

# Tarea 5 Aprendizaje no supervisado

Universidad Autonoma de Nuevo León  
Alanis Mares, Victor Hugo

July 8, 2024

## 1 Introducción

Se pretende analizar un conjunto de series de tiempo con diferentes indicadores macroeconomicos para implementar un sistema de clasificacion no supervisada.

## 2 Descripción de los datos

Los datos a trabajar cuentan con 16 variables de las cuales 11 son referentes a la serie de tiempo multivariada y 5 son categoricas. Para propósitos de este estudio de clusterización se utilizarán las 11 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: Indice de grasa corporal promedio
- Polio: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para la Polio
- Diphtheria: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para tetanos
- 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 analisis de características se decidió omitir ciertas variables del conjunto, para más información al respecto referirse a este repositorio

## 3 Metodología

Para este analisis se usara un algoritmo de gas neuronal creciente (o GNS por sus siglas en ingles) partiendo de la libreria `neupy` para la aplicacion del algoritmo, `numpy` para algebra matricial necesaria y `pandas` para manejo de datos.

## 4 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.

## 5 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 de 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 función de los errores acumulados de cada nodo, en dirección 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 estén 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 decidida 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

## 6 Resultados

Tras aplicar el algoritmo de clusterizacion se llevo 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:

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

## 7 Conclusiones

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.

## 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

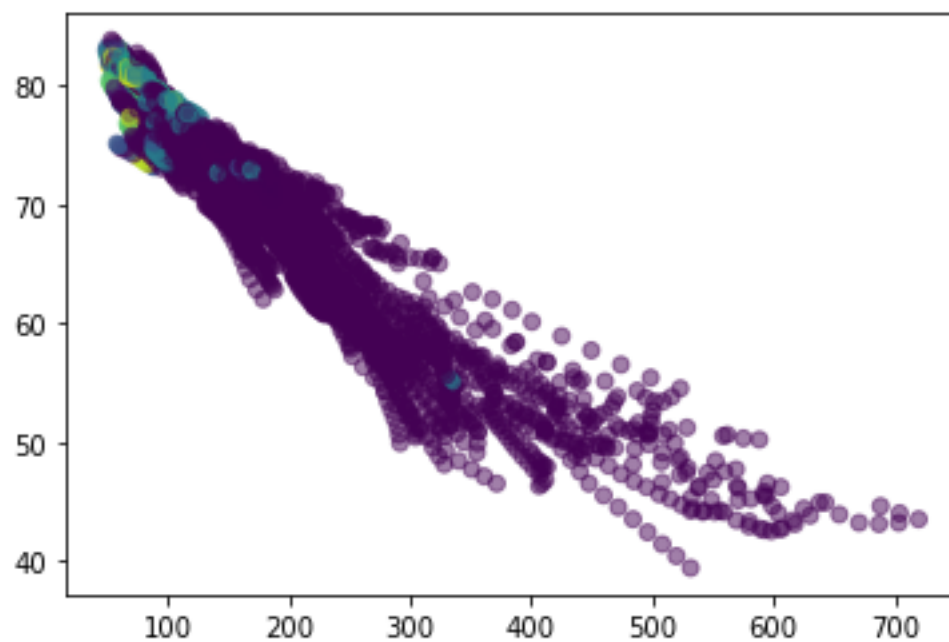


Figure 1: Clusterizacion en el año 2000

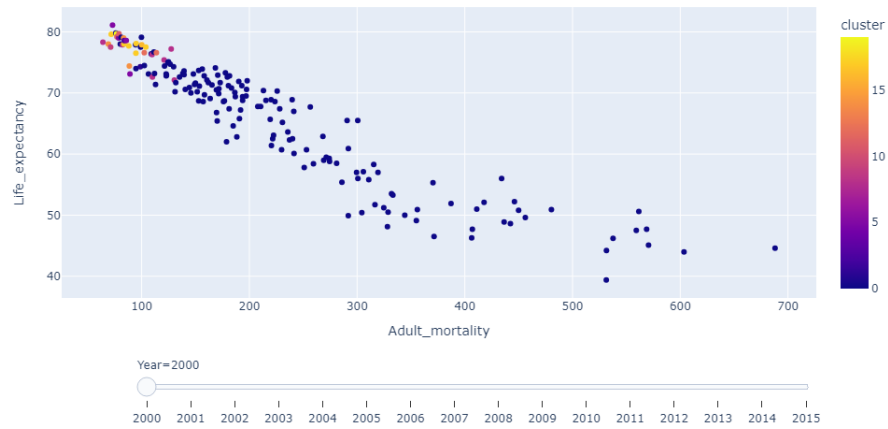


Figure 2: Clusterizacion en el año 2000

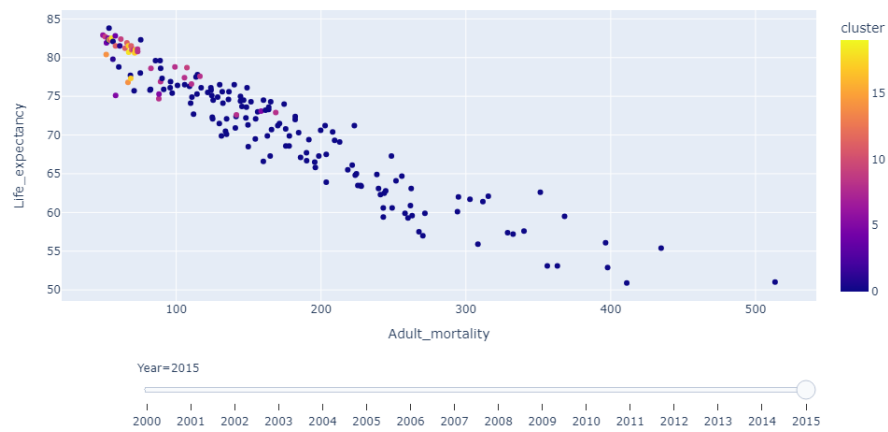


Figure 3: Clusterizacion en el año 2015