Análisis de datos médicos con aprendizaje no supervisado

José de Jesús Gutiérrez Aldrete
Ernesto Ignacio Borbón Martínez
Luis Felipe Villaseñor Navarrete

10 de junio de 2021

1. Introducción

Mucho se ha hablado acerca de los factores de riesgo que pueden agravar los padecimientos de la COVID-19 y que aumentan su mortalidad. Estos incluyen la edad, el sexo, el peso y otras enfermedades como la diabetes o la insuficiencia renal. Si bien el peso está ampliamente relacionado con el régimen alimenticio, aplicamos técnicas de agrupamiento para dividir los países según su régimen alimenticio y los comparamos con el promedio de la tasa de mortalidad por grupo para así descubrir si hay categorías de alimentos que afecten la mortalidad de la COVID-19, ya sea que la reduzca o que la aumente.

2. Pregunta de investigación

- ¿El régimen alimenticio afecta a la mortalidad de la COVID-19?
- ¿Qué alimentos están asociados a una mayor mortalidad y a una menor mortalidad de la COVID-19?

3. Metodología

3.1. Búsqueda de base de datos médicos

Se entró al sitio Kaggle y se seleccionó la tabla Food_Supply_kcal_Data de la base de datos *COVID-19 Healthy Diet Dataset*.

La motivación de esta base de datos es proteger la salud a través de la dieta saludable. Con ella es posible recopilar información sobre los patrones de dieta de países con una tasa de infección por COVID más baja.

La tabla elegida contiene el porcentaje de consumo de energía en kilocalorías proveniente de 23 diferentes grupos de alimentos en 170 países. Las últimas columnas incluyen recuentos de casos de obesidad, desnutrición y COVID-19 como porcentajes de la población total para fines de comparación.

Los datos de la ingesta calórica provienen Organización de las Naciones Unidas para la Alimentación y la Agricultura, y algunos grupos de alimentos incluidos son alcohol, endulzantes, especias, pescado y productos vegetales.

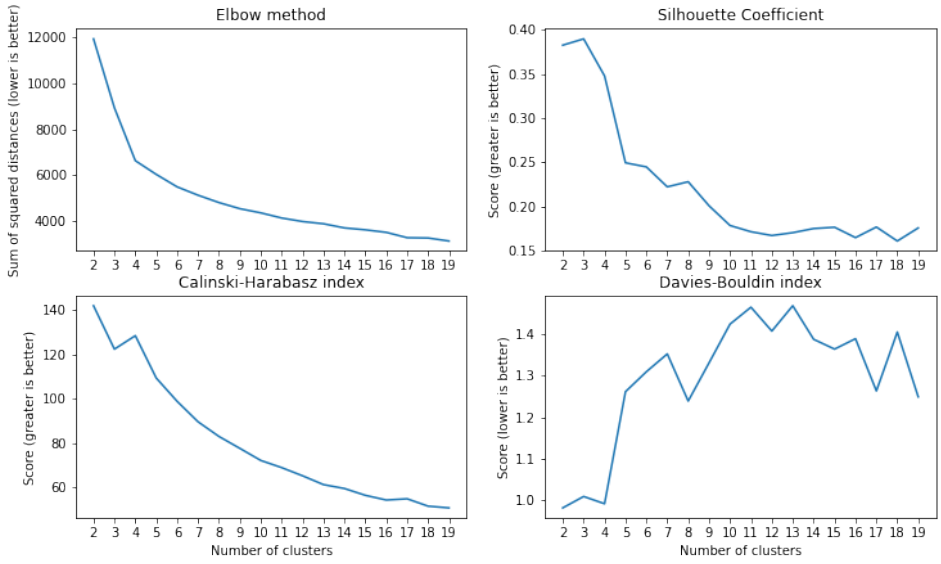
3.2. Búsqueda de patrones con modelos de agrupamiento

Para el desarrollo de este proyecto se emplearon tres métodos de agrupamiento distintos, K-Medias, Mini Batch K-Means y Birch, y 4 métodos para encontrar el número de clusters óptimo: método del codo, coeficiente de Silhouette, Calinski-Harabasz index y Davies-Bouldin index.

4. Resultados

4.1. K-Medias:

Mediante la comparación de los metódos de selección del número de grupos, encontramos que el número óptimo de grupos es entre 3 y 4. Nosotros decidimos usar k=4.



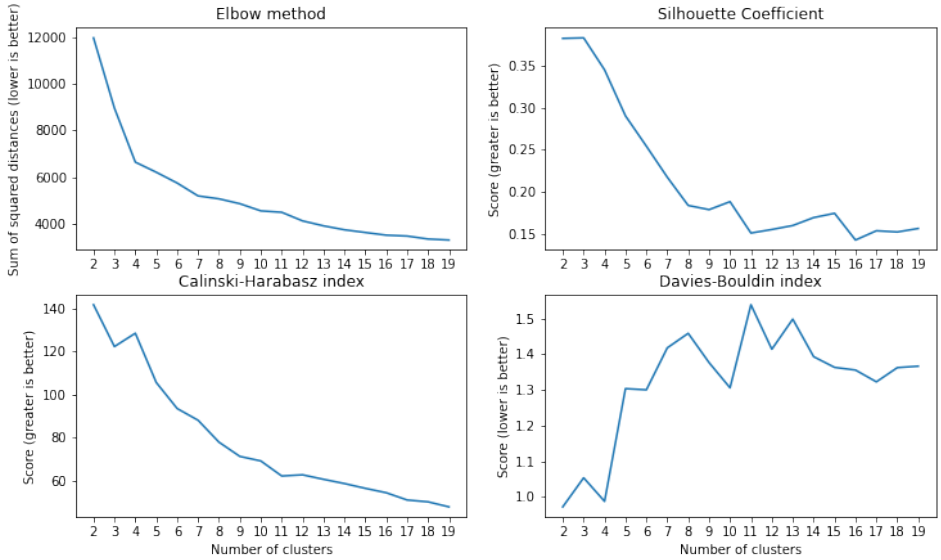
Considerando al grupo con mayor promedio de tasas de mortalidad de COVID como el grupo con la dieta menos sana y al grupo con menor promedio de tasas de mortalidad de COVID como el grupo con la dieta más sana, la dieta menos sana es aquella cuya ingesta calórica tiene más bebidas alcohólicas, nueces, estimulantes y productos de origen animal, incluyendo leche, carne, pescado, mariscos, vísceras, grasas animales, y también son los que consumen más aceites vegetales a pesar de que consumen menos productos de origen vegetal incluyendo cereales, legumbres, especias y raíces almidonadas.

Por otro lado, la ingesta calórica del grupo con la dieta más sana tiene menos productos de origen animal como leche y carne, menos alcohol y más productos de origen vegetal como cereales, legumbres, especias y sobre todo raíces almidonadas. Algunos países dentro del grupo con la peor dieta son Albania, Antigua y Barbuda y Argentina. Ejemplos dentro del grupo con la dieta más sana son Angola, Benín y Camerún.

	Animal Products	Animal fats	Aquatic Products, Other	Cereals - Excluding Beer	...	Obesity	Confirmed	Deaths	Recovered
cluster									
0	4.250028	0.439592	0.000000	29.979125	...	9.672222	0.425322	0.009426	0.376568
1	8.642136	0.911848	0.007486	21.366796	...	20.642857	1.914895	0.039459	1.672368
2	14.583105	2.397295	0.000908	14.472840	...	25.308772	3.668133	0.068939	2.317887
3	3.784839	0.263044	0.000000	17.663967	...	9.855556	0.137557	0.002051	0.107419

4.2. Mini Batch K-Means

Al comparar los distintos métodos de selección del número óptimo de grupos está entre 2 y 4, debido a que encontramos que como que 4 grupos se repetía mas como se muestra en la siguiente figura, elegimos 4.



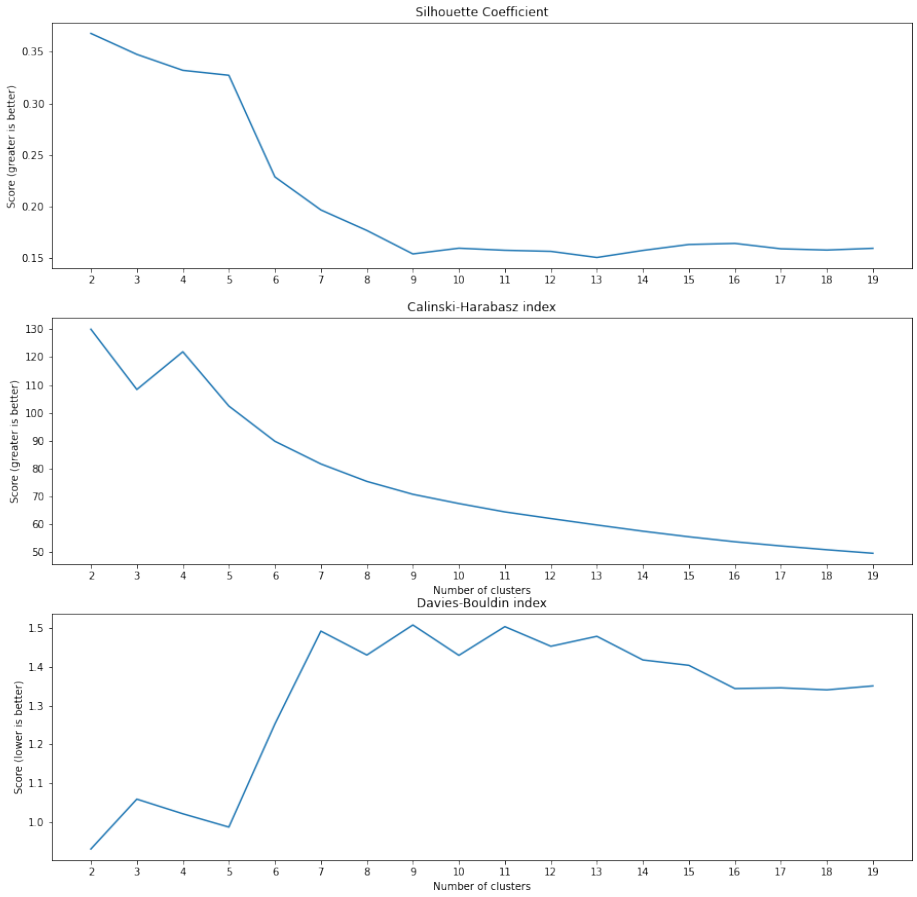
Esta agrupación mostró resultados muy similares a los anteriores. Por ejemplo, hay un grupo para los países desarrollados con mucha obesidad, consumo calórico de alcohol y muertes por covid (como Estados Unidos), y otro grupo para los países en vías de desarrollo (buena parte de Latinoamérica) cuyo ingesta calórica destaca en productos acuáticos, endulzantes y especias. Sin embargo, los países más subdesarrollados (africanos y asiáticos) fueron agrupados de manera distinta, y esta vez uno de esos grupos aparenta tener menos muertes.

	Animal Products	Animal fats	Aquatic Products, Other	Cereals - Excluding Beer	...	Obesity	Confirmed	Deaths	Recovered
cluster									
0	14.583105	2.397295	0.000908	14.472840	...	25.308772	3.668133	0.068939	2.317887
1	8.787452	0.938828	0.007763	21.419517	...	20.857407	1.907301	0.038333	1.663142
2	4.250028	0.439592	0.000000	29.979125	...	9.672222	0.425322	0.009426	0.376568
3	3.878215	0.255080	0.000000	17.891905	...	10.355000	0.334655	0.008662	0.287441

4.3. Birch

El número óptimo de grupos se encuentra entre 2 y 5. Consideramos que el mejor fue 5. Cabe recalcar que al emplear este método no se pudo utilizar el método del codo, por lo que la comparación para obtener el número de grupos óptimos fue con 3 modelos.

Este método nos entrega que el grupo con más saludable tiene un porcentaje de .004 % de muertes COVID-19 mientras que el grupo menos saludable tiene un porcentaje de .069 % de muertes por COVID-19, se puede apreciar que los grupos menos saludables (0,2) son muy similares pero a diferencia de los otros modelos anteriores este al emplear 5 grupo los grupos más saludables. El grupo nuevo es básicamente una nueva partición de los países subdesarrollados.



	Animal Products	Animal fats	Aquatic Products, Other	Cereals - Excluding Beer	...	Obesity	Confirmed	Deaths	Recovered
cluster									
0	14.472485	2.381618	0.000893	14.475661	...	25.277586	3.621572	0.068914	2.378794
1	4.486097	0.468620	0.000000	30.015774	...	10.840000	0.492520	0.011569	0.430030
2	8.761783	0.914596	0.007763	21.217557	...	20.303704	1.871499	0.035932	1.534843
3	3.139531	0.211038	0.000000	21.117281	...	9.568750	0.294507	0.009116	0.260081
4	4.217550	0.254025	0.000000	11.229725	...	7.300000	0.256631	0.004735	0.194559

5. Videos

De José de Jesús Gutiérrez Aldrete.

5.1. Inspiración

- STC Computer Science and Engineering. (2019, 25 de agosto). *Reducción de dimensiones — Análisis del componente principal — PCA —Machine Learning— Python* [video]. Youtube. `https://acortar.link/XQQJP`.
  Lo seleccioné porque desde la miniatura vi que hacían el código paso a paso y quería saber

cómo lo explicaban.

- Kindson The Tech Pro. *How to Perform K Means Clustering in Python(Step by Step* [video]. Youtube. Consultado del 9 de junio de 2021 de `https://acortar.link/PE0b7`.
  Lo seleccioné porque el tema era similar al que usaré para mi video.
- Sundog Education with Frank Kane. *XGBoost: How it works, with an example.* [video]. Youtube. Consultado del 9 de junio de 2021 de `https://acortar.link/MByLr`.
  Lo seleccioné porque combinaba teoría y tutorial.

5.2. Mi video

`https://youtu.be/R3ukE6NyNjQ`

6. Conclusiones

A través del análisis realizado después de aplicar las diferentes técnicas de agrupamiento, podemos concluir que el régimen alimenticio sí afecta a la mortalidad de la COVID-19.

En general, aquellos países que consumen más alcohol y productos de origen animal (como carne y leche) y que consumen menos productos de origen vegetal (como cereales, legumbres y en ciertos casos frutas) son los que reportan mayores tasas de mortalidad de COVID-19, mientras que los países que reportan menor mortalidad de la COVID-19 ocurre lo contrario. De lo anterior se puede concluir que una dieta sana contra la COVID-19 es baja en alcohol y productos de origen animal y alta en productos de origen vegetal.

Los métodos de agrupamiento podrían dar otros resultados (incluso más completos) si se incluyeran las calorías consumidas en lugar de la proporción.