Tarea 6 Aprendizaje no supervisado

Universidad Autonoma de Nuevo León
Alanis Mares, Victor Hugo

18 de Julio, 2024

1 Introducción

Se pretende analizar un conjunto de series de tiempo con diferentes indicadores macroecónomicos para implementar un sistema de prediccion supervisada.

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 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.

2.2 Preprocesamiento

Basado en un análisis de características se decidió omitir ciertas variables del conjunto, para más informacion al respecto referirse a este repositorio

3 Metodología

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.

4 Resultados

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 1, 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.

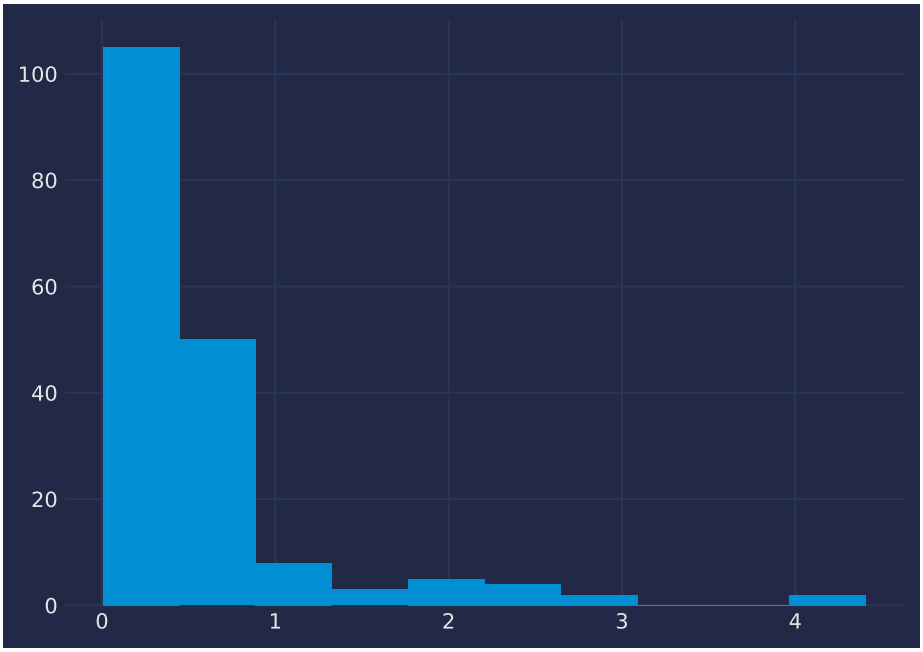

Figure 1: Histograma de error medio absoluto

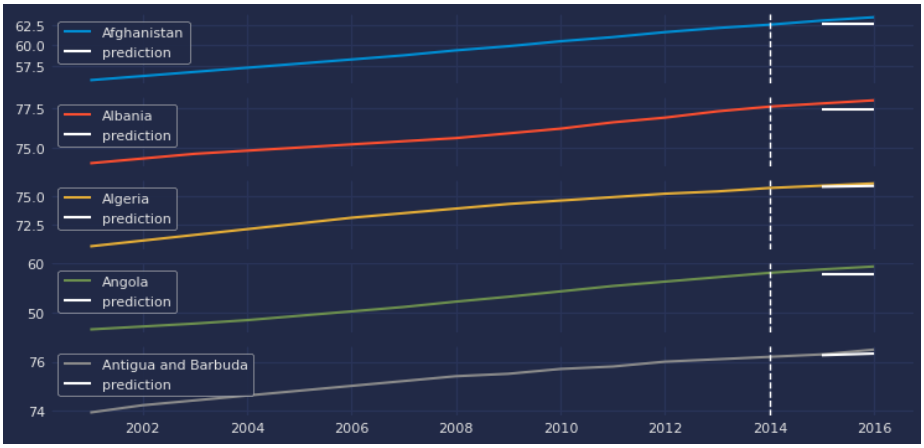

Figure 2: Predicciones comparadas con el conjunto de validación

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