este proyecto tiene como objetivo desarrollar modelos predictivos para enfermedades respiratorias en las diferentes provincias de Cuba, utilizando técnicas avanzadas de Machine Learning. La información analizada proviene de los registros hospitalarios de los policlínicos de la isla, segmentados por provincias.

El conjunto de datos proporcionado abarca el período comprendido entre el 19 de septiembre de 2022 y el 15 de diciembre de 2023. Los datos están organizados por días, lo que permite un análisis detallado de la evolución de los ingresos hospitalarios. Sin embargo, se identificaron importantes problemas de datos faltantes, obteniendo inicialmente solo 142 registros de un total esperado de 454. En determinados periodos, la información es extremadamente escasa, lo que dificulta significativamente el entrenamiento de los modelos. Por ejemplo, en noviembre de 2022 solo se dispone de un día de datos y en diciembre del mismo año, únicamente dos días.

Estado del Arte

Se llevó a cabo una revisión bibliográfica de múltiples artículos científicos relacionados con el análisis de series temporales aplicadas a la salud, con un enfoque particular en enfermedades respiratorias. Los estudios revisados permitieron identificar los modelos más utilizados y con mejores resultados en contextos similares: ARIMA, XGBoost y Redes Neuronales Recurrentes (RNN).

Detalles de Implementación y Experimentos

Metodología

El equipo de trabajo se dividió en grupos para analizar cada uno de los tres modelos mencionados. Los resultados iniciales fueron insatisfactorios (detallados en las secciones correspondientes a cada modelo), lo que motivó la necesidad de mejorar la calidad del conjunto de datos mediante la ampliación y generación de información adicional.

Generación de Datos

Dado el alto nivel de datos faltantes, se exploraron distintos métodos para completar la información. El cliente sugirió el uso de **Interpolación**, la cual permite estimar valores intermedios de manera continua. Sin embargo, este método mostró limitaciones, ya que genera tendencias artificiales de crecimiento o decrecimiento constantes, las cuales no reflejan la realidad en escenarios donde las fluctuaciones pueden ser abruptas.

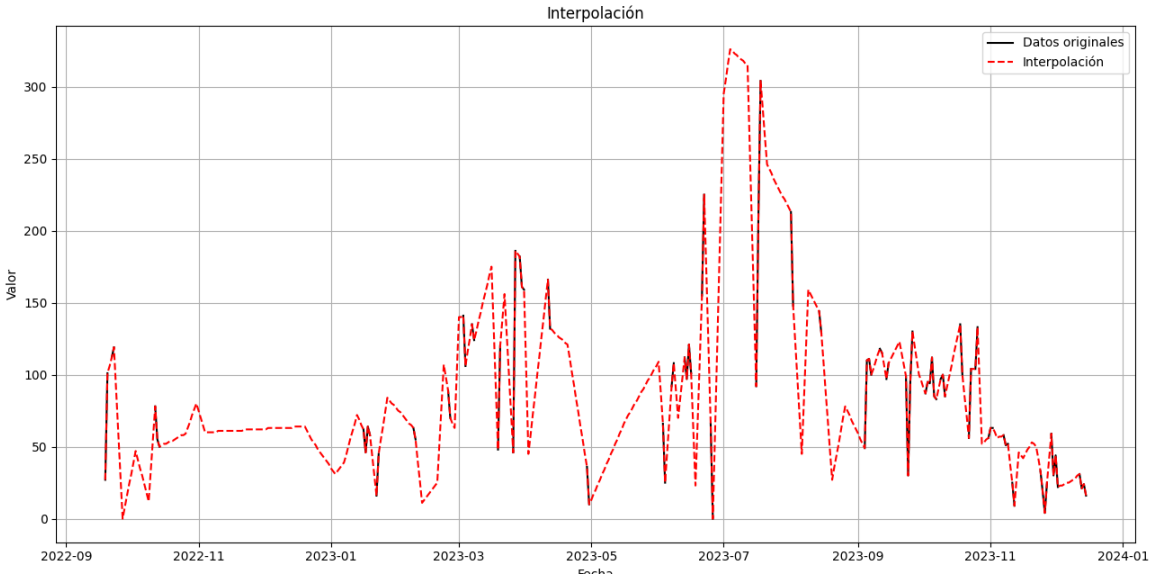

Figura 1: Datos completados con Interpolación.

Posteriormente, se probó el método de **Regresión Lineal**, obteniendo igualmente resultados insatisfactorios debido a la naturaleza no lineal de los datos, los cuales presentan numerosos picos y variaciones bruscas.

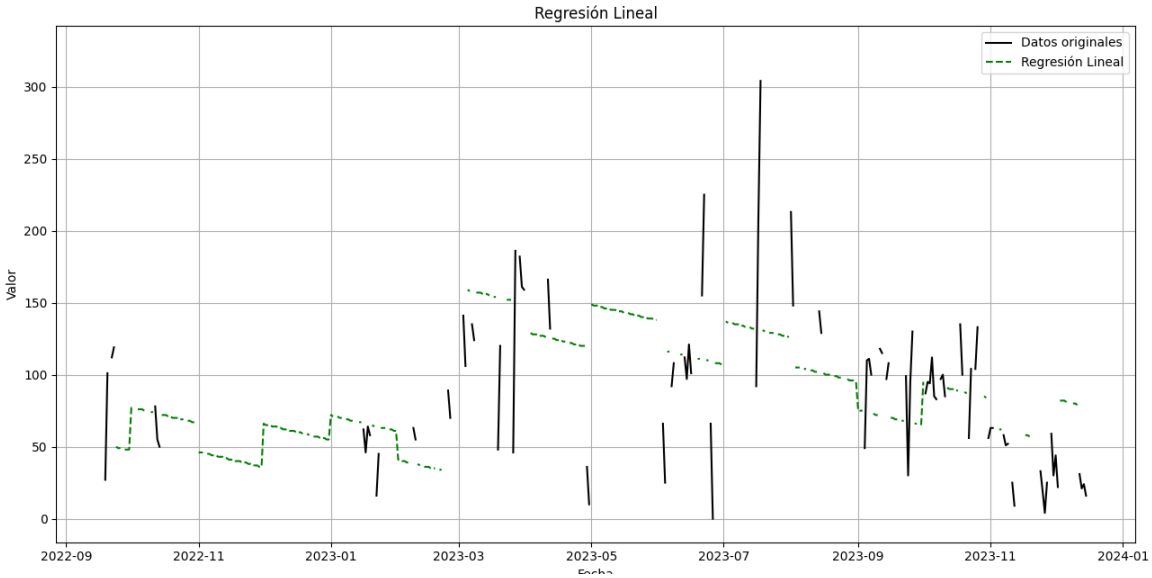

Figura 2: Datos generados con Regresión Lineal.

Finalmente, se optó por **Random Forest** como método de generación de datos, obteniendo resultados más realistas. Esta técnica captó de manera más precisa la estructura de la serie temporal, reproduciendo picos y variaciones de manera más dinámica. A pesar de ello, es fundamental destacar que, aunque estos datos permiten mejorar el entrenamiento de los modelos, siguen siendo valores generados y no reales. No obstante, tras un análisis **cualitativo**, se concluyó que **Random Forest** era el método más adecuado para la generación de datos en este contexto.

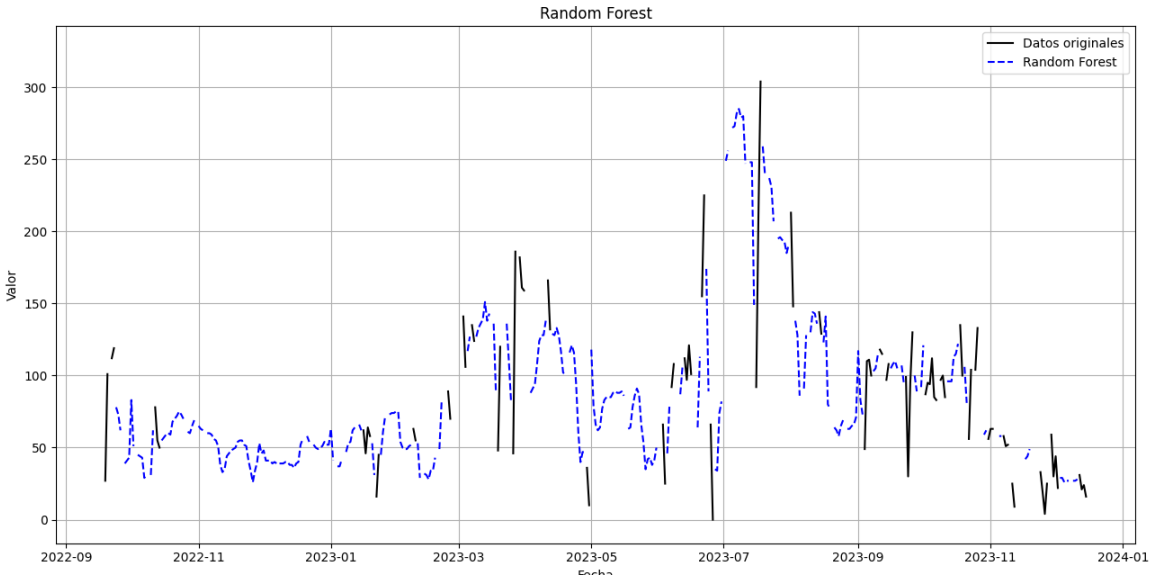

Figura 3: Datos generados con Random Forest.

Evaluación de Modelos

Redes Neuronales Recurrentes (RNN)

Las **RNN** han sido ampliamente utilizadas en estudios previos de predicción en el ámbito de la salud y series temporales. Para su implementación en este proyecto, se prestó especial atención a la selección del hiperparámetro **time steps**, el cual define la cantidad de días previos considerados para predecir el siguiente valor.

Se evaluaron diferentes configuraciones de **time steps**, y se determinó que:

- Valores inferiores a 7 no lograban capturar adecuadamente los patrones temporales.
- Valores cercanos a 30 reducían la capacidad de predicción a corto plazo.
- **El valor óptimo fue 10**, permitiendo capturar patrones semanales y, en cierta medida, variaciones mensuales.

Resultados Iniciales

Debido a la gran cantidad de datos faltantes, los resultados iniciales fueron poco satisfactorios (Provincia de ejemplo Pinar del Río):

- **R^2**: 0.4999
- **RMSE**: 43.3

Después de la generación de datos con Random Forest, los resultados mejoraron notablemente:

- **R^2**: 0.938
- **RMSE**: 15.4

Además, al incrementar el conjunto de entrenamiento (0.9) y reducir el de prueba (0.1), se logró obtener valores de **R^2 de hasta 0.96** .

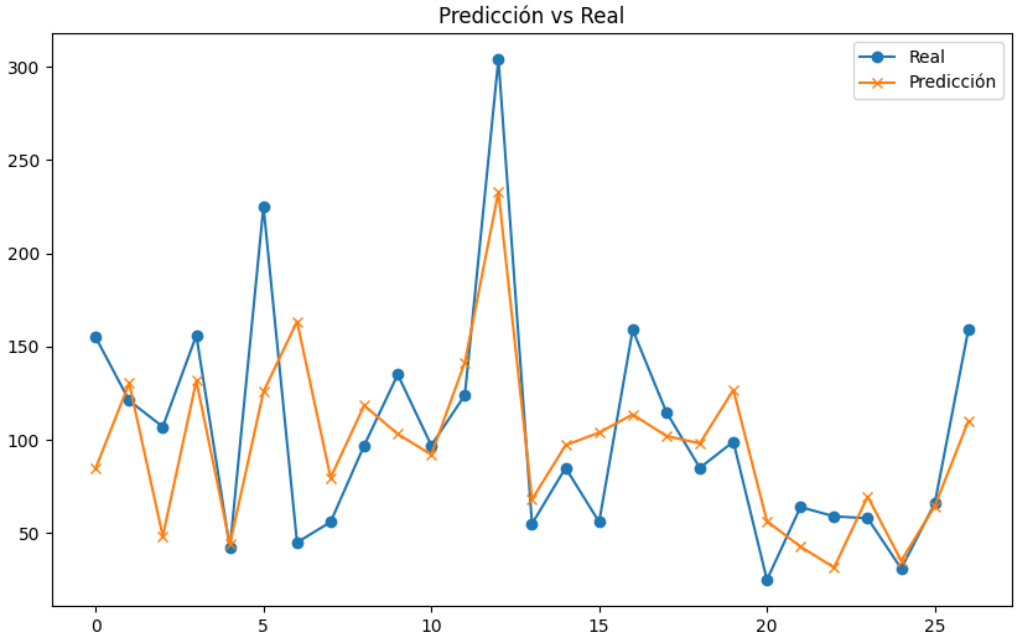

Figura 4: Predicción con datos sin aumentar. Pinar del Río. RNN.

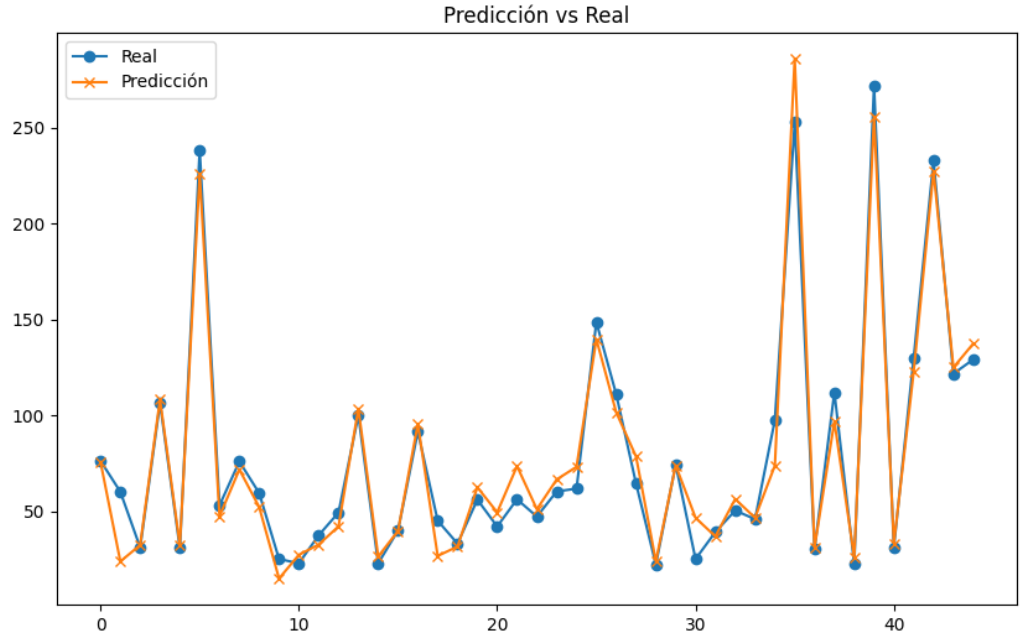

Figura 5: Predicción con datos aumentados. Pinar del Río. RNN.

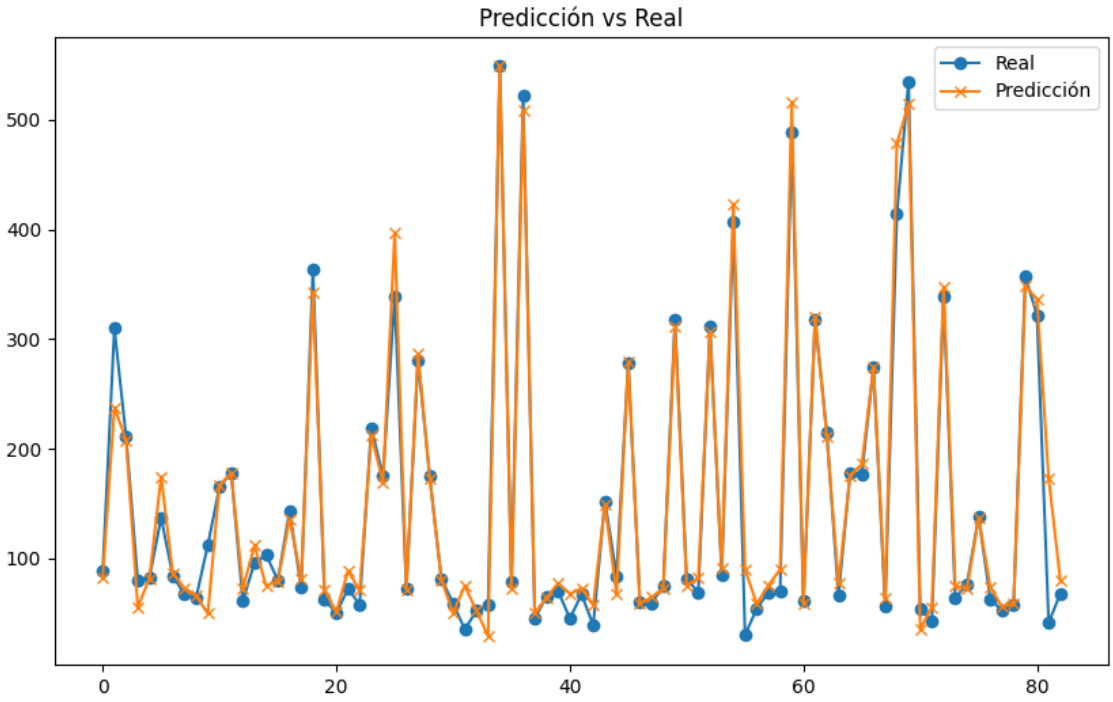

Figura 6: Predicción con datos aumentados. La Habana, RNN. R2: 0.96

Validación y Overfitting

Para descartar problemas de sobreajuste, se realizaron múltiples pruebas, incluyendo:

- Variación en la proporción de entrenamiento/prueba.
- Validación cruzada, **K-Fold**.
- Análisis de curvas de aprendizaje.
- Ajuste de hiperparámetros.

Para el ejemplo de predicción de La Habana (R2: 0.964), los valores de K-Fold fueron los siguientes:

- Fold 1: RMSE=30.3309, R2=0.9521
- Fold 2: RMSE=26.4124, R2=0.9541
- Fold 3: RMSE=23.1644, R2=0.9404

- Fold 4: RMSE=24.4993, R2=0.9575
- Fold 5: RMSE=26.0905, R2=0.9449

Resultados de Validación Cruzada:

- RMSE Promedio: 26.0995
- R2 Promedio: 0.9498

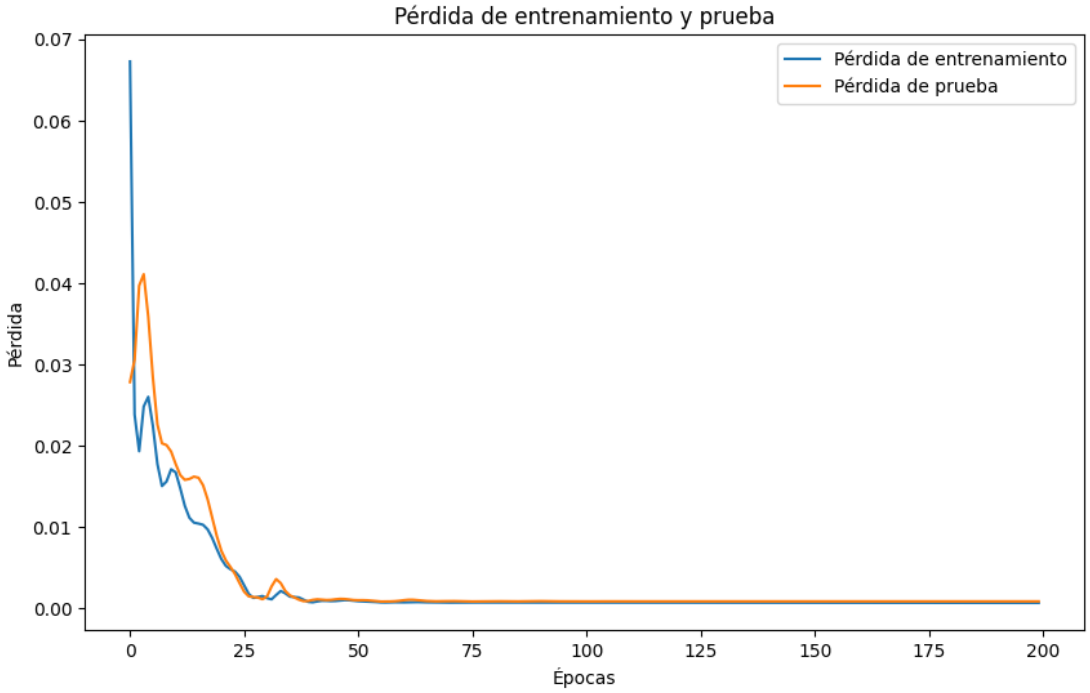

Figura 7: Análisis de curvas de Aprendizaje. La Habana. RNN.

Los resultados indicaron que el modelo se mantenía estable con un **R^2 siempre superior a 0.9** en todas las validaciones diferentes, lo que sugiere una adecuada generalización del aprendizaje.

XGBoost

Dado que **XGBoost** es uno de los modelos más utilizados en el análisis de series temporales, se decidió utilizarlo en este proyecto. Inicialmente, los datos faltantes fueron completados mediante la **media de los valores registrados**.

Los resultados obtenidos para la provincia de Pinar del Río fueron:

- **RMSE**: 28.10
- **MSE**: 789.85
- **MAE**: 19.54

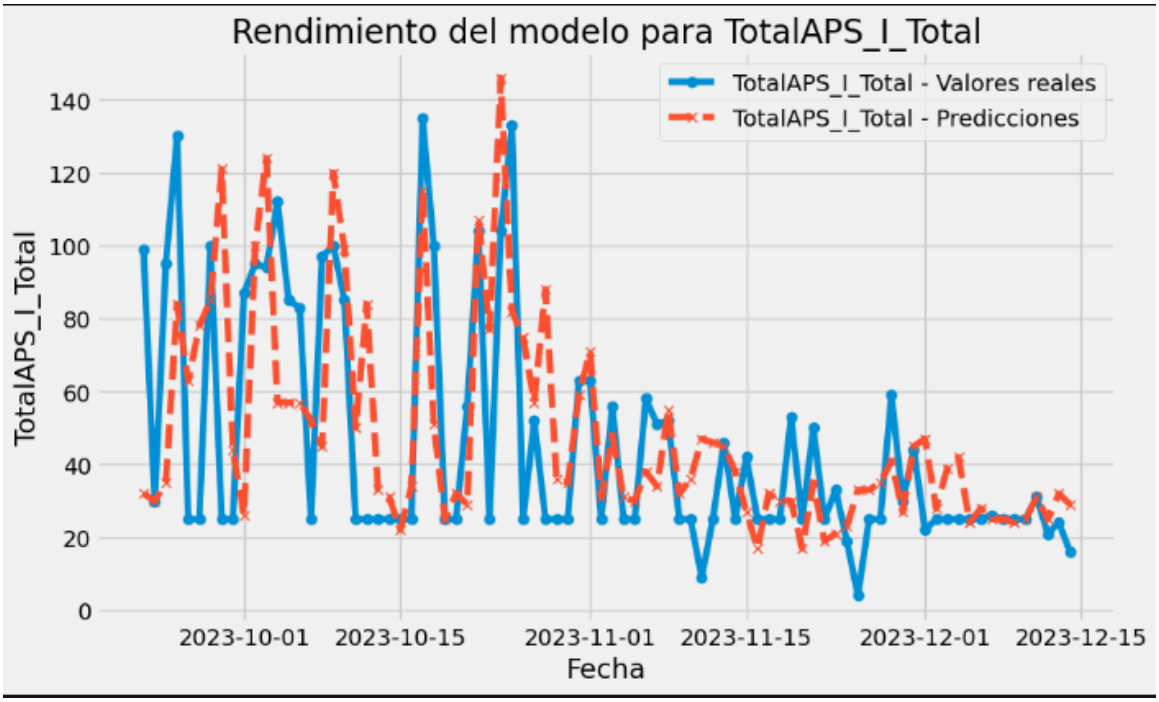

Figura 8: Predicción con datos generados mediante media. XGBoost. Pinar del Río.

Tras la generación de datos con Random Forest, se obtuvo una ligera mejora, pero a manera general los resultados para el resto de las provincias se mantuvieron similares:

- **RMSE**: 25.34
- **MSE**: 642

- **MAE**: 19.64

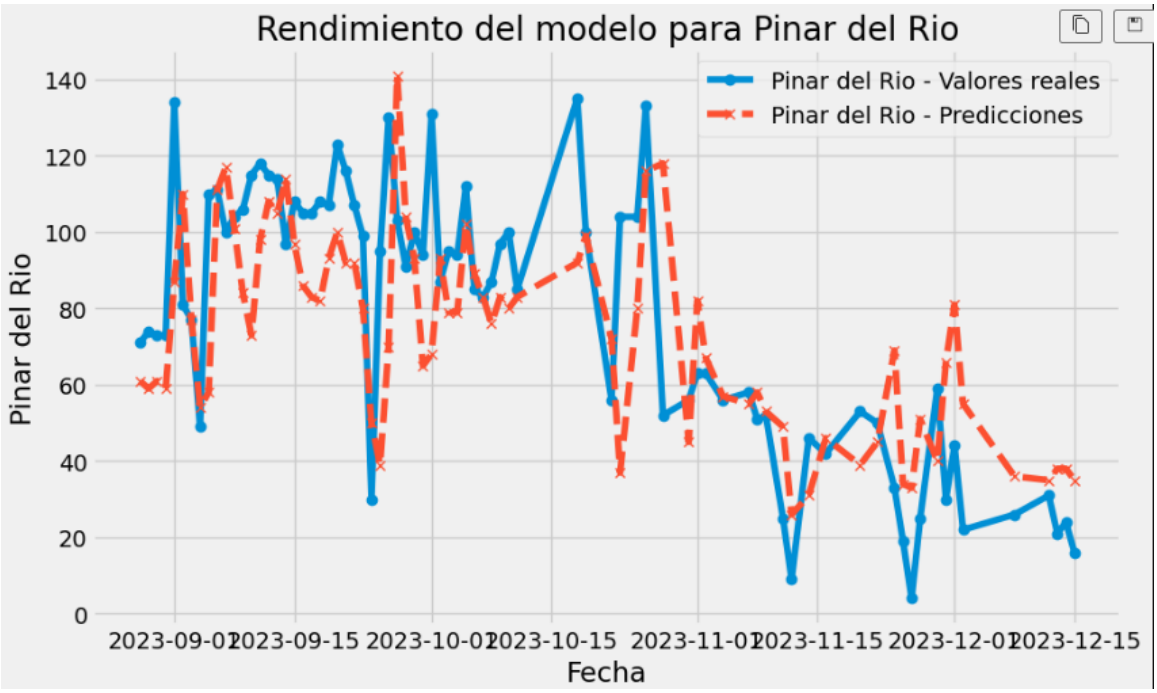

Figura 9: Predicción con datos generados mediante Random Forest. XGBoost. Pinar del Río.

Los gráficos de predicción mostraron diferencias en los valores generados, reflejando los efectos de los distintos métodos de generación de datos.

ARIMA

El modelo ARIMA (AutoRegressive Integrated Moving Average) es una herramienta estadística ampliamente utilizada en la predicción de series temporales. Su capacidad para modelar tendencias, estacionalidades y componentes de ruido lo hace idóneo para una variedad de aplicaciones, incluyendo la predicción de ventas, análisis financiero y control de calidad. En este informe, se describe el proceso de implementación del modelo ARIMA, la selección de hiperparámetros utilizando auto_arima, y la evaluación de los resultados obtenidos.

Fundamentos del Modelo ARIMA

ARIMA combina tres componentes principales:

AutoRegresivo (AR): Utiliza la relación entre una observación y un número de observaciones pasadas.

Integrado (I): Aplica diferenciación para convertir una serie no estacionaria en estacionaria.

Media Móvil (MA): Modela el error de predicción como una combinación lineal de errores previos.

Ecuación de Arima:

$$Y_t = \phi_0 + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

Para determinar el orden óptimo del modelo ARIMA (p, d, q), se emplean gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF), además de criterios de información como AIC y BIC.

Preparación de los Datos

Se identificó que los datos no estaban equiespaciados en el tiempo, ya que las entradas correspondían a fechas aleatorias, lo que generó una distribución irregular de observaciones en cada mes.

El hecho de que los datos no estén uniformemente espaciados en el tiempo afecta la aplicabilidad de ARIMA, dado que este modelo asume intervalos de tiempo constantes entre observaciones. Para mitigar este problema, se utilizó random forest para aumentar los datos.

Tabla Comparativa - AutoARIMA. Datos no Aumentados.

Provincia	**MAE**	**MSE**	**RMSE**	**MAPE (%)**	**NMSE**
Pinar del Rio	50.788120	3116.626374	55.826753	-	2.991508
Artemisa	34.425523	1370.488377	37.020108	-	6.953710
La Habana	95.407571	11346.815229	106.521431	-	17.777240
Mayabeque	36.781495	1636.667272	40.455745	-	10.716257
Matanzas	15.659169	398.042387	19.951000	86.47521	6.327168
Villa Clara	54.973408	4034.028331	63.514001	-	1.339105
Cienfuegos	8.390317	110.032492	10.489637	-	1.008320
Sancti Spiritus	14.810543	388.158822	19.701747	-	0.662101
Ciego de Ávila	48.657051	3016.018973	54.918294	-	2.976254
Camagüey	48.482672	2940.567314	54.226998	-	8.330199
Las Tunas	32.488591	1699.586761	41.226045	-	1.036261
Holguin	62.477623	4927.030856	70.192812	-	2.322323
Granma	49.918573	3443.817983	58.684052	-	1.356636
Santiago de Cuba	102.292822	12596.768564	112.235327	-	3.562211
Guantánamo	23.614520	736.415454	27.136976	-	1.938690
Isla de la Juventud	4.583840	27.087905	5.204604	-	30.080324

Tabla Comparativa - ARIMA. Datos no Aumentados.

Provincia	**MAE**	**MSE**	**RMSE**	**MAPE (%)**	**NMSE**
Pinar del Rio	50.788120	3116.626374	55.826753	-	2.991508
Artemisa	20.353874	574.537131	23.969504	-	2.915139
La Habana	57.367539	4270.585660	65.349718	-	6.690796
Mayabeque	24.672553	714.518512	26.730479	-	4.678388
Matanzas	24.492937	721.150066	26.854237	-	11.463196
Villa Clara	60.402410	4896.865289	69.977606	-	1.625525
Cienfuegos	13.449177	225.580514	15.019338	-	2.067184
Sancti Spiritus	31.369058	1238.003837	35.185279	-	2.111721
Ciego de Ávila	61.156317	4573.242734	67.625755	-	4.512946
Camagüey	19.863965	522.228376	22.852317	-	1.479397
Las Tunas	32.488591	1699.586761	41.226045	-	1.036261
Holguin	52.146951	3768.618076	61.389071	-	1.776313
Granma	49.897348	3441.541706	58.664655	-	1.355739
Santiago de Cuba	54.860379	4383.771074	66.210053	-	1.239676
Guantánamo	16.269390	400.031095	20.000777	-	1.053123
Isla de la Juventud	2.043586	4.847096	2.201612	-	5.382558

Selección de Hiperparámetros con auto_arima

Para encontrar la mejor combinación de hiperparámetros (p, d, q), se utilizó la función auto_arima de la librería pmdarima. Esta herramienta realiza una búsqueda automatizada basada en criterios de información (AIC/BIC) y pruebas estadísticas para seleccionar el modelo más adecuado. Se consideraron los siguientes parámetros:

seasonal=False: Dado que no se observó una estacionalidad clara en los datos.

stepwise=True: Para acelerar la búsqueda utilizando heurísticas inteligentes.

maxiter=50: Para garantizar la convergencia del modelo.

out_of_sample_size=10: Para evaluar el desempeño fuera de muestra.

Evaluación del Modelo

Se evaluaron los modelos obtenidos utilizando múltiples métricas de error:

MAE (Mean Absolute Error): Media de los errores absolutos.

RMSE (Root Mean Squared Error): Raíz cuadrada del error cuadrático medio, útil para penalizar grandes desviaciones.

MAPE (Mean Absolute Percentage Error): Error absoluto medio relativo al valor real, expresado en porcentaje.

RRSE (Relative Root Squared Error): Proporciona una comparación con un modelo de referencia.

Tabla Comparativa - AutoARIMA. Datos aumentados.

Provincia	**MAE**	**MSE**	**RMSE**	**MAPE (%)**	**NMSE**
Pinar del Rio	50.831496	3099.850338	55.676300	118.056924	2.975406
Artemisa	17.113455	396.784468	19.919450	49.594493	2.013241
La Habana	35.388293	1545.761087	39.316168	49.559195	2.421769
Mayabeque	9.090837	164.124535	12.811110	91.576935	1.074623
Matanzas	9.855378	130.566662	11.426577	54.310031	2.075450
Villa Clara	63.817831	5613.377569	74.922477	40.711314	1.863373
Cienfuegos	8.327402	122.446646	11.065561	22.026482	1.122081
Sancti Spiritus	34.619617	1578.323114	39.728115	51.060764	2.692220
Ciego de Ávila	69.216009	5628.460564	75.023067	98.950926	5.554252
Camagüey	17.347852	406.139780	20.152910	163.082097	1.150535
Las Tunas	34.903809	1970.664514	44.392167	147.072576	1.201541
Holguin	43.817711	2636.177490	51.343719	73.298819	1.242545
Granma	48.297528	3723.018198	61.016540	44.916425	1.466622
Santiago de Cuba	51.288417	3827.663700	61.868115	56.331048	1.082416
Guantánamo	17.734751	469.016052	21.656778	76.874849	1.234733
Isla de la Juventud	1.416332	3.266338	1.807301	-	3.627173

Tabla Comparativa - ARIMA

Provincia	**MAE**	**MSE**	**RMSE**	**MAPE (%)**	**NMSE**
Pinar del Rio	50.831496	3099.850338	55.676300	118.056924	2.975406
Artemisa	15.080441	356.323239	18.876526	43.775989	1.807946
La Habana	24.171838	886.412693	29.772684	49.344475	1.388757
Mayabeque	10.001121	167.145096	12.928461	113.892115	1.094401
Matanzas	10.520305	157.662562	12.556375	52.429996	2.506159
Villa Clara	66.319031	6024.186224	77.615631	42.566482	1.999742
Cienfuegos	8.797764	116.237939	10.781370	27.433875	1.065186
Sancti Spiritus	27.408835	1090.361685	33.020625	38.743252	1.859881
Ciego de Ávila	69.216009	5628.460564	75.023067	98.950926	5.554252
Camagüey	21.006912	581.757430	24.119648	195.364759	1.648034
Las Tunas	35.025818	1972.070075	44.407996	48.759865	1.202398
Holguin	42.943204	2545.555940	50.453503	73.802970	1.199831
Granma	51.047196	4201.142422	64.816220	41.166754	1.654972
Santiago de Cuba	53.892988	4323.543519	65.753658	58.913639	1.222645
Guantánamo	16.889140	425.649878	20.631284	79.503004	1.120567
Isla de la Juventud	0.908952	1.097582	1.047656	-	1.218833

Para intentar mejorar los resultados obtenidos por ARIMA se aumentaron los datos de manera que aparecieran espaciados por la misma cantidad de tiempo (un día).

También se probó usar el modelo ARIMA con los datos aumentados pero acumulados por diferentes períodos de tiempo para intentar mejorar las predicciones.

AutoML

Se utilizó H2O AutoML con el objetivo de identificar modelos de aprendizaje automáticoque ofrecieran mejor desempeño en nuestro conjunto de datos.

Se pudo comprobar que los modelos que obtuvieron las mejores métricas de desempeño para este AutoML fueron Gradient Boosting Machine (GBM) y Extremely Randomized Trees (XRT) como se puede corroborar en los Tops de modelos por provincia.

	model_id	rmse	mse	mae
1	XRT_1_AutoML_1_20250130_212900	105.11917439114944	11050.040824676887	57.35279201853345
2	DRF_1_AutoML_1_20250130_212900	108.11532803245724	11688.924155565834	58.358692700235466
3	GBM_5_AutoML_1_20250130_212900	112.19925827874977	12588.673558301594	56.87985328425636
4	GBM_2_AutoML_1_20250130_212900	134.823922644989	18177.49011738197	71.07233251610451
5	GBM_4_AutoML_1_20250130_212900	136.4136765544869	18608.69115111217	71.17759593817024
6	GBM_grid_1_AutoML_1_20250130_212900_model_1	138.66375781310686	19227.63773085195	76.98868414250602
7	GBM_3_AutoML_1_20250130_212900	140.52860772371054	19748.289588764517	72.62396705367162
8	GLM_1_AutoML_1_20250130_212900	150.3755642841109	22612.810333764777	89.66053448907604
9	DeepLearning_1_AutoML_1_20250130_212900	151.07790764082458	22824.534177129524	97.673311981819
10	GBM_1_AutoML_1_20250130_212900	825.0119415130133	680644.7036390718	528.6109808271982

Figura 10: Top de modelos de H2O para La Habana.

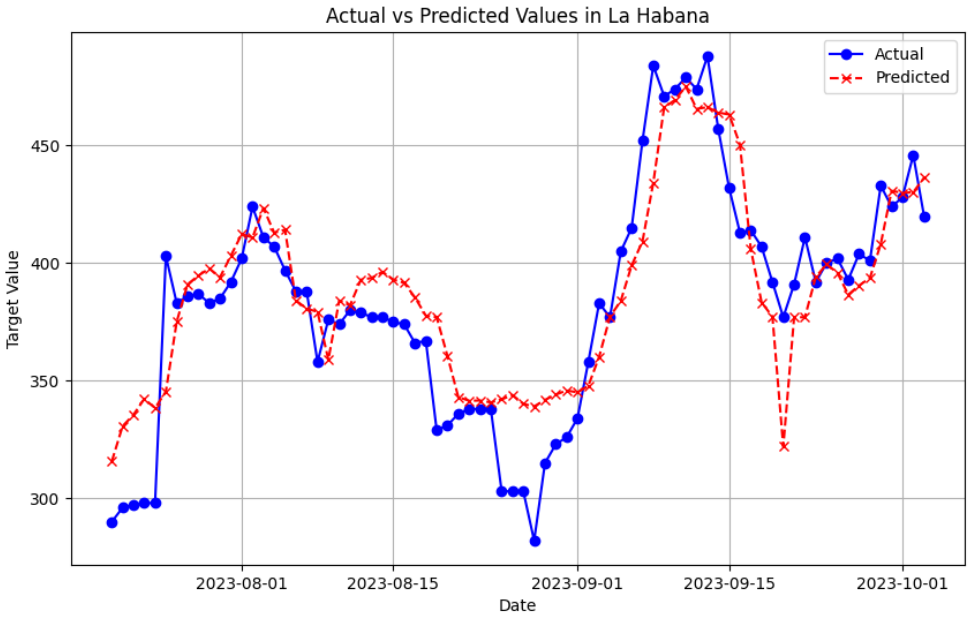

Figura 11: Predicción del modelo más efectivo según H20 AutoML.

H2O AutoML se enfoca principalmente en modelos de árboles de decisión (GBM, XGBoost, DRF), modelos lineales y Deep Learning estándar (MLP/DNN) por lo que no prueba Redes Neuronales Recurrentes(RNN) de forma nativa, explicando esto último el hecho de que en el leaderboard no aparece RNN cuando manualmente se han obtenidos mejores resultados que los allí presentes.

Conclusiones

En este estudio, se implementaron y evaluaron diferentes modelos de Machine Learning para la predicción de enfermedades respiratorias en Cuba. Se logró una mejora significativa en la calidad de los datos mediante la generación de datos con Random Forest, lo que permitió optimizar el rendimiento de los modelos utilizados.

Se determinó que el uso de RNN fue el resultado más efectivo, con 10 *time_steps* proporciona el mejor equilibrio entre captura de patrones temporales y capacidad de generalización obteniéndose de esta manera las mejores predicciones.

Como se ha explicado a lo largo del informe uno de los principales problemas encontrados es la falta de datos para el entrenamiento de los modelos (en ciertos meses, más del 90 % de los datos fueron generados), por ello se recomienda obtener un mayor conjunto de datos de entrenamiento para futuros trabajos, teniendo énfasis en la consistencia y continuidad de los datos, preferiblementa cantidades mayores a dos años para lograr una correcta predicción y captura de patrones.

Además, se recomienda continuar explorando enfoques híbridos y estrategias de optimización para mejorar la confiabilidad de las predicciones en futuros estudios.