Asignatura	Datos del Alumno		Fecha
Técnicas de inteligencia Artificial	Apellidos:	**Martínez Cerrato**	14/06/2021
	Nombre:	**Shirley Claudette**	

Actividad 3: Análisis libre de un catálogo de datos

Resumen

El desarrollo de esta actividad busca el llevar a cabo la unión de un dataset principal que brinda información del número de muertes por país a causa del virus covid-19 junto con un dataset a elegir que permita obtener conclusiones del comportamiento del virus y que nos ayude a verificar, validar y explorar una hipótesis planteada a través de técnicas de inteligencia artificial.

1. Objetivo a analizar

Revisión de datos principales

La elección de datos se lleva a cabo con base a los criterios indicados en las rubricas de desarrollo del concurso. Previo al inicio del desarrollo de la actividad es necesario realizar una revisión inicial de los datos con los cuales se va a trabajar, de esta manera podemos tener una referencia del tipo de información con la cual podemos complementar el dataset inicial, y se pueden tener criterios en la selección de un dataset complementario. Por esta razón se realiza la revisión de los datos presentes inicialmente y se tiene lo siguiente:

Para el dataset principal se tienen los atributos día, mes, año, los cuales corresponden a la fecha en que fueron tomados los datos, también se poseen los atributos número de casos, muertes y acumulado de casos de covid-19 en los últimos catorce días, todos estos atributos son de tipo numérico. Los atributos categóricos del dataset principal son: nombre de país, nombre de continente, id de país y código de país. Con base a esto se busca un dataset que permita relacionar la información descrita anteriormente con el fin de plantear y probar una hipótesis y poder llevar a cabo un análisis con base a técnicas de inteligencia artificial.

Elección de datos complementarios

La elección de datos se lleva a cabo con base a los criterios indicados en las rubricas del concurso, el dataset se seleccionó de la fuente: https://apps.who.int/gho/data/node.main.INADEQUATEWSH?lang=en. Este dataset trata sobre los niveles de exposición a aguas inseguras en diferentes países. La hipótesis a probar se basa en validar si en los países donde las personas están expuestas a un mayor porcentaje de aguas consideradas inseguras presentan más casos positivos de covid-19.

Los atributos del dataset complementario son tres, el país (country), el cual es de tipo categórico, la tasa de mortalidad por exposición a aguas inseguras y el número de muertes por aguas inseguras, ambas de tipo numérico.

Limpieza de los datos

Posterior a la selección de los datos a usar para complementar la información del dataset principal se realiza la limpieza de los datos, dicha tarea se realiza en dos partes, la limpieza del dataset principal y la limpieza del dataset complementario.

1. Limpieza de dataset principal:

Para llevar a cabo la limpieza del dataset principal se tienen en cuenta las variables que pueden ser eliminadas sin llevar a cabo perdida de información en la totalidad del dataset. Para ello se opta por eliminar las columnas "dateRep", "geoId" y "countryterritoryCode", pues aportan información redundante que puede ser extraída de otras columnas, aparte los algoritmos de clustering no entienden el formato de fecha, ya que necesitan datos numéricos para funcionar. Posterior a la remoción de los campos mencionados anteriormente, se realiza el tratamiento de los datos referentes a los campos vacíos, y se determina que el mejor tratamiento a efectuar es la eliminación de dichos datos, dado que representan un bajo porcentaje en la totalidad de los datos para tratar de completar la información faltante y se presentan casos en los cuales no se tiene información referente a un país o a un continente, por lo que pueden generar incongruencias en los resultados finales.

	day	month	year	cases	deaths	countriesAndTerritories	popData2019	continentExp	Cumulative_number_for_14_days_of_COVID-19_cases_per_100000
0	14	12	2020	746	6	Afghanistan	38041757	Asia	9.013779
1	13	12	2020	298	9	Afghanistan	38041757	Asia	7.052776
2	12	12	2020	113	11	Afghanistan	38041757	Asia	6.868768
3	11	12	2020	63	10	Afghanistan	38041757	Asia	7.134266
4	10	12	2020	202	16	Afghanistan	38041757	Asia	6.968658
...	...	...	...	...	...	...	...	...	...
59016	7	4	2020	0	0	Zimbabwe	14645473	Africa	0.047796
59017	6	4	2020	0	0	Zimbabwe	14645473	Africa	0.047796
59018	5	4	2020	0	0	Zimbabwe	14645473	Africa	0.047796
59019	4	4	2020	1	0	Zimbabwe	14645473	Africa	0.054624
59020	3	4	2020	0	0	Zimbabwe	14645473	Africa	0.054624

Figura 1. Dataset principal posterior a la limpieza de datos.

2. Limpieza de dataset complementario:

Posterior a la selección del dataset que se integrará al principal se realiza la limpieza de este, con el fin de tener la información más útil del dataset previo a su unión, para ello se realizaron algunas modificaciones en el mismo. Inicialmente el dataset brindaba información referente a sexo masculino, femenino y ambos sexos, por lo que se mantuvieron solo los datos correspondientes a ambos sexos. Se realizó la corrección de algunos valores numéricos que no se encontraban insertados correctamente en el dataset, lo que generaría errores al efectuar algún análisis en los datos debido a la diferencia de tipo en los mismos, por lo que se corrigieron dichos errores. Se realizó el ajuste en el parámetro que se tenía común en los dos datasets que era el nombre del país, ya que, en el dataset complementario se evidenció que algunos países diferían en la forma en la que se escribían con respecto al dataset principal, por lo que al momento de efectuar un "join" de los datos estos presentarían error en la agrupación. Finalmente se eliminaron columnas que no aportaban información necesaria para el análisis a desarrollar.

	Country	mortality_rate_ unsafe_wash_services	wash_deaths
0	Afghanistan	13.9	4824
1	Albania	0.2	5
2	Algeria	1.9	758
3	Angola	48.8	14065
4	Antigua_and_Barbuda	0.1	0
...	...	...	...
178	Venezuela	1.4	439
179	Vietnam	1.6	1533
180	Yemen	10.2	2814
181	Zambia	34.9	5793
182	Zimbabwe	24.6	3965

Figura 2. Dataset complementario posterior a la limpieza de datos.

Unión de datos

Posterior a la limpieza efectuada en cada uno de los datasets que usaremos, se realizó la integración de los mismos, para ello se usó la función “merge”, lo que nos permite realizar un inner de los datos usando una variable común, que en nuestro caso será la variable “countriesAndTerritories” y “Country”, de esta manera podemos unificar los datos que componen los dos datasets. Finalmente se elimina una de las dos columnas mencionadas previamente pues se presenta la misma información y es redundante.

```
dataset_tmp = dataset_principal.merge(dataset_complementario, left_on='countriesAndTerritories', right_on='Country', how='inner')

dataset_tmp.drop('Country', axis= 1, inplace=True)
```

Figura 3. Unión de los dataset principal y complementario y eliminación de columan redundante de nombre de país.

Finalmente se evidencia un aumento en el total de columnas presente en nuestro dataframe final con el cual efectuaremos todos los métodos de inteligencia artificial.

```
print(dataset_principal.shape)
```

```
(59021, 9)
```

Figura 4. Dimensión del dataframe principal previa a integración con dataset complementario.

Distribución de valores del Dataset Final

Luego de haber limpiado y unificado ambos datasets, procedemos a visualizar como está distribuida la información para el caso de las variables numéricas.

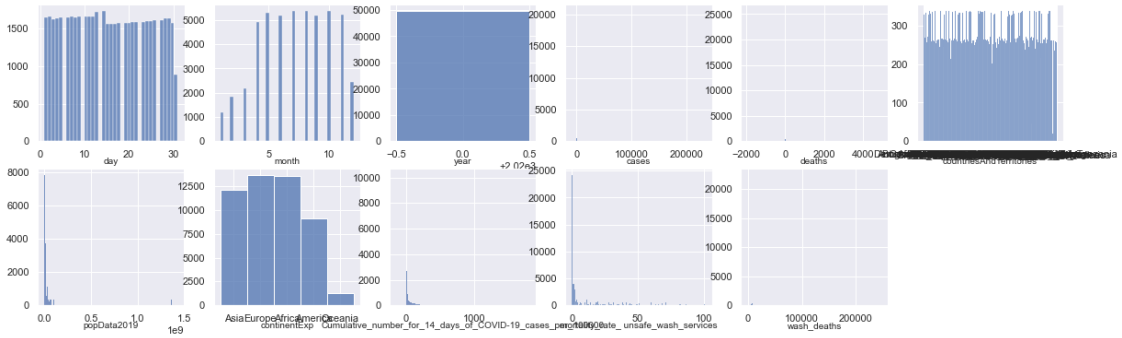

Figura 5. Distribución de frecuentas Atributos numéricos de datset unificado

Como se puede observar en las figuras 5 y 6 los únicos atributos con valores balanceados corresponden mayormente a los relacionado con fecha, que son los atributos día y mes, para el caso del atributo año, se puede observar que no posee ninguna variación, es decir, el valor es el mismo para todo el conjunto de datos. Las variables categóricas del conjunto de datos (país y continente) también presentan una distribución uniforme en sus valores.

Para los atributos relacionados a las estadísticas de covid-19 y aguas inseguras se puede observar que los valores de los datos tienden a concentrarse mayormente en cantidades relativamente pequeñas, presentando varios outliers, especialmente para el caso de los atributos relacionados al covid-19.

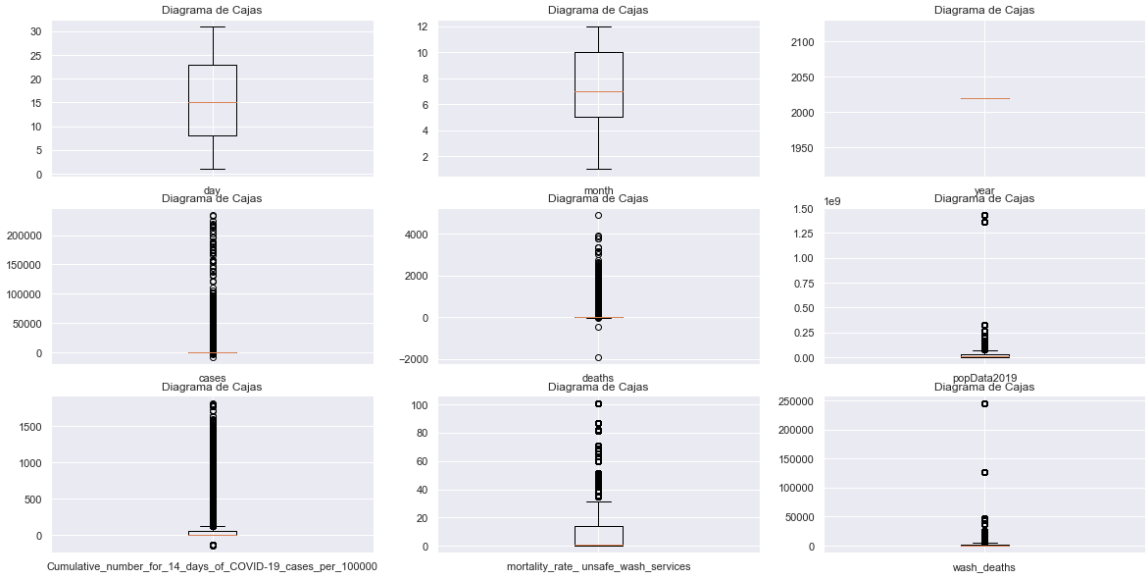

Figura 6. Diagrama de Cajas para atributos numéricos Dataset final

Transformación del dataset Unificado

Partiendo del dataframe final (integración del dataframe principal junto con dataset complementario) se procese a transformar las variables categóricas “countries” y “continent”, ya que los algoritmos de clustering sólo pueden procesar datos numéricos, para llevar a cabo esta tarea utilizaremos el método de one-hot encoding, el cual es un proceso con el cual las variables categóricas son transformadas a variables numéricas para que puedan ser procesadas por diferentes algoritmos de machine learning, luego de realizar esta transformación, nuestro dataset unificado posee la siguiente forma, la cual se aprecia en la figura.

	day	month	year	cases	deaths	popData2019	Cumulative_number_for_14_days_of_COVID-19_cases_per_100000	mortality_rate_unsafe_wash_services	wash_deaths
0	14	12	2020	746	6	38041757	9.013779	13.9	4824
1	13	12	2020	298	9	38041757	7.052776	13.9	4824
2	12	12	2020	113	11	38041757	6.868768	13.9	4824
3	11	12	2020	63	10	38041757	7.134266	13.9	4824
4	10	12	2020	202	16	38041757	6.968658	13.9	4824
...	...	...	...	...	...	...	...	...	...
49579	7	4	2020	0	0	14645473	0.047796	24.6	3965
49580	6	4	2020	0	0	14645473	0.047796	24.6	3965
49581	5	4	2020	0	0	14645473	0.047796	24.6	3965
49582	4	4	2020	1	0	14645473	0.054624	24.6	3965
49583	3	4	2020	0	0	14645473	0.054624	24.6	3965

49584 rows × 189 columns

Figura 7. Dataframe posterior a generar One Hot Encoding.

Después de transformar los datos con one-hot encoding se procede a realizar la normalización del conjunto de datos, este paso es importante ya que al normalizar los valores de todos los atributos del dataset, los rangos de valores finales se establecen entre 0 y 1, lo cual evita que los algoritmos de clustering alteren las distancias entre las cada una de las instancias.

La formula final que se utilizo para normalizar los datos de dataset fue la siguiente:

```
dataset_norm_cluster = (dataset_tmp - dataset_tmp.min())/(dataset_tmp.max()-dataset_tmp.min())
```

2. Algoritmos de Clustering

Una vez transformados los datos, aplicamos técnicas de clustering que nos permita agrupar las instancias del dataset. Antes de seleccionar una técnica de clustering en concreto, se probarán diferentes técnicas de clustering, en base a los resultados obtenidos se seleccionará la mejor.

Primeramente, descartamos los algoritmos de MeanShift y Agglomerative Clustering debido a que su escalabilidad con conjuntos grandes de datos puede generar ciertos problemas si no se cuenta con hardware potente.

K-means: El primer algoritmo probado fue k-means, este algoritmo posee la particularidad que necesita que se especifique el número de clustering antes de ejecutar el algoritmo. Para saber qué número de cluster es el más propicio, se utiliza la técnica de codo de Jambú. A continuación, se visualiza el resultado de la técnica.

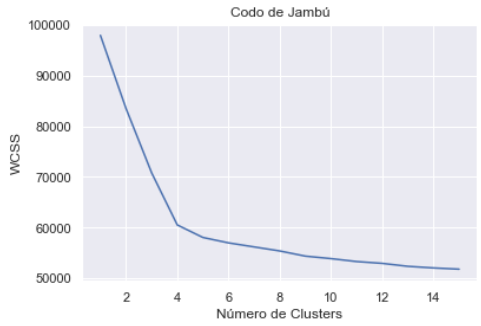

Figura 8. Gráfica de codo de Jambú.

Como se puede observar a partir del cuarto cluster deja de haber una disminución drástica de WCSS (Within-Cluster Sum of Square), por lo tanto la cantidad optima de número de clusters es 4.

DBSCAN: Con este algoritmo se debe de tener en cuenta el parámetro épsilon, el cual es el umbral de distancia máximo aceptado para que dos puntos sean considerados del mismo cluster. Con respecto a este algoritmo se probaron diferentes valores de épsilon y n_samples y siempre se obtuvo el mismo resultado, el cual es un solo cluster.

GaussianMixture Cluster: Para este algoritmo de clustering también se debe especificar el número de clusters previamente, gracias a la técnica de codo de jambú aplicada anteriormente y a la técnica de análisis de componentes principales, se puede saber que el número de clústeres a utilizar para este algoritmo son 4.

Figura 9. Comparación algoritmo técnicas de clustering

Como se puede observar en la figura 9, el algotimo de lcustering que mejor clasificó las instancias del dataset es el algoritmo de k-means, por lo tanto, se utilizará el resultado de su agrupamiento para probar la hipótesis inicialmente planteada.

3. Análisis de Datos

En base a los clusters obtenidos anteriormente, se procede a observar la relación entre el porcentaje de población por cada 100,000 habitantes expuestos a aguas inseguras con el porcentaje acumulado de casos de covid-19 en los últimos 14 días por cada 100,000 habitantes. Si la relación de estas variables es directamente proporcional, entonces la distribución de los puntos de cada instancia en la gráfica final, tomará una forma de regresión lineal. A continuación, se muestran los resultados para cada clúster:

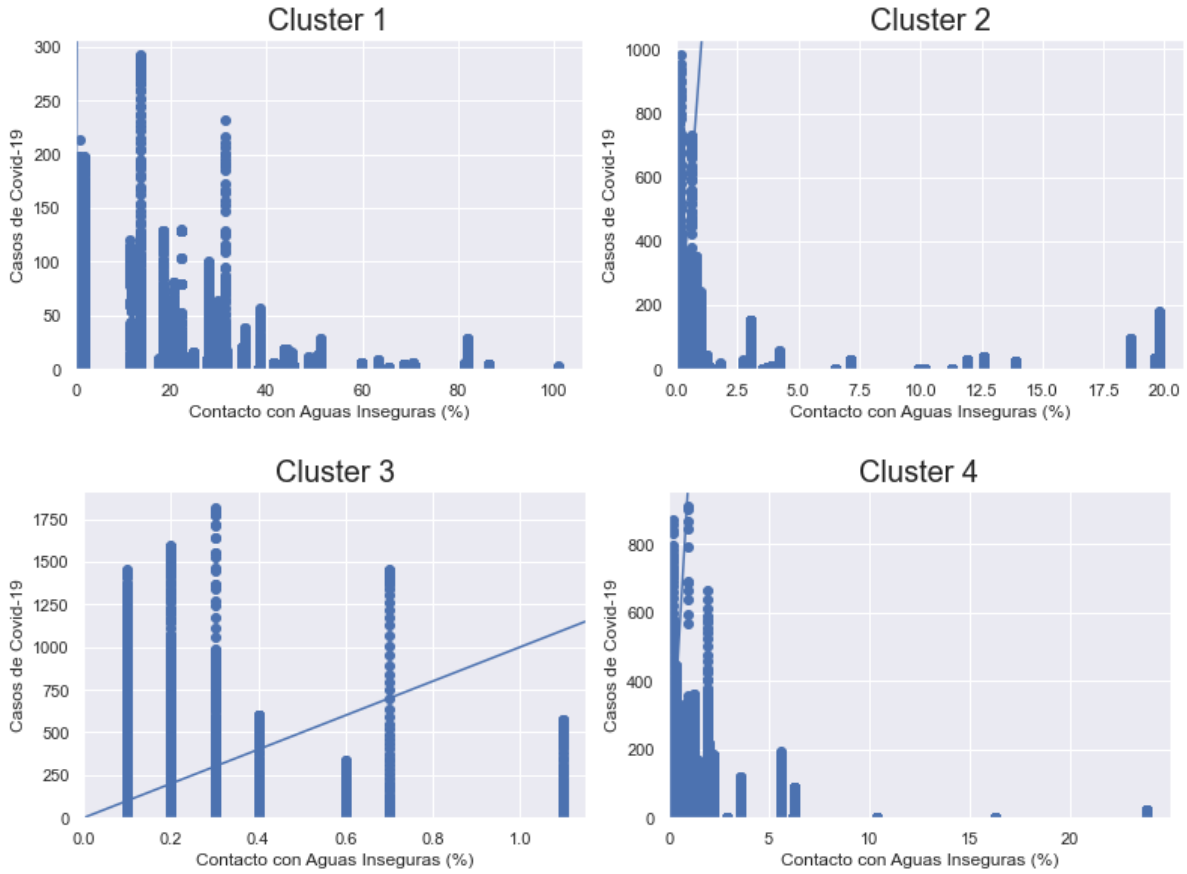

Figura 10. Comparación de resultados por cada clúster

4. Resultados

Se observa que en el clúster número 3 se observa que, aunque el porcentaje de contacto con aguas inseguras es más bajo que el presente en los demás clústeres, el número de casos acumulados de casos de covid-19 es más alto que en los demás clústeres.

A partir del clúster número 4 podemos observar que la mayor parte de casos de covid-19 acumulados se encuentra en lugares que presentan un porcentaje de contacto con aguas inseguras por debajo del 5%.

En el clúster número 1 no se presenta una tendencia entre el porcentaje de personas con contacto con aguas inseguras y el número de casos de covid-19 presentes.

Llevando a cabo el análisis de datos es posible observar que no se presenta una relación directa entre el porcentaje de contacto con aguas inseguras y el número de casos de covid-19 por país, lo que permite inducir que la cantidad de variables que se consideran en el estudio no brinda información que permita soporten la hipótesis planteada inicialmente.