Proyecto Final Inteligencia Analítica

Michael Stiven Bustos Aguirre

1. Entendimiento del negocio

Descripción del Negocio

La OMS: Organización Mundial de la Salud es la autoridad directiva y coordinadora de la acción sanitaria en el sistema de las Naciones Unidas. Es la organización responsable de desempeñar una función de liderazgo en los asuntos sanitarios mundiales, configurar la agenda de las investigaciones en salud, establecer normas, articular opciones de política basadas en la evidencia, prestar apoyo técnico a los países y vigilar las tendencias sanitarias mundiales.

Descripción del problema

En vista de la pandemia declarada en el año 2020 y con los datos del progreso mundial de la vacunación contra COVID-19, se requiere definir países con la misma tendencia de vacunación.

Diseño de solución

Empresa	Problema	Tipo de Minería	Tarea Analítica	Métodos	Evaluación
OMS	Progreso mundial de vacunación contra COVID-19: Objetivo; Definir países con la misma tendencia de vacunación	Descriptivo	Clustering	K-meas, Jerarquico (Cobweb), Expectation-Maximization.	Separabilidad y Cohesión

2. Entendimiento de los datos

Progreso mundial de vacunación contra COVID-19: Objetivo; definir países con la misma tendencia de vacunación.

Ciclo de los datos

- **Generación de los datos:**

Los datos se recopilan diariamente del repositorio *Our World in Data GitHub* para covid-19, se fusionan y se cargan.

- **Almacenamiento de los datos:**

Los datos son almacenados en la página web kaggle.com mediante el siguiente enlace: https://www.kaggle.com/gpreda/covid-world-vaccination-progress

- **Modificación de los datos:**

Los datos son modificados por Gabriel Preda, quien es el que pone a disposición de los datos recopilados de Our World in Data sobre vacunas. Además, realiza la recopilación, fusión y carga diaria de la versión actualizada, según lo mantiene Our World in Data en su repositorio de GitHub.

Diccionarios de datos

Variable	Descripción	Tipo
country	Este es el país para el que se proporciona la información de vacunación.	Categórica
iso_code	Código ISO del país.	Categórica
date	Fecha para la entrada de datos; para algunas de las fechas solo tenemos las vacunas diarias, para otras, solo el total (acumulativo).	Date
total_vaccinations	Este es el número absoluto de inmunizaciones totales en el país.	Numéricas
peolpe_vaccinated	Una persona, según el esquema de inmunización, recibirá una o más (normalmente 2) vacunas; en un momento determinado, el número de vacunaciones puede ser mayor que el número de personas.	Numéricas
people_fully_vaccinated	Este es el número de personas que recibieron el conjunto completo de inmunización de acuerdo con el esquema de inmunización (normalmente 2); en un momento determinado, puede haber un cierto número de personas que recibieron una vacuna y otro número (menor) de personas que recibieron todas las vacunas del esquema.	Numéricas
daily_vaccinations_raw	Para una determinada entrada de datos, el número de vacunaciones para esa fecha / país.	Numéricas
daily_vaccinations	Para una determinada entrada de datos, el número de vacunaciones para esa fecha / país.	Numéricas
total_vaccinations_per_hundred	Relación (en porcentaje) entre el número de vacunaciones y la población total hasta la fecha en el país.	Numéricas
people_vaccinated_per_hundred	Relación (en porcentaje) entre la población inmunizada y la población total hasta la fecha en el país.	Numéricas
people_fully_vaccinated_per_hundred	Relación (en porcentaje) entre la población totalmente inmunizada y la población total hasta la fecha en el país.	Numéricas
daily_vaccinations_per_million	Relación (en ppm) entre el número de vacunaciones y la población total para la fecha actual en el país	Numéricas
vaccines	Tipo de vacunas utilizadas en el país (hasta la fecha)	Categórica
source_name	Fuente de la información (autoridad nacional, organización internacional, organización local, etc.).	String
source_website	Sitio web de la fuente de información	String

Reglas de calidad

Variable	Rango valido
country	N/A
iso_code	N/A
date	1/01/2020 – 20/03/2021
total_vaccinations	Mayor o igual a 0
peolpe_vaccinated	Mayor o igual a 0
people_fully_vaccinated	Mayor o igual a 0
daily_vaccinations_raw	Mayor o igual a 0
daily_vaccinations	Mayor o igual a 0
total_vaccinations_per_hundred	0 - 200
people_vaccinated_per_hundred	0 - 100
people_fully_vaccinated_per_hundred	0 - 100
daily_vaccinations_per_million	0 - 1000000
vaccines	N/A
source_name	N/A
source_website	N/A

3. Preparación de los datos

a. Integración de los datos

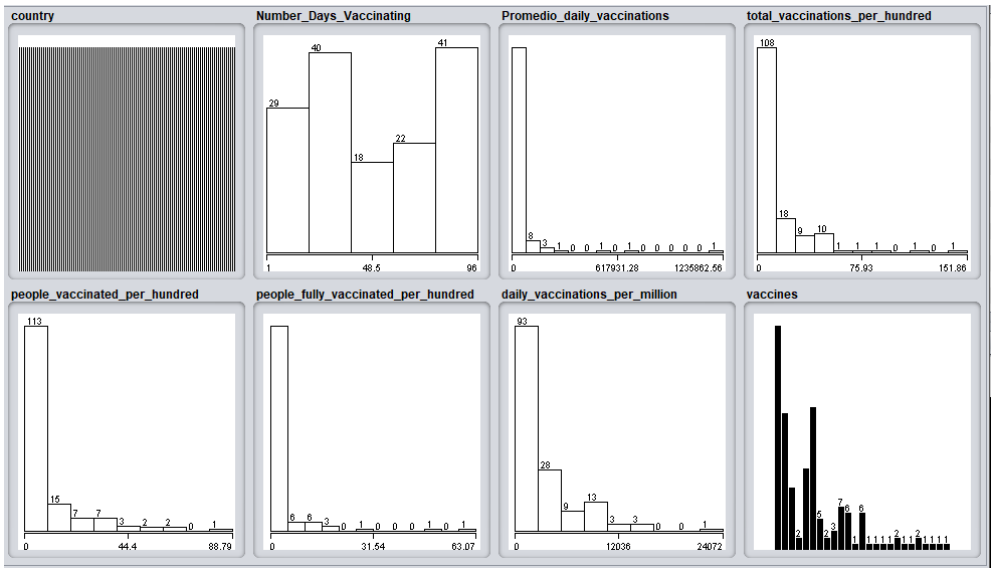
Se crea el archivo con los datos integrados en CSV y ARFF

b. Eliminar variables irrelevantes y redundantes

Se proceden a eliminar las siguientes variables: total_vaccinations, people_vaccinated, people_fully_vaccinated y Promedio_daily_vaccinations_raw, debido a que son variables redundantes, ya que se cuentan con esas mismas variables en porcentaje, lo cual permite una mejor interpretación y análisis de los datos.

c. Descripción estadística de los datos

Se realiza la descripción estadística de los datos por medio del software Weka, obteniendo los siguientes gráficos



d. Limpieza de datos

Como no se tiene datos atípicos no hay necesidad de hacer limpieza de datos, pero si agregan en las casillas con datos faltantes el valor de 0 ya que la sabana inicial de datos tenía valores vacíos que correspondían a un valor de 0.

e. Creación de nuevas variables

La sabana de datos inicial es modificada de tal manera que se cuenta con el registro de 150 países y se agregan nuevas variables como el número de días que lleva vacunando, (Number_Days_vaccinating) basados en las entradas de las fechas y el promedio de vacunas diarias (Promedio_daily_vaccinations).

f. Análisis de correlaciones

Con el método de PCA se obtiene la siguiente matriz de correlaciones

Correlation matrix

1	0.28	0.41	0.38	0.36	0.33	-0.51	0.12	-0.2	0.05	0.27	0.43	-0.05	0.12	-0.07	-0.22	-0.11	0.12	0.11	-0.08	-0.08	0.09	0.04	0	0.11	-0.06	0.04	0.13	-0.05	0.07	0.12
0.28	1	0.1	0.09	0.09	0.07	-0.13	-0.11	-0.07	-0.01	0.03	-0.03	0.04	-0.02	0.04	-0.05	-0.02	0.42	-0.02	-0.02	-0.02	-0.01	0.33	-0.03	0.01	-0.02	0.02	0.03	-0.02	0.03	0.77
0.41	0.1	1	0.91	0.85	0.81	-0.13	0.17	-0.14	-0.06	0.17	-0.04	-0.08	0.11	-0.06	-0.14	-0.03	-0.04	0.16	-0.05	-0.05	0.03	-0.04	-0.06	-0.04	-0.05	0.23	-0.03	-0.05	0.23	0.08
0.38	0.09	0.91	1	0.76	0.75	-0.11	0.14	-0.13	-0.05	0.26	-0.02	-0.08	0.1	-0.06	-0.13	-0.03	-0.05	0.1	-0.05	-0.05	0.04	-0.04	-0.05	-0.05	-0.05	0.23	-0.03	-0.05	-0.05	0.08
0.36	0.09	0.85	0.76	1	0.59	-0.22	0.22	-0.11	-0.04	0.03	0.01	-0.04	0.15	-0.05	-0.09	-0.02	-0.03	0.24	-0.03	-0.03	0.02	-0.03	-0.04	-0.03	-0.03	0.21	-0.01	-0.03	-0.03	0.1
0.33	0.07	0.81	0.75	0.59	1	-0.05	0.23	-0.12	-0.05	0.15	-0.07	-0.09	0.13	-0.05	-0.16	0.07	-0.05	0.02	-0.06	-0.06	0.03	-0.04	-0.06	-0.04	-0.05	0.02	-0.05	-0.06	0.14	0.09
-0.51	-0.13	-0.11	-0.11	-0.22	-0.05	1	-0.23	-0.15	-0.07	-0.17	-0.24	-0.1	-0.07	-0.08	-0.12	-0.11	-0.05	-0.11	-0.05	-0.05	-0.05	-0.05	-0.07	-0.05	-0.05	-0.07	-0.05	-0.05	-0.05	-0.05
0.12	-0.11	0.17	0.14	0.22	0.23	-0.23	1	-0.11	-0.05	-0.13	-0.18	-0.08	-0.05	-0.06	-0.09	-0.08	-0.03	-0.08	-0.03	-0.03	-0.03	-0.03	-0.05	-0.03	-0.03	-0.05	-0.03	-0.03	-0.03	-0.03
-0.2	-0.07	-0.14	-0.13	-0.11	-0.12	-0.15	-0.11	1	-0.03	-0.08	-0.11	-0.05	-0.03	-0.04	-0.06	-0.05	-0.02	-0.05	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02
0.05	-0.01	-0.06	-0.05	-0.04	-0.05	-0.07	-0.05	-0.03	1	-0.04	-0.05	-0.02	-0.01	-0.02	-0.03	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.27	0.03	0.17	0.26	0.03	0.15	-0.17	-0.13	-0.08	-0.04	1	-0.13	-0.06	-0.04	-0.04	-0.07	-0.06	-0.03	-0.06	-0.03	-0.03	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04	-0.03	-0.03	-0.03	-0.03
0.43	-0.03	-0.04	-0.02	0.01	-0.07	-0.24	-0.18	-0.11	-0.05	-0.13	1	-0.08	-0.05	-0.06	-0.09	-0.09	-0.03	-0.09	-0.03	-0.03	-0.03	-0.03	-0.05	-0.03	-0.03	-0.05	-0.03	-0.03	-0.03	-0.03
-0.05	0.04	-0.08	-0.08	-0.04	-0.09	-0.1	-0.08	-0.05	-0.02	-0.06	-0.08	1	-0.02	-0.03	-0.04	-0.04	-0.02	-0.04	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
0.12	-0.02	0.11	0.1	0.15	0.13	-0.07	-0.05	-0.03	-0.01	-0.04	-0.05	-0.02	1	-0.02	-0.03	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
-0.07	0.04	-0.06	-0.06	-0.05	-0.05	-0.08	-0.06	-0.04	-0.02	-0.04	-0.06	-0.03	-0.02	1	-0.03	-0.03	-0.01	-0.03	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01
-0.22	-0.05	-0.14	-0.13	-0.09	-0.16	-0.12	-0.09	-0.06	-0.03	-0.07	-0.09	-0.04	-0.03	-0.03	1	-0.05	-0.02	-0.05	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02
-0.11	-0.02	-0.03	-0.03	-0.02	0.07	-0.11	-0.08	-0.05	-0.02	-0.06	-0.09	-0.04	-0.02	-0.03	-0.05	1	-0.02	-0.04	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
0.12	0.42	-0.04	-0.05	-0.03	-0.05	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	1	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.11	-0.02	0.16	0.1	0.24	0.02	-0.11	-0.08	-0.05	-0.02	-0.06	-0.09	-0.04	-0.02	-0.03	-0.05	-0.04	-0.02	1	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
-0.08	-0.02	-0.05	-0.05	-0.03	-0.06	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
-0.08	-0.02	-0.05	-0.05	-0.03	-0.06	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.09	-0.01	0.03	0.04	0.02	0.03	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.04	0.33	-0.04	-0.04	-0.03	-0.04	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0	-0.03	-0.06	-0.05	-0.04	-0.06	-0.07	-0.05	-0.03	-0.01	-0.04	-0.05	-0.02	-0.01	-0.02	-0.03	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.11	0.01	-0.04	-0.05	-0.03	-0.04	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
-0.06	-0.02	-0.05	-0.05	-0.03	-0.05	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1	-0.01	-0.01	-0.01	-0.01	-0.01
0.04	0.02	0.23	0.23	0.21	0.02	-0.07	-0.05	-0.03	-0.01	-0.04	-0.05	-0.02	-0.01	-0.02	-0.03	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1	-0.01	-0.01	-0.01	-0.01
0.13	0.03	-0.03	-0.03	-0.01	-0.05	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1	-0.01	-0.01	-0.01
-0.05	-0.02	-0.05	-0.05	-0.03	-0.06	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1	-0.01	-0.01
0.07	0.03	0.23	-0.05	-0.03	0.14	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1	-0.01
0.12	0.77	0.08	0.08	0.1	0.09	-0.05	-0.03	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1

Se observa una correlación mayor a 0.8 entre la variable total_vaccinations_per_hundred con people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred y daily_vaccinations_per_million. Esto es debido a que son variables que aumentan correspondientemente una en función de la otra, ya que son variables que llevan el registro de la vacunación.

g. Reducción de variables

Según el método PCA se requiere realizar al menos una reducción a 19 variables, ya que con esos se obtiene una varianza acumulada de 0.79, aunque para este modelo no se realizara una reducción de variables debido a la complejidad de los nuevos componentes.

h. Balanceo de datos

No se requiere balanceo ya que no se tiene variable objetivo y el método a realizar es un clustering.

i. Transformación de tipo de datos según el método

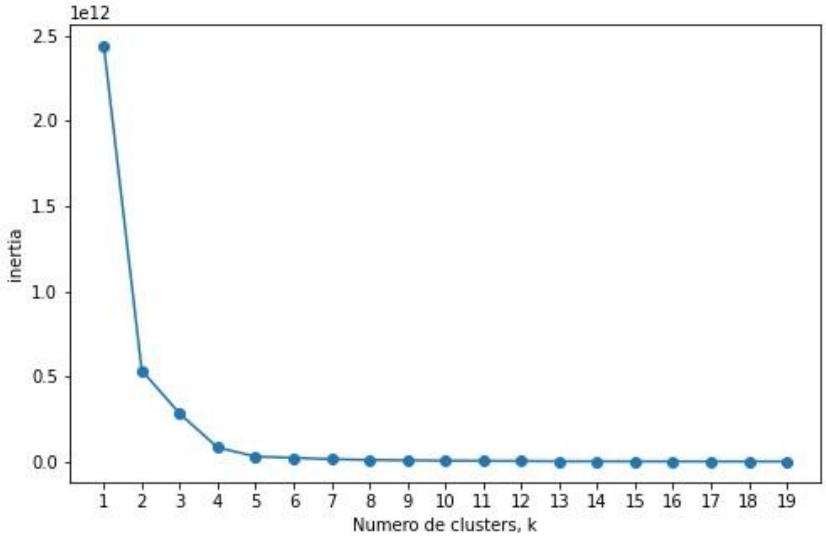
Para realizar el modelo de Clustering se requiere realizar la trasformación de las variables categóricas a números, pero esta transformación se realizará en la fase de modelamiento por medio de Python.

4. Modelamiento

El modelamiento de clustering se realizó por el método de Kmeans por medio de código en Python y en el siguiente enlace se puede evidenciar el código realizado:

https://colab.research.google.com/drive/1zP2b32QRtP76U-0uPZajhZ83utsfqQJg?usp=sharing

Por medio de la gráfica obtenida con el método del codo se identificó que el número de clústers que representaba el cambio brusco en la cohesión o inercia es de 4 clústers, por lo cual ese es el número de clústers ideal para realizar el clustering.



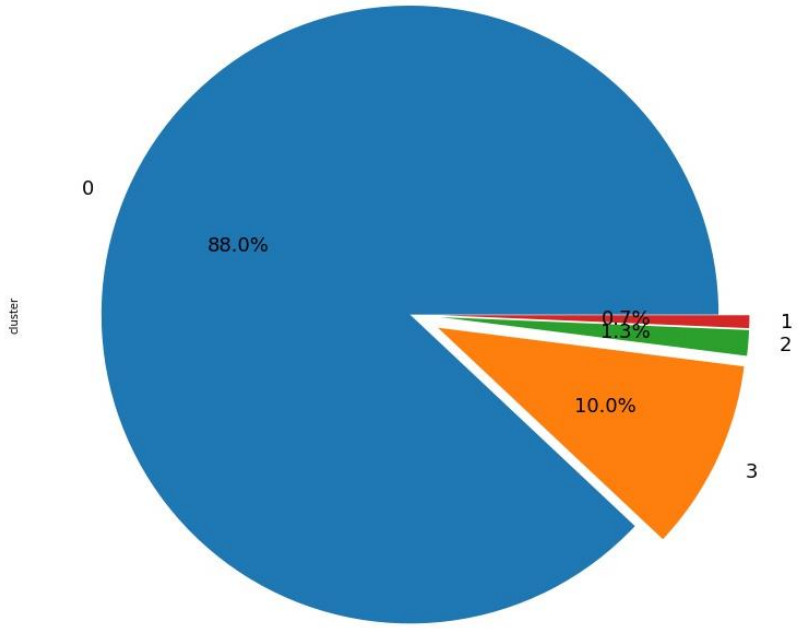
Para realizar el aprendizaje del modelo no se tuvieron en cuenta las variables "country", ya que el objetivo del modelo es clasificar los 150 países de acuerdo con sus datos de vacunación y "vaccines", ya que esta variable categórica no representaba un cambio significativo en las medidas de evaluación y simplemente complicaban la lectura de los resultados del modelo.

- **Descripción de los perfiles asignados**

Clústers	Perfil	Nombre
0	Países que llevan aproximadamente 45 días vacunando, con una vacunación promedio diaria de 7130, un total de 12% de vacunas, un 9% de personas vacunadas y un 3% de personas inmunizadas. Además, diariamente por cada millón de personas se vacunan 2926 personas.	Países con un proceso de vacunación regular
1	Países que llevan aproximadamente 90 días vacunando, con una vacunación promedio diaria de 1235863, un total de 35% de vacunas, un 23% de personas vacunadas y un 13% de personas inmunizadas. Además, diariamente por cada millón de personas se vacunan 7341 personas.	Países con el mejor proceso de vacunación
2	Países que llevan aproximadamente 77 días vacunando, con una vacunación promedio diaria de 619441, un total de 4% de vacunas, un 1% de personas vacunadas y un 0% de personas inmunizadas. Además, diariamente por cada millón de personas se vacunan 1026 personas.	Países con el peor proceso de vacunación

3	Países que llevan aproximadamente 77 días vacunando, con una vacunación promedio diaria de 128652, un total de 24% de vacunas, un 16% de personas vacunadas y un 7% de personas inmunizadas. Además, diariamente por cada millón de personas se vacunan 4572 personas.	Países con proceso de vacunación decente

En la siguiente gráfica y tabla se puede ver la distribución de los 4 clústers obtenidos, donde se observa que el 88% de los países cuentan con un proceso de vacunación regular, el 10% de los países cuentan con un proceso de vacunación decente, el 1.3%, es decir, solo dos países cuentan con el peor proceso de vacunación y el 0.7%, es decir, solamente un país cuenta con el mejor proceso de vacunación contra el Covid-19.



Clústers	Países	Nombre
0	132	Países con un proceso de vacunación regular
1	1	Países con el mejor proceso de vacunación
2	2	Países con el peor proceso de vacunación
3	15	Países con proceso de vacunación decente

5. Evaluación

El modelo de Kmeans fue evaluado usando tres medidas de evaluación diferentes que se pueden observar en la siguiente Tabla.

Índice	Valor obtenido	Valor optimo
Inercia	82454568512	Valores pequeños representan una buena cohesión del clustering
Davies and Bouldin	0.33	Valores pequeños indican una buena estructura del clustering
Silhouette	0.84	Valores grandes indican una buena estructura del clustering

Aunque la inercia represento un valor demasiado alto, indicando que los clustering no tienen una buena cohesión, el índice de Davies and Bouldin obtuvo un valor bajo y el índice de Silhouette obtuvo un valor alto, lo cual indica los que los clustering tienen una buena estructura.

6. Implantación

En el punto 5 del código se guarda el modelo diseñado para el clustering realizado.

7. Conclusiones

- Los 4 clústers obtenidos fueron modelos con datos recopilados hasta el 23 de marzo del 2021, por lo cual, estos pueden presentar variaciones debido a las actualizaciones que se realizan a la base de datos de acuerdo con el progreso de vacunación mundial contra el Covid-19.

- Estados unidos es el país con el mejor proceso de vacunación hasta la fecha, con 90 días de vacunación cuenta con el 23% de su población vacunada y el 15.54% inmunizada.

- Según el modelo, India y China cuentan con el peor proceso de vacunación. Aunque, también se evidencia que los datos recopilados de China estan incompletos, por lo cual la falta de datos del proceso de vacunación de China lo clasificaron junto a la India, la cual solo cuenta con el 2.52% de las personas vacunadas y un 0.52% inmunizadas en sus 64 días de vacunación

- Colombia con 31 días de vacunación cuenta un 2.01% de su población vacunada y un 0.11% inmunizada, lo cual clasifica al país en el clúster de países con un proceso de vacunación regular. En este clúster se encuentra el 88% de los países.