

TC2034 Proyecto de Aprendizaje No Supervisado

Equipo 6

Adrián Landaverde Nava A01745052

Naomi Padilla Mora A01745914

Sabrina Nicole Rodríguez Salgado A01745197

Índice

Introducción al conjunto de datos (dataset) seleccionado	2
Métodos de clustering	5
Método visto en clase	6
K-MEANS	6
Métodos de investigación	7
Affinity Propagation	7
Spectral Clustering	9
DBSCAN	12
Comparación de todos los métodos implementados	14
Conclusiones	15
Posibles mejoras de cada método y en general	15
Conclusiones generales	15
Conclusiones individuales	16
Adrián Landaverde Nava	16
Naomi Padilla Mora	16
Sabrina Nicole Rodríguez Salgado	17
Referencias	18

Introducción al conjunto de datos (dataset) seleccionado

Para el desarrollo de este proyecto se utilizó el dataset *Life Expectancy (WHO)*, esta base de datos es un archivo .csv *Life Expectancy Data.csv* compuesto por 22 columnas y 2938 filas. Este dataset se encuentra disponible en la plataforma Kaggle. Las variables de estudio o *features* que componen la base de datos son las siguientes:

Variable	Descripción	Tipo
Country	País. 193 países.	Cualitativa - object
Year	Año de estudio.	Cuantitativa - int
Status	Estado de desarrollo. Developing: En desarrollo, Developed: Desarrollado.	Cualitativa - object
Life expectancy	Esperanza de vida, edad promedio a la que se espera que llegue cada habitante según el país.	Cuantitativa - float
Adult Mortality	Tasa de mortalidad para ambos sexos (de entre 15 a 60 años por cada mil habitantes).	Cuantitativa - float
infant deaths	Cantidad de infantes fallecidos por cada mil habitantes.	Cuantitativa - int
Alcohol	Consumo de alcohol (litros) per cápita.	Cuantitativa - float
percentage expenditure	Porcentaje de gastos en salud como porcentaje del PIB per cápita (del total del PIB per cápita).	Cuantitativa - float
Hepatitis B	Cobertura de vacunación contra la Hepatitis B entre niños de 1 año.	Cuantitativa - float
Measles	Casos confirmados de sarampión por cada mil habitantes.	Cuantitativa - int
BMI	Índice de masa corporal promedio de toda la población.	Cuantitativa - float

under-five deaths	Muertes en menores de 5 años por cada mil habitantes.	Cuantitativa - int
Polio	Cobertura de vacunación con el Polio entre niños de 1 año.	Cuantitativa - float
Total expenditure	Porcentaje de gasto general del gobierno en el sector salud (de su total de gastos).	Cuantitativa - float
Diphtheria	Cobertura de vacunación contra la difteria entre niños de 1 año.	Cuantitativa - float
HIV/AIDS	Muertes por mil nacidos vivos VIH/SIDA (0-4 años)	Cuantitativa - float
GDP	Producto Interno Bruto Per Cápita (USD).	Cuantitativa - float
Population	Población en el país.	Cuantitativa - float
thinness 1-19 years	Porcentaje de prevalencia de delgadez en niños y adolescentes de 10 a 19 años.	Cuantitativa - float
thinness 5-9 years	Porcentaje de prevalencia de delgadez en niños de 5 a 9 años.	Cuantitativa - float
Income composition of resources	Índice de Desarrollo Humano en términos de composición de ingresos de los recursos (índice que va de 0 a 1).	Cuantitativa - float
Schooling	Número de años de escolaridad.	Cuantitativa - float

Tabla 1: Variables de estudio presentes en el dataset.

Justificación

La esperanza de vida (*Life expectancy* en inglés) constituye el tiempo promedio que una persona puede esperar vivir, dicho de otra forma, es la media de años que vive la población en un determinado lugar en cierto periodo. Esta medida varía dependiendo de factores como edad, sexo, origen étnico y ubicación geográfica (Bezy, 2022). Igualmente, la esperanza de vida suele servir como indicador de la salud general de una comunidad, además de que afecta el tipo de inversión que debe hacer un gobierno en planeación y provisión de servicios de salud (Utah Department of Health, 2021). Descubrir las condiciones o las políticas que contribuyen en cada país a un alza en la esperanza de vida

podría ser crucial para mejorar la salud global, esta es la razón por la que se escogió este dataset. Este dataset está centrado en esperanza de vida, pero tiene más datos relacionados con esta medida, por lo que consideramos que todos éstos, incluyendo esperanza de vida, podrían tratarse como *features*. Con datos de factores que influyen en la salud y esperanza de vida en los países alrededor del mundo, es buena idea identificar patrones en aquellos países que tienen estadísticas similares, con el fin de saber qué es lo que realmente contribuye a una mejor salud y cómo dichos países han podido, o no han podido, llegar a ella. Nos pareció que agrupar países con distintos métodos de *clustering* nos podría dar información valiosa sobre la salud general de cada país y cada agrupación. Por lo tanto, al usar todas estas variables relacionadas a la salud, se pueden encontrar patrones sobre las decisiones que toma cada país relacionados a sus temas de salud y cómo se relaciona con la esperanza de vida.

Preparación de datos

Liga del notebook de limpieza:

<https://colab.research.google.com/drive/1G4-OmJcKIWUdpAdvo0LDszKURGGczyQN?usp=sharing>

Posterior a la selección de la base de datos y la familiarización con cada una de las variables se realizó un procesamiento de datos que nos permitiría continuar con la aplicación de los métodos de clustering una vez que este finalizara. Comenzamos analizando el tipo de variable, concentrándose en las variables cualitativas - *object*, ya que para llevar a cabo la implementación de los métodos es necesario que todas las variables sean cualitativas. Por ello, se identificó a las variables *Country* y *Status* como *object*, mientras que la variable *Country* se dejará como variable cualitativa ya que esta permitirá saber de qué país estamos hablando y no es necesario “renombrarla”, la variable *Status* se renombró como *Developing: 0 , Developed: 1*.

Debido a que el dataset cuenta con 2938 filas que contienen información de 193 países desde el año 2000 hasta el 2015, se decidió dejar solo una fila para cada país. Para que esto fuera posible se promediaron los datos de cada una de las variables por país, dando por resultado un dataset de 22 columnas por 193 filas. Una vez realizado esto, se buscó la cantidad de datos nulos en cada una de las variables, obteniendo que *Life expectancy* 10 nulos, *Adult Mortality* 10 nulos, *Alcohol* 2 nulos, *Hepatitis B* 9 nulos, *BMI* 4 nulos, *Total expenditure* 2 nulo, *GDP* 30 nulos, *Population* 48 nulos, *thinness 1-19 years* 4 nulos, *thinness 5-9 years* 4 nulos, *Income composition of resources* 17 nulos, *Schooling* 13 nulos.

Antes de darle un tratamiento a estos datos nulos por cada variable, decidimos eliminar las variables que consideramos no serían útiles para el enfoque planteado en el análisis. Las variables eliminadas fueron las siguientes:

- *Year*: Ya que se realizó el promedio de los datos ya no era necesario.
- *Alcohol*: No se considera de valor para el análisis y es información difícil de encontrar para sustituir los valores nulos.
- *Total expenditure*: Es muy similar al *percentage expenditure* y este no contiene nulos.
- *thinness 1-19 years*: No se considera de valor para el análisis y es información difícil de encontrar para sustituir los valores nulos.

- *thinness 5-9 years*: No se considera de valor para el análisis y es información difícil de encontrar para sustituir los valores nulos.
- *Income composition of resources*: No se considera de valor para el análisis y es información difícil de encontrar para sustituir los valores nulos.
- *Schooling*: No se considera de valor para el análisis y es información difícil de encontrar para sustituir los valores nulos.

Ahora, ya que solo contamos con las variables que se consideran relevantes para el análisis, se importó el dataset actualizado como un archivo csv *archivo1.csv* para rellenar los datos nulos restantes de manera manual. Observando que las variables que los contenían son información fácil de obtener por medio de una investigación en la red. Se decidió usar las bases de datos *DataBank* de la página *The World Bank*. Al analizar qué países eran los que contenían valores nulos en cada variable notamos que 13 países se repetían dentro de estas variables y que no se encontraba disponible la información que necesitábamos en su mayoría. Por lo tanto, se decidió eliminar las filas de estos 13 países, siendo los siguientes según su índice, Cook Islands, Democratic People's Republic of Korea, Dominica, Marshall Islands, Monaco, Nauru, Niue, Palau, Saint Kitts and Nevis, San Marino, The former Yugoslav republic of Macedonia, Tuvalu y Venezuela (Bolivarian Republic of). Obteniendo un dataset con 180 países y 15 variables. De tal manera que la preparación de los datos fue la misma para todos los métodos implementados.

Tras realizar la sustitución de los datos nulos, se creó un nuevo notebook con la base de datos actualizada y limpia *life.csv*. Se comprobó el tipo de cada variable y que ninguna contara con ningún valor nulo. Se cargó el archivo *World_Countries_Generalized_.shp*, el cual es un *shapefile* que contiene diversas columnas, entre las más importantes *geometry*, donde podemos encontrar las geometrías de cada uno de los países para poder modelarlos más adelante e ilustrar los resultados de la situación. Finalmente, se agregó a la base de datos *life.csv* la columna *geometry* del shapefile.

Se estandarizaron las variables para que todas tuvieran el mismo peso en el análisis. El rango de valores que abarcaba cada variable era muy distinto en varios casos, por ejemplo, comparando los valores que toman *infant deaths* y *Population*. Luego, para reducir las dimensiones del conjunto de datos, se usó un Análisis de Componentes Principales (PCA) con 2 componentes, el cual fue necesario en el caso de Spectral Clustering, y además se usó como las nuevas dimensiones sobre las que se haría el análisis y la graficación de cada método.

Métodos de clustering

Todos los métodos de aprendizaje no supervisado se realizaron en el mismo notebook.

Liga del notebook de métodos de aprendizaje no supervisado:

https://colab.research.google.com/drive/1dSIOuPSdZPSdGYksFhYVz_Tc4p7Te5wu?usp=sharing

Mapa con los resultados de los métodos:

<https://e6lifenosupervisado.netlify.app/>

Método visto en clase

K-MEANS

El algoritmo de K-Means agrupa datos tratando de separar muestras en n clusters de igual varianza y minimizando la inercia, o distancia entre puntos dentro del cluster (scikit-learn, 2022). Requiere la especificación del número de clusters a usar. El algoritmo inicia con un varios centroides seleccionados aleatoriamente, los cuales funcionan como el punto de partida para cada cluster, luego se hacen cálculos iterativos para optimizar las posiciones de dichos centroides (Garbade, 2018). El algoritmo termina cuando los centroides se han estabilizado o cuando se ha alcanzado el número máximo de iteraciones. Entre las ventajas del método es que es relativamente simple de implementar, funciona bien con conjuntos de datos grandes, adapta fácilmente nuevas observaciones y generaliza a clusters de distintos tamaños y formas (Google Developers, 2021). Sin embargo, entre sus desventajas están que hay que escoger el número de clusters, que tiene problemas para agrupar datos de distinta densidad, y también tiene problemas con los *outliers*, que pueden desplazar los centroides.

- a) Persona encargada del método: *Adrián Landaverde Nava, Naomi Padilla Mora, Sabrina Nicole Rodríguez Salgado*
- b) Implementación y entrenamiento

Para escoger el número de clusters k en este método, se usó el método del codo, el cual funciona graficando la SSE (*Sum of Squared Error*) para cada número distinto de clusters en un rango especificado, todo esto es equivalente a usar la inercia para determinar el número de clusters. Con este método se detectó una disminución en el SSE a partir de $k=4$, por lo que este fue el número de clusters empleados. Se usó un número máximo de iteraciones de 300. Para esto se conservó la dimensionalidad de los datos, para ver cómo se comportaba el algoritmo con las 14 dimensiones. Resultó que el modelo tuvo un buen rendimiento, como se detalla abajo.

- c) Evaluación del modelo entrenado.

El modelo pudo dividir los grupos adecuadamente, con un *Silhouette Coefficient* de 0.34381192387449017, y como se aprecia en la Figura 1. Puede notarse que incluso usando las 14 dimensiones del dataset, este algoritmo obtuvo clusters bien definidos y que tienen sentido a la hora de verificar los resultados manualmente.

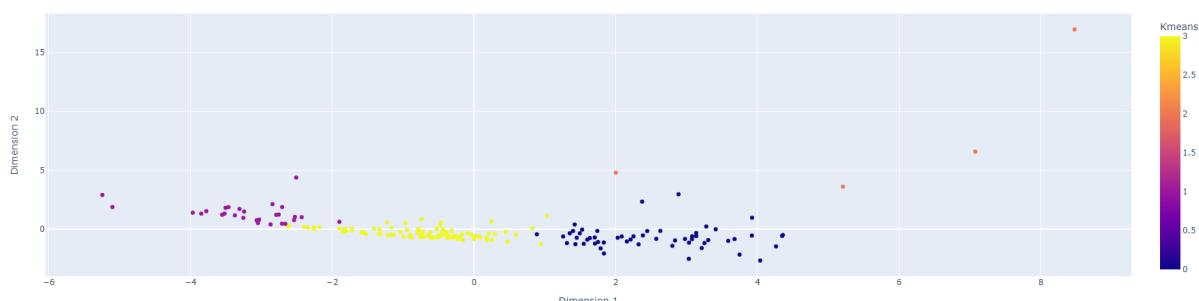


Figura 1: Gráfico de dispersión de la clasificación con K-Means

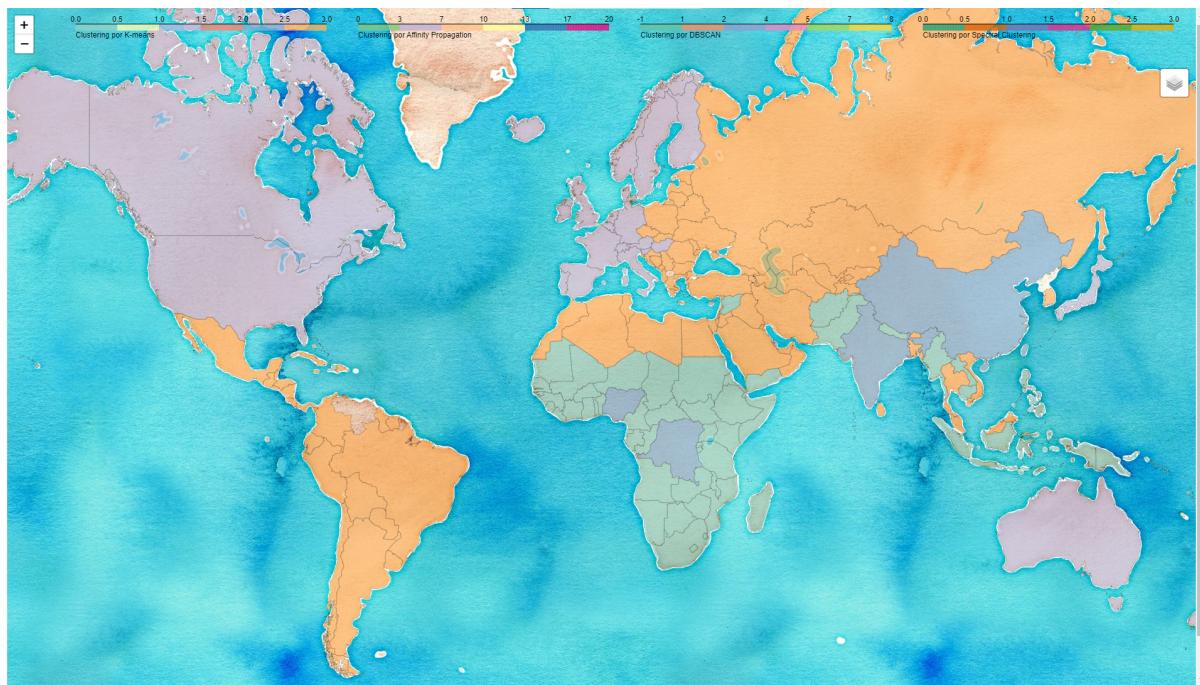


Figura 2: Mapa de la clasificación de los países con K-Means

En la Figura 2 pueden visualizarse los 4 distintos grupos de países a los que llegó el algoritmo en cuanto a condiciones de salud.

- d) Conclusiones del método con la base de patrones utilizada (Particularidades del método con el dataset utilizado y Comportamiento para predecir).

Pudo verse que, si bien los datos tenían estaban acumulados con diferentes densidades, K-means pudo llegar a una agrupación correcta desde el punto de vista teórico, tomando en cuenta el *Silhouette Coefficient* y también a la hora de evaluar los resultados empíricamente. Puede verse cómo se agruparon las regiones de Latinoamérica, Europa oriental, y el norte de África en un sólo cluster, por ejemplo. En general, el método se desempeñó bien con el conjunto de datos y alcanzó a hacer clusters bien definidos.

Métodos de investigación

Affinity Propagation

Este algoritmo, a diferencia de muchos algoritmos de clustering, se caracteriza por poder funcionar sin tener que especificar un número de clusters. Esto debido a que se basa en el cálculo de matrices, que, a grandes rasgos, hace que cada punto envíe información a los otros puntos sobre el parecido de cada uno de estos datos con otros. Estos cálculos se iteran hasta que se llega a un “consenso” sobre los valores iguales, y los datos se agrupan de acuerdo a los puntos que tengan los mismos valores (Maklyn, 2019)

- a) Persona encargada del método: Adrián Landaverde Nava
- b) Implementación y entrenamiento

Entre las características de este algoritmo de aprendizaje no supervisado incluyen que los datos de entrenamiento pueden ser datos de muchas dimensiones, pero con datos estandarizados para que las distancias entre los puntos no afecten el resultado final. Por lo tanto, se usó el dataset con todas las dimensiones de los datos estandarizados. Este método define la cantidad de clusters ‘automáticamente’ con base en los cálculos matriciales que realiza, por lo tanto, al final obtuvo 21 clusters. Por lo tanto, sólo se probaron diferentes números de iteraciones para este método, desde 200 (la cantidad por default) hasta 10,000. Sin embargo, en todos los casos se obtuvo el mismo resultado. Por lo tanto se dejaron las condiciones por default.

c) Evaluación del modelo entrenado.

Para evaluar el modelo se usaron diferentes técnicas, ya que es un método de aprendizaje no supervisado, no existe un target real con el qué comparar los resultados. La primera evaluación consistió en realizar una reducción de las dimensiones de las variables a 2 dimensiones y hacer un plot de cada una de las categorías de los datos, como se muestra en el gráfico de la figura 3.

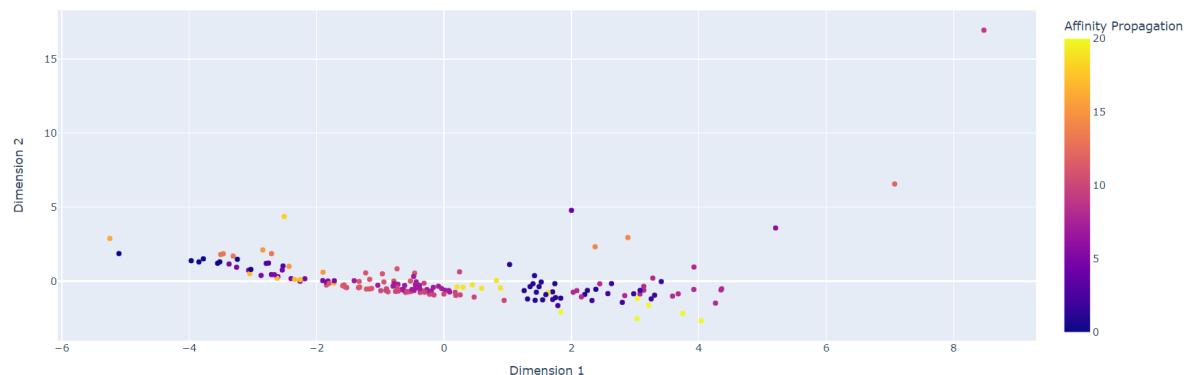


Figura 3: Gráfico de dispersión de la clasificación con Affinity Propagation

Asimismo, otro método usado fue el de realizar un mapa de los países seleccionados y graficarlos de manera que se mostraran con un color diferente con base en su cluster. En la figura 4 se muestra el clustering de los países con base en este método. Aunque sólo pueden observarse 6 colores diferentes en el mapa, en realidad se incluyen los 21 clusters formados, pero algunos clústeres están coloreados del mismo color. Este mapa puede ser estudiado a detalle en el siguiente link <https://e6lifenosupervisado.netlify.app/> seleccionando la capa de “Affinity propagation”

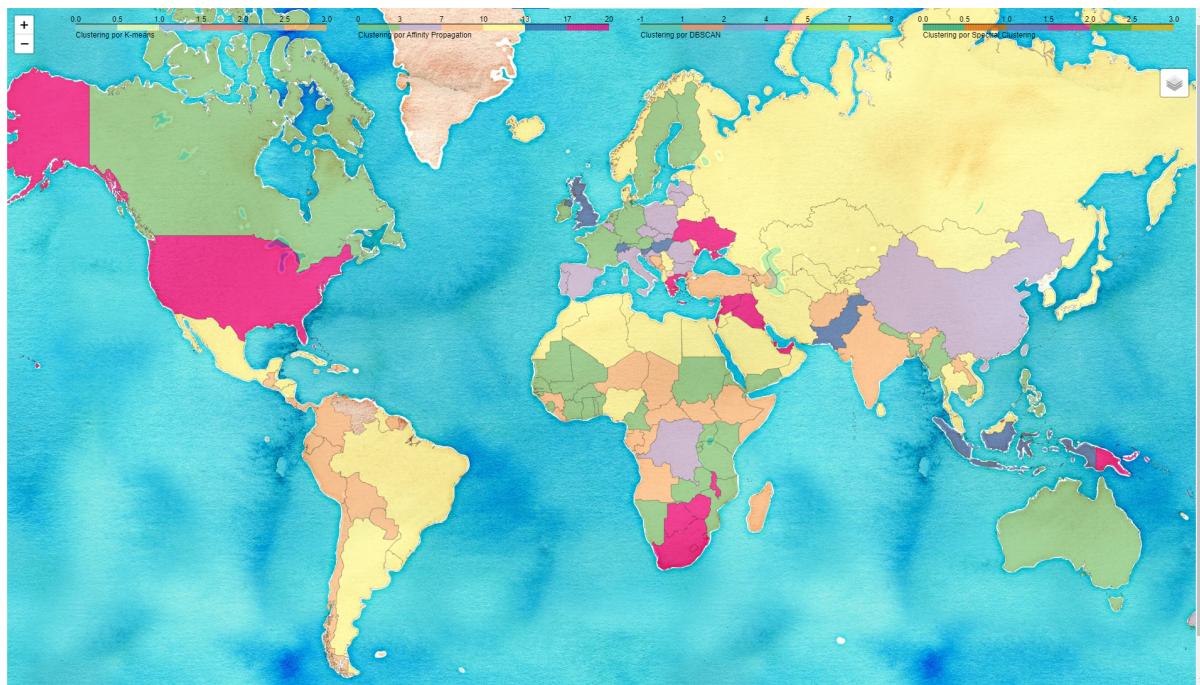


Figura 4: Mapa de la clasificación de los países con Affinity Propagation

Finalmente, otra medida para calificar el método se usó la métrica de silhouette, misma que calcula la distancia interclusters de los datos. Usando esta medida se obtuvo un *score* de 0.2518, mismo que se encuentra como un puntaje medio en comparación de los otros métodos.

- d) Conclusiones del método con la base de patrones utilizada Particularidades del método con el dataset utilizado y Comportamiento para predecir.

En cuanto a este método, se puede concluir que es capaz de analizar datos muy complejos en muchas dimensiones. Sin embargo, al ser muy complejo, también ocasiona que los datos se agrupen en clusters muy pequeños, por lo que, aunque sí se agrupan países con variables muy parecidas, el algoritmo es tan complejo que sólo los agrupa si sus variables son muy parecidas. Esto hace que se tengan clusters de hasta 27 elementos, y otros clusters de sólo 1 elemento. Por lo tanto, este método es capaz de encontrar patrones muy similares en los datos pero no patrones muy generales. Aunque el mapa es muy bueno para comparar los resultados, dado que no se tienen tantos colores, parecería que se mezclan ciertos grupos, por lo que también se vuelve difícil encontrar diferencias entre los clusters.

Spectral Clustering

El Spectral Clustering es un algoritmo de agrupamiento que en muchos casos ha funcionado mejor que muchos algoritmos de agrupamiento tradicionales. Este consiste en tratar cada punto de datos como un nodo gráfico y, por lo tanto, transforma el problema de agrupamiento a uno de partición de gráficos. Si la matriz de afinidad es la matriz de adyacencia de un gráfico, este método se puede usar para encontrar la partición del gráfico. La matriz de afinidad se construye utilizando una función de kernel, como el kernel gaussiano, con distancia euclidiana. La matriz de afinidad también puede ser configurada

por el usuario o se puede utilizar la matriz de conectividad *k-nearest neighbors* (Scikit-learn, 2022). Para poder aplicar este método, es necesario realizar una disminución de dimensiones, explícitamente a 2D (GeeksforGeeks, 2019).

- a) Persona encargada del método: *Naomi Padilla Mora*
- b) Implementación y entrenamiento

Para poder implementar el método a la base de datos se necesita conocer el número de clusters a realizar. Comúnmente, el valor por default suele ser ocho clusters sin embargo, para obtener resultados precisos y eficientes se utilizó el método del codo para identificar el número de clusters necesarios, que en este caso, resultó ser 4 clusters. De igual forma, fue sumamente relevante que al igual que con los otros métodos se tuvo que realizar la estandarización de los datos. Para el método en específico, se requiere utilizar la *X_scaled_pca*, ya que la Análisis de Componentes Principales (PCA) es lo que permite hacer la reducción de 15 a 2 dimensiones, que como se ha mencionado es el paso principal para poder aplicar el método de manera eficiente, no solo para moldearlo.

Para el parámetro *assing_labels*, que es el método a implementar para la forma en la que se asignan los datos a los clusters, se decidió utilizar *kmeans* porque suele ser una de las más comunes y presenta cierto grado de sensibilidad que puede favorecer a los resultados, además que al realizar pruebas con los diversos métodos fue el que presentó mejores resultados. Por ejemplo, al realizarse las pruebas, se probó aplicar el método *discretize*, el cual presenta menor nivel de sensibilidad, sin embargo este no clasificaba a ningún país en el cluster 3. Además, en este caso no se utiliza la matriz de afinidad puesto que el método *kmeans* del *assing_labels* no la requiere.

- c) Evaluación del modelo entrenado

El modelo generó los clusters con las siguientes agrupaciones: cluster 0 contiene 145 países, cluster 1 contiene 1 país, cluster 2 contiene 5 países y el cluster 3 contiene 29 países. Como se puede observar, tanto en estas cifras como en la figura 5, los clusters no se encuentran compuestos de manera equilibrada, sin embargo dentro de cada cluster podemos apreciar una agrupación *eficiente* a simple vista. Para corroborar el rendimiento de clasificación del algoritmo Spectral Clustering se utilizó el *Silhouette Coefficient*, con el cual se obtuvo un valor de 0.35388461627917484, implicando que el rendimiento es bueno.

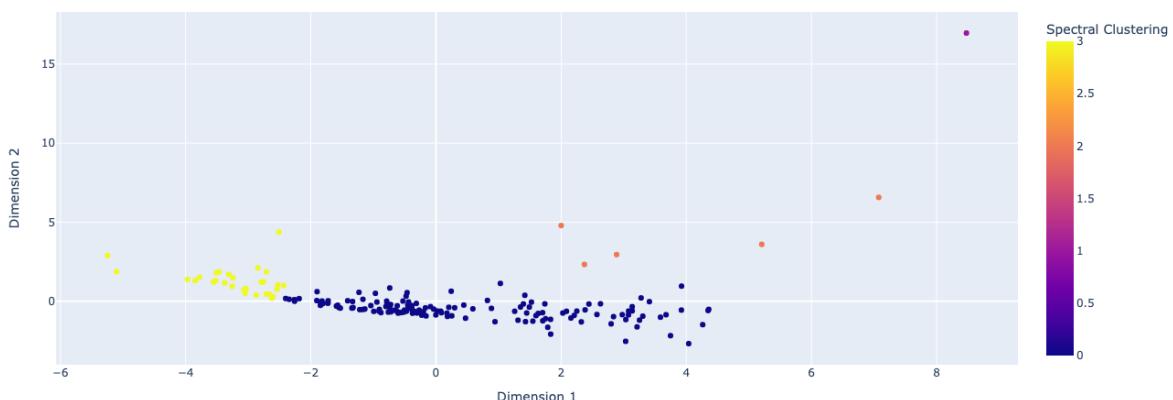


Figura 5: Gráfico de dispersión de la clasificación con Spectral Clustering

Para poder apreciar de manera más clara la clasificación realizada por el método, se observa la figura 6. En esta podemos apreciar, de manera más clara, que la clasificación es bastante acertada. El que sean 4 clusters ofrece dos perspectivas. Por un lado, hay una muy buena diferenciación de clusters con tan solo los colores que permite reconocer la eficiencia de clasificación en el grupo naranja, donde podemos ubicar a países con las mejores condiciones en calidad de vida, incluyendo en el sector salud. Se observa cómo agrupa a potencias como Estados Unidos, Canadá, Japón, Australia y los países más fuertes de Europa. Por otra parte, vemos que la mayoría de países se concentra en un solo cluster, el cluster 0, demostrando la desventaja de solo tener 4 grupos, ya que asimila que las condiciones son muy similares entre todos países porque ya no hay más clusters para separarlos. Colocando en el mismo grupo a la mayoría de los países del continente Asiático, Africano y Latinoamérica, de lo cual, podemos reconocer que no es tan exacto.

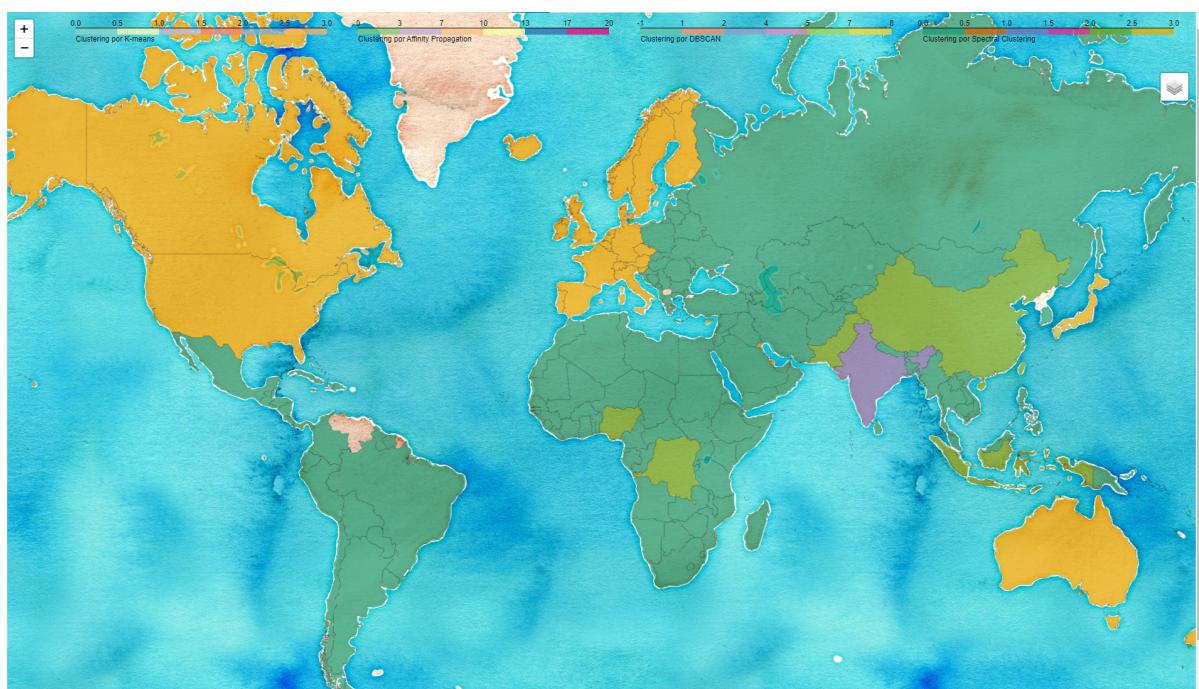


Figura 6: Mapa de la clasificación de los países con Spectral Clustering

- d) Conclusiones del método con la base de patrones utilizada (Particularidades del método con el dataset utilizado y Comportamiento para predecir)

La implementación del algoritmo Spectral Clustering suele ser bastante eficiente y es considerada uno de los mejores métodos de clasificación en el aprendizaje no supervisado. Incluso en esta problemática, tuvo un gran desempeño según observamos el puntaje obtenido en el *Silhouette Coefficient*. Sin embargo, es muy importante remarcar que método que se seleccione en el `assing_labels` implica en gran medida el desempeño del modelo. Es decir, en la implementación de este, se aplicó el método *Kmeans* por ser uno de los más utilizados y fue el que ofreció mejores resultados, y se obtuvieron buenos resultados. Pero, este método se caracteriza por su sensibilidad al momento de clasificar y por eso podemos explicar la gran concentración de países en el cluster 0.

DBSCAN

El DBSCAN (Density-Based Spatial Clustering of Applications with Noise o Agrupamiento Espacial Basado en Densidad de Aplicaciones con Ruido) identifica regiones en las que haya puntos acumulados (densidad) en contraste con regiones poco densas, o con puntos muy separados. Lo que hace es básicamente identificar puntos “ruido” y excluirlos, para luego agrupar el resto de los datos, los cuales están más densamente agrupados desde un inicio, esto se hace detectando *core points*, puntos directamente alcanzables y *outliers* (Versloot, 2022).

- a) Persona encargada del método: *Sabrina Nicole Rodríguez Salgado*
- b) Implementación y entrenamiento

Para funcionar, este algoritmo necesita dos parámetros principales: *epsilon* y “*minPts*”, el número mínimo de muestras cerca de un punto para que éste se pueda considerar un *core point*. Un punto p es un *core point* si al menos “*minPts*” puntos están a una distancia *epsilon* de él (este número de puntos incluye a p). El método no toma como parámetro el número de clusters, éste se calcula automáticamente dependiendo de *epsilon* y *minPts*. Sin embargo, para el cálculo del valor óptimo de *epsilon* se debe seguir el método del codo, se calcula la distancia promedio entre cada punto y sus k vecinos más cercanos. El valor óptimo de *epsilon* se encuentra en el punto con mayor curvatura de la gráfica. Para hacer lo anterior se escogió un valor de 11 vecinos, para tomar en cuenta un número razonable de vecinos cercanos para cada punto y que cada respectiva distancia tomara en cuenta más muestras. Esto dio un valor óptimo del parámetro de: 1.1105896604679124. A partir de aquí, quedaba seleccionar el otro parámetro: *minPts*, el cual se eligió de forma empírica, visualizando los resultados en una gráfica como la de la figura 7. Dado que DBSCAN busca hallar los puntos “ruido”, se usó un *minPts* de 3, lo cual maximizaba el número de agrupaciones de manera que tuvieran sentido y minimizaba el número de puntos de ruido, el cual siempre era muy alto.

- c) Evaluación del modelo entrenado.

El método dio como resultado 9 clusters, aparte de la agrupación de los puntos de ruido. Como se mencionó, el DBSCAN siempre detectó muchos puntos de este tipo y en general hubo clusters de tamaños muy variados, el cluster 0 contuvo 68 puntos, mientras que el cluster 1 contuvo sólo 3. El rendimiento de este algoritmo no fue tan satisfactorio por el número excesivo de puntos de ruido: 62, lo cual representa más de una tercera parte de las instancias en el conjunto de datos. A pesar de usar el método del codo para estimar *epsilon* y de experimentación con varios intentos para estimar *minPts*, el algoritmo clasificó como ruido una gran parte de los datos, lo que hace que se descarten dichos puntos y que no se pueda obtener más información a partir de esto. Una desventaja del aprendizaje no supervisado es que frecuentemente requiere validación por parte de una persona, y en el caso de este algoritmo todo esto fue necesario para ver si la asignación de clusters y puntos de ruido tenía sentido.

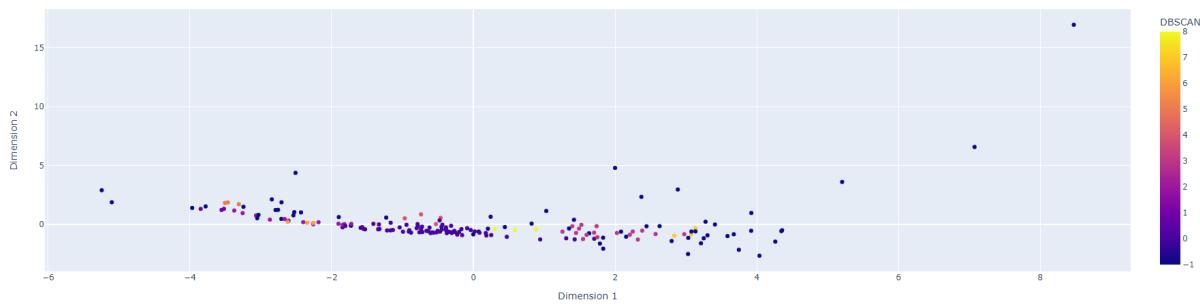


Figura 7: Gráfico de dispersión de la clasificación con DBSCAN

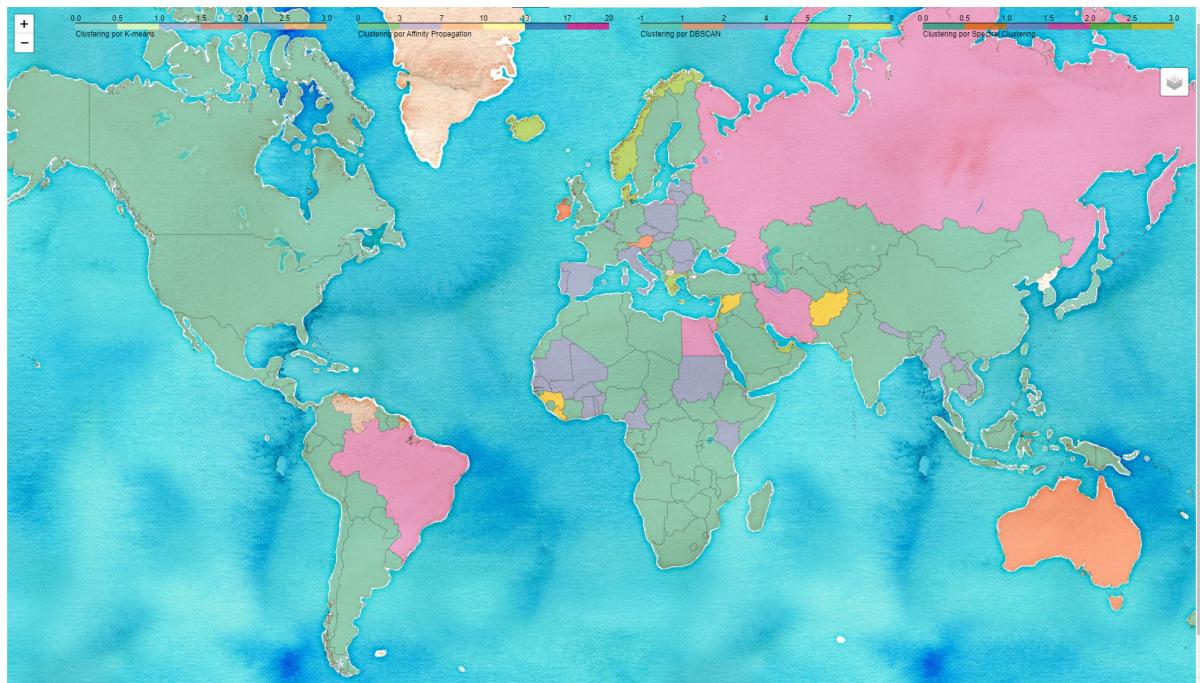


Figura 8: Mapa de la clasificación de los países con DBSCAN

Como métrica de evaluación se usó el *Silhouette Coefficient*, el cual tuvo un valor de -0.002850664119779683. Recordando que este número puede tomar valores de -1 a 1, y mientras más alto, mejor definidos están los clusters, el DBSCAN tuvo un rendimiento regular. También hay que notar que los puntajes de este coeficiente que estén alrededor de 0 indican clusters que se sobreponen, o sea que están “encimados”, lo cual se alcanza a ver claramente en la Figura 7. Otro aspecto importante es que este coeficiente suele ser más alto para clusters convexos, más que para otro tipo de clusters, como son los basados en densidad, que es como los que hace DBSCAN. De cualquier forma se detectó demasiado ruido. En el mapa también puede apreciarse la gran cantidad de puntos ruido, representados por verde.

- d) Conclusiones del método con la base de patrones utilizada (Particularidades del método con el dataset utilizado y Comportamiento para predecir).

El uso del DBSCAN se recomienda principalmente en datasets que tengan mucho ruido, pero en vista de los resultados anteriores, es probable que este método no sea el óptimo a

emplear en este conjunto de datos. Probablemente sería más útil usar este método si hubieran muchas más observaciones (lo cual no es posible en este caso, tratándose de países del mundo), donde la detección de ruido podría ser mucho más útil. Igualmente, hay que recordar que DBSCAN tiende a formar un número bajo de clusters y a detectar fácilmente ruido, en el contexto de agrupar países de acuerdo a sus estadísticas de salud tal vez no sea lo más adecuado, ya que lo que se quiere es encontrar la máxima información posible respecto a estas agrupaciones de países y sus similitudes.

Comparación de todos los métodos implementados

Los algoritmos implementados para la clasificación del dataset *Life Expectancy* presentaron un rendimiento de grado regular a bueno. Esto quiere decir, que algunos métodos se comportaron mejor que otros. Basándonos en el puntaje de desempeño obtenido en el parámetro *Silhouette Coefficient*, el algoritmo Spectral Clustering es el de mejor puntaje. Seguido por el K-Means. Sin embargo, comparando ambos métodos, podemos observar que en la clasificación de los países, según la gráfica de dispersión y el mapa, el K-Means tiene un mejor desempeño. Sobre todo destacando que son menos los países que se concentran en un solo cluster, haciendo menos general la clasificación de estos, cosa que en el Spectral Clustering vemos como un gran grupo de países por las similitudes más generales. Además, es importante destacar que estos métodos comparten ciertas similitudes, como que ambas implementaron el método del codo para el número de clusters. Recordando que el Spectral Clustering usa el método *kmean* para asignar sus etiquetas de clasificación, método que proviene del mismo algoritmo K-Means tradicional.

El DBSCAN tuvo el rendimiento más deficiente, ya que detectó demasiado ruido y, por lo tanto, excluyó parte importante del dataset. A pesar de que se usó un método analítico, el método del codo, para estimar uno de los parámetros del método y de que se experimentó con varios valores para el otro parámetro, aún así el método no se comportó de la mejor manera. Lo más probable es que este método no sea adecuado para el problema y el conjunto de datos en cuestión, pero fue bueno implementarlo en contraste con el resto de los algoritmos empleados.

Una consideración destacable entre todos los métodos, es que el Clustering es el único que realiza su procesamiento con los datos que ya pasaron por la reducción de dimensiones. Mientras que el resto, K-Means, Affinity Propagation y DBSCAN, utilizan los datos escalados. Esto se debe a que los últimos tienen la capacidad de procesar datos de más de 2 dimensiones, que en este caso contábamos con 15 dimensiones. En cambio, el Spectral Clustering está diseñado para únicamente dos dimensiones debido al corte gráfico que debe emplear en la clasificación de datos.

Conclusiones

Posibles mejoras de cada método y en general

En general podría decirse que un área de oportunidad es usar variables que tengan una relación más directa con la salud. En el dataset que elegimos, aunque sí está relacionado a la salud, hay algunas variables que están relacionados de manera indirecta y otras de manera directa. Por lo tanto, una área de oportunidad sería buscar más variables que estén directamente relacionadas con la salud y así se tenga un estudio más amplio sobre la agrupación de los países y poder realizar la abstracción de todos estos datos tan complejos

Para mejorar el método K-Means, podría probarse usar las *features* de dimensiones reducidas, es decir, a las que se les aplicó el PCA. A pesar de que el algoritmo de K-Means tuvo un buen rendimiento comparado con los otros métodos, incluso con 14 dimensiones del dataset, sería buena idea hacer más pruebas para evaluar los resultados con menos dimensiones. Esto porque se sabe que K-Means se vuelve menos eficiente a medida que aumenta el número de dimensiones. En el caso de Affinity Propagation, este método, una posible mejora es la de usar menos variables o hacer una mejor reducción de las dimensiones del dataset. Esto con el propósito de que el método sea menos complejo y así no se creen muchos clusters con pocos datos, sino que se creen reglas generalizadas.

Una de las posibles mejoras en el Spectral Clustering sería hacer más pruebas con los diversos métodos disponibles para el parámetro *assing_labels*, los cuales pueden ser el *kmeans*, *discretize* o *cluster_qr*, ya que cada uno cuenta con diversas características que pueden beneficiar o no a los resultados según la base de datos. Además, la implementación de otros parámetros como pueden ser el *affinity*, que se encarga de las diversas opciones en las que se puede construir la matriz de afinidad, la cual en este caso no se utilizó porque ya se contaba con el método de Affinity Propagation. Otro de los parámetros *n_neighbors* que puede complementar al parámetro anterior, *affinity='nearest_neighbors'*, dando el número de n vecinos que se requiere implementar.

En el caso de DBSCAN, algo que se recomienda en algunos casos para aplicar este método es eliminar puntos de ruido. Esto tal vez hubiera mejorado el rendimiento general de este algoritmo, pero hubiera implicado remover del dataset instancias útiles de países, lo cual no era el punto de este análisis. De cualquier manera se hubiera podido experimentar aún más con los dos parámetros que toma el DBSCAN, para ver si lo anterior podría mejorar los resultados.

Conclusiones generales

Para la evaluación de los algoritmos se usó el *Silhouette Coefficient*, que toma valores de -1 a 1, y un valor mayor significa que los clusters están mejor definidos, tomando en cuenta la inercia y la distancia media entre una muestra y todos los otros puntos en el cluster más cercano. Usando esta medida numérica se tuvo que el método de *Spectral Clustering* fue el que mejor rendimiento tuvo, con un puntaje de 0.35388461627917484. En segundo lugar quedó el algoritmo de *K-Means* con 0.34381192387449017, seguido *Affinity Propagation*, con un *silhouette score* de 0.25183584181868046. Por último, DBSCAN fue el método que

peor rendimiento obtuvo: -0.002850664119779683. El puntaje de este último fue negativo incluso, pero aún así se encontró en el punto medio entre -1 y 1, lo que indica que pudo hacer agrupaciones, pero éstas no estuvieron bien definidas.

Sin embargo, debido a que en la implementación del *Spectral Clustering* se utilizó el método *K-Means* en el parámetro *assing_labels*, que como ya se ha mencionado presenta un grado importante de sensibilidad, es posible que esto beneficiara en gran medida el puntaje obtenido por el método. Por ello y realizando la comparativa entre métodos, al observar los mapas de la figura 2 y figura 6, concluimos que el mejor método de clasificación para el set *Life Expectancy* es el método de aprendizaje no supervisado *K-Means*. El cual está compuesto por 4 clusters y se observa una mejor clasificación sobre las condiciones de salud entre los 180 países analizados.

Conclusiones individuales

Adrián Landaverde Nava

El método que yo implementé fue Affinity Propagation. Me pareció interesante que este método no necesita el número de clusters como hiper parámetro, sino que debido a los cálculos matemáticos que hace, identifica el número de clusters que debe tener. Asimismo, una ventaja de este, es que era capaz de funcionar adecuadamente en dimensiones superiores, por lo que nuestro set de datos al tener muchas variables pudo ser agrupado eficientemente usando este método. Sin embargo, un aspecto no tan favorable de este método, es que suele ser muy complejo, es por esto que creó tantos clusters, y de estos, hubo algunos que tenían muy pocos datos, o incluso sólo 1. Me gustó mucho poder comparar todos estos métodos visualmente, ya que cada uno puede agrupar los datos de manera muy distinta, por lo que se pueden obtener diferentes grupos dadas las necesidades que se tengan para agrupar los datos. Finalmente, un aspecto más que me pareció muy importante para visualizar los resultados fue que pudimos realizar un mapa interactivo con el que se puedan visualizar los datos de una mejor manera, ya que de esta forma se pueden interactuar con los datos y desplegar los datos de manera visual y comparar de una manera más eficiente los métodos

Naomi Padilla Mora

En la realización de este proyecto, implementé el método Spectral Clustering el cual, considero que es un método muy eficiente para la solución de problemas de clasificación. Considero que de realizar una investigación más extensa sobre sus funciones y parámetros el algoritmo puede tener un gran desempeño en este y otros datasets. Además, este método, en la cuestión de código, la mayor parte del procesamiento lo hace de manera interna, considero que sería enriquecedor presentar algunos de estos procesos de manera gráfica para tener mayor comprensión del comportamiento del método. Considero que la deficiencia del método se encuentra en la elección del etiquetado, ya que se seleccionó un parámetro sensible y que de hacer más pruebas se podrían tener resultados y puntajes más acertados aunque no sea el mejor puntaje. Me pareció muy interesante las similitudes que tiene con el método de clasificación tradicional *K-Means*, ya que incluso el número de clusters necesarios es el mismo, obtenido por el método del codo. Algo que destaco mucho

del proyecto, es que nos permite explorar cada uno de los métodos con el objetivo de desarrollar la mejor solución posible. Me parece que el dataset seleccionado fue una gran elección para el enfoque que se dió a este proyecto pero considero que se hubieran obtenido resultados más interesantes si se agregan otras variables más relacionadas a los indicadores de salud en los países. Aún así, el proceso de selección, preprocesamiento y limpieza de los datos también fue una parte fundamental para mi compresión del proyecto e implementación de decisiones. El realizar la representación gráfica y a su vez, la reducción de dimensiones me pareció muy interesante y considero que me permitió obtener una mejor comprensión de los métodos y su desempeño de clasificación.

Sabrina Nicole Rodríguez Salgado

El método que yo implementé fue el DBSCAN. Me pareció interesante implementar este método porque tiene un enfoque ligeramente distinto al resto de los algoritmos, pues busca identificar y excluir el ruido y únicamente quedarse con las regiones más densamente agrupadas. A pesar de que todos los algoritmos hacen esencialmente lo mismo, agrupar, cada uno tiene una forma diferente de hacerlo, dándole prioridad a ciertos aspectos. Por las características del conjunto de datos y de aquello que tratábamos de visualizar con estos métodos, me parece que el DBSCAN no fue el algoritmo más adecuado para usar en este problema. Este algoritmo en lugar de darnos la información que buscábamos acerca de la agrupación de países en términos de salud, prácticamente excluyó una tercera parte de los datos asumiendo que eran ruido, y todo esto ocurrió a pesar de hacer varios intentos para mejorar los resultados, por ejemplo con diferentes estimaciones del parámetro *minPts*. De todas formas, creo que fue muy útil para demostrar que el algoritmo (clustering, en este caso) que se use debe escogerse con cuidado, y de acuerdo con las características del conjunto de datos y del problema a resolver. Aprendí que no es indistinto usar cualquier método, sino que debe escogerse estratégicamente. Me gustó que una vez más pudimos decidir el conjunto de datos a usar y cómo lo íbamos a usar, creo que esto hace que nos intereseamos e involucremos más con el proyecto. Lo que menos me gustó fue que no hubo componentes relativos a asociación, me hubiera gustado aprender un poco más de esta parte del aprendizaje no supervisado, aparte del clustering.

Referencias

- Bezy, J. (2022). *Life expectancy*. Recuperado de
<https://www.britannica.com/science/life-expectancy>
- Freeman, T. et al. (2020). *Why do some countries do better or worse in life expectancy relative to income? An analysis of Brazil, Ethiopia, and the United States of America.* Recuperado de
<https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-020-01315-z>
- Garbade, M. (2018). *Understanding K-means Clustering in Machine Learning*. Recuperado de
<https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>
- Google Developers. (2021). *K-Means Advantages and Disadvantages*. Recuperado de
<https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages>
- Gupta, A. (2019). *ML | Spectral Clustering*. Recuperado de:
<https://www.geeksforgeeks.org/ml-spectral-clustering/>
- Kumar, V. (2021). *Tutorial for DBSCAN clustering in Python*. Recuperado de
https://machinelearningknowledge.ai/tutorial-for-dbscan-clustering-in-python-sklearn/#Finding_the_Optimal_value_of_Epsilon
- Maklyn C. (2019). *Affinity Propagation Algorithm Explained*. Recuperado de:
<https://towardsdatascience.com/unsupervised-machine-learning-affinity-propagation-algorithm-explained-d1fef85f22c8>
- Rajarshi, K. (2016). *Life Expectancy (WHO)*. Recuperado de:
<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>
- Scikit-learn. (2022). *Clustering*. Recuperado de
<https://scikit-learn.org/stable/modules/clustering.html>
- Scikit-learn. (2022). *Spectral Clustering*. Recuperado de
<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html#sklearn.cluster.SpectralClustering>
- The World Bank. (2022). *World Development Indicators*. Recuperado de:
<https://databank.worldbank.org/reports.aspx?source=world-development-indicators#>
- Utah Department of Health. (2021). *Important Facts for Life Expectancy at Birth*. Recuperado de
https://ibis.health.utah.gov/ibisph-view/indicator/important_facts/LifeExpect.html

Versloot, C. (2022). *Performing DBSCAN clustering with Python*. Recuperado de
<https://github.com/christianversloot/machine-learning-articles/blob/main/performing-dbscan-clustering-with-python-and-scikit-learn.md>