ANÁLISIS DE LA RELACIÓN DE LAS MUERTES CAUSADAS POR EL COVID 19 CON VARIABLES SOCIO DEMOGRÁFICAS Y EL VIRUS DEL VIH

PRESENTADO A:

OSCAR GARCIA GARCIA

AUTORES:

ANDRÉS FELIPE LEAL MORA

JUAN DAVID ESCOBAR ESCOBAR

JUAN MANUEL BAUTISTA CORREA

WILLIAM RAMIRO RIOS HENAO

UNIVERSIDAD INTERNACIONAL DE LA RIOJA

MÁSTER EN VISUALIZACION Y PROCESAMIENTO DE DATOS MASIVOS

TÉCNICAS DE INTELIGENCIA ARTIFICIAL

MARZO 04 DE 2022

TABLA DE CONTENIDO

1. Descripción del objetivo

Para este análisis se ha determinado utilizar varios conjuntos de datos que se han compilado, los cuales parten del número de casos de contagio de covid 19 en cada uno de los países del mundo, junto con los números de muertos asociados a este virus en el año 2020 (Data Europa, 2020). Teniendo en cuenta que no hay mucha información hasta el momento sobre cuáles pueden ser las variables que tengan mayor relación con las muertes a causa del covid 19, se determinó complementar el conjunto de datos con el número aproximado de habitantes por país, el promedio de edad y el área de cada territorio en el mismo periodo de tiempo (United Nations Population Division, 2020).

Adicional a la información anterior se incluyeron dos variables relacionadas con el virus del VIH, las muertes a causa de esta enfermedad en cada país y el porcentaje de cobertura de la terapia médica en el año 2020 (World Health Organization, 2021). Con estas dos nuevas variables se analizará la relación que puede o no existir entre ambos virus o cualquiera de las variables sociodemográficas incluidas. En total el conjunto de datos final tiene diez variables, tres categóricas y siete numéricas, con 213 observaciones.

Teniendo como punto de partida este grupo de datos, se plantearon dos problemas: el primero determinar si existen grupos de países que tengan características similares y si se agrupan dependiendo de variables sociodemográficas y los indicadores relacionados con el Covid19. El segundo es determinar si el conjunto de atributos cuantitativos tiene alguna incidencia sobre las muertes de personas portadoras del virus del VIH y su respectiva relevancia.

2. Técnicas utilizadas

El primer modelo desarrollado se hizo a través del algoritmo de clustering jerárquico, el cual se seleccionó debido a que no se tenía certeza de cuantos cluster se podrían formar y por la precisión que se puede obtener al ir particionando los datos en cada iteración, hasta maximizar las medidas de similitud entre los clusters. En el segundo mólelo se probó un algoritmo de regresión que busca explicar a través de una relación lineal con las diferentes variables numéricas del conjunto de datos, el número de muertes producidas por

el virus del VIH. Con este método se pretende encontrar la combinación de atributos que mejor explique la variable objetivo y obtener una predicción lo más precisa posible.

3. Descripción del proceso

3.1. Preprocesado de datos: se realizó un ajuste de los dos dataset utilizados en la respuesta al problema planteado, mediante el uso de varias herramientas ofimáticas con el fin de eliminar, ajustar o corregir datos inconsistentes. Seguidamente, se consolidaron ambos en un solo fichero. Luego, en Python se realizó el cambio del tipo de datos sobre las variables en las que se consideró era necesario, así como la corrección de los valores nulos encontrados.

3.2. Descripción del dataset: el set de datos consolidado consta de 1704 instancias, con 10 atributos, uno de éstos corresponde al de interés o atributo de salida de tipo cuantitativo. Sin embargo, en razón a que el análisis está relacionado con la problemática del Covid-19, se seleccionaron los datos correspondientes al año 2020, anualidad sobre la cual están disponibles ambos dataset utilizados. Es así, que ahora se dispone de 213 instancias para el análisis. A continuación, se realiza una descripción somera de cada uno de los atributos contenidos en el dataset:

- PAIS: Nombre de los países incluidos.
- CONTINENTE: Nombre del del continente al que pertenece cada país.
- PERIODO: Año correspondiente a cada registro.
- CASOS_COVID: Numero de contagios reportado por cada país.
- MUERTES_COVID: Número de muertes reportado por cada país.
- MUERTES_VIH: Número de muertes reportado por cada país.
- POBLACION: Total habitantes por cada país a corte del 2020.
- AREA_KM2: Área total de cada país.
- PROMEDIO_EDAD: Promedio de edad de los habitantes de cada país.
- COVERTURA_TERAPIA_VIH: Porcentaje de cobertura.

3.3. Descripción estadística: se calcularon los estadísticos descriptivos básicos, presentados en la **tabla No.1**. En esta se observan los respectivos valores

mínimos, máximos, la media, desviación estándar típica muestral, y los percentiles al 0.25, 0.50 y 0.75.

	count	mean	std	min	25%	50%	75%	max
PERIODO	213.0	2.020000e+03	0.000000e+00	2020.00	2020.000000	2.020000e+03	2.020000e+03	2.020000e+03
CASOS_COVID	213.0	3.356943e+05	1.419300e+06	1.00	1947.000000	1.763800e+04	1.471500e+05	1.625675e+07
MUERTES_COVID	213.0	7.571953e+03	2.847139e+04	0.00	35.000000	2.730000e+02	2.552000e+03	2.991770e+05
MUERTES_VIH	213.0	2.816479e+03	8.580443e+03	0.00	0.000000	9.900000e+01	1.100000e+03	8.300000e+04
POBLACION	213.0	3.639588e+07	1.415709e+08	348.00	786552.000000	6.825445e+06	2.549988e+07	1.439324e+09
AREA_KM2	213.0	4.744327e+05	1.786603e+06	0.44	62.200000	2.576700e+02	7.696300e+02	1.637687e+07
PROMEDIO_EDAD	213.0	3.158685e+01	9.502146e+00	15.00	24.000000	3.200000e+01	4.000000e+01	7.000000e+01
COVERTURA_TERAPIA_VIH	213.0	3.823005e+01	3.413387e+01	0.00	0.000000	4.400000e+01	6.900000e+01	9.300000e+01

Tabla No.1: Estadísticos del dataset.

3.4. Variables calculadas: para un análisis adecuado de la situación del Covid-19, se calcularon las tasas de mortalidad y de letalidad, las cuales están definidas de la siguiente manera:

$$Tasa\ de\ Mortalidad = \frac{Número\ de\ fallecidos\ por\ Covid}{Población\ total} * 100.000$$

$$Tasa\ de\ Letalidad = \frac{Número\ de\ fallecidos\ por\ Covid}{Cantidad\ de\ contagios\ de\ Covid} * 100$$

3.5. Normalización: Se realizó el proceso basado en la media y la desviación estándar muestral, mediante el uso de la función preprocessing.scale() de la librería sklearn de Python.

3.6. Outliers: Se detectaron de manera visual, mediante el apoyo de gráficas de dispersión, algunas observaciones anormales o atípicas, en los atributos calculados. Éstos fueron eliminados del dataset al superar el umbral definido por: mean ± 3*std

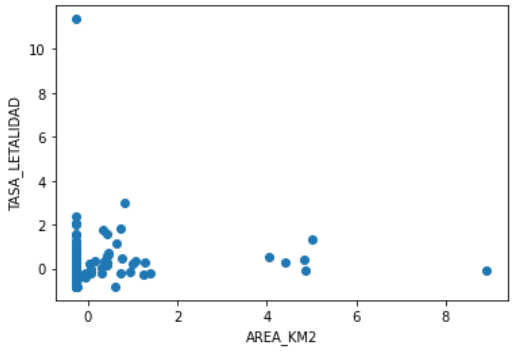

Figura No.1: Diagrama de dispersión del atributo calculado tasa de letalidad vs área.

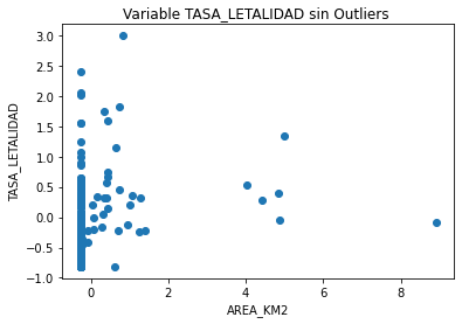

Figura No.2: Diagrama de dispersión del atributo calculado tasa de letalidad vs área, sin outliers.

3.7. Clústering Jerárquico: Se aplicó este algoritmo mediante la función sch.dendrogram() de la librería scipy.clúster.hierarchy de Python. Para la elaboración del dendrograma correspondiente se utilizaron los atributos calculados TASA_LETALIDAD, TASA_MORTALIDAD, y los atributos originales AREA_KM2 y PROMEDIO_EDAD, que son los relacionados directamente con la problemática general del Covid-19.

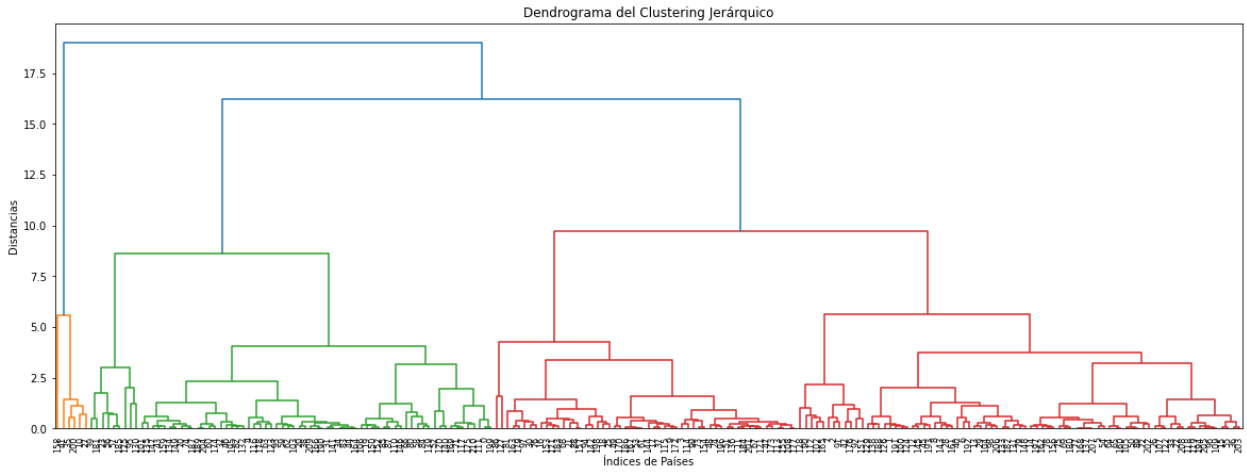

Figura No.3: Dendrograma del clústering jerárquico aplicado al dataset del periodo 2020.

Como resultado de la aplicación del método, se obtuvieron tres clústeres conformados por 6, 72 y 133 observaciones.

Medidas descriptivas para el clúster número 1:

	count	mean	std	min	25%	50%	75%	max
AREA_KM2	6.0	1.000774e+07	3182905.952	7682300.000	8541982.500	9120465.000	9328013.250	1.637687e+07
TASA_LETALIDAD	6.0	2.924000e+00	1.238	1.769	2.037	2.772	3.158	5.151000e+00
TASA_MORTALIDAD	6.0	4.122800e+01	38.905	0.329	10.712	33.876	72.903	9.038500e+01

Medidas descriptivas para el clúster número 2:

	count	mean	std	min	25%	50%	75%	max
AREA_KM2	72.0	212504.952	516807.192	10.12	95.810	265.635	838.918	2267050.000
TASA_LETALIDAD	72.0	2.467	1.943	0.00	1.236	1.980	3.076	9.116
TASA_MORTALIDAD	72.0	9.591	18.217	0.00	0.596	2.024	7.011	88.382

Medidas descriptivas para el clúster número 3:

	count	mean	std	min	25%	50%	75%	max
AREA_KM2	133.0	193284.235	589486.814	0.44	51.000	230.170	566.730	2973190.000
TASA_LETALIDAD	133.0	1.438	1.038	0.00	0.671	1.431	2.088	4.710
TASA_MORTALIDAD	133.0	30.732	34.961	0.00	2.436	17.946	48.854	154.889

Claramente se observa que los países pertenecientes al clúster 1, son los que presentan una mayor media en la tasa de letalidad, es decir, la relación entre la cantidad de personas fallecidas a causa del virus SARS-Cov-2 y la cantidad de contagiados por el virus. Sin embargo, el resultado es lógico, pues son los países que tienen una mayor extensión geográfica, en los que fue dispendiosa la labor del proceso de vacunación para evitar muertes. Confidencialmente, estos países, en su mayoría, no presentan datos sobre las muertes a causa de VIH, que es el tema de interés del presente estudio.

Por consiguiente, el análisis se centrará en los países que hacen parte del clúster 2, los cuales presentan una media de la tasa de letalidad mayor que los del tercer clúster.

3.8. Correlación: Los atributos finales utilizados para las etapas siguientes del presente análisis son: MUERTES_VIH como la variable objeto o de salida, POBLACION, AREA_KM2, PROMEDIO_EDAD, COVERTURA_TERAPIA_VIH, TASA_LETALIDAD y TASA_MORTALIDAD. A continuación, se relaciona el valor del parámetro de correlación para cada uno de los seis atributos considerados para el análisis de predicción:

```
POBLACION              0.487482
COVERTURA_TERAPIA_VIH  0.394585
PROMEDIO_EDAD         -0.332945
TASA_MORTALIDAD       -0.220224
TASA_LETALIDAD        -0.120605
AREA_KM2               0.052066
```

3.9. Modelo de Regresión Lineal: Se calcularon cuatro modelos de regresión lineal, para cada uno de los atributos seleccionados, mediante la el método ols() de la librería statsmodels de Python. Se presentan a continuación, las figuras correspondientes a cada modelo lineal calculado, en contraste con las gráficas

de dispersión de la variable a predecir MUERTES_VIH, en función de cada uno de los atributos predictores utilizados.

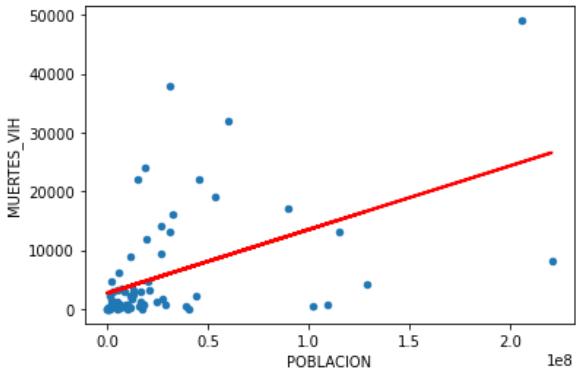

Figura No.4: Gráfica de dispersión datos de muertes por VIH en función de la población y modelo lineal No.1 calculado.

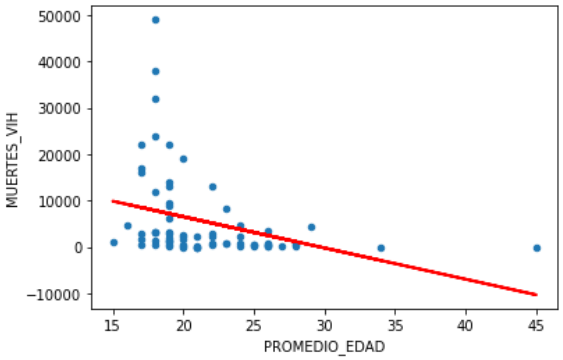

Figura No.5: Gráfica de dispersión datos de muertes por VIH en función del promedio de edad y modelo lineal No.2 calculado.

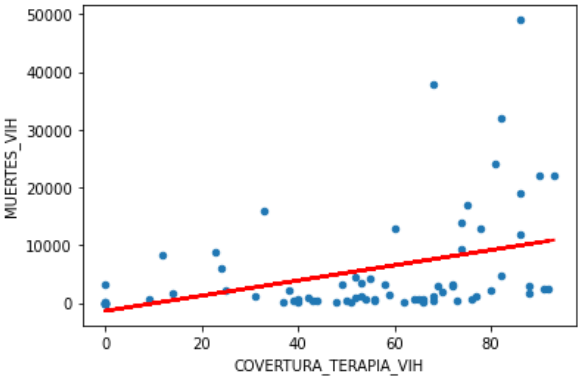

Figura No.6: Gráfica de dispersión datos de muertes por VIH en función de la cobertura de terapia VIH y modelo lineal No.3 calculado.

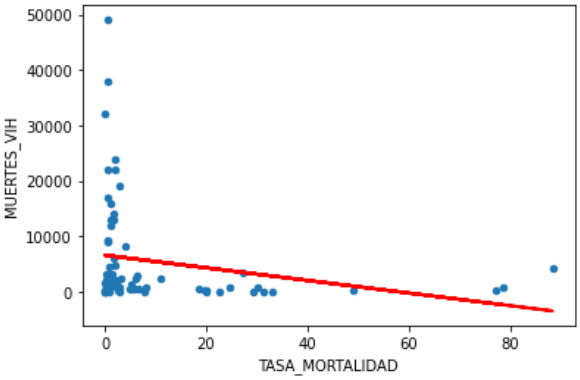

Figura No.7: Gráfica de dispersión datos de muertes por VIH en función de la tasa de mortalidad y modelo lineal No.4 calculado.

	R-Squared	**Coeficiente**	**P-Valor**	**Intervalo Confianza [0.025 0.975]**	**Error Promedio**
Modelo 1 Población	0.238	0.0001	0.000	6.2e-05 0.000	1.5000
Modelo 2 Prom. Edad	0.111	-673.4740	0.004	-1128.155 -218.793	1.6199
Modelo 3 Cobertura Terapia	0.156	131.7935	0.001	58.633 204.954	1.5786
Modelo 4 Tasa Mortalidad	0.048	-113.6594	0.063	-233.669 6.351	1.6758

Tabla No.2: Parámetros para evaluación de los modelos de regresión lineales.

En la **tabla No.2** se muestran algunos de los parámetros que proporcionan una medida de validez de cada modelo lineal. El error promedio se calculó de la siguiente manera: error_promedio = RSE/MUERTES_VIH_mean, donde RSE es la desviación típica de la suma de los cuadrados de la diferencia entre el valor real y la predicción de muertes por VIH.

3.10. Regresión Lineal Múltiple: En razón a que los modelos lineales creados, no tienen un buen ajuste con los datos, partiendo del atributo "POBLACION", que presentó el mayor valor para el coeficiente de correlación, se construirá un modelo de regresión lineal múltiple, combinando las otras variables que obtuvieron un mayor nivel de correlación.

Se procedió entonces a utilizar el modelo anterior y añadir una variable, para calcular el modelo de regresión lineal múltiple y verificar si el error promedio residual de suma de los cuadrados residuales disminuye. La combinación de las variables que genere el menor error será el modelo de regresión lineal múltiple escogido.

	R-Squared	Coeficiente	P-Valor	Intervalo Confianza [0.025 0.975]	Error Promedio
Modelo 5 Población + Cobertura Terapia	0.368	0.0001 120.9413	0.000 0.000	6.01e-05 0.000 57.018 184.865	1.3755
Modelo 6 Población + Cobertura Terapia + Promedio Edad	0.413	0.0001 99.9838 -448.1462	0.000 0.003 0.026	5.94e-05 0.000 35.266164.701 -839.749 -56.543	1.3353

Tabla No.3: Parámetros para evaluación de los modelos de regresión lineal múltiple.

Con base en la **tabla No.3**, se observa que el parámetro R2, aumenta su valor considerablemente. El coeficiente de la variable predictora POBLACION, es prácticamente cero y el coeficiente de la variable predictora PROMEDIO_EDAD es negativo. Los P-valores de todas las variables, son prácticamente cero. Sin embargo, el intervalo de confianza al 95%, contiene el cero para el parámetro POBLACION. El error promedio se redujo, pero continúa siendo significativamente alto. El modelo con las tres variables predictoras es un poco mejor, en las métricas de R2 y del error promedio, pero no lo suficiente para considerarse como un modelo que se ajuste al comportamiento del atributo MUERTES_VIH en función de las variables predictoras.

4. Discusión de resultados

Los resultados obtenidos a través de la ejecución de la metodología de clustering jerárquico permitió identificar tres grupos de países con un coeficiente cofenético de 0,6, teniendo un desempeño regular. En el primer clúster se encuentran los países (seis) con mayor área y tasa de mortalidad a causa del covid 19 con 41,2 muertes por cada cien mil habitantes. En el segundo clúster se encuentran los países más jóvenes (setenta y dos), con una media de edad de 21 años y la tasa de letalidad más alta de las tres agrupaciones con 2,4 muertes por cada 100 contagios. Por último, en el tercer clúster se identifican los países (ciento treinta y tres) con menor área y tasa de letalidad (1,4 muertes por cada 100 contagios).

Para la ejecución del modelo de regresión lineal, se seleccionó el segundo clúster, debido a que presenta la mayor tasa de letalidad y por lo tanto es probable que exista una relación más estrecha, entre las variables seleccionadas para el desarrollo del problema planteado y las muertes de personas portadoras del virus VIH.

Los modelos de regresión lineal fueron elaborados a partir de los atributos que presentaron un mayor nivel de correlación con la variable de salida (número de habitantes, edad promedio, cobertura de terapia de VIH y tasa de mortalidad del covid 19). Sin embargo, las métricas de evaluación de cada modelo (R2, P-valor, intervalo de confianza, error promedio de predicción), evidenciaron que los atributos seleccionados no se ajustan a la distribución de los datos de muertes por VIH.

Finalmente, se aplicó el modelo de regresión lineal múltiple, con el fin de comprobar si se obtenía un aumento en la eficiencia, mediante la combinación de las variables predictoras. Métricas como el R2 y el error promedio de predicción, mejoraron un 73% y 10% respectivamente. No obstante, estos modelos tampoco son suficientemente adecuados para describir la distribución de los datos de muertes por VIH.

BIBLIOGRAFÍA

Data Europa. (14 de 12 de 2020). Obtenido de https://data.europa.eu/data/datasets/covid-19-coronavirus-data?locale=en

United Nations Population Division. (2020). *worldometers*. Obtenido de https://www.worldometers.info/world-population/population-by-country/

World Health Organization. (01 de 10 de 2021). Obtenido de https://apps.who.int/gho/data/view.main.23300?lang=en

TABLA DE VALORACIÓN INDIVIDUAL

	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	