Instituto Tecnológico de Monterrey
Campus Guadalajara

Análisis de datos médicos con aprendizaje no supervisado

Modelación del aprendizaje con inteligencia artificial
Grupo 301

Alejandra Velasco Zárate A01635453
Francisco Javier Chávez Ochoa A01641644
Laura Merarí Valdivia Frausto A01641790

Junio 2023

1. Introducción

En las últimas décadas, la comunidad científica y médica ha mostrado un gran interés en analizar con técnicas modernas de aprendizaje computacional la enorme cantidad de datos disponibles relacionados con la salud. La promesa está en que es posible diseñar nuevos tratamientos o elaborar nuevas medicinas si aprovechamos la información disponible en dichas bases de datos, la cual es dificil de extraer sin modelos matemáticos complejos que sólo una computadora podría manejar.

2. Objetivo

Aplicar técnicas de aprendizaje no supervisado a datos médicos, mediante el uso de datos adquiridos de la página web Kaggle.

3. Descripción de los datos

El Centro para Políticas y Promoción de la Nutrición del USDA recomienda una pauta de ingesta de dieta diaria muy simple: 30 % de granos, 40 % de vegetales, 10 % de frutas y 20 % de proteínas, pero ¿estamos realmente comiendo con el estilo de alimentación saludable recomendado por estas divisiones de alimentos y saldos? En este conjunto de datos se combinaron datos de diferentes tipos de alimentos, la tasa de obesidad y desnutrición de la población mundial, y el recuento global de casos de COVID-19 en todo el mundo para obtener más información sobre cómo un estilo de alimentación saludable podría ayudar a combatir el virus Corona. Y a partir del conjunto de datos, se recopila información sobre los patrones de dieta de los países con una tasa de infección por COVID más baja y ajustar nuestra propia dieta en consecuencia.

Cada una de las variables que conforman la base datos nos indican el índice de la ingesta de diversos alimentos, tales como los vegetales, frutas, alimentos de origen animal, cereales, etc. También se ha agregado la tasa de obesidad y desnutrición (también en porcentaje) para comparar. El final de los conjuntos de datos también incluye los casos confirmados/muertes/recuperados más actualizados (también en porcentaje de la población actual de cada país).

La información que se utilizó para generar la base de datos es la siguiente:

- Los datos para las cantidades de suministro de diferentes grupos de alimentos, los valores nutricionales, la obesidad y los porcentajes de desnutrición se obtienen del sitio web de la FAO de la Organización de las Naciones Unidas para la Agricultura y la Alimentación . Para ver los tipos específicos de alimentos incluidos en cada categoría de los datos de la FAO, eche un vistazo al último conjunto de datos `https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset`

- Los datos del conteo de población de cada país provienen del sitio web PRB de la Oficina de Referencia de Población.
- Los datos de COVID-19 confirmados, muertes, casos recuperados y activos se obtuvieron del sitio web CSSE del Centro Johns Hopkins de Ciencia e Ingeniería de Sistemas
- La información sobre las pautas de ingesta dietética del Centro de Políticas y Promoción de la Nutrición del USDA se puede encontrar en ChooseMyPlate.gov

4. Métodos de agrupamiento

4.1. K-Means

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática.

El algoritmo consta de tres pasos:

1. **Inicialización:** una vez escogido el número de grupos, k, se establecen k centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.
2. **Asignación objetos a los centroides:** cada objeto de los datos es asignado a su centroide más cercano.
3. **Actualización centroides:** se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

El algoritmo k-means resuelve un problema de optimización, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster (K-means, s.f.).

4.2. Mezclas Gaussianas

El modelo de agrupamiento de Mezclas Gaussianas es un algoritmo de aprendizaje automático no supervisado utilizado para la agrupación de datos. Este modelo asume que los datos provienen de una combinación de varias distribuciones Gaussianas, distribuciones normales, donde cada distribución Gaussiana representa una clase. El objetivo del modelo de agrupamiento de Mezclas Gaussianas es estimar los parámetros de las distribuciones Gaussianas subyacentes

para describir los datos observados. Estos parámetros incluyen la media, la varianza y los pesos de cada distribución Gaussiana.

El proceso de funcionamiento del modelo de agrupamiento de Mezclas Gaussianas se puede resumir en los siguientes pasos:

1. **Inicialización:** Selecciona de manera aleatoria los parámetros iniciales para las distribuciones Gaussianas.
2. **Asignación de pertenencia:** Calcula la probabilidad de que cada punto de datos pertenezca a cada una de las distribuciones Gaussianas.
3. Actualización de parámetros: Utilizando los resultados de la asignación de pertenencia, se actualizan los parámetros de las distribuciones Gaussianas para maximizar la verosimilitud de los datos.

Se repiten los pasos 2 y 3 iterativamente hasta que se alcance un criterio de convergencia. Este algoritmo es útil para identificar grupos o patrones en los datos cuando la estructura subyacente de los grupos puede seguir una distribución Gaussiana ('Algoritmo de mezclas gaussianas o Gaussian Mixture Model (GMM)', s. f.).

4.3. Spectral Clustering

La agrupación espectral es una técnica que tiene sus raíces en la teoría de grafos y que se utiliza para identificar comunidades de nodos en un grafo basándose en las aristas que los conectan. El método es flexible y permite agrupar también datos no gráficos.

La agrupación espectral utiliza información de los valores propios (espectro) de matrices especiales construidas a partir del grafo o del conjunto de datos. El algoritmo de clustering espectral es un método no supervisado que utiliza la información del espectro de similitud de los datos para agruparlos. Su funcionamiento se puede resumir en los siguientes pasos:

1. **Construcción de una matriz de similitud:** Se calcula una matriz de similitud que representa las relaciones entre los puntos de datos. Esto puede hacerse utilizando medidas de distancia.
2. **Construcción de la matriz Laplaciana:** A partir de la matriz de similitud, se construye la matriz Laplaciana, que captura la estructura de vecindad de los datos.
3. **Cálculo de los autovectores y autovalores:** Se calculan los autovectores y autovalores de la matriz Laplaciana. Los autovectores correspondientes a los autovalores más pequeños capturan la estructura de agrupamiento de los datos.

4. **Reducción de dimensionalidad:** Si el espacio de datos es de alta dimensión, se puede reducir su dimensionalidad utilizando los autovectores correspondientes a los autovalores más pequeños.

5. **Aplicación de un algoritmo de clustering:** Finalmente, se utiliza un algoritmo de clustering, para agrupar los datos en función de las coordenadas obtenidas en la etapa anterior.

5. Resultados de métodos de agrupamiento e información relevante

5.1. K-Means

En primera instancia se graficaron los datos usando el algoritmo de K-Means utilizando 10, ($k = 10$) clústers. Veáse Figura 4.

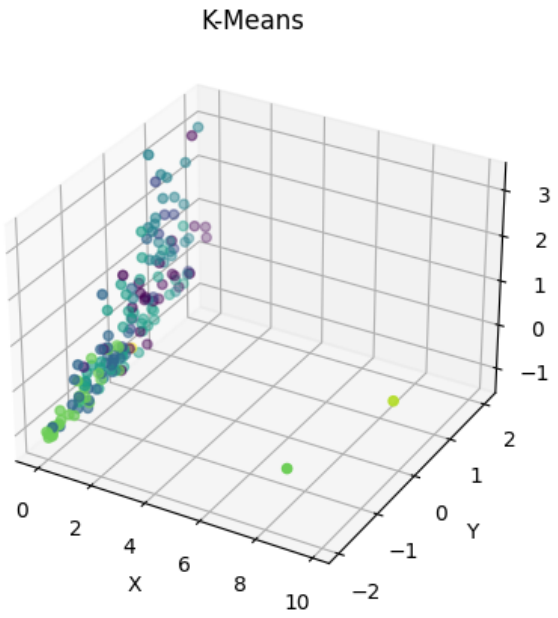


Figura 1: Datos con K-Means, k=10

etiquetas								
0	27.945238	22.054441	25.248718	0.439579	0.687995	5.427486	0.104628	3.558411
1	17.652261	32.348891	10.991304	0.134735	0.479143	0.452775	0.009141	0.357459
2	9.929673	40.069812	7.719231	0.323054	0.840496	0.124381	0.002212	0.099511
3	21.309779	28.690655	23.434483	0.209117	0.951162	1.934106	0.042573	1.723708
4	20.048300	29.956300	4.900000	1.035900	0.474100	0.156227	0.002841	0.136855
5	16.547015	33.453592	25.376923	0.124908	0.946700	2.323950	0.050801	1.933671
6	12.225020	37.771960	11.440000	0.097640	1.616200	0.282695	0.004369	0.232298
7	27.893543	22.107478	21.491304	0.224043	1.003330	0.748220	0.007419	0.480647
8	16.604600	33.401600	13.133333	0.033333	1.222133	0.358174	0.005437	0.341465
9	11.072700	38.909100	4.800000	0.548600	3.865600	0.126135	0.001675	0.093013

Figura 2: Datos promediados por grupos con K-Means, k=10

Después se aplicaron dos métodos para ver cuál es el número óptimo de clústers: Elbow Method y Silhouette Method.

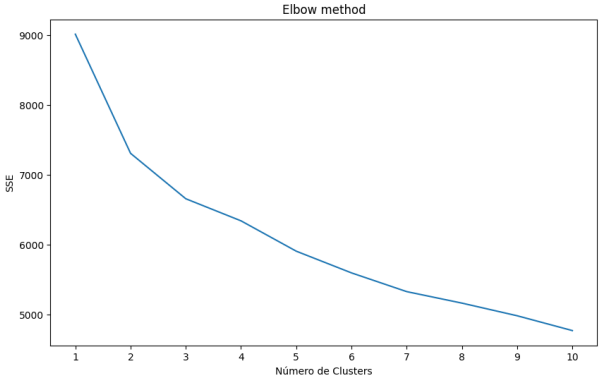


Figura 3: Elbow Method, técnica para determinar clusters

El método del codo para determinar el número óptimo de clusters implica ejecutar el algoritmo de clustering para diferentes valores de k, calcular el SSE para cada k y encontrar el punto en el gráfico donde se produce un cambio significativo en la disminución del SSE. Ese punto es considerado como el número óptimo de clusters. Como se observa en la imagen, el número óptimo de clusters es 2 porque se observa el cambio significativo en la disminución del SSE.

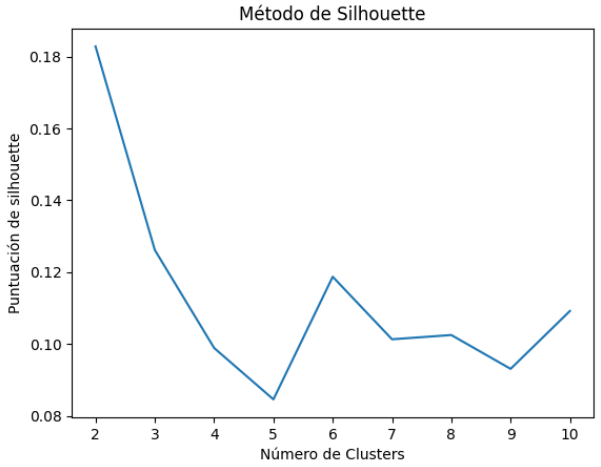


Figura 4: Silhouette Method, técnica para determinar clusters

Complementando Elbow method, el método de Silhouette para determinar el número óptimo de clusters implica calcular el coeficiente de Silhouette, que evalúa la coherencia de asignación de los clusters al comparar la distancia promedio intra-cluster con la distancia inter-cluster, para diferentes valores de k, y seleccionar el valor de k que maximice el valor promedio del coeficiente de Silhouette, indicando una mejor calidad de la agrupación. En este caso, en la

grafica se obserba que el óptimo número de clusters es 2.

Ambas técnicas concluyron que 2 es el número optimo de clusters, por lo que se aplico K-Means con $k = 2$ y este fue el resultado:

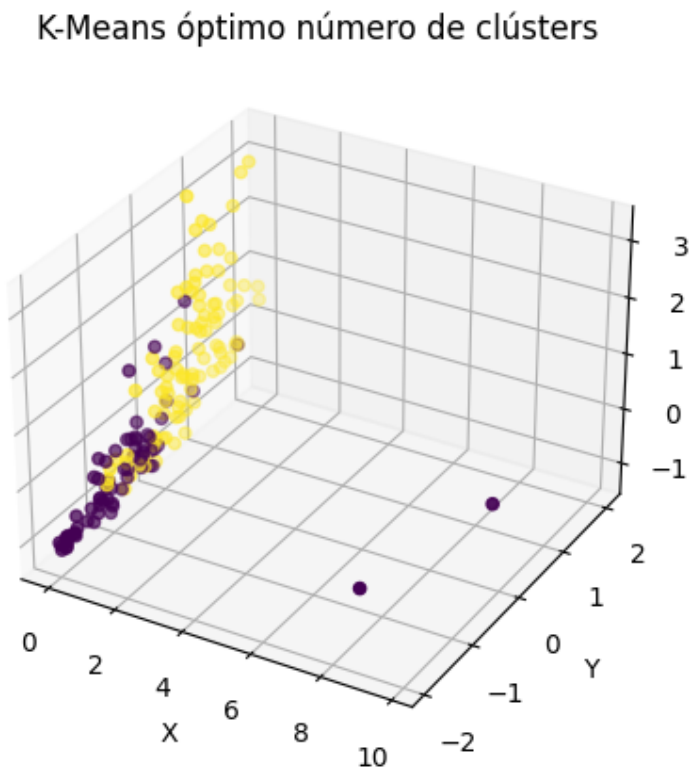


Figura 5: Gráfica con k óptimo, k=2

etiquetas								
0	14.160529	35.839766	10.869118	0.227625	0.861476	0.306410	0.005951	0.253563
1	25.345558	24.654594	24.307368	0.296603	0.843991	3.271196	0.063705	2.325689

Figura 6: Datos con clusters óptimos, k=2

	Animal Products_x	Vegetal Products_x	Obesity_x	Alcoholic Beverages_y	Fruits - Excluding Wine_y	Confirmed_y	Deaths_y	Recovered_y
etiquetas								
0	23.066765	26.933715	21.979032	0.207056	0.930234	1.295854	0.023438	1.066156
1	12.672261	37.327827	9.964286	0.212386	0.870427	0.309665	0.006463	0.253600
2	27.354622	22.645376	25.057778	0.420549	0.718691	5.198118	0.103144	3.508433

Figura 7: Datos con clusters óptimos, k=3

Al analizar el primer agrupamiento, cuando se utilizaron 10 clusters, observamos que se podía observar una pequeña diferencia entre los promedios de las variables Deaths, Recovered y Confirmed,(las cuales son las variables que más nos interesan en este modelo), no había una tendencia clara de que estas variables influyeran en el clusters formados.

Sin embargo una vez que analizamos el número de cluster óptimos, el cual nos arrojó que eran alrededor de 2, hicimos los modelos con 2 y 3 clusters, en los cuales ya se nota una diferencia muy clara en el agrupamiento, donde podemos observar que agrupó los países tomando en cuenta indicadores de cómo le fue en la pandemia tales como número de muertos, numero de contagiados y numere de recuperados; y al mismo tiempo,los países se agruparon por la calidad de su alimentación, los países que consumen más vegetales, más frutas; menos alcohol, alimentos de origen animal y que tenían menos obesidad, fueron los países que tenían los mejores números en la pandemia, demostrando así que la alimentación jugó un papel importante a la hora de determinar los afectados por la pandemia del Covid-19.

Analizando las labels y verificando los paises que estan en cada grupo, observamos que la mayoría de países africanos y asiáticos, que tienen buena alimentación en general, fueron agrupados en el cluster con buenos indicadores en la pandemia, por otro lado, países que tienen mala alimentación, como Estados Unidos, Canadá, México, España, Italia, Francia; fueron agrupados en el cluster de países con mal manejo de pandemia.

5.2. Mezclas Gaussianas

Al igual que en el modelo de agrupamiento de K-Means, se evaluó el modelo de agrupamiento de Mezclas Gaussianas con 10 clusters y se graficó para observar la distribución de los datos.

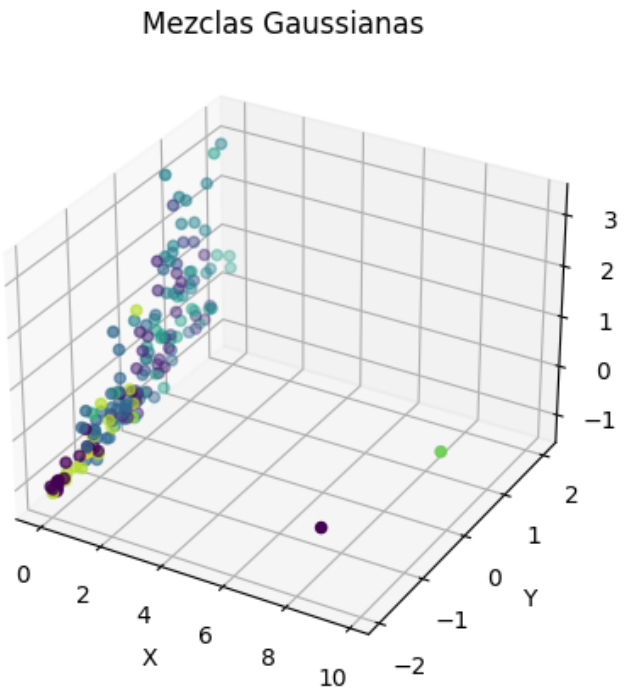


Figura 8: Gráfica con Mezclas Gaussianas, k=10

etiquetas	Animal Products_x	Vegetal Products_x	Obesity_x	Alcoholic Beverages_y	Fruits - Excluding Wine_y	Confirmed_y	Deaths_y	Recovered_y
0	27.177560	22.822387	20.053333	0.288953	0.786053	1.231303	0.037162	1.045015
1	21.951024	28.049724	25.152381	0.139043	0.920081	2.648793	0.048952	2.266664
2	20.291122	29.710066	20.615625	0.168734	0.944981	0.899510	0.015869	0.688727
3	20.048300	29.956300	4.900000	1.035900	0.474100	0.156227	0.002841	0.136855
4	26.701781	23.297700	25.340741	0.384952	0.750741	5.973449	0.111939	4.030382
5	29.743093	20.257333	24.853333	0.541893	0.590073	3.515343	0.073038	2.113173
6	12.301327	37.697355	13.809091	0.231600	1.326000	0.184354	0.002659	0.148200
7	12.808580	37.191806	8.414286	0.200440	0.646843	0.248733	0.005350	0.209777
8	10.459140	39.539020	8.680000	0.368880	1.990120	0.147852	0.003225	0.110121
9	29.739800	20.255000	7.900000	0.036700	0.636300	3.078743	0.009982	2.690573

Figura 9: Datos con Mezclas Gaussianas, k=10

Para este método de agrupamiento, las técnicas del método de codo y Silhouette no funcionan, por lo que se utilizó AIC (Akaike Information Criterion). Es un

valor numérico que combina la bondad del ajuste y la complejidad del modelo, considerandola capacidad de ajuste y de generalización. En el contexto de Mezclas Gaussianas, se utiliza para seleccionar el número óptimo de componentes o clusters en el modelo. Se busca minimizar el valor del AIC, ya que un valor menor indica un mejor equilibrio entre ajuste y complejidad del modelo (Bevans, 2020).

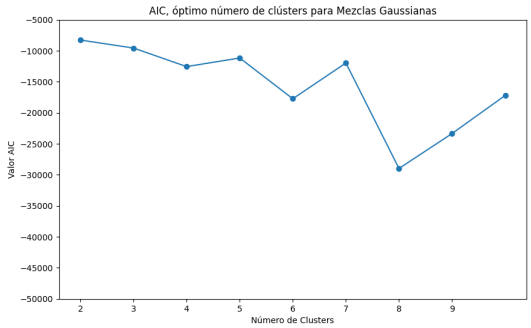


Figura 10: AIC Valores, técnica para determinar clusters

Observando la Figura 10, se llega a la conclusión que el óptimo número de clusters para el modelo de agrupamiento de mezclas gaussianas es de 8. Por lo que se evaluó este modelo de agrupamiento pero ahora con 'n_components' = 2.

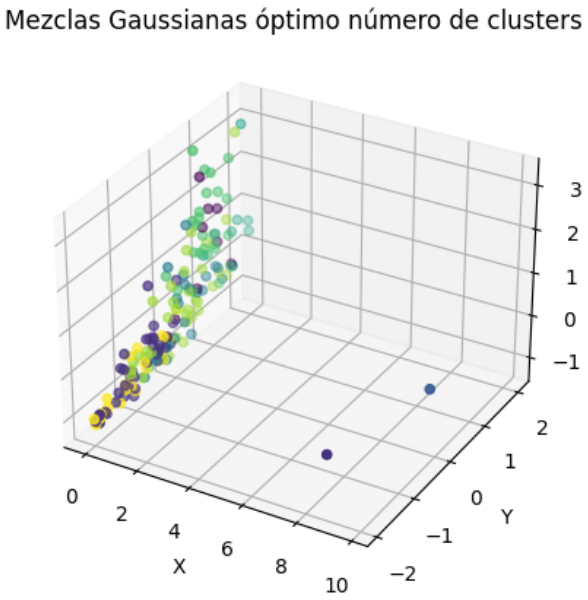


Figura 11: Gráfica con k óptimo, k=8

etiquetas	Animal Products_x	Vegetal Products_x	Obesity_x	Alcoholic Beverages_y	Fruits - Excluding Wine_y	Confirmed_y	Deaths_y	Recovered_y
0	25.155934	24.844305	24.806452	0.295403	0.848673	3.310052	0.064584	2.348558
1	14.732030	35.268144	10.590000	0.231190	0.854756	0.339496	0.006432	0.282384

Figura 12: Datos con clusters óptimos, k=2

etiquetas	Animal Products_x	Vegetal Products_x	Obesity_x	Alcoholic Beverages_y	Fruits - Excluding Wine_y	Confirmed_y	Deaths_y	Recovered_y
0	15.080529	34.919951	15.741176	0.111518	0.684865	0.829726	0.014975	0.708952
1	26.832027	23.168208	24.150000	0.332577	0.833419	3.612893	0.071419	2.511113
2	14.962909	35.036847	10.641176	0.353747	1.141903	0.219937	0.003584	0.181159

Figura 13: Datos con clusters óptimos, k=3

Al analizar el primer agrupamiento, cuando se utilizaron 10 clusters, observamos que se podía observar una pequeña diferencia entre los promedios de las variables Deaths, Recovered y Confirmed, (las cuales son las variables que más nos interesan en este modelo), realmente no son muy claras las diferencias entre cada una de las clases formadas.

Sin embargo, una vez que analizamos el número de cluster óptimos, el cual nos arrojó que eran alrededor de 2, hicimos los modelos con 2 y 3 clusters, en los cuales ya se nota una diferencia muy clara en el agrupamiento. En ambos se ve que entre mayor sea el índice de casos confirmados de COVID, mayor será las muertes y personas recuperadas, lo cual nos hace sentido.

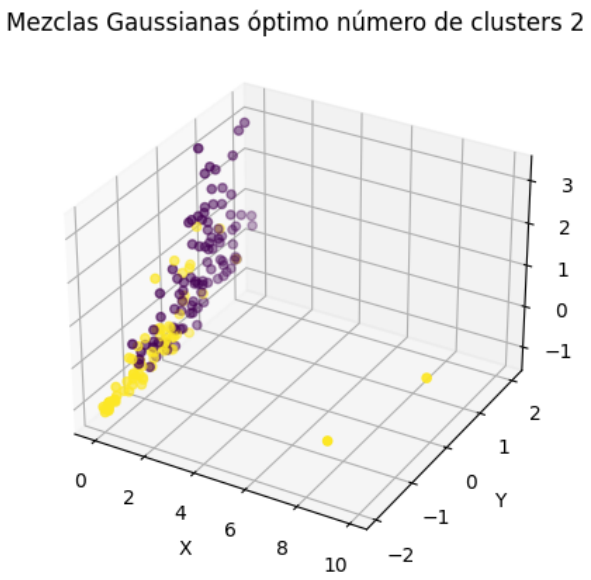


Figura 14: Datos con clusters óptimos, k=3

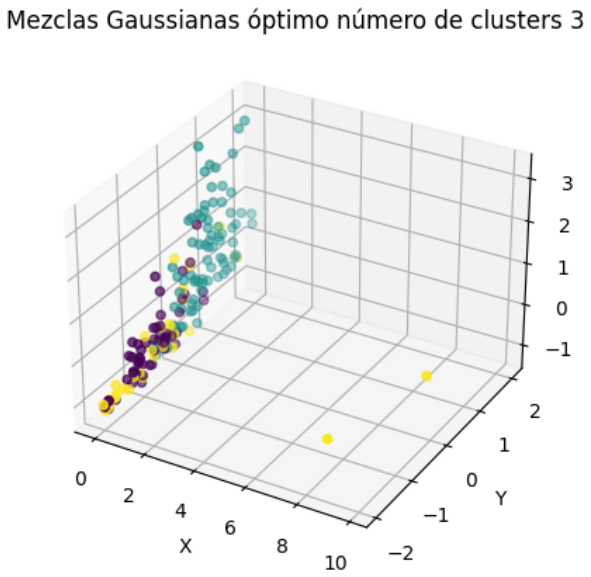


Figura 15: Datos con clusters óptimos, k=3

Otro patrón que se observa es que entre mayor sea el índice de consumo de productos vegetales y menor el consumo de productos animales, el índice de casos confirmados es menores y en consecuencia el índice de muertes y personas recuperadas es menor. También el índice de obesidad juega un papel importante, ya que, entre menor sea este, menor es el índice de casos confirmados, muertes y recuperados.

Explorando a más detalles los datos, observamos que la mayoría de países que se categorizan en aquellas clases que tiene un menor índice en casos confirmados, muertes y recuperados, son aquellos que se encuentran en el continente Africano o Asiático, por ejemplo, Afghanistan, Angola, Bangldesh, Benín, Yemen, Zambia, etc.

5.3. Spectral Clustering

Primeramente se evaluó el método de agrupamiento de Spectral Clustering con $k = 10$:

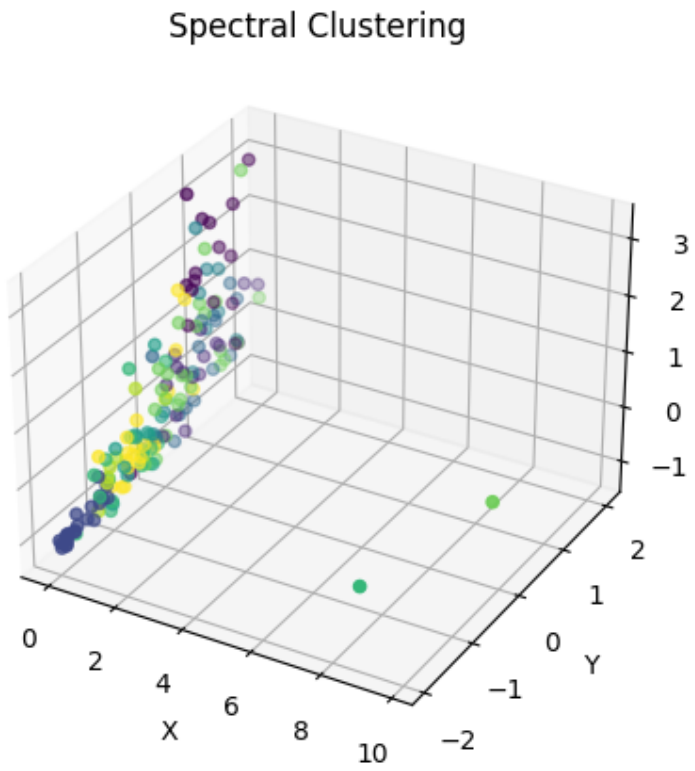


Figura 16: Datos con Spectral Clustering, k=10

etiquetas	Animal Products_x	Vegetal Products_x	Obesity_x	Alcoholic Beverages_y	Fruits - Excluding Wine_y	Confirmed_y	Deaths_y	Recovered_y
0	27.052325	22.948306	22.900000	0.253544	0.838475	3.091939	0.055812	2.383332
1	18.599704	31.400607	20.339286	0.189011	0.839111	1.202931	0.030501	1.044220
2	27.427440	22.573607	22.580000	0.195173	1.068747	1.065192	0.010189	0.661791
3	8.551061	41.447565	7.256522	0.269570	1.109800	0.185041	0.002788	0.158293
4	25.933861	24.065689	25.483333	0.309961	0.772583	4.954481	0.106621	2.723957
5	28.010625	21.988600	24.525000	0.498400	0.719263	8.043833	0.129844	7.205541
6	28.594645	21.405464	25.090909	0.577845	0.590073	4.133580	0.093653	2.395907
7	14.779925	35.220150	26.900000	0.077988	0.970638	2.415808	0.040243	1.979614
8	15.563422	34.437850	9.366667	0.160144	0.649728	0.269471	0.006689	0.215397
9	22.511806	27.489372	16.227778	0.319433	0.815589	0.279721	0.003454	0.249948

Figura 17: Datos con Spectral Clustering, k=10

Como se observa en la Figura 17, en primera estancia no se puede obtener grandes conclusiones con 10 clusters, por lo que se llevó a cabo el método de

Silhouette para determinar el número óptimo de clusters.

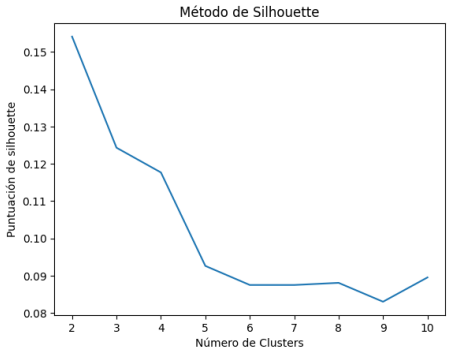


Figura 18: Silhouette Method, técnica para determinar clusters

En la Figura 18, la gráfica de Silhouette demuestra claramente que el número óptimo de clusters es 2, ya que tiene la mayor puntuación del coeficiente de Silhouette. Por consiguiente, se evaluó el método de agrupamiento de Spectral Clustering con 2 clusters.

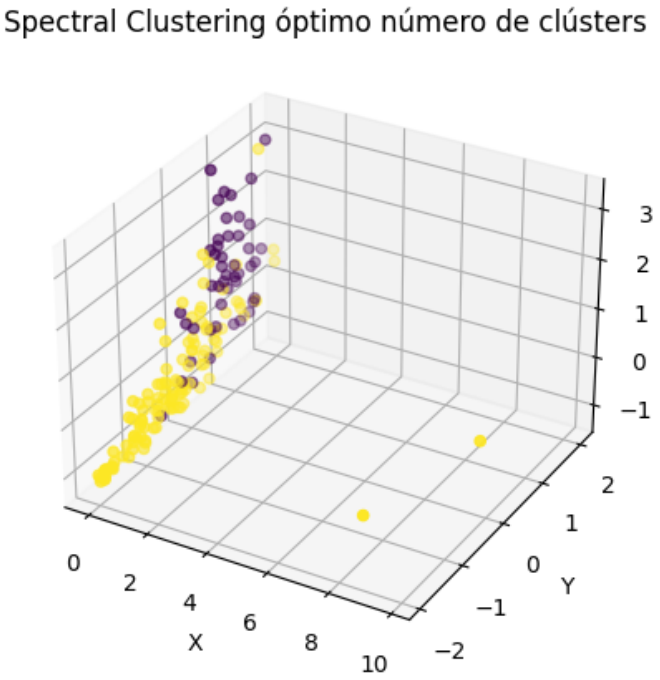
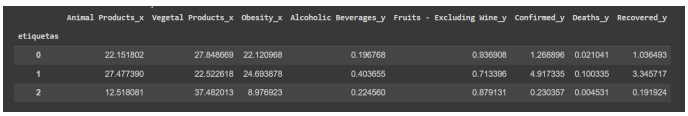


Figura 19: Gráfica con k óptimo, k=2

etiquetas	Animal Products_x	Vegetal Products_x	Obesity_x	Alcoholic Beverages_y	Fruits - Excluding Wine_y	Confirmed_y	Deaths_y	Recovered_y
0	16.693525	33.306807	14.911881	0.203213	0.931430	0.684606	0.011796	0.546038
1	27.172548	22.827468	24.874194	0.373085	0.720727	4.233134	0.084922	2.952144

Figura 20: Datos con clusters óptimos, k=2



etiquetas	Animal Products_x	Vegetal Products_x	Obesity_x	Alcoholic Beverages_y	Fruits - Excluding Wine_y	Confirmed_y	Deaths_y	Recovered_y
0	22.151802	27.848669	22.120968	0.196768	0.936908	1.268896	0.021041	1.036493
1	27.477390	22.522618	24.693878	0.403655	0.713396	4.917335	0.100335	3.345717
2	12.518081	37.482013	8.976923	0.224560	0.879131	0.230357	0.004531	0.191924

Figura 21: Datos con clusters óptimos, k=3

Al momento de querer realizar los agrupamientos con 10 clusters no se pueden observar patrones o comportamientos para cada una de las clases que se generaron, ya que la diferencias entre el promedio de los índices de cada característica no es muy notable.

Sin embargo, una vez que analizamos el número de cluster óptimos, el cual nos arrojó que eran alrededor de 2, hicimos los modelos con 2 y 3 clusters, los cuales nos permitieron observar a detalle las diferencias entre cada clase. Hace sentido que en ambos se ve que entre mayor sea el índice de casos confirmados de COVID, mayor será las muertes y personas recuperadas.

Un patrón que se observa es que entre mayor sea el índice de consumo de productos vegetales y menor el consumo de productos animales, el índice de casos confirmados es menores y en consecuencia el índice de muertes y personas recuperadas es menor. También el índice de obesidad juega un papel importante, ya que, entre menor sea este, menor es el índice de casos confirmados, muertes y recuperados.

Explorando a más detalles los datos, observamos que la mayoría de los países que se categorizan en aquellas clases que tiene un menor índice en casos confirmados, muertes y recuperados, son aquellos que se encuentran en el continente Africano o Asiático, por ejemplo, Afghanistan, United Republic of Tanzania, Bangldesh, Zambia, Algeria, Benín, Yemen, etc.

6. Video

6.1. Alejandra Velasco

Video:
Algoritmo de Agrupamiento: Modelo de Mezclas Gaussianas
→ https://youtu.be/sFdJlZ-TPZ8

Videos de apoyo:

- StatQuest: K-means clustering
  → https://www.youtube.com/watch?v=4b5d3muPQmA
- Modelo de mezcla gaussian
  → https://www.youtube.com/watch?v=q71Niz856KE
- Mezclas gaussianas Fundamentos de Machine Learning 9
  → https://acortar.link/OZS8zS
  (Se utilizó un acortador de link, ya que estaba demasiado largo).

¿Por qué seleccionaste los videos que indicas en tu reporte?
Seleccioné los tres videos porque me pareció muy interesante la manera en la que explicaban los distintos métodos de agrupamiento. A veces entender estos métodos suele ser complicado y puede ser díficil explicarlo en poco tiempo, pero estos videos lograban hacer concisa la información para que no te perdieras o te distrajeras. Entonces decidí adoptar esta idea al transmitir mis ideas pero de una manera un poco mas interactiva y aterrizada. Principalmente me basé en el segundo video, ya que la manera de explciar el método de Meclas Gaussianas me pareció muy interesante y muy bueno. Después de ver ese video no me quedó ninguna duda con este método y eso fue lo que intenté hacer con mi video.

6.2. Laura Valdivia

Video: **K-medias**
→ https://youtu.be/NsVgqY1YfuA
Videos de apoyo:

- A Gentle Introduction to Machine Learning
  → https://www.youtube.com/watch?v=vP06aMoz4v8&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=4
- k-means cluster paso a paso — Machine Learning para novatos
  → https://www.youtube.com/watch?v=T-Q49u5EJ7w&t=708s
- ALGORITMOS DE AGRUPAMIENTO — no. 3 Curso Aprendizaje no Supervisado con Python
  → https://www.youtube.com/watch?v=cmBkWwq3yZU

¿Por qué seleccionaste los videos que indicas en tu reporte?
Anteriormente ya había recurrido a uno de esos canales con el fin de aclarar dudas que tenía respecto a las clases. Realmente me gusta mucho sus forma tan creativa y dinámica con la que explican las cosas. Son muy pocos aquellos que logran explicar los temas de tal manera que los entienda muy bien, además de que no utilizan un vocabulario técnico, lo cual facilita el entender de que están hablando. Además de que el apoyo visual (imágenes, animaciones, etc.) es de gran utilidad. Esto es lo que trate de replicar en mi video, de tal manera que un corto lapso de tiempo quede claro el tema.

6.3. Francisco Chávez

Video: **Algoritmo de Agrupamiento: K-Medias**
→ `https://youtu.be/dKazMSqO6NQ`
Videos de apoyo:

- StatQuest: K-means clustering
  → `https://www.youtube.com/watch?v=4b5d3muPQmA`
- Machine Learning Tutorial Python - 13: K Means Clustering Algorithm
  → `https://youtu.be/EItlUEPCIzM`
- K-means clustering: how it works
  → `https://youtu.be/_aWzGGNrcic`

¿Por qué seleccionaste los videos que indicas en tu reporte?
Estos videos los encontre gracias al recomendador de youtube cuando busque videos sobre k-means por el alto número de vistas que tenían, entonces supuse que eran buenos y efectivamente lo son, creo que cubren el tema de manera completa y tienen buenas animaciones que te ayudan a visualizar lo que está sucediendo durante el método de k-means

7. Conclusiones

Trabajar en este proyecto de bases de datos médicos utilizando técnicas de aprendizaje no supervisado fue una experiencia enriquecedora que nos permitió consolidar los conocimientos adquiridos durante el curso de 5 semanas. A través de la aplicación de algoritmos de clustering y análisis de datos, pudimos explorar y comprender mejor la estructura subyacente de los datos médicos, identificando patrones y grupos relevantes.

Además, esta experiencia destacó el potencial de la inteligencia artificial en el campo de la medicina. La capacidad de extraer información significativa y encontrar patrones ocultos en grandes conjuntos de datos médicos puede tener un impacto significativo en el diagnóstico, tratamiento y seguimiento de enfermedades. La inteligencia artificial puede ayudar a los profesionales de la salud

a identificar subgrupos de pacientes, predecir resultados clínicos, descubrir correlaciones y proporcionar recomendaciones personalizadas de tratamiento.

En conclusión, este proyecto fue una oportunidad valiosa para consolidar los conocimientos en aprendizaje no supervisado y reconocer el potencial de la inteligencia artificial en la medicina. La combinación de tecnología y experiencia médica puede llevar a avances significativos en la atención médica, mejorando la precisión y eficiencia en el diagnóstico y tratamiento de enfermedades.

8. Referencias

Bevans, R. (2020). Akaike Information Criterion — When How to Use It (Example). *Scribbr*. Recuperado de https://www.scribbr.com/statistics/akaike-information-criterion/

Algoritmo de mezclas gaussianas o Gaussian Mixture Model (GMM). (s. f.). *Ediciones Eni*. Recuperado de https://www.ediciones-eni.com/open/mediabook.aspx?idR=1922ca92926a43a880845e

El algoritmo k-means aplicado a clasificación y procesamiento de imágenes. (s. f.). *Uni video*. Recuperado de https://www.unioviedo.es/compnum/laboratorios$_p y/kmeans/kmeans.html$

Fleshman, W. (2019). Spectral Clustering. *Towards Data Science*. Recuperado de https://towardsdatascience.com/spectral-clustering-aba2640c0d5b