Modelos de Inteligencia artificial

Samuel X. Pimienta, samuelpiro@unisabana.edu.co

Juan D. Ramirez Hernandez, est.juan.ramirez3@unimilitar.edu.co

Resumen—Este documento tiene como objetivo analizar el comportamiento y evaluar los resultados de múltiples tipos de redes neuronales para la imputación de datos. En particular, se busca reconstruir dos tipos de señales biomédicas: una señal de electrocardiografía (EKG) y una señal de frecuencia respiratoria. Las redes neuronales implementadas pertenecen a diferentes metodologías, entre ellas una red autorregresiva de tipo *MLPRegressor*, *ARIMA*, *Fourier y polinomial*. Finalmente, se realiza una comparación entre los distintos modelos y su aplicación en señales periódicas.

***Abstract*--This document aims to analyze the behavior and evaluate the performance of multiple types of neural networks for data imputation. Specifically, it seeks to reconstruct two types of biomedical signals: an electrocardiography (EKG) signal and a respiratory frequency signal. The implemented neural networks belong to different methodologies, including an autoregressive network of the *MLPRegressor* type, *ARIMA*, *Fourier and polynomial*. Finally, a comparison is made between the different models and their application to periodic signals.**

I. Introducción

La imputación de datos en señales biomédicas, como electrocardiogramas (ECG) y frecuencia respiratoria, es un desafío crítico en el análisis clínico debido a la alta tasa de perdida que puede afectar la precisión de los diagnósticos y modelos predictivos[1]. En este contexto, diversas técnicas de imputación han sido desarrolladas para abordar este problema. Este documento se centra en evaluar y comparar tres enfoques distintos: una red MLPRegressor (autorregresiva), el modelo ARIMA[2] para series temporales, y la reconstrucción Fourier[3].

La selección del método de imputación adecuado es crucial, ya que diferentes técnicas pueden ofrecer mejores aproximaciones a los valores verdaderos dependiendo del tipo de datos y del patrón de pérdida de respuesta. La imputación múltiple, por ejemplo, ha demostrado ser efectiva en análisis clínicos al proporcionar estimadores cercanos a los parámetros reales mediante la generación de múltiples conjuntos de datos completos. Sin embargo, este estudio se enfoca en comparar la eficacia de los modelos MLPRegressor, ARIMA y Fourier en señales periódicas biomédicas, evaluando su capacidad para preservar características clínicamente relevantes y minimizar errores[4].

La evaluación de estas técnicas se realizará considerando métricas como la precisión en la predicción y la distribución, así como la capacidad de mantener las distribuciones marginales y conjuntas de los datos originales. Este análisis contribuirá a

identificar el enfoque más adecuado para imputar datos faltantes en señales biomédicas periódicas, lo cual es esencial para garantizar diagnósticos precisos y modelos predictivos confiables en la práctica clínica.

II. Modelos de Imputación de Datos

Los modelos implementados son un modelo de regresión llamado MLPRegressor, una red neuronal llamada ARIMA, un modelo de interpolación aplicando la transformada de Fourier y un modelo polinomial.

TABLA I
MODELOS IMPLEMENTADOS Y METODOLOGÍAS

Modelo de imputación	Metodología
MLPRegressor	Auto regresivo
ARIMA	Auto regresivo
Interpolación FT	Interpolación
Polinomial	Interpolación

A. MLPRegressor

Las siglas MLP hacen referencia a *multilayer perceptron*, siendo conocido en español como regresión con perceptrón multicapa se basa en la arquitectura de una red neuronal artificial (RNA) de tipo perceptrón multicapa (MLP) y es utilizada para la estimación de funciones no lineales. A diferencia de los modelos lineales, un MLP puede aproximar relaciones complejas entre variables de entrada y salida mediante la combinación de transformaciones no lineales y pesos ajustables.

La base matemática de esta metodología se encuentra en la función de activación no lineal, la propagación de la información hacia adelante (*forward propagation*), el cálculo del error mediante una función de costo y la actualización de los parámetros mediante retro propagación (*backpropagation*).

En este caso la red es un conjunto de múltiples capas de perceptrones, el perceptrón puede ser visto como la aproximación más pequeña de una red neuronal, esta pequeña pieza puede ser descrita de la siguiente manera:

$$f(w) = \sum_{i=1}^{n} (W_i * X_i) + b \quad (1)$$

Donde:

- n es el número total de conexiones de entrada
- W es el peso de la i-ésima entrada
- X es el valor de la i-ésima entrada
- b es el sesgo de la neurona

La arquitectura será de únicamente dos capas, la primera capa será de entrada y tendrá un total de 32 neuronas, la segunda capa será de la de salida y contara con 16 neuronas, se cumplirán un total de 1000 épocas o iteraciones, además de esto la función de activación utilizada en cada neurona será una función *relu* la cual mantiene los valores en un rango de 0 a x siendo x el valor máximo del set de datos la cual hace referencia a la siguiente expresión:

$$F(x) = \max(0, x) \quad (2)$$

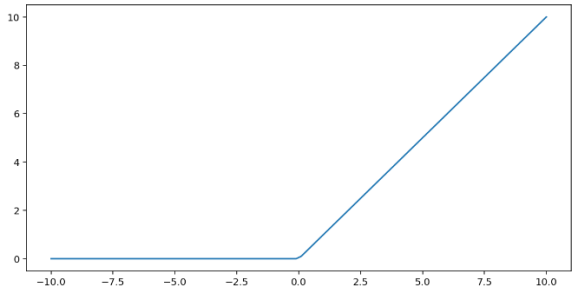

Fig. 1. Función Relu. [5]

1) Implementación

La implementación de este modelo y de los posteriores se realizó en el lenguaje de programación de alto nivel Python, utilizando una librería de *sklearn* y para los cálculos del error se utilizaron dos tipos de mediciones, estas fueron el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). Finalmente, para una comprobación de precisión al momento de reconstruir la señal realizamos una separación del 10% de la señal original de forma aleatoria, de esta manera podemos comparar la señal original y la señal reconstruida mediante la imputación de datos.

a) Error cuadrático medio (MSE)

Mide el error al cuadrado, el error entre el valor esperado y el valor obtenido.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

Donde:

- n es el número total de datos del set de datos
- y_i es el valor real del set de datos
- $\hat{y}_i$ es el valor predicho por la red neuronal

b) Coeficiente de determinación (R^2)

mide la precisión de un modelo estadístico para predecir un resultado.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (4)$$

Donde:

- n es el número total de datos del set de datos
- y_i es el valor real del set de datos
- $\hat{y}_i$ es el valor predicho por la red neuronal
- $\bar{y}_i$ es el valor medio de la variable dependiente

2) Resultados

Al momento de implementar este código obtuvimos resultados bastante negativos, la ventaja de este modelo es la rapidez en la cual puede ser implementado además de que la capacidad de maquina necesaria para entrenar he implementar este modelo debido a que la ecuación de perceptrón simple, aunque se encuentre en múltiples capaz permite un cálculo rápido y preciso. Dentro de las desventajas de este modelo es que precisamente con la arquitectura implementada no obtuvimos buenos resultados en la imputación de datos, teorizamos en este trabajo que la razón de esto es la gran variabilidad de las señales EKG las cuales, aunque son periódicas cuentan con picos de amplitud variable.

TABLA II
RESULTADOS DEL MÉTODO DE REGRESIÓN

Método estadístico	Resultado
MSE	0.2010
R^2	-0.0097

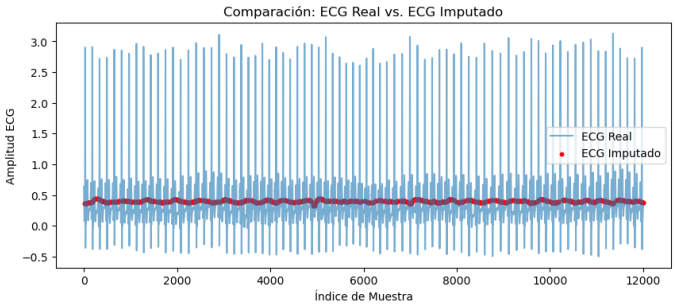

Fig. 2. Señal real vs señal reconstruida método autorregresivo.

La señal azul representa la señal original y completa de electrocardiografía y la señal de color rojo referencia la señal reconstruida con el modelo. El error cuadrático medio (MSE) obtenido, de 0.2010, es relativamente elevado considerando que

la escala de los datos varía entre -0.5 y 3. Por otro lado, el coeficiente de determinación (R^2) de -0.0097 indica que el modelo no logra capturar prácticamente ninguna variabilidad de la variable objetivo. Un valor negativo de R^2 implica que el desempeño del modelo es inferior al de una estrategia trivial, como predecir siempre la media o un valor constante. De hecho, si se hubieran imputado todos los valores faltantes utilizando la media del EKG, se habría obtenido un R^2 igual o superior a cero, lo que resalta las limitaciones del enfoque actual.

B. ARIMA

El modelo ARIMA (*Autoregressive Integrated Moving Average*) o media móvil integrado autorregresivo es un modelo de análisis y pronóstico de series temporales que combina tres elementos: auto regresión (AR), integración (I) y media móvil (MA). El componente autorregresivo (p) modela la relación entre el valor actual y los valores previos de la serie, mientras que la integración (d) representa la diferenciación necesaria para lograr la estacionariedad. Por su parte, el componente de media móvil (q) modela la dependencia entre el valor actual y los errores previos. El modelo ARIMA (p, d, q) se construye estimando estos parámetros, generalmente a partir de criterios de información como *AIC* y *BIC*, seguido de la estimación de parámetros mediante máxima verosimilitud. Además, puede generalizarse a extensiones como SARIMA, que incorpora estacionalidad, y ARIMAX, que integra variables exógenas para mejorar la capacidad predictiva.

1) Implementación

En este caso el modelo ARIMA contara con los siguientes parámetros para mejorar la predicción:

- $p = 2$
- $i = 0$
- $q = 1$

En canto a la predicción del error utilizaremos las mismas mediciones que en el método anterior. Utilizaremos la misma estrategia de eliminar un 10% de los datos de la señal original esto con la finalidad de verificar la capacidad de imputación de datos de este modelo.

2) Resultados

Los resultados obtenidos en este modelo fueron mucho más acertados sobre todo en los picos de mayor amplitud, sobre todo en comparación con el modelo anterior de regresión MLP. El modelo ARIMA requiere a su vez una baja capacidad de cómputo debido a los cálculos rápidos que realiza y la nula necesidad de utilizar épocas, esto crea una predicción rápida y directa, el mayor inconveniente que se puede obtener con este modelo es la necesidad de escoger los parámetros de funcionamiento como lo son los anteriormente mencionados p, i y q, para solucionar este inconveniente existen modelos como el auto ARIMA, el cual mediante múltiples iteraciones selecciona valores óptimos para estas tres características, con la clara desventaja de necesitar una mayor capacidad de computo y un mayor tiempo de ejecución.

TABLA III
RESULTADOS DEL MÉTODO DE ARIMA

Método estadístico	Resultado
MSE	0.3711
R^2	-0.8645

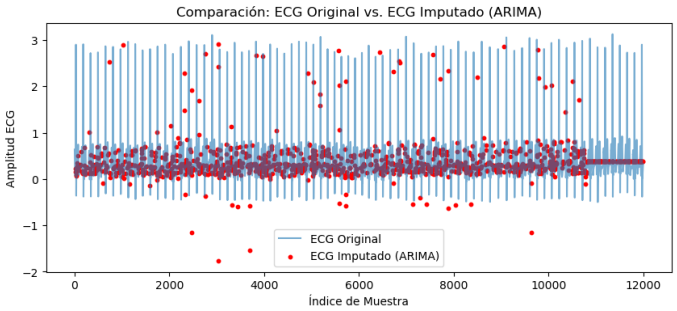

Fig. 3. Señal real vs señal reconstruida método ARIMA.

Vemos una clara mejora en la imputación de datos en la señal, en este caso el modelo busca recrear con mayor exactitud los datos, mejorando en la variabilidad de los datos e intentando imitar los valores de mayor amplitud en la señal. Este modelo muestra un mejor desempeño que el anterior pero no cuenta con una gran calidad de precisión en la imputación de los datos.

C. Fourier

La regresión basada en series de Fourier es una forma de analizar la periodicidad o los comportamientos cíclicos mediante la descomposición de la señal en combinaciones lineales de senos y cosenos de distintas frecuencias; esto con la idea de que cualquier función periódica puede ser expresada como una combinación de senos y cosenos con diferentes frecuencias, por lo que se puede capturar la estacionalidad de la serie sin que sea necesario que haya variables categóricas o indicadores específicos parar cada periodo:

$$y_t = \beta_0 + \sum_{k=1}^{K} \left[\beta_{1k} \cos\left(\frac{2\pi k t}{T}\right) + \beta_{2k} \sin\left(\frac{2\pi k t}{T}\right) \right] + \varepsilon_t \quad (5)$$

Donde:

- β_0 es un término independiente
- K es el número de armónicos utilizado
- T es el periodo de estacionalidad
- β_{1k} *y* β_{2k} son los coeficientes de los términos de Fourier
- ε_t es el error aleatorio

1) Implementación

Para la implementación de este modelo se utiliza nuevamente la librería *sklearn* de *Python*, los valores β son estimados por el modelo, así que únicamente los únicos datos que tenemos que asignarle al modelo es el periodo de estacionalidad y el número

de armónicos utilizados. El valor T será asignado por cada uno de los datos del arreglo y el valor de los armónicos será asignado en 10.

2) Resultados

Los resultados del modelo fueron muy similares a el primero mostrado, mostrando así el mismo error en la predicción y la imputación de datos, teniendo una muy baja precisión al momento de recrear los datos que se encuentran en los picos de amplitud más amplios. Los resultados de la tabla IV muestran la mala predicción que tuvo este modelo en este set de datos con los parámetros especificados.

TABLA IV
RESULTADOS DEL MÉTODO DE FOURIER

Método estadístico	Resultado
MSE	0.1996
R^2	-0.0029

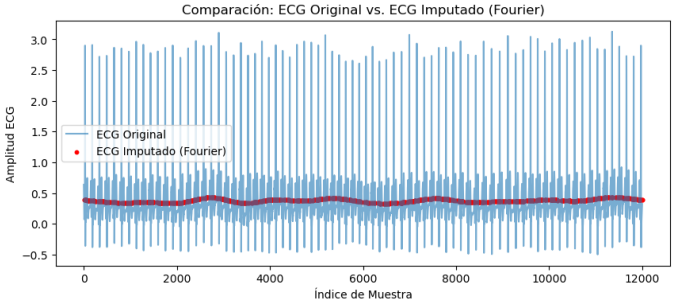

Fig. 4. Señal real vs señal reconstruida método Fourier.

Para explorar en mejor medida este modelo y verificar realmente su funcionalidad realizamos una prueba adicional, en donde el valor de los armónicos fue asignado a 5 y fua asignada una ventana de tiempo para el análisis de los datos en este caso la ventana deslizante de datos fue establecida en un valor de 200 datos, esta ventana nos permitirá analizar de forma separada los datos y le permitirá al modelo prestar más atención a la correlación de los datos en segmentos mas pequeños de la señal. Los resultados obtenidos muestran una notable mejora en la predicción e imputación de los datos.

TABLA V
RESULTADOS DEL MÉTODO DE FOURIER CON VENTANA DESLIZANTE

Método estadístico	Resultado
MSE	0.1682
R^2	0.1549

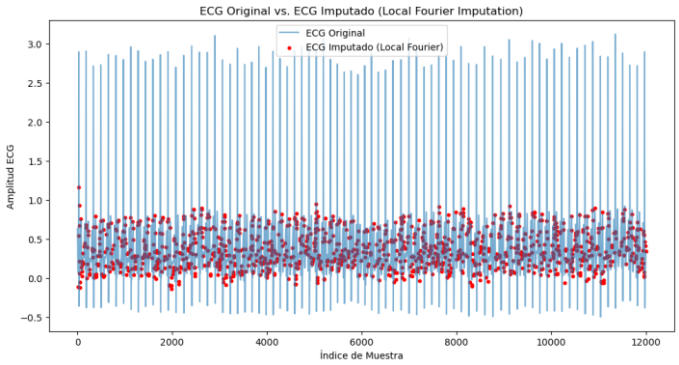

Fig. 5. Señal real vs señal reconstruida método Fourier con ventana deslizante.

El resultado de la imputación de los datos teniendo el 10% faltante como los otros métodos anteriores es claramente mejor que el método de Fourier sin aplicar la ventana deslizante, esto demuestra que en este caso en específico, la ventana permite que la encuentre mayor relación entre los datos, aun así este modelo tiene un problema similar a los modelos anteriores y es que la predicción de los datos en las amplitudes mas altas de la señal no es correcta, el modelo se centra principalmente en los datos cerca a la media del set de datos.

D. Polinomial

La técnica de interpolación polinómica se puede aplicar como técnica para imputar datos y consiste en calcular los valores faltantes en una serie temporal a partir de la construcción de un polinomio que ajuste todos los datos conocidos, de forma que se considere la complejidad de las variaciones observadas en la serie. De hecho, la diferencia principal entre este método y las técnicas más simples de imputación de datos (por ejemplo, la técnica de imputación por media o la interpolación lineal), es que se intenta analizar de forma más clara las tendencias y variaciones más importantes mediante la forma polinómica, mientras que con las técnicas simples anteriores se puede perder información relevante.

1) Implementación

La base matemática de la técnica de interpolación polinómica reside en encontrar un polinomio de grado n que pase a través de un conjunto de puntos observados de manera que se garantice que la representación de los datos sea continua y sea diferenciable. En este caso el grado seleccionado para el acoplamiento del método fue de grado 3 además de esto se estipula que la red realice esta interpolación hacia delante del modelo y hacia atrás, de una manera similar a como se realiza en el modelo MLPRegressor.

2) Resultados

Los mejores resultados fueron obtenidos mediante este modelo demostrando la gran adaptabilidad del modelo para este tipo de señales periódicas, este no únicamente cuenta con valores de *MSE* bajos y valores de R^2cercanos a 1, lo cual nos demuestra numéricamente el excelente comportamiento del modelo, sino que también entrega mejores resultados en el principal punto débil de los demás modelos, y es que en los

puntos de mayor amplitud el modelo se acerca de una forma precisa.

TABLA VI
RESULTADOS DEL MÉTODO POLINOMIAL

Método estadístico	Resultado
MSE	0.0002
R^2	0.9991

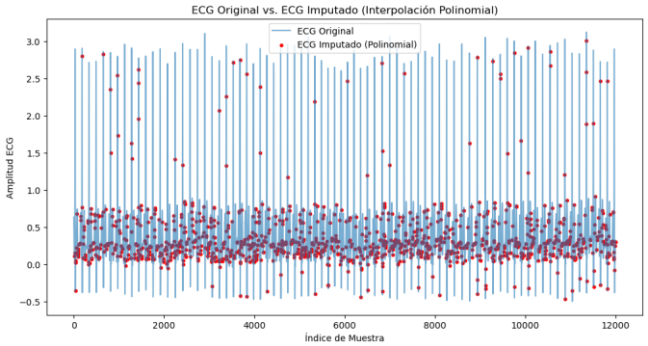

Fig. 6. Señal real vs señal reconstruida método Polinomial 10%.

El modelo mostro tan buenos resultados que decidimos realizar pruebas adicionales para validar el comportamiento en la predicción e imputación de los datos en la señal, en este caso buscábamos observar como se comportaba el mismo modelo con un set de datos al cual se le fueron eliminando periódicamente un mayor porcentaje de los datos, la prueba inicial se realizo como en todos los modelos eliminando un 10% de los datos, la segunda prueba se realizo eliminando un 30% de los datos, un 50% de los datos y por ultimo eliminando un total de 70% de los datos de la señal original.

TABLA VII
RESULTADOS DEL MÉTODO POLINOMIAL 30%

Método estadístico	Resultado
MSE	0.0008
R^2	0.9958

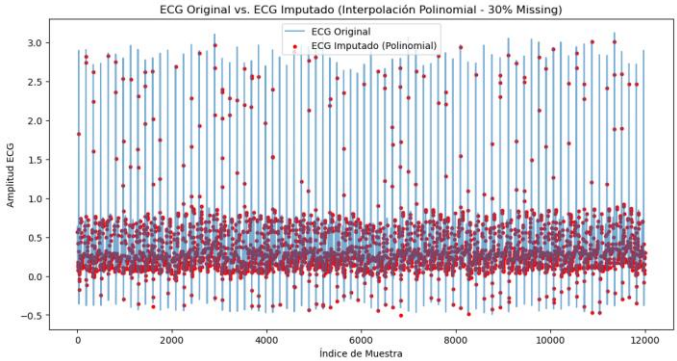

Fig. 7. Señal real vs señal reconstruida método Polinomial 30%.

TABLA VIII
RESULTADOS DEL MÉTODO POLINOMIAL 50%

Método estadístico	Resultado
MSE	0.0116
R^2	0.9360

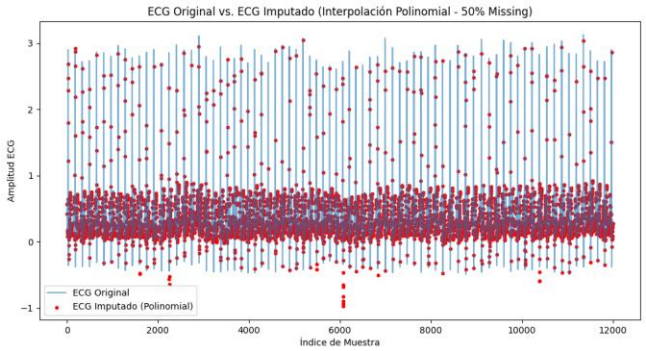

Fig. 8. Señal real vs señal reconstruida método Polinomial 50%.

TABLA IX
RESULTADOS DEL MÉTODO POLINOMIAL 70%

Método estadístico	Resultado
MSE	0.0711
R^2	0.6125

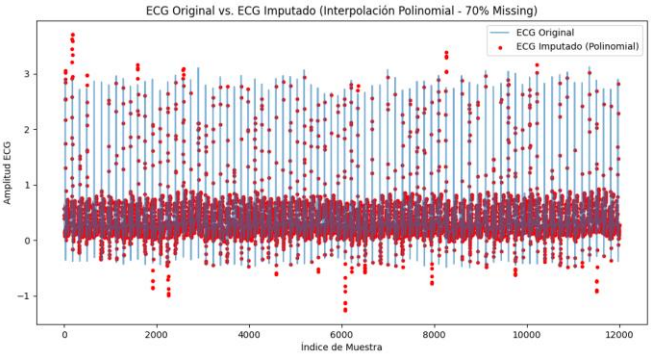

Fig. 9. Señal real vs señal reconstruida método Polinomial 70%.

E. Conclusiones

Las conclusiones que podemos obtener al implementar estas redes neuronales y metodologías de predicción e imputación de datos a este set de datos en especifico muestran claramente que en los parámetros escogidos por cada modelo el método polinomial demostró los mejores resultados no solo en cuanto a predicción e imputación sino también con la adaptabilidad a la perdida de datos que se podrían llegar a obtener en este tipo de señales periódicas, para la comparación mantendremos los resultados obtenidos con el 10% de datos eliminados del grupo de datos originales.

TABLA X
COMPARACIÓN DE LOS RESULTADOS

Método estadístico	MLPRegressor	ARIMA	Interpolación FT	Polinomial
MSE	0.2010	0.3711	0.1682	0.0002
R^2	-0.0097	-0.8645	0.1549	0.9991

En los resultados podemos observar que, entre los modelos estudiados, el modelo polinomial es el que ofrece el mejor resultado, ya que cuenta con el menor error cuadrático medio (*MSE* = 0.0002) y para el que se obtiene un coeficiente de determinación R^2 = 0.9991, lo que indica que el ajuste a los datos es casi perfecto. A diferencia del resto de los métodos,

ARIMA y MLPRegressor, que ofrecen un resultado mucho menor (R^2 negativo o cerca de cero) y, por tanto, no modelan de forma satisfactoria la relación entre las variables. De hecho, el modelo polinomial realiza el ajuste con menor capacidad computacional que las redes neuronales, aspectos que lo convierten en una opción eficiente y veraz para este problema.

Adicional a esto podemos realizar una comparación del modelo Polinomial al momento de imputar datos, pero esta vez eliminando cada vez mas datos de la señal original.

TABLA XI
COMPARACIÓN DE LOS RESULTADOS MÉTODO POLINOMIAL

Método estadístico	10%	30%	50%	70%
MSE	0.0002	0.0008	0.0116	0.0711
R^2	0.9991	0.9958	0.9360	0.6125

Los resultados muestran que el modelo polinómico logra un gran rendimiento en situaciones de poca pérdida de datos; sin embargo, la tasa de exactitud va disminuyendo conforme va aumentando el porcentaje de perdida de datos. Por ejemplo, para una pérdida del 10 % el error cuadrático medio (*MSE*) es mínimo (0.0002) y el coeficiente de determinación (R^2) permanece en 0.9991, lo que indica un excelente ajuste del modelo.

Por el contrario, cuando la pérdida de los datos es del 50 % o el 70 %, el *MSE* será notablemente más alto (0.0116 y 0.0711, respectivamente) y el coeficiente R^2 será de 0.6125, lo que indicará que el modelo ha perdido la capacidad de reconstrucción adecuada de los datos. En resumen, la interpolación polinómica, desde una perspectiva estadística, es una herramienta de interpolación de alta densidad para situaciones donde la pérdida de datos existentes pasa a una situación de pérdida de datos moderada, pero su rendimiento disminuye rápidamente cuando el porcentaje de pérdida de datos existente sobrepasa el 30% del set de datos, por lo que se debe buscar adecuadamente alternativas complementarias para mejorar aún más la capacidad de imputación en este modelo.

B. Referencias

[1] S. P. Heilbroner, C. Carter, D. M. Vidmar, E. T. Mueller, M. C. Stumpe, y R. Miotto, “LIFE: A Deep Learning Framework for Laboratory Data Imputation in Electronic Health Records”, *medRxiv*, p. 2023.10.31.23297843, nov. 2023, doi: 10.1101/2023.10.31.23297843.

[2] R. de Arce Ramón Mahía Dpto Economía Aplicada, “MODELOS ARIMA Mayo 2001”.

[3] “(PDF) AN IMPROVED PROCEDURE FOR FOURIER REGRESSION ANALYSIS”. Consultado: el 16 de marzo de 2025. [En línea]. Disponible en: https://www.researchgate.net/publication/338235894_AN_IMPROVED_PROCEDURE_FOR_FOURIER_REGRESSION_ANALYSIS

[4] S. Vujović *et al.*, “Sparse Analyzer Tool for Biomedical Signals”, *Sensors (Basel)*, vol. 20, núm. 9, p. 2602, may 2020, doi: 10.3390/S20092602.

[5] “Redes neuronales: Funciones de activación | Machine Learning | Google for Developers”. Consultado: el 15 de marzo de 2025. [En línea]. Disponible en: https://developers.google.com/machine-learning/crash-course/neural-networks/activation-functions?hl=es-419