PIA Aprendizaje Automatico

Universidad Autonoma de Nuevo León
Alanis Mares, Victor Hugo

19 de Julio, 2024

1 Introducción

Se pretende analizar la evolución de la expectativa de vida a lo largo de 15 años en diferentes países tomando como referencia un conjunto de series de tiempo de diferentes indicadores macroeconómicos para implementar un sistema de clasificación no supervisada y posteriormente una predicción con base en un algoritmo supervisado.

2 Descripción de los datos

Los datos a trabajar cuentan con 16 variables, de las cuales once son referentes a la serie de tiempo multivariada y cinco son categóricas. Para propósitos de este estudio se tomarán en cuenta las siguientes 11 variables numéricas:

- Year: Año de observación.
- Under_five_deaths: Cantidad de muertes en infantes menores de 5 años por cada 1000 habitantes.
- Adult_mortality: Cantidad de muertes en adultos por cada 1000 habitantes.
- Alcohol_consumption: Litros anualizados de consumo de alcohol per capita.
- Hepatitis_B: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para la Hepatitis B.
- BMI: Índice de grasa corporal promedio.
- Polio: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para la Polio.
- Diptheria: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para tétanos.
- Incidents_HIV: Casos de VIH por cada 1000 habitantes.

- GDP_per_capita: Producto interno bruto per capita.
- Schooling: Cantidad de años promedio que un adulto (25+) ha pasado estudiando.
- Life_expectancy: expectativa de vida al nacimiento.

Cabe mencionar, que el conjunto original de datos presentaba un total de 21 variables, mismas que fueron reducidas a las 16 previamente mencionadas, el proceso de selección de variables será explicada de forma más extensa en la sección "Preprocesamiento".

2.1 Origen de los datos

Los datos fueron obtenidos de un registro público de conjuntos de datos, los cuales provienen de estudios de múltiples hospitales, a continuación se encuentra el enlace al conjunto de datos.

3 Preprocesamiento

3.1 Omisión de variables previo a estudio

Dado que se realizará un análisis de series de tiempo multivariado para propósitos de estudiar el impacto de las variables, no se tomarán en cuenta 5 variables en específico:

- 1: Country.
- 2: Region.
- 3: Year.
- 4: Economy_status_Developed.
- 5: Economy_status_Developing.

Primeramente, el año no constituye a una variable de la cual que se pueda eliminar debido a que se hará un estudio de series de tiempo, dicho estó se omitirá completamente de esta parte de selección de variables dado que podría considerarse válida "por defecto". En contraste, las otras 4 variables constituyen a variables categóricas, las cuales no cumplen con una función que pueda considerarse útil para propósitos de nuestro análisis de series de tiempo, dado que se busca hacer una clasificación con base en las propiedades intrínsecas de las series de tiempo, no de variables categóricas por lo cual se omitirán de este paso.

3.2 Análisis de relación de variables

3.2.1 Análisis de regresión F

Utilizando un pre-procesador se busca encontrar las variables más relevantes con respecto a la variable respuesta, para esto usaremos un selector basado en las K mejores variables utilizando una regresión F.

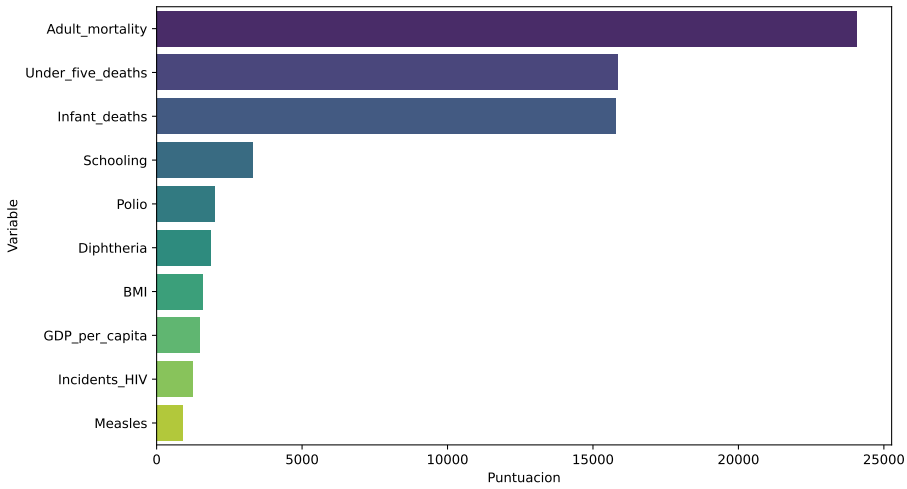

Figure 1: Top 10 variables ordenadas por su puntuación F

Se puede apreciar en la Figura 1 como la mortalidad adulta es precisamente uno de los mejores indicadores con respecto a la expectativa de vida, lo cual tiene sentido dado que un país con una gran mortalidad en adultos de todas las edades no contaría con una expectativa de vida alta; sin embargo, vamos a se realizarán más pruebas para confirmar qué variables pueden ser relevantes, empezando con el análisis del p-valor.

3.2.2 Análisis de p-valor

Variable	p-valor
Infant_deaths	0
Under_five_deaths	0
Adult_mortality	0
Polio	0
Schooling	0
Diptheria	0
BMI	0
GDP_per_capita	0
VIH	0
Measles	0
Thinnes_ten_nineteen_years	0
Thinness_five_nine_years	0
Hepatitis_B	0
Alcohol_consumption	0
Population_mln	.15

Bajo la misma línea, se puede apreciar cómo otro indicativo importante para determinar la expectativa de vida (la cual hasta cierto punto es un indicativo de la salud general de un país) es precisamente la mortalidad tanto en adultos como en infantes.

Algo a destacar es el hecho de que todos los p-valores que se obtuvieron son significativamente bajos, con excepción de la población, esto podría ser un buen indicativo de que la población tal vez no sea la mejor variable a estudiar para este análisis dado a su gran variación entre países.

3.2.3 Análisis de correlación

Al emplear un analisis de correlación con respecto a la expectativa de vida se obtuvieron los resultados visibles en la Figura 2.

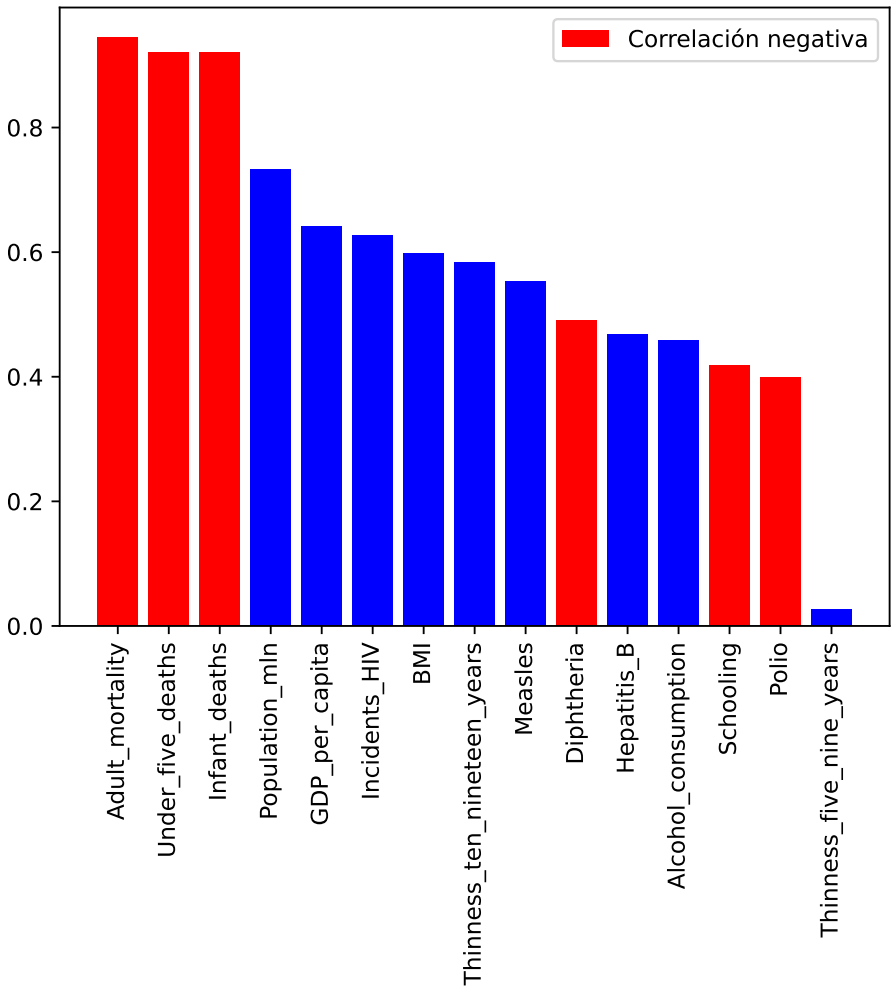

Figure 2: Variables ordenadas por su valor de correlación con respecto a la expectativa de vida

Se puede apreciar como la variable "Thinness_five_nine_years" tiene la correlación más baja siendo practicamente 0.

3.2.4 Análisis de información mutua

Al emplear un analisis de información mutua se obtuvieron los resultados visibles en la Figura 3.

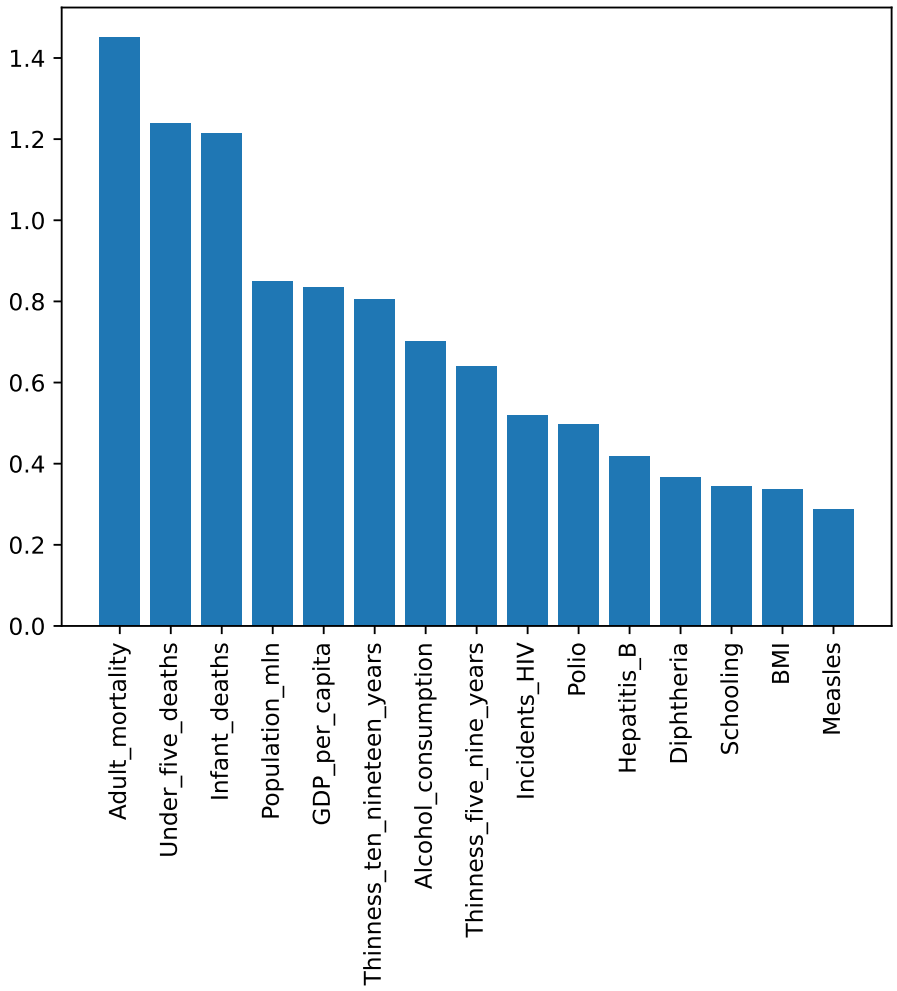

Figure 3: Variables ordenadas por su valor de información mutua

Se puede apreciar como la variable "Thinness_ten_nineteen_years" junto con "Alcohol_consumption" son las más bajas.

3.3 Selección final de variables

Basado en los análisis previos se han omitido las siguientes variables:

Population: No se considera que la población total sea un indicador efectivo dado que las variaciones en la misma no necesariamente tienen relación con la expectativa de vida, tanto como se tiene con el espacio disponible (i.e: Un país puede tener una muy buena calidad de vida pero tener poca población

puramente por su territorio disponible)

Thiness_five_nine_year y Thinness_ten_nineteen_years: Toda la información que nos puede presentar esta variable podría ser explicado con otras dos variables que sí se van a incluir en el modelo (BMI y Under_five_years), esto se puede apreciar en el análisis de correlación dado que las variables mencionadas obtuvieron un índice de correlación muy alto

Infant_deaths: Parecido a las variables previamente mencionadas, la información relevante se encuentra en Under_five_deaths dado que es una edad donde una muerte sería más inesperada.

Measless: Estudiar el grado de inmunidad a una enfermedad puede ser un buen indicador del nivel de salud del país, sin embargo el sarampión no representa tanto un buen indicador con respecto al polio, esto se puede apreciar dado que en prácticamente todas las pruebas de relevancia de variable el polio obtuvo un mejor valor que el sarampión

4 Metodología no supervisada

Para este analisis se usará un algoritmo de gas neuronal creciente (o GNS por sus siglas en inglés) partiendo de la librería `neupy` para la aplicación del algoritmo, `numpy` para álgebra matricial necesaria y `pandas` para manejo de datos.

4.1 Descripción del algoritmo GNS

Citado por primera vez en 1995, el algoritmo GNS parte de vectores con la misma dimensionalidad y busca generar un grafo el cual explique la distribución de los puntos al mismo tiempo que minimiza el error.

4.2 Fundamento matemático

Partiendo de 2 puntos aleatorios conectados entre sí, se sigue el siguiente algoritmo en cada iteración con cada vector ingresado por el conjunto datos:

- 1: Se calcula la distancia entre el vector ingresado y los dos nodos más cercanos
- 2: El error del nodo más cercano es sumado al error del vector (el cual inicializa en 0)
- 3: El nodo más cercano y todo nodo conectado a él es movido en función de los errores acumulados de cada nodo, en dirección del vector ingresado.
- 4: La variable "edad" (Con la cual todo arista es inicializado en 0) de cada arista ya existente se aumenta en 1.
- 5: En caso de que el nodo más cercano y el segundo nodo más cercano estén conectados entre sí su arista es igualada a 0 , en caso contrario se crea una arista entre ambos nodos.

- 6: Se comprueba la edad de cada arista, en caso de que sea mayor a una variable previamente decidida se elimina la arista.
- 7: En caso de que la iteración actual sea multiplo de una variable previamente decidida se agrega un nodo al azar al sistema conectando el nodo con el error más grande y su vecino más cercano, después se elimina la arista que conecta a ambos nodos.
- 8: Se decrementa el error acumulado de todos los nodos en el sistema por un factor constante
- 9: Se repite el proceso hasta que el algoritmo haya llegado a las iteraciones previamente definidas

4.3 Resultados no supervisados

Tras aplicar el algoritmo de agrupamiento se llegó a 20 grupos identificados, mismos que fueron graficados a lo largo del tiempo, la gráfica general de todos los puntos de datos agrupados se ve de la siguiente forma:

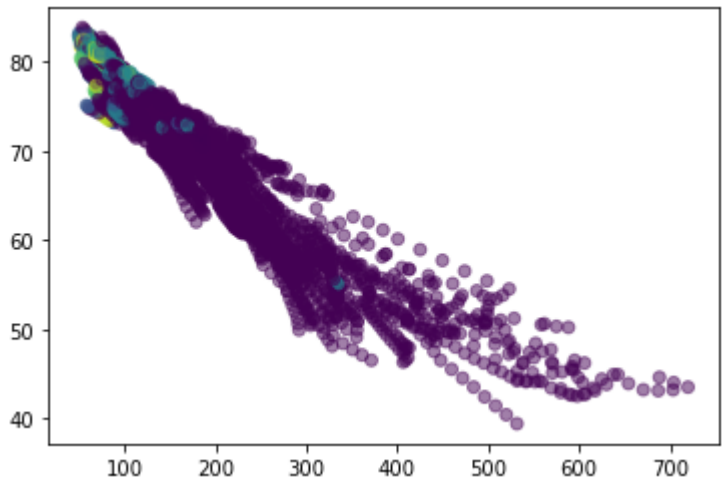

Figure 4: Agrupamiento de todos los puntos estudiados.

Como se puede apreciar en la figura 4, realmente no existe mucha distinción debido a la gran densidad de datos, por eso mismo se procedió a dividir los datos por año para observar si se podía obtener un mejor entendimiento de los grupos.

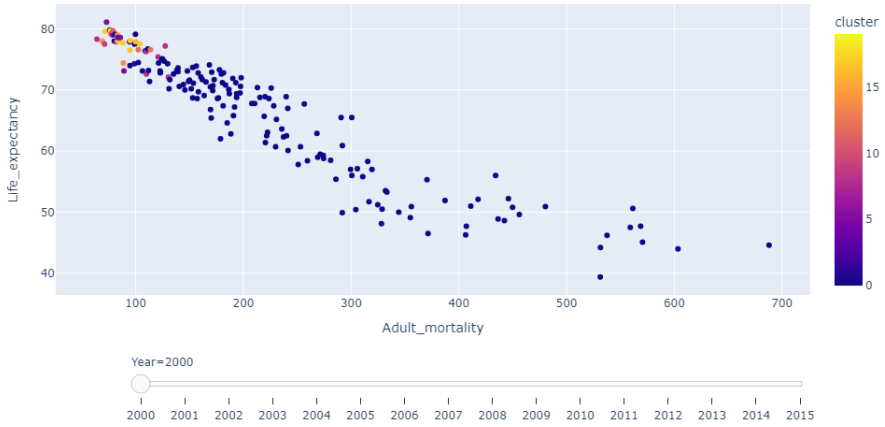

Figure 5: Agrupamiento en el año 2000

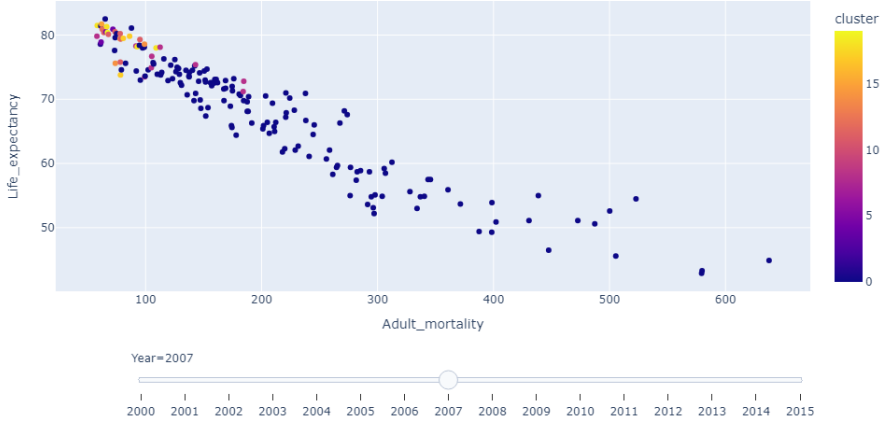

Figure 6: Agrupamiento en el año 2007

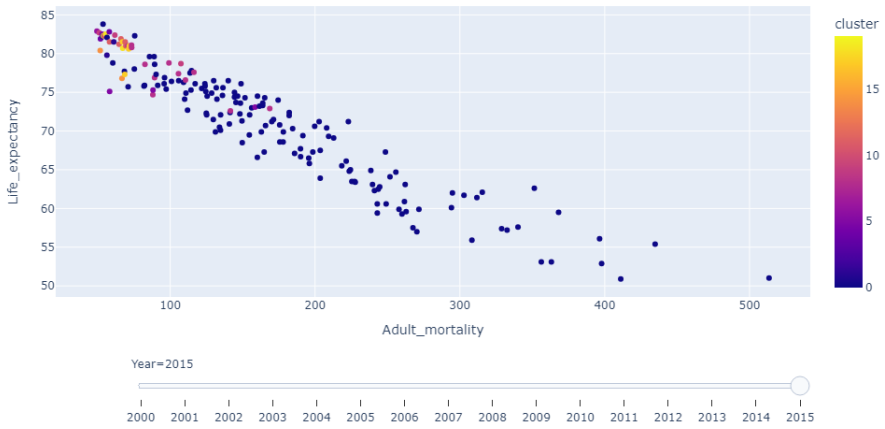

Figure 7: Agrupamiento en el año 2015

4.4 Conclusiones del algoritmo no supervisado

Se aprecia cómo el agrupamiento se mantiene relativamente constante al pasar los años y cómo el algoritmo parece haberse concentrado en la parte de los datos con alta expectativa de vida y clasificando el resto de datos dentro del mismo cluster.

5 Metodología supervisada

Para este análisis se tomará como variable objetivo la expectativa de vida para cada país, se utilizará un regresor de potenciación de gradiente apoyado de la librería `skforecast`. La razón principal de elección de la librería previamente mencionada es su método `ForecasterAutoregMultiSeries`, el cual implementa un regresor de potenciación de gradiente multivariado, este método recibe dos argumentos:

- Un conjunto de series de tiempo referentes a nuestra variable objetivo.
- Un conjunto de series de tiempo referentes a variables exógenas a la variable objetivo, las cuales se presuponen afectan a la variable objetivo.

Debido a la baja cantidad de observaciones por serie de tiempo, se utilizará un regresor que sólo toma en cuenta las últimas 5 observaciones de la serie de tiempo.

5.1 Resultados del algoritmo supervisado

Al aplicar el algoritmo previamente mencionado y calculando el error medio absoluto entre el conjunto de prueba y el conjunto de validación, se obtiene que en promedio el modelo esta desviado de las observaciones reales por .5 años, es decir 6 meses, sin embargo, como se puede apreciar en la figura 4, este promedio podría estar sesgado dado que la mayoría de los errores medios absolutos son notablemente bajos, mientras que contamos con algunos valores atípicos. Se anexa tambien las predicciones concatenadas a las primeras 5 series de tiempo del conjunto.

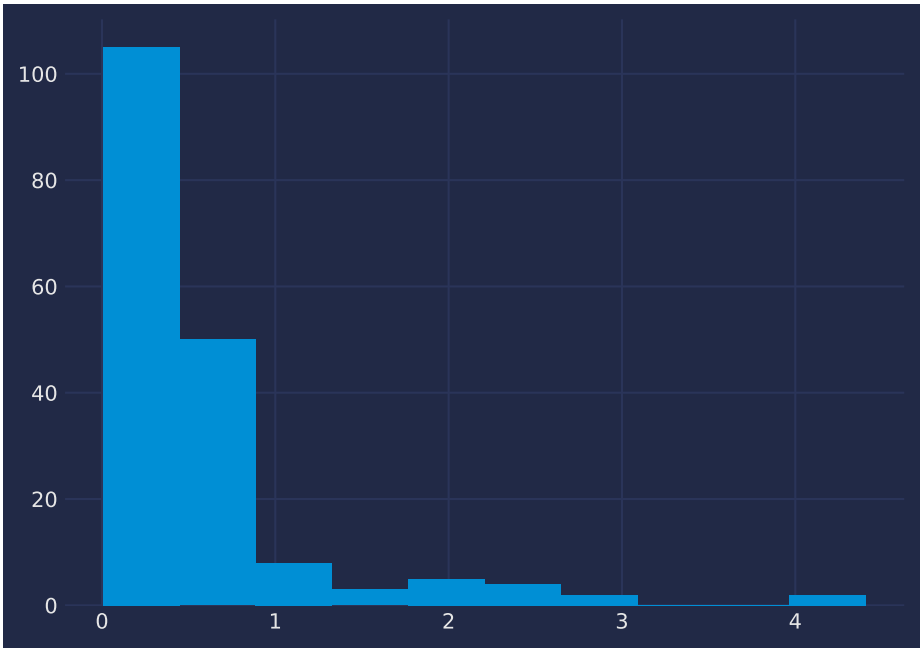

Figure 8: Histograma de error medio absoluto

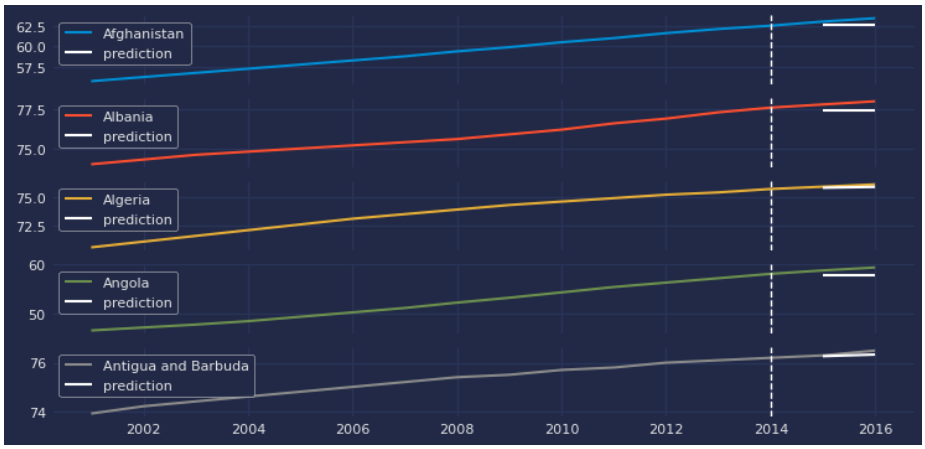

Figure 9: Predicciones comparadas con el conjunto de validación

5.2 Métricas para algoritmo supervisado

Para evaluar el modelo y proceder con la optimización de hiper parámetros, se procedió a analizar los siguientes tipos de métricas:

- Error promedio al cuadrado.
- Error promedio absoluto.
- Error promedio absoluto porcentual.
- Error promedio cuadrado logarítmico.

Después de optimizar los hiperparametros se consiguieron las siguientes métricas de desempeño:

Métrica	Promedio	Máximo	Mínimo
Promedio al cuadrado	0.810385	19.811893	0.000090
Promedio absoluto	0.560838	4.408434	0.007003
Promedio absoluto porcentual	0.008592	0.078813	0.000097
Promedio cuadrado logarítmico	2.219828e-04	6.929335e-03	1.693964e-08

Al evaluar el gráfico de caja (Figura 10) de la métrica “Promedio absoluto” podemos apreciar que se tienen muchos valores atípicos, sin embargo, la gran mayoría de los datos están contenidos en un error entre 0 y 1, lo cual considerando las bajas observaciones y la variable respuesta medida es un error aceptable, siendo esto referente a máximo un año de la esperanza de vida dado los estimadores macroeconómicos usados.

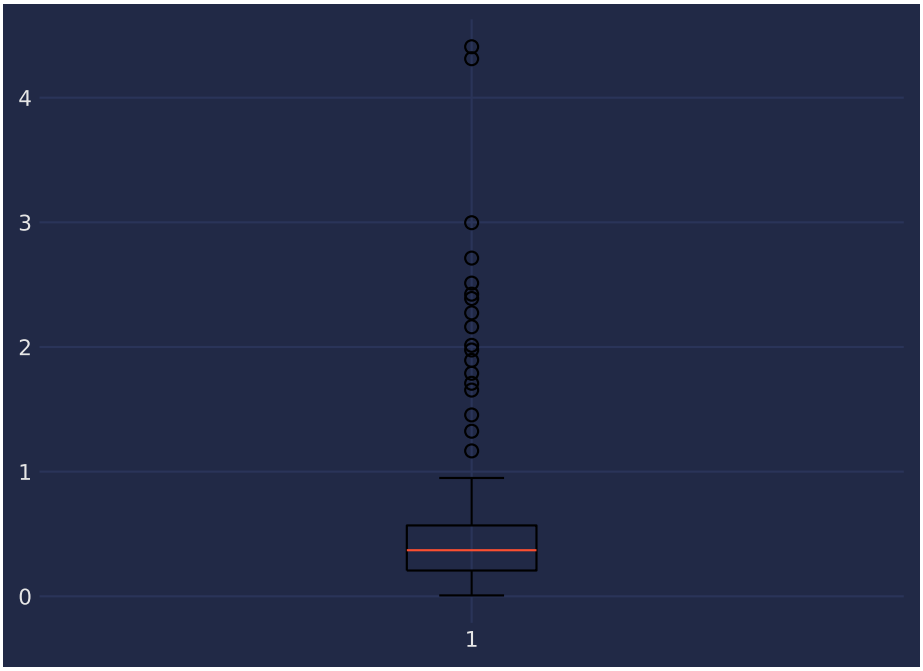

Figure 10: Gráfico de caja del error promedio absoluto del modelo

6 Conclusiones

Ambos enfoques, tanto supervisado como no supervisado, plantean conclusiones diferentes pero relevantes, el modelo supervisado dice que dado las variables macroeconómicas adecuadas se puede llegar a predecir la expectativa de vida de un país aún contando con pocas observaciones, esto se sustenta en que, aunque el modelo supervisado se entrenó solo con 10 observaciones por país, eso fue suficiente para lograr un error promedio de 6 meses con respecto al conjunto de validación.

Dejando como aprendizaje que la expectativa de vida no es algo ajeno a las decisiones a gran escala del gobierno del país, esto si bien podría parecer obvio, no se debe dejar pasar por alto dado que, aunque no sea difícil ver que la mortalidad infantil afecta a la calidad de vida, la relación entre la misma y el grado de inmunidad a una enfermedad como la polio no es tan fácil de ver a primera instancia.

Mientras tanto, los métodos no supervisados al no ser efectivos permiten recordar que la falta de respuesta es una respuesta en si misma. Al no ser capaz de agrupar los diferentes países por sus condiciones macroeconómicas, aunque tengan expectativa de vida similar, permite ver que si bien estas variables constituyen un indicador suficiente para su predicción no representan toda la infor-

mación relevante como para decir que el modelo esta completo.
Las principales conclusiones que se pueden observar aquí es que este es un estudio que definitivamente tiene futuro sin embargo, para que el análisis sea congruente y permita tener un mayor entendimiento se deben cumplir dos cosas:

- 1.- Se debe continuar extrayendo los indicadores macroeconómicos para tener un conjunto de información más grande a futuro.
- 2.- Se debería considerar agregar más variables de campos no estudiados. Sin ir muy lejos en el conjunto de datos no se incluye ninguna variable de percepción de calidad de vida.

Por último, se considera que este análisis es importante debido a una razón muy específica, los datos provienen de años previos a la pandemia. Esto quiere decir que de cierta manera el análisis provee una visión de las variables relevantes en un estudio de predicción de expectativa de vida, mientras que al mismo tiempo sirve de recordatorio que en los estudios siempre habrá variables que no se puedan controlar y que no se puedan predecir. No importa qué tanta información pasada se hubiera tenido ni qué tantas variables significativas se hubieran incluido, el modelo no hubiera sido capaz de predecir el impacto de la pandemia en la expectativa de vida en lo más mínimo y precisamente por eso se decidió hacer este estudio.

References

[1] Bernd Fritzke (1995) A Growing Neural Gas Network Learns Topologies.

[2] Halil Ertan (2022) Multivariate Time Series Clustering Using Growing Neural Gas and Spectral Clustering

[3] Joaquín Amat Rodrigo, Javier Escobar Ortiz (2022) Modelos de forecasting globales: modelado de múltiples series temporales con machine learning