# PIA Aprendizaje Automatico

Universidad Autonoma de Nuevo León
Alanis Mares, Victor Hugo

19 de Julio, 2024

## 1 Introducción

Se pretende analizar un conjunto de series de tiempo con diferentes indicadores macroeconómicos para implementar un sistema de clasificación no supervisada y posteriormente una prediccion 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 categoricas. Para propositos de este estudio de clusterizacion se utilizaran las once variables numericas, las cuales son:

- Year: Año de observacion.
- 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 por 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.

### 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 el análisis que se tiene pensado hacer es un análisis de series de tiempo multivariado para propósitos de estudiar el impacto de las variables no se tomaran 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 podamos deshacernos debido a que se hará un estudio de series de tiempo, debido a esto se omitirá completamente de esta parte de selección de variables dado que podría considerarse es válida "por defecto". En contraste las otras 4 variables constituyen a variables categóricas, las cual probablemente en un futuro reporte se usaran para analizar los resultados obtenidos pero al menos en primera instancia no cumplen con una función que pueda considerarse útil para propósitos de nuestro análisis de series de tiempo por lo cual se omitirán de este paso.

Cabe destacar que para las 5 variables serán excluidas únicamente de esta parte del estudio (i.e la selección de variables ) sin embargo en futuros análisis las variables no serán omitidas a menos que se dé una justificación específica para no usarlas.

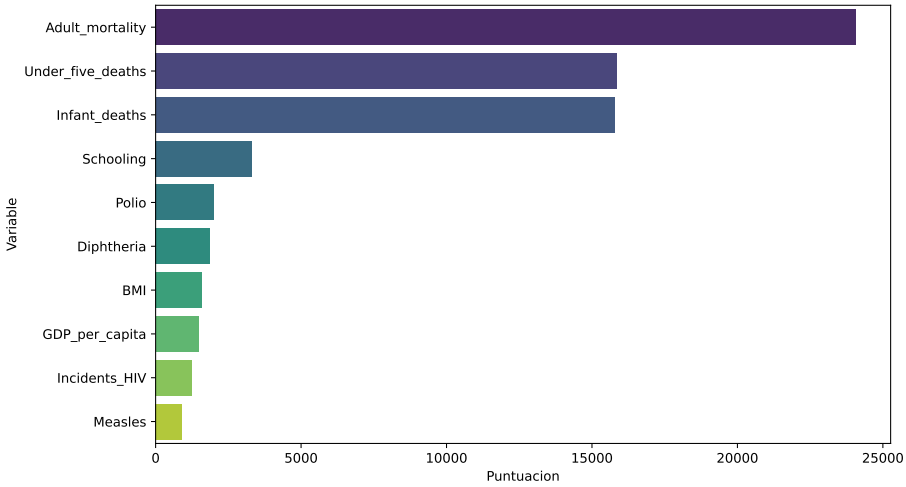


Figure 1: Top 10 variables ordenadas por su puntuación F

## 3.2 Análisis de relación de variables

Utilizando un preprocesador se busca encontrar las variables más relevantes con respecto a nuestra variable respuesta, para esto usaremos un selector basado en las K mejores variables utilizando una regresión f.

Se puede apreciar como La mortalidad adulta es precisamente uno de los mejores indicadores/más correlacionado con respecto a la expectativa de vida, lo cual tiene sentido, sin embargo, vamos a continuar con más pruebas para confirmar que variables pueden ser relevantes, empezando con el analisis 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 podemos apreciar como 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 que encuentro relevante 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.

### 3.3 Selección final de variables

Basado en los análisis previos he decidido deshacerme de las siguientes variables:

Population: No considero 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_year y Thinness_ten_nineteen_years: Creo que toda la información que nos puede presentar esta variable podría ser explicado con otras dos variables que si se van a incluir en el modelo (BMI y death 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 considero 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 considero que el sarampión no representa tanto un buen indicador con respecto al polio, esto se puede

apreciar dado que en prácticamente todas las pruebas el polio obtuvo un mejor valor que el sarampión

# 4 Metodología no supervisada

Para este analisis se usara un algoritmo de gas neuronal creciente (o GNS por sus siglas en inglés) partiendo de la librería `neupy` para la aplicacion del algoritmo, numpy para algebra 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 distribucion de los puntos al mismo tiempo que minimiza el error.

## 4.2 Fundamento matematico

Partiendo de 2 puntos aleatorios conectados entre si se sigue el siguiente algoritmo en cada iteracion 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 el es movido en funcion de los errores acumulados de cada nodo, en direccion del vector ingresado.
- 4: La variable "edad" de cada arista ya existente se aumenta en 1.
- 5: En caso de que el nodo más cercano y el segundo nodo más cercano esten conectados entre si 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 el arista.
- 7: En caso de que la iteracion actual sea multiplo de una variable previamente decida se agrega un nodo al azar al sistema conectando el nodo con el error más grande y su vecino más cercano, despues se elimina el 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 clusterizacion se llego a 20 clusteres identificados mismos que fueron graficados a lo largo del tiempo, la grafica general de todos los puntos de datos clusterizada se ve de la siguiente forma:

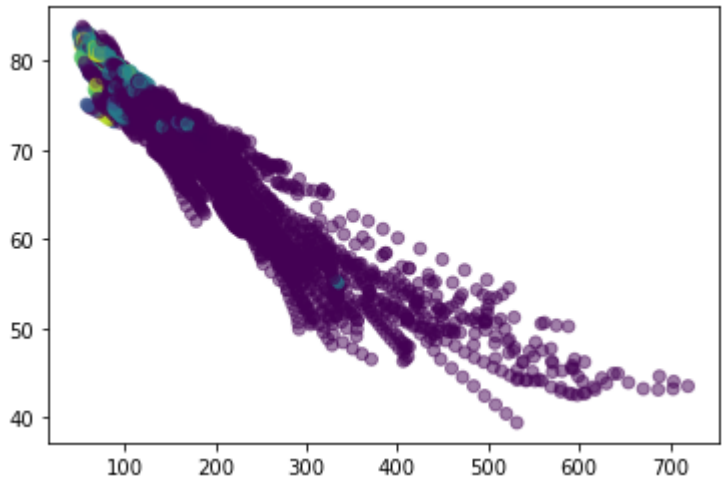


Figure 2: Clusterizacion en el año 2000

Como se puede apreciar en la figura 1 realmente no existe mucha distincion debido a la gran densidad de datos, por eso mismo se procedio a dividir los datos por año para ver si se podia obtener un mejor entendimiento de los clusteres

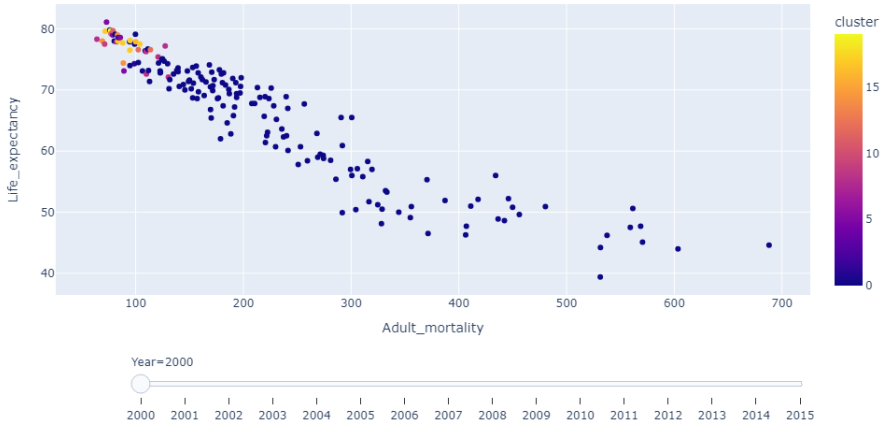


Figure 3: Clusterizacion en el año 2000

### 4.4 Conclusiones del algoritmo no supervisado

Se aprecia como la clusterizacion se mantiene relativamente constante al pasar los años y como 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 Metodologia supervisada

Para este análisis se tomará como variable objetivo la expectativa de vida para cada país, se utilizara un regresor de potenciación de gradiente apoyándonos 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 nuestra variable objetivo.

Debido a la baja cantidad de observaciones por serie de tiempo se utilizara un regresor que solo toma en cuenta las ultimas 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 nuestro modelo esta desviado de las observaciones reales por .5 años, osease 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.

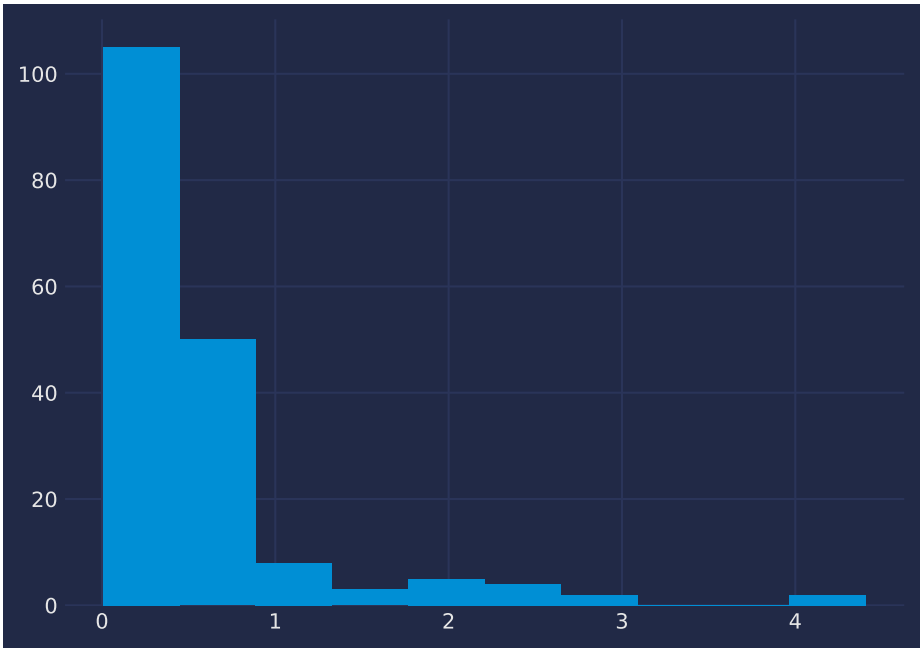


Figure 4: Histograma de error medio absoluto

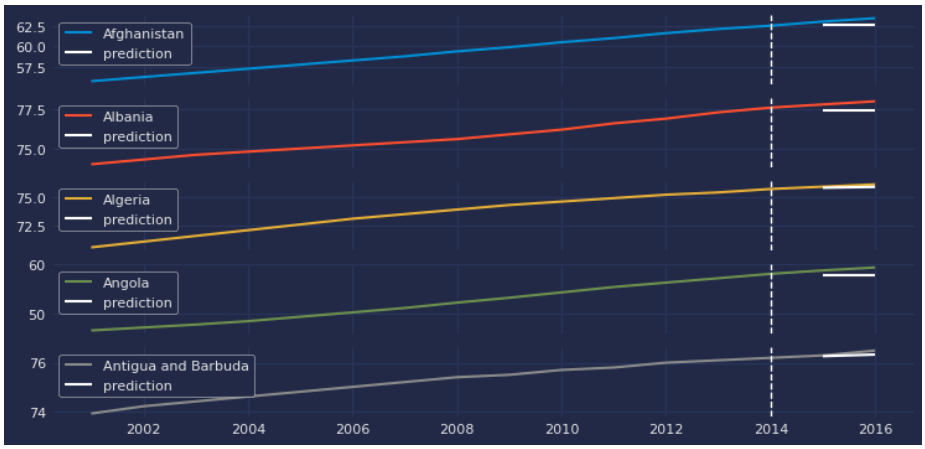


Figure 5: Predicciones comparadas con el conjunto de validación

## 5.2 Metricas 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.

Despues 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 grafico de caja (Figura 6) de la métrica “Promedio absoluto” podemos apreciar como 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 considero que es un error aceptable, siendo esto referente a máximo un año de la esperanza de vida dado los estimadores macroeconómicos usados.

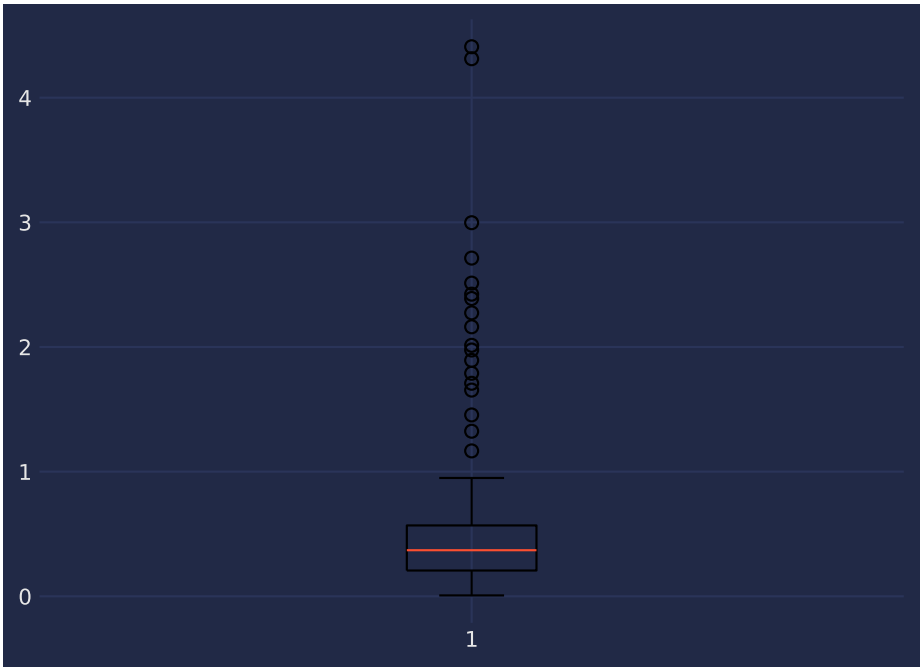


Figure 6: Grafico de caja del error promedio absoluto del modelo

# 6 Conclusiones

Creo que ambos enfoques, tanto supervisado como no supervisado plantean conclusiones diferentes pero relevantes, el modelo supervisado nos dice que dado las variables macroeconómicas adecuadas se puede llegar a predecir la expectativa de vida de un país aun contando con pocas observaciones, esto se sustenta en que aunque el modelo supervisado se entreno 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 a primera instancia 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 nos 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, nos permite ver que si bien estas variables constituyen un indicador suficiente para su predicción no representan toda la información relevante como para decir que

nuestro modelo esta completo. Las principales conclusiones que puedo observar aquí es que este es un estudio que definitivamente tiene futuro sin embargo para que el análisis se congruente y nos 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 considero que este análisis es importante debido a una razón muy específica, se hizo antes de la pandemia. Con esto me refiero a que de cierta manera este análisis provee una visión de las variables relevantes en un estudio de predicción de expectativa de vida, mientras que al mismo tiempo nos recuerda que en nuestros estudios siempre habrá variables que no se puedan controlar y que no se puedan predecir. No importa que tanta información pasada hubiera tenido ni que tanta variables significativas hubiera incluido no creo hubiera sido capaz de predecir el impacto de la pandemia en la expectativa de vida en lo más mínimo y precisamente por eso me decidí a 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