UNIVERSIDAD INTERNACIONAL DE LA RIOJA

Máster Universitario en Visual Analytics & Big Data

TÉCNICAS DE INTELIGENCIA ARTIFICIAL

ACTIVIDAD GRUPAL

Presentado a:

PROF. ÓSCAR GARCÍA

Presentado por:

JULIANA ANDREA AMÉZQUITA ABELLO

CARLOS FERNANDO CHÁVEZ

LAURA MILENA HERRERA RIVERA

Marzo 04 de 2022

TABLA DE CONTENIDO

1 CONTEXTO

Para esta actividad se utilizaron los datos del Portal Oficial de Datos Europeos y del Portal de la Organización Mundial de la Salud (WHO). El primer dataset contiene información acerca la evolución del Covid-19 en todos los países, principalmente del año 2020. Incluye información sobre la cantidad de contagios, muertes, población, etc. para cada uno de los países. Se puede acceder a este dataset en el siguiente link. Adicionalmente se utilizó un segundo dataset extraído de la plataforma de WHO, el cual contiene información sobre saneamiento e higiene, específicamente sobre el porcentaje de población que vive en hogares que tienen una instalación de lavado de manos con agua y jabón. Dicho dataset contiene información de algunos países desde el año 2000. En este caso se trabajará con los datos del año 2020 de los países en los que está disponible la información de este indicador, es decir, 55 países. Se puede acceder a este fichero desde el siguiente link.

Ambos datasets han sido unificados, depurados y limpiados a efectos de cumplir con el objetivo propuesto. Si se desea acceder a la información del dataset resultante, se puede descargar de este repositorio público de GitHub.

2 OBJETIVO

A través de dos algoritmos de aprendizaje no supervisado, se estudiará la manera en cómo se agrupan los países de acuerdo al porcentaje de población que vive en hogares que tienen una instalación de lavado de manos con agua y jabón respecto a la cantidad de muertes y contagios por Covid-19 durante el año 2020. También, se intentará responder si un alto o bajo porcentaje de acceso a instalaciones de lavado de manos en el hogar se relaciona con la cantidad de contagios y muertes Covid-19 en un país.

3 ANÁLISIS EXPLORATORIO DE DATOS

3.1 Caracterización del dataset

Inicialmente se tiene un dataset con 15302 instancias y 10 variables, no se tienen valores nulos, pues ya se ha hecho la limpieza durante la depuración de los datos (ver Figura 1).

Como su nombre lo indica la variable “date” es de tipo objeto y representa la fecha en que fue tomado el registro y está expresada en año-mes-día. Las variables “day” y “month”, ambas de tipo entero, indican el día y el mes en que fue tomado el registro. Las variables “deaths” y “cases”, ambas de tipo entero, representan el número de muertes y los casos totales del día. “Country” y “Continent”, ambas de tipo objeto, indican el país y el continente en los que cual fue tomado el registro. El atributo “population” de tipo flotante representa el tamaño de la población de cada país. La variable “handwashing_facilities” es de tipo entero y representa el porcentaje

de población de cada país que vive en hogares que tienen una instalación de lavado de manos con agua y jabón. Finalmente, la variable “comulative_days” de tipo flotante, indica el número de casos acumulados para 14 días de pacientes con Covid-19 por cada 100.000 habitantes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15302 entries, 0 to 15301
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   date                   15302 non-null  object
 1   day                    15302 non-null  int64
 2   month                  15302 non-null  int64
 3   cases                  15302 non-null  int64
 4   deaths                 15302 non-null  int64
 5   country                15302 non-null  object
 6   population             15302 non-null  float64
 7   continent              15302 non-null  object
 8   comulative_days        15302 non-null  float64
 9   handwashing_facilities 15302 non-null  int64
dtypes: float64(2), int64(5), object(3)
memory usage: 1.2+ MB
```

Figura 1: Información dataset

3.2 Primer Análisis de Correlaciones

Continuando con el análisis exploratorio, se hace un primer Análisis de Correlación de Pearson para conocer la relación entre los atributos. Como se ve en la Figura 2 el Coeficiente de Pearson varía entre 1 y -1. Cuando el valor del coeficiente se acerca a 1 indica que hay una relación directa entre esas dos variables y cuando el valor del coeficiente es cercano a -1 indica que hay una relación inversa. De igual manera, si el coeficiente es cero indica que no existe una relación entre esas dos variables.

Para ver esta relación en nuestro dataset se hace un mapa de calor y como se observa en la Figura 2, los “casos” y las “muertes”, como era de esperarse, tienen una relación directa (0.9) y en menor medida la variable “población” guarda una relación media con los casos y las muertes, del 0.66 y 0.64 respectivamente. Las variables “handwashing_facilities” y “comulative_days” guardan una relación cercana a cero respecto a todas las otras variables, lo que indicaría que no existe relación alguna.

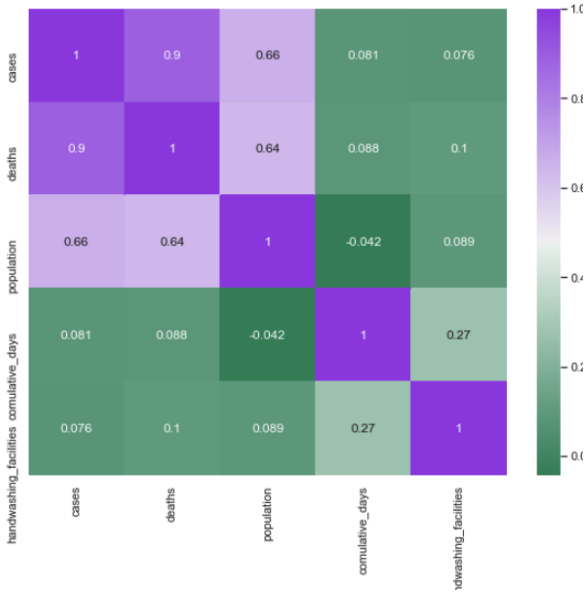

Figura 2: Primer Análisis de correlaciones

3.3 Generación de nuevas variables (operaciones entre variables)

Para poder hacer comparaciones entre los países, se procede a sacar los indicadores poblacionales, porque si bien los números de contagios y muertes, en términos absolutos, muestran la realidad de lo que está pasando en cada país, no permite saber qué tan crítica es la situación en un país comparado con otro. El hecho de que la cantidad de contagios y muertes en un país sea mayor que en otro, no significa que la situación del segundo sea menos grave que la del primero, toda vez que realizar análisis de términos absolutos no sirven de mucho cuando se trata de hacer comparaciones. Por lo anterior, se quiso llevar los datos absolutos a indicadores calculados sobre una base poblacional idéntica para cada país y por cada millón de habitantes, lo cual permite hacer análisis comparativos objetivos. De esta manera se agregaron 2 variables más al dateset: “total_cases_per_million” y “total_deaths_per_million” (Figura 3).

	country	population	continent	handwashing_facilities	deaths	cases	total_cases_per_million	total_deaths_per_million
0	Afghanistan	3.804176e+07	Asia	38	1971	49273	1295.234602	51.811487
1	Algeria	4.305305e+07	Africa	85	2596	92102	2139.267519	60.297697
2	Angola	3.182530e+07	Africa	27	369	16180	508.400565	11.594549
3	Armenia	2.957728e+06	Europe	95	2503	148682	50268.990252	846.257668
4	Bangladesh	1.630462e+08	Asia	58	7047	490485	3008.258280	43.220886
5	Belize	3.903510e+05	America	90	195	9291	23801.655433	499.550405

Figura 3: Generación de nuevas variables

3.4 Segundo Análisis de correlaciones

Se hace un segundo análisis de correlaciones para ver la relación que tienen las variables con las nuevas que hemos calculado en el punto anterior. En este caso se observa una relación media entre la variable “handwashing_facilities” respecto a “total_cases_per_million” y “total_detaths_per_million” (0.552 y 0.51 respectivamente). Como era de esperarse existe una alta relación entre las dos variables nuevas que se han creado (0.86). Ver Figura 4.

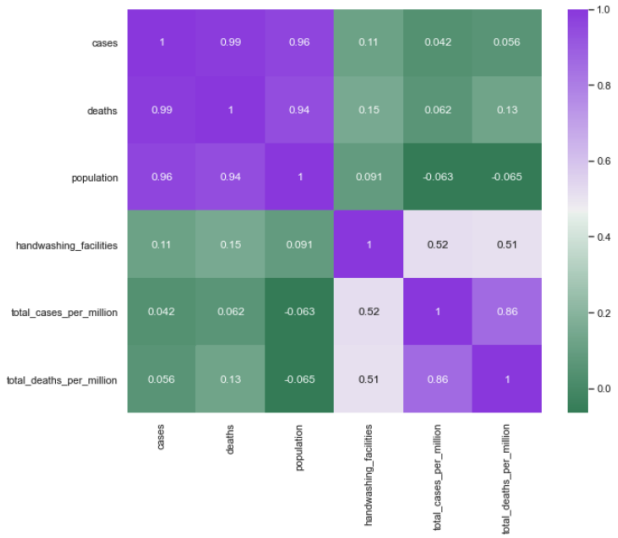

Figura 4: Segundo Análisis de correlaciones

3.5 Análisis comparativo de indicadores entre países

La Figura 5 muestra en orden descendente para cada uno de los países de la base de datos, el porcentaje de la variable “handwashing_facilities”, y cómo este porcentaje se relaciona con la cantidad de casos Covid-19 por millón de habitantes presentados durante el año 2020.

Lo primero que se observa es que casi la mitad de los países cuentan con más de un 70% de adecuación de dichas instalaciones, pero la gran mayoría de la otra mitad cuentan con instalaciones bastante precarias y en algunos casos casi nulas (menos del 50%).

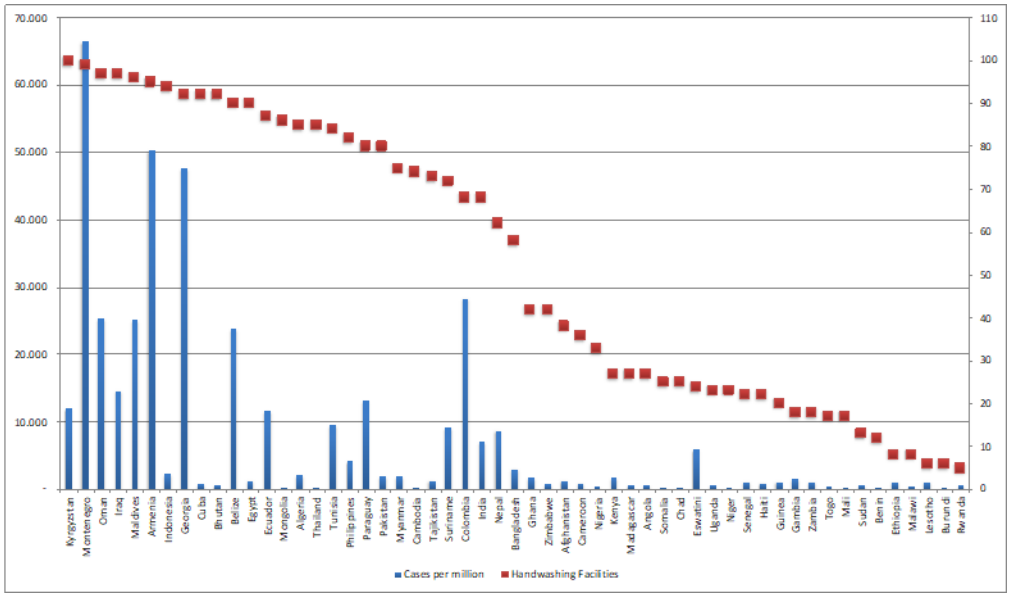

Figura 5: Casos por millón vs Porcentaje de instalación de lavado de manos con agua y jabón en el hogar

La Figura 5 también muestra que existe una gran variabilidad en cuanto a la cantidad de casos Covid-19 por millón de habitantes y que claramente los contagios (al menos así parece), y también como se había visto en el análisis de correlaciones, no tienen una relación directa con el porcentaje de “handwashing_facilities”, ya que existen países como Montenegro, Armenia y Georgia, que a pesar de tener un grado de instalaciones superior a 90%, también tienen un indicador alto de casos por millón, respecto a los demás países. También existen muchos países africanos que, a pesar de carecer casi por completo de instalaciones adecuadas para el lavado de manos, registran cifras casi nulas de casos Covid.

La Figura 6 muestra también en orden descendente para cada uno de los países de la base de datos, el porcentaje de “handwashing_facilities” y cómo ese porcentaje se relaciona con la cantidad de muertes generadas por contagios de Covid-19 durante el año. En este caso, se puede observar también que definitivamente la cantidad de muertes por millón de habitantes, no guarda mucha relación con el porcentaje

de “handwashing_facilities”, ya que, por ejemplo, países como Montenegro, Armenia, Ecuador y Colombia que tienen un nivel de instalaciones superior al 68%, también registran un número elevado de muertes por millón de habitantes, comparados con los demás países. Así mismo, muchos países africanos que no cuentan casi con instalaciones adecuadas de lavado de manos, tampoco registran números importantes de fallecimientos.

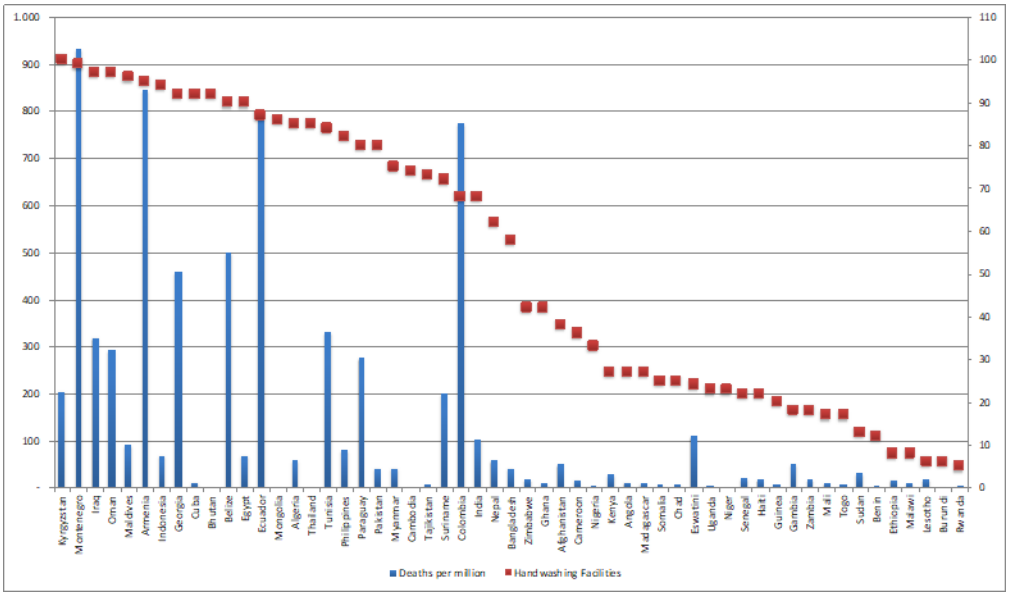

Figura 6: Muertes por millón vs Porcentaje de instalación de lavado de manos con agua y jabón en el hogar

La Figura 7 muestra de nuevo en orden descendente el porcentaje de “handwashing_facilities” por país y cómo este se relaciona con el nivel poblacional de los países de la base de datos. En este caso también se observa que el nivel de instalaciones de lavado no guarda una relación directa con el nivel de población ya que existen países con poblaciones pequeñas, pero con un gran avance en cuanto a sus instalaciones de lavado y países como Etiopía con poblaciones altas, pero con niveles de instalaciones de lavado inadecuadas, casi inexistentes.

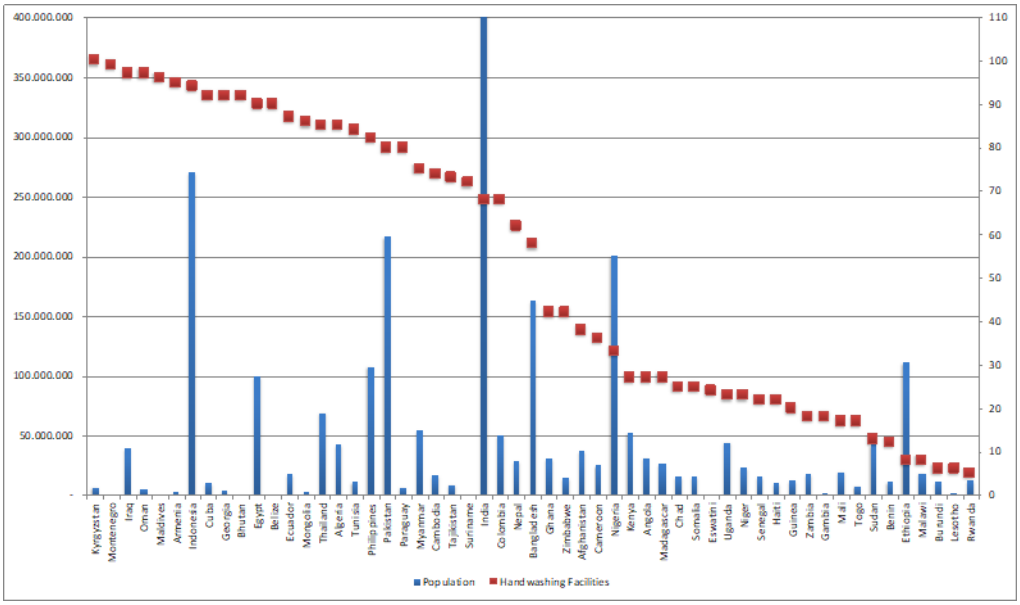

Figura 7: Población vs Porcentaje de instalación de lavado de manos con agua y jabón en el hogar

4 ALGORITMO K-MEANS

K-Means es un algoritmo de agrupamiento, en el cual el modelo va asignando instancias a cada grupo (clúster) de acuerdo a su distancia con un centroide, es decir, cada observación o instancia del dataset se asignará a un grupo o clúster de acuerdo a la cercanía que ésta tenga con el valor medio del grupo. Entre más cercano al valor medio, indica que esa instancia tiene más características similares con las otras observaciones del mismo grupo.

En nuestro caso, el algoritmo agrupará los países de acuerdo a las siguientes variables: “population”, “deaths”, “cases”, “handwashing_facilities”, “total_cases_per_million” y “total_deaths_per_million”. La decisión de tomar estas variables se estableció de acuerdo al resultado del análisis de correlaciones que se hizo en la sección anterior, en el cual se tomaron solo aquellas variables que guardaban alguna relación.

4.1 Método del codo

Antes de aplicar K-Means es necesario configurar el número de clústers (k) que queremos que el algoritmo haga. Este paso es muy importante porque si se llegara a determinar una cantidad equivocada de clusters (muchos o pocos), el resultado de las agrupaciones entregadas por el algoritmo podría ser bastante malo.

Para determinar lo anterior, existen varias técnicas, en nuestro caso utilizaremos “el método del codo”, el cual nos orienta sobre el número óptimo de grupos que debemos configurar en el algoritmo. Esta técnica, traza la inercia como una función del número de clusters k. La curva generalmente presenta un punto de quiebre o inflexión (“el codo”), que es la señal que nos indica el número óptimo de clústers que deberíamos usar (Géron, 2020). En nuestro caso, la técnica del codo indica que el valor óptimo de k que deberíamos utilizar es 4 (ver Figura 8).

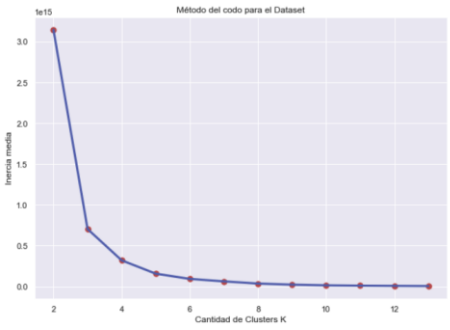

Figura 8: Aplicación Método del Codo

4.2 Ejecución Algoritmo K-Means

Se instancia la clase Kmeans de la librería Scikit Learn y se configuran los siguientes hiperparámetros:

- n-clusters = 4. Indica el número de clústeres a formar, así como el número de centroides a generar. Se define como 4, de acuerdo a lo que se indicó en el método del codo.
- init = 'k-means++'. Es el método de inicialización; con este parámetro se define el lugar donde deberían estar los centroides del modelo. Este hiperparámetro utiliza una métrica de rendimiento, denominada "inercia" que es la distancia cuadrática media entre cada instancia y su centroide más próximo. El algoritmo ejecuta n_init veces y mantiene el modelo con la inercia más baja (Gerón, 2020). Por defecto el hiperparámetro es 'k-means++' que permite seleccionar automáticamente y de manera óptima los centroides.
- random_state = 42. Determina la generación de números aleatorios para la inicialización del centroide.

Luego de instanciar la clase KMeans y configurar los hiperparámetros a utilizar se entrena el modelo y se hacen las predicciones sobre nuestro dataset (ver Figura 9).

```
from sklearn.cluster import KMeans

## Number of clusters, in this case 4 clusters
kmeans = KMeans(n_clusters = 4, init = 'k-means++', random_state = 42)
y_kmeans = kmeans.fit_predict(data_modelo)
y_kmeans1=y_kmeans+1
cluster = pd.DataFrame(y_kmeans1)
# Adding cluster to the Dataset
data_modelo['cluster'] = cluster
#Mean of clusters
kmeans_mean_cluster = pd.DataFrame(round(data_modelo.groupby('cluster').mean(),1))

## Identificación de los valores promedio (centroides)
kmeans_mean_cluster
```

Figura 9: Ejecución algoritmo K-Means

4.3 Resultado Algoritmo K-Means

El modelo agrupa los resultados en 4 grupos diferentes de países: el Grupo 1 contiene 40 países, el Grupo 2 solamente 1 país, el Grupo 3 contiene 4 países y el Grupo 4 contiene 10 países (ver Figura 10). No siempre la forma en que el modelo agrupa los resultados obedece a una razón sencilla de visualizar porque es habitual que existan siempre aspectos que influyeron en cada una de las variables pero que no guardan una relación evidente con lo que debería de suceder.

Para el caso particular de la base de datos analizada que son los Casos Covid, las muertes y el grado de instalaciones de lavado de manos por país, existen aspectos como el nivel de pruebas realmente realizadas, el grado de reportabilidad y la confiabilidad de la información, que pueden también estar afectando los resultados al momento de hacer comparaciones.

```
Países incluídos en el grupo 1
['Afghanistan', 'Angola', 'Armenia', 'Belize', 'Benin', 'Bhutan', 'Burundi', 'Cambodia', 'Cameroon', 'Chad',
'Cuba', 'Ecuador', 'Eswatini', 'Gambia', 'Georgia', 'Ghana', 'Guinea', 'Haiti', 'Iraq', 'Kyrgyzstan', 'Lesoth
o', 'Madagascar', 'Malawi', 'Maldives', 'Mali', 'Mongolia', 'Montenegro', 'Nepal', 'Niger', 'Oman', 'Paragua
y', 'Rwanda', 'Senegal', 'Somalia', 'Suriname', 'Tajikistan', 'Togo', 'Tunisia', 'Zambia', 'Zimbabwe']
Países incluídos en el grupo 2
['India']
Países incluídos en el grupo 3
['Bangladesh', 'Indonesia', 'Nigeria', 'Pakistan']
Países incluídos en el grupo 4
['Algeria', 'Colombia', 'Egypt', 'Ethiopia', 'Kenya', 'Myanmar', 'Philippines', 'Sudan', 'Thailand', 'Ugand
a']
```

Figura 10: Resultado ejecución algoritmo K-Means

5 ALGORITMO DE AGRUPAMIENTO JERÁRQUICO POR AGLOMERACIÓN

En este caso, se decide aplicar otro algoritmo para comprobar su funcionamiento y el comportamiento de los resultados con nuestra base de datos. Se decide tomar este algoritmo ya que en teoría escala bien cuando se tienen números grandes de instancias o grupos como es nuestro caso. Este algoritmo puede organizar de diversas formas los grupos. Particularmente construye una jerarquía de manera ascendente utilizando cualquier distancia por pares, los cuales van siendo mezclados mientras se sube en la jerarquía (Géron, 2020).

5.1 Dendrograma

Los resultados de este algoritmo generalmente se muestran en un dendrograma, que para nuestro caso es el que se muestra en la Figura 11.

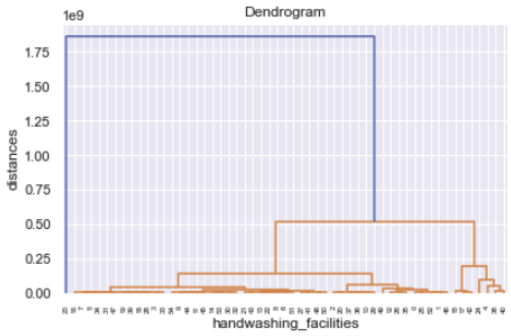

Figura 11: Dendrograma algoritmo de Agrupamiento Jerárquico por Aglomeración

5.2 Ejecución Algoritmo de Agrupamiento Jerárquico por Aglomeración

Se instancia la clase 'AgglomerativeClustering' de la librería Scikit Learn y se configuran los siguientes hiperparámetros:

- n-clusters = 4. Indica el número de clústers a formar, así como el número de centroides a generar. Se define como 4, de acuerdo a lo que se indicó en el método del codo.
- linkage = 'ward'. Representa el criterio de vinculación a utilizar. Determina qué distancia usar entre los grupos. El algoritmo minimiza esta distancia entre los pares. Se utiliza 'ward' que minimiza la varianza de los grupos fusionados.

- affinity = 'euclidean'. Se refiere a la métrica que utiliza el modelo para calcular el vínculo o la distancia de los pares. En nuestro caso se utiliza la distancia Euclidiana, ya que el hiperparámetro linkage fue configurado como 'ward' y éste solo acepta “euclidean”

Luego de instanciar la clase ‘AgglomerativeClustering’ y configurar los hiperparámetros a utilizar se entrena el modelo y se hacen las predicciones sobre nuestro dataset (ver Figura 12).

```
from sklearn.cluster import AgglomerativeClustering

# creando los grupos
hc = AgglomerativeClustering(n_clusters=4, affinity = 'euclidean', linkage = 'ward')
y_hc = hc.fit_predict(data_jerarquico)
```

Figura 12: Ejecución algoritmo de Agrupamiento Jerárquico por Aglomeración

5.3 Resultado Algoritmo de Agrupamiento Jerárquico por Aglomeración

Aunque el funcionamiento del Algoritmo de agrupamiento jerárquico por aglomeración es diferente al del algoritmo K-Means, al final el resultado es muy parecido y solo difiere en el agrupamiento de uno de los grupos (ver Figura 13).

```
**Name of countries in cluster 0**
['Afghanistan', 'Algeria', 'Angola', 'Armenia', 'Belize', 'Benin', 'Bhutan', 'Burundi', 'Cambodia', 'Cameroon', 'Cha
d', 'Colombia', 'Cuba', 'Ecuador', 'Eswatini', 'Gambia', 'Georgia', 'Ghana', 'Guinea', 'Haiti', 'Iraq', 'Kenya', 'Kyrg
yzstan', 'Lesotho', 'Madagascar', 'Malawi', 'Maldives', 'Mali', 'Mongolia', 'Montenegro', 'Myanmar', 'Nepal', 'Niger',
'Oman', 'Paraguay', 'Rwanda', 'Senegal', 'Somalia', 'Sudan', 'Suriname', 'Tajikistan', 'Thailand', 'Togo', 'Tunisia',
'Uganda', 'Zambia', 'Zimbabwe']
**Name of countries in cluster 1**
['Egypt', 'Ethiopia', 'Philippines']
**Name of countries in cluster 2**
['Bangladesh', 'Indonesia', 'Nigeria', 'Pakistan']
**Name of countries in cluster 3**
['India']
```

Figura 13: Resultado algoritmo de Agrupamiento Jerárquico por Aglomeración

6 CONCLUSIONES

A simple vista, pareciera que los criterios con los que el modelo agrupó los resultados están basados en el nivel de densidad poblacional de los países de la base de datos, ya que se observó lo siguiente:

- Los países contemplados en el Grupo 1 tienen una densidad poblacional de menos de 40 millones de habitantes.
- El país contemplado en el Grupo 2 (que es solamente uno, India), tiene una densidad poblacional altísima, de más de 1.000 millones de habitantes.
- Los países contemplados en el Grupo 3, tienen una densidad poblacional entre 160 y 1.000 millones de habitantes .
- Los países contemplados en el Grupo 4, tienen una densidad poblacional entre 40 millones y 160 millones de habitantes.

En todo el análisis deben tenerse en cuenta otras consideraciones que van más allá de los números de la base de datos y tienen que ver con aspectos como:

- Nivel de confiabilidad de los datos reportados por los países durante la pandemia.
- Cantidad de pruebas Covid-19 realmente efectuadas en cada uno de los países.
- Exposición real ante el Covid-19 que tienen los países considerados en la base de datos, toda vez que son países que quizá no tuvieron tanto tráfico de personas que entraron y salieron, propagando el virus.

Por los resultados obtenidos, no existe una relación directa (al menos fuerte) entre el porcentaje de "handwashing_facilities" de los países de la base de datos, con la cantidad de contagios Covid-19, ni tampoco con la cantidad de muertes por Covid-19. Tampoco existe una relación directa de dichas instalaciones, con el nivel poblacional de los países considerados en la base de datos.

El algoritmo K-Means no tiene un comportamiento muy bueno cuando los grupos tienen diámetros muy diferentes porque este está enfocado en asignar instancias a grupos de acuerdo a la distancia al centroide (Géron, 2020). Por lo anterior, al probar con otro algoritmo, el "Algoritmo de agrupamiento jerárquico por aglomeración", aunque su funcionamiento es diferente, al final el resultado es muy parecido.

7 REFERENCIAS

data.europa.eu. (2020, diciembre 14). Europa.eu. https://data.europa.eu/data/datasets/covid-19-coronavirus-data/?locale=es

Geron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2a ed.). O'Reilly Media.

GHO | By category | Handwashing with soap - Data by country. (2021). https://apps.who.int/gho/data/node.main.WSHHYGIENE?lang=en

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn: Machine learning without learning the machinery. *GetMobile Mobile Computing and Communications*, *19*(1), 29–33. https://doi.org/10.1145/2786984.2786995

8 EVALUACIÓN GRUPO

	Sí	No	A veces
Todos los miembros se han integrado al trabajo del grupo	X		
Todos los miembros participan activamente	X		
Todos los miembros respetan otras ideas aportadas	X		
Todos los miembros participan en la elaboración del informe	X		
Me he preocupado por realizar un trabajo cooperativo con mis compañeros	X		
Señala si consideras que algún aspecto del trabajo en grupo no ha sido adecuado		X	