ANALÍTICA II - APRENDIZAJE NO SUPERVISADO - CASO DE ESTUDIO: ANÁLISIS DE LAS CARACTERÍSTICAS DE LAS DEFUNCIONES OCURRIDAS EN EL VALLE DEL CAUCA PARA EL PERÍODO 2021 -2022

Cesar Iván Ávila Díaz. cesar.avila@udea.edu.co
Marcelo Lemus. lemus@udea.edu.co
María Gabriel Pérez Barrios. mgabriel.perez@udea.edu.co
Verónica Andrea Morales González. veronica.moralesg@udea.edu.co

Repositorio GitHub: https://github.com/veronica1908/T2_APRENDIZAJE_NO_SUPERVISADO

RESUMEN

El presente proyecto se sumerge en el análisis de un conjunto de datos que abarca el período de cinco años, desde 2018 hasta 2022, con un enfoque en la cantidad de defunciones según el lugar en donde habitan, y otros aspectos como el tipo de muerte. Este caso de estudio analiza las características de las personas fallecidas en el departamento del Valle del Cauca durante los años 2021 y 2022. Considera siete variables categóricas y dos variables numéricas para aplicar tres algoritmos de aprendizaje no supervisado e identificar cuál de ellos genera un mejor agrupamiento (clústers) de las características de las variables de la base de datos.

A lo largo de este proyecto, se aplicarán técnicas avanzadas de aprendizaje no supervisado, como algoritmos de clustering y reducción de dimensionalidad, para analizar y visualizar los datos. Además, se presentarán conclusiones y recomendaciones basadas en los hallazgos obtenidos, con el objetivo de proporcionar información útil para aquellos que trabajan en la formulación de políticas públicas, así como para cualquier persona interesada en comprender la relación entre la educación y la salud.

PALABRAS CLAVE

I. CASO DE ESTUDIO

El panorama de la educación y la salud es fundamental en el desarrollo de cualquier sociedad. La calidad de la educación y el bienestar de la población están intrínsecamente ligados, y la información precisa sobre la relación entre estos dos factores es esencial para la formulación de políticas efectivas. En este contexto, el análisis de datos se ha convertido en una herramienta poderosa para obtener información valiosa que respalde la toma de decisiones en ámbitos tan críticos como la educación y la salud pública.

Se ha elegido la base de datos "CANTIDAD DEFUNCIONES POR NIVEL EDUCATIVO 2018 - 2022" del sitio de datos abiertos en Colombia [1]. De esta base se pretende extraer las variables más representativas para caracterizar los diferentes casos de fallecimiento de las personas en el departamento del Valle del Cauca, considerando tres algoritmos de aprendizaje no supervisado para generar los *clústers* de acuerdo con la base de datos y sus características, permitiendo obtener como resultado, diferentes grupos que clasifican las características de los casos de fallecidos.

II. DISEÑO DE LA SOLUCIÓN DEL CASO DE ESTUDIO

Para desarrollar este caso de estudio se tuvo en cuenta los siguientes pasos:

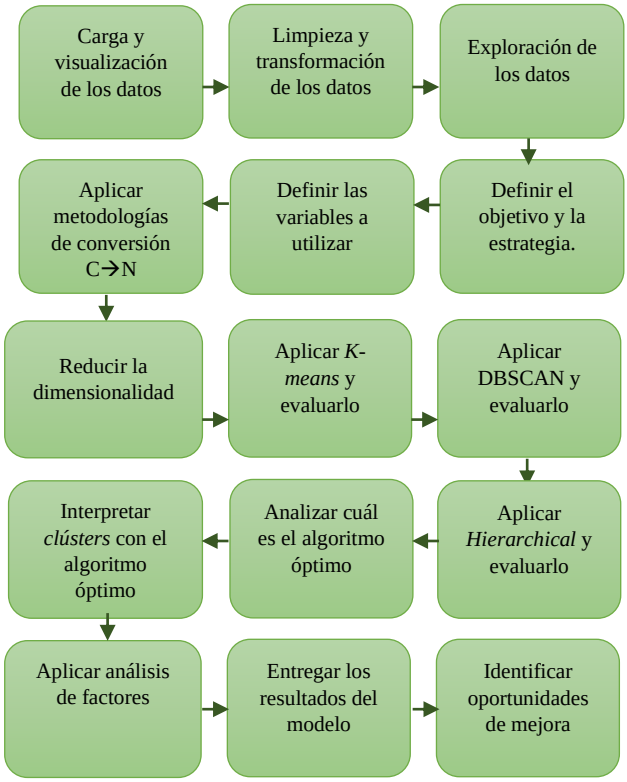


Ilustración 1. Solución teórica propuesta.

III. DESARROLLO DEL CASO DE ESTUDIO

a. Carga y visualización de los datos: Se realiza la carga de los datos en el repositorio en Git hub y alternativamente en Colab para agilizar la elaboración del código.

b. Limpieza y transformación de los datos: Se verifica una a una las 31 variables de la base de datos inicial con 5307 características. Se unifican las categorías, se hace tratamiento de datos nulos y se descartan las variables N°, Hora de defunción y Último curso fallecido, quedando hasta aquí con una base de dimensiones (2424, 29).

```
RangeIndex: 5308 entries, 0 to 5307
Data columns (total 31 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   N°                              4167 non-null   float64
 1   Departamento defunción          5308 non-null   object
 2   Municipio defunción             5308 non-null   object
 3   Área defunción                  5305 non-null   object
 4   Sitio defunción                 5308 non-null   object
 5   Otro sitio de defunción         54 non-null     object
 6   Nombre institución de defunción 5308 non-null   object
 7   Tipo de defunción               5308 non-null   object
 8   Fecha de Defunción              5308 non-null   object
 9   Hora de defunción               4236 non-null   object
 10  Sexo                            5308 non-null   object
 11  Estado civil                    5257 non-null   object
```

Ilustración 2. Información de la base de datos inicial

c. Exploración de los datos: Una vez que obtuvimos los datos, realizamos un análisis exploratorio de los mismos. Esta etapa implicó: Resumir estadísticamente las variables clave para comprender la distribución de los datos y detectar posibles valores atípicos y crear visualizaciones, como gráficos de barras e histogramas, para representar las características de los datos.

Teniendo en cuenta la exploración rápida de los datos para cada variable, tenemos el siguiente análisis:

Todos los datos de defunciones de la base son de personas que vivían y murieron en Colombia, del departamento del Valle del Cauca, principalmente en el municipio de Guadalajara de Buga con el 99%, sin embargo sus municipios de residencia variaban entre Buga con el 73% y Guacari con un 7%.

La principal área de defunción ha sido la cabecera municipal con el 96% de los casos en diferentes barrios, 2% en área rural dispersa y 1% en un centro poblado. En cuanto al sitio específico de defunción, el 64% ha fallecido en el hospital, el 30% en su casa, 5% en vías públicas y el otro 1% en otros lugares.

Las instituciones de salud en las que se han presentado mayor cantidad de fallecidos son la Fundación Hospital San José con el 74% de los casos, 6% en la E.S.E. Hospital Divino Niño y hay un 15% sin información. Para el tipo de defunción, se ha tenido un 100% de defunciones de tipo No fetal, la mayoría de los fallecidos han sido del sexo Masculino con el 55%, el 25% de fallecidos estaban viudos, el 23% estaban casados, el 22% solteros.

La mayoría de los fallecidos eran mayores de 50 años, su nivel educativo era principalmente básica primaria (55%) y en segundo lugar básica secundaria con 15%. El 30% se ocupaba de su hogar y se tiene un 42% sin esta información de ocupación. En cuanto a rasgos físicos y cultura, la mayoría tenía "ninguna" clasificación (98%), solo un 2% era afrocolombiano y menos del 1% indígena.

La mayoría de los fallecidos contaban con seguridad social: 48% Contributivo y 48% subsidiado, la EPS era principalmente Nueva EPS con el 31% y luego EMSSANAR E.S.S. con el 17%. La muerte natural por enfermedad con un 65% es la que más se presentó. Se tiene un 25% de los casos sin información.

Cuando se quiere ver la información de fallecidos por enfermedad y su EPS, se tiene que la mayoría pertenecían a La Nueva EPS, y EMSSANAR.

Cuando se observa la información por manera de morir y su distribución de edades, se puede decir que la muerte por enfermedad predomina en personas mayores de 50 años, mientras que para el resto de maneras predominan edades inferiores a 50 años.

Los datos son principalmente de los años 2021 y 2022 y los meses con mayor cantidad de fallecidos fueron los meses de Julio y enero con 11% cada uno y luego Junio con el 10% de los casos. Tanto en 2021 como en 2022, la mayoría de fallecidos fueron del sexo masculino.

Tanto en 2021 como en 2022, la mayoría de muertes fueron de manera natural por enfermedad, sin embargo hay una buena proporción sin información en el 2022. En 2021 se dio una porción de homicidios y en 2022 se cataloga como muerte no natural. Se observa que en todos los meses del año 2022 se presentó reducción en la cantidad de muertes en comparación con el año 2021. En la mayoría de meses domina la muerte por enfermedad, en todos está presente la muerte por homicidio y una baja proporción por accidentes.

d. Definir el objetivo y la estrategia: Con la disponibilidad de información y verificando que la gran mayoría de las variables son categóricas, se define la variable "Probable manera de muerte" como la que clasifica principalmente el caso de estudio y a partir de allí, se hace una selección de variables categóricas complementarias como características de cada evento y las cuales, de acuerdo con sus categorías, podrían convertirse fácilmente en numéricas.

e. Definir las variables a utilizar: Teniendo en cuenta lo anterior, se eligen las variables relacionadas en la ilustración 3, de las cuales siete son categóricas y dos son numéricas (edad y año). Se consideró el año como variable numérica dado que, en la exploración de los datos, se observó diferencia entre los años 2021 y 2022 contrastado con diferentes variables como mes y probable manera de muerte.

	mes	Área defunción	Sexo	Probable manera de muerte	Seguridad social	Estado civil	Nivel educativo	Edad	año
1579	November	CABECERA MUNICIPAL	FEMENINO	HOMICIDIO	CONTRIBUTIVO	SOLTERO(A)	NINGUNO	35	2019
2752	January	CABECERA MUNICIPAL	MASCULINO	NATURAL (ENFERMEDAD)	CONTRIBUTIVO	SIN INFORMACIÓN	MEDIA ACADÉMICA	80	2021
2753	January	CABECERA MUNICIPAL	FEMENINO	NATURAL (ENFERMEDAD)	SUBSIDIADO	VIUDO(A)	BÁSICA SECUNDARIA	70	2021
2754	January	CABECERA MUNICIPAL	MASCULINO	NATURAL (ENFERMEDAD)	CONTRIBUTIVO	SOLTERO(A)	BÁSICA PRIMARIA	81	2021
2756	January	CABECERA MUNICIPAL	MASCULINO	NATURAL (ENFERMEDAD)	CONTRIBUTIVO	CASADO(A)	BÁSICA SECUNDARIA	77	2021

Ilustración 3. Variables seleccionadas

f. Aplicar metodologías de conversión de variables categóricas a variables numéricas: Se utilizó el método de codificación de las categorías de cada variable, principalmente de acuerdo con la cantidad de características dentro de cada categoría teniendo en cuenta el análisis exploratorio de los datos, en donde el mayor número representa la categoría con mayor cantidad de datos, a excepción de las variables sexo que se codificó similar a *dummies*, la variable nivel de educación que se asignó la numeración por la naturaleza que representa justamente ese nivel de educación y la variable probable manera de muerte, la cual se codificó en orden de gravedad de acuerdo con consultas aleatorias de la forma en que lo considera la criminalística colombiana.

Con esta conversión de categorías se revisa la matriz de correlación y solo se identifica una correlación baja entre la edad y el estado civil.

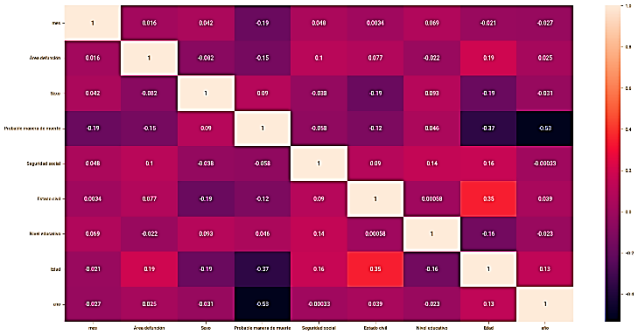


Ilustración 4. Matriz de correlación.

Se grafican las variables para verificar su distribución y se observa solo distribución más o menos normal para las variables Nivel educativo y Edad.

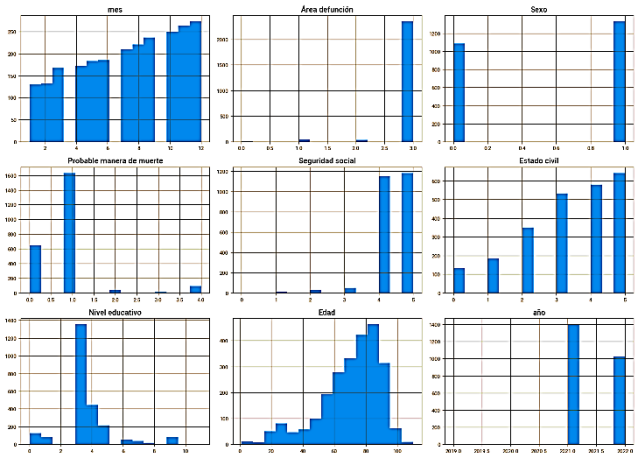
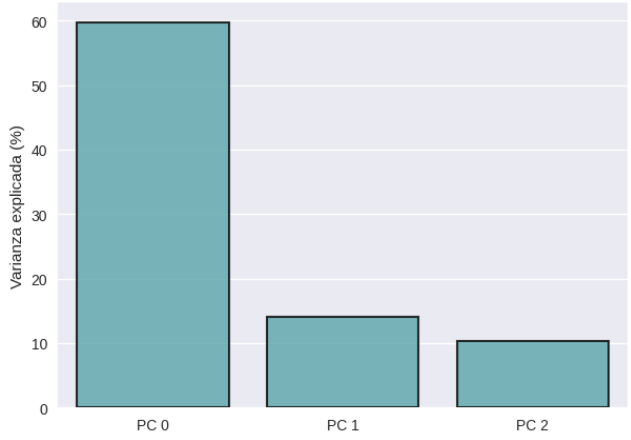


Ilustración 5. Distribución de las nuevas variables numéricas

g. **Reducir la dimensionalidad:** Para reducir aún más la dimensionalidad, se realizó un tratamiento de datos atípicos para las variables nivel educativo, año y edad (*Ilustración 8*) y finalmente se realiza el escalado para las variables numéricas para proceder a realizar la aplicación de los algoritmos considerando unas dimensiones de la base de datos de (2283, 9). Se aplica el PCA para que explique el 84% de la varianza y se definen tres componentes.



h. *Ilustración 6. Varianza explicada (84%)*

El componente 1 explica el 59.8% de la varianza.

	Varianza explicada (%)
0	59.804379
1	14.068341
2	10.377526

Ilustración 7. Distribución de los componentes

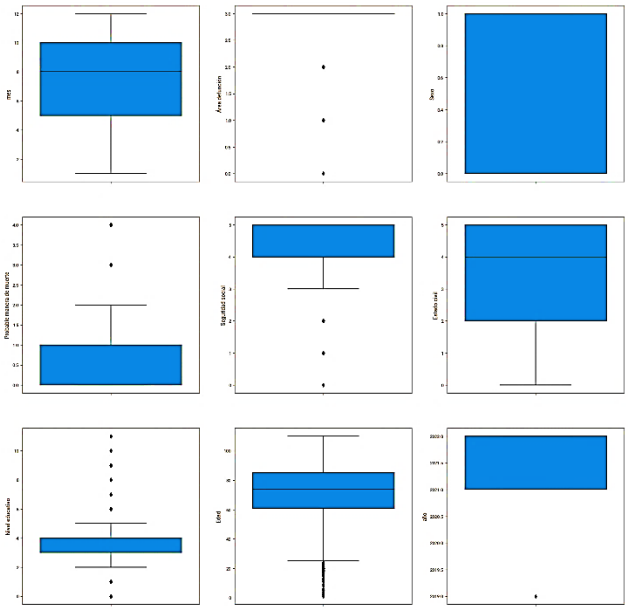


Ilustración 8. Identificación de datos atípicos

i. **Aplicar *K-means* y evaluarlo:** Aquí se considera el *dataset* escalado y el *dataset* reducido, obteniendo las siguientes métricas para cada uno al aplicar el método de silueta para definir el número óptimo de *clústers.*

Para el *dataset* escalado se obtuvo una definición de tres *clústers.*

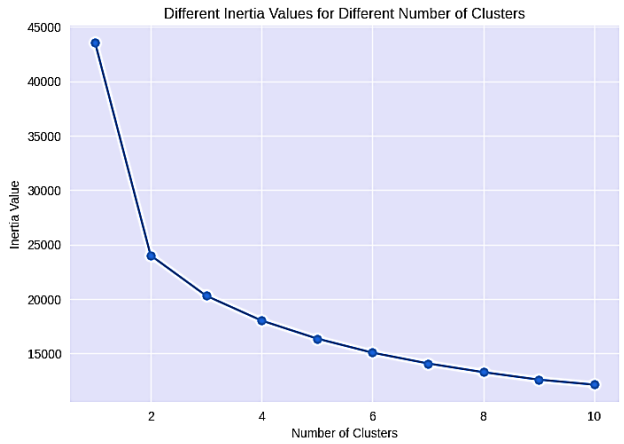


Ilustración 9. Valor de la inercia para diferentes tipos de clústers en dataset escalado.

El resultado de las métricas parael *K-means* en el *dataset* escalado es:

```
K-MEANS D.E.
Inertia:  20307.56647987564
Silhouette Score:  0.2509019854780971
Calinski harabasz score:  1302.310275870395
```

Ilustración 10. Métricas K-means dataset escalado

Aquí la alta inercia indica *clusters* no tan compactos, la puntuación de silueta indica que no están tan bien definidas al ser bajo y el bajo puntaje *Calinski-Harabasz* sugiere *clusters* no tan bien separados.

Para el *dataset* reducido se tiene igualmente selección de tres *clústers*

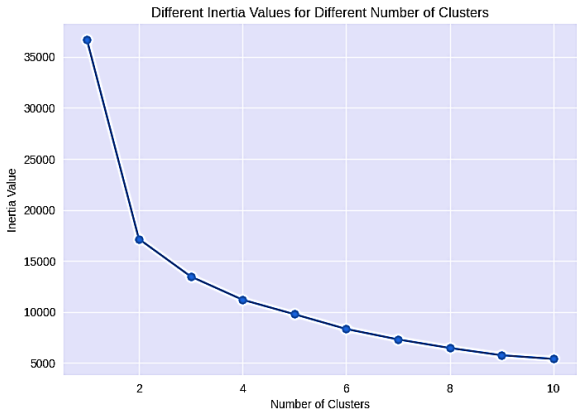


Ilustración 11. Valor de la inercia para diferentes tipos de clústers en dataset reducido.

```
K-MEANS D.R.
Inertia:  13479.166705923515
Silhouette Score:  0.33355561721437565
Calinski harabasz score:  1960.0367992411645
```

Ilustración 12. Métricas K-means dataset reducido

Estos resultados indican que el modelo de *K-Means* NO ha logrado formar *clusters* suficientemente compactos y bien definidos en el conjunto de datos, aunque tiene un mejor puntaje en comparación con *Dataset* Escalado y por lo tanto se procede a graficarlo:

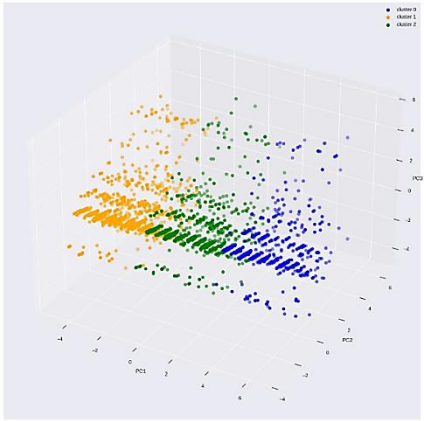


Ilustración 13. Gráfico de clústers K-means Dataset reducido

j. Aplicar DBSCAN y evaluarlo: Para este algoritmo, se aplica en el *dataset* original y también en el *dataset* reducido. Para el original, se obtuvo una curvatura máxima en 2277.

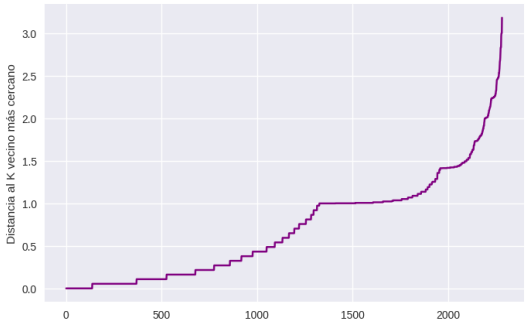


Ilustración 14. Curvatura máxima DBSCAN en dataset original

Se aplicó la selección de hiperparámetros de muestras mínimas y se obtuvo lo siguiente:

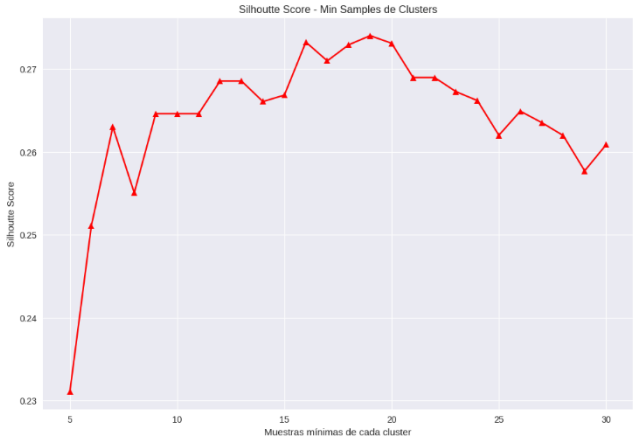


Ilustración 15. Muestras mínimas dataset original

```
DBSCAN
DBSCAN(eps=2.8366918457571466, min_samples=19, n_jobs=-1)
```

Ilustración 16. Muestras mínimas y eps para dataset original

eps: Esto significa que cualquier punto que esté a la distancia indicada en unidades de otro punto se considera parte del mismo *cluster*.
min_muestras=19: El parámetro min_samples especifica el número mínimo de muestras que deben estar en la vecindad de un punto para que ese punto sea considerado un "punto central". En este caso, se ha establecido que al menos 19 muestras deben estar dentro de la vecindad de un punto para que ese punto sea considerado un "punto central".
Finalmente se calculan las métricas:

```
DBSCAN D.O.
Silhouette Score:  0.27401473659011294
Calinski harabasz score:  31.06441990973646
```

Ilustración 17. Métricas DBSCAN dataset original

Se obtuvo un *Silhouette Score* más bajo en comparación con *K-means* en el *dataset* reducido, pero levemente mejor en comparación con el *dataset* escalado de *K-means*. En cuanto a *Calinski harabasz score* es muchísimo menor.

Se aplica la predicción de los clústers y se observa que hay dos etiquetas o clases (0 y -1). DBSCAN está llevando todos los datos a la clase cero y los valores de ruido y/o atípicos a la clase -1. Esto puede explicar el resultado bajo de *Calinski harabasz score*, al tener solo un *cluster* comparado con *K-means* donde teníamos 3, por lo tanto no puede compararse tanto el resultado entre algoritmos con respecto a *Calinski harabasz score*, sino solamente con el índice de silueta que resultó mejor para *K-means*. Hasta aquí podría decirse que DBSCAN no es tan útil para nuestra base de datos. Sin embargo, vamos a utilizar otros valores de epsilon más bajos para verificar si se forman otros tipos de *clusters*.

Se aplica entonces nuevamente DBSCAN para el *dataset* reducido:

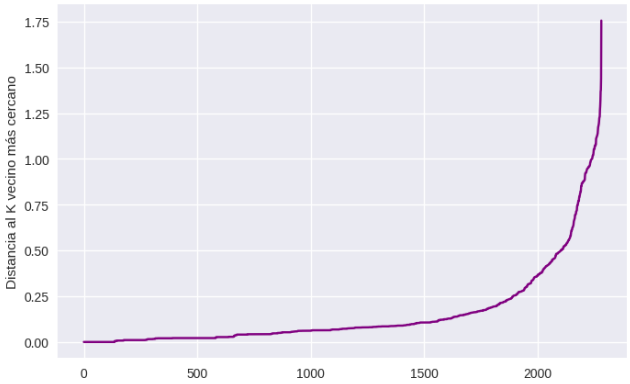


Ilustración 18. Curvatura máxima DBSCAN en dataset reducido

Se observa una gráfica diferente a la inicial con una curvatura de 2251 y el valor de epsilon es menor. Un valor epsilon mucho más grande implica que el algoritmo DBSCAN considera puntos como cercanos incluso si están a una distancia considerable entre sí. Esto puede llevar a la formación de *clusters* más grandes y menos densos, ya que puntos que están más separados aún pueden ser considerados parte del mismo *cluster*.
Se aplicó la selección de hiperparámetros de muestras mínimas y se obtuvo lo siguiente:

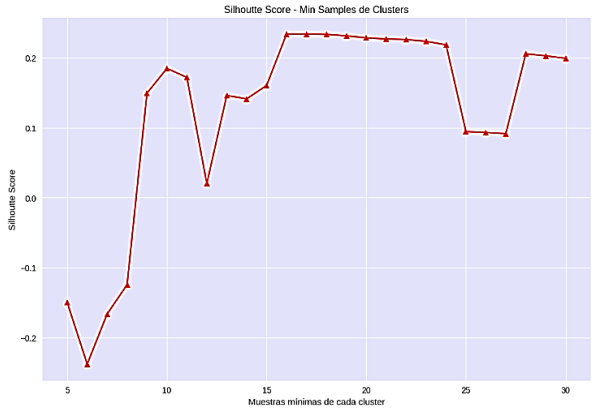


Ilustración 19. Muestras mínimas dataset reducido

```
DBSCAN
DBSCAN(min_samples=16)
```

Ilustración 20. Muestras mínimas en dataset reducido

Para este *dataset* se reduce en tres muestras mínimas y el valor eps es más bajo (1.05).

Finalmente se calculan las métricas:

```
DBSCAN D.R.
Silhouette Score:  -0.14986081158806078
Calinski harabasz score:  105.03330618500573
```

Ilustración 21. Métricas DBSCAN dataset reducido

Se observa que el silhoette *score* disminuyó por debajo de cero, el *Calinski harabasz score* es muy bajo también, y en la predicción de clústers aumenta hasta 18 etiquetas. Graficamos las 18 etiquetas para este algoritmo.

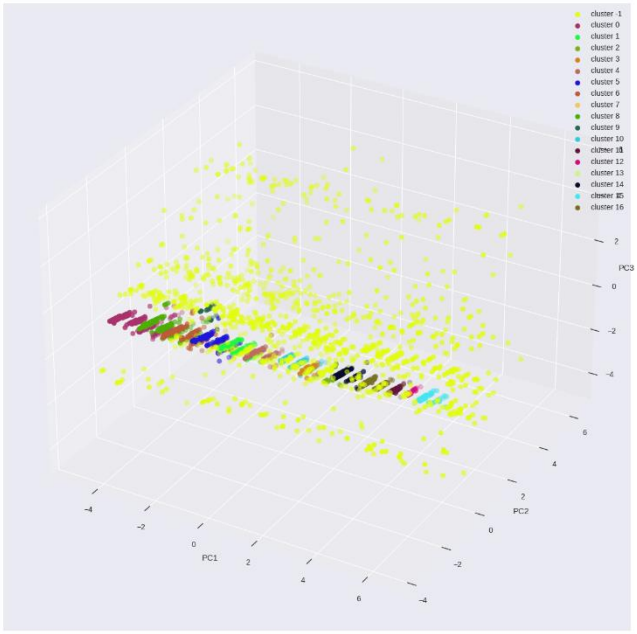


Ilustración 22. Gráfico de clústers DBSCAN Dataset reducido

Al observar gráficamente y ver el número de clústers formados, puede verse que el algoritmo no es útil, ya que generó demasiadas clases, lo cual no es efectivo para el agrupamiento de la base de datos y por lo tanto, se concluye que este algoritmo no es óptimo para nuestra base de datos

k. Aplicar *Hierarchical* y evaluarlo: Se usó el dendrograma aplicado a la variable “Probable manera de muerte” para observar el número de *clústers* y se eligen tres.

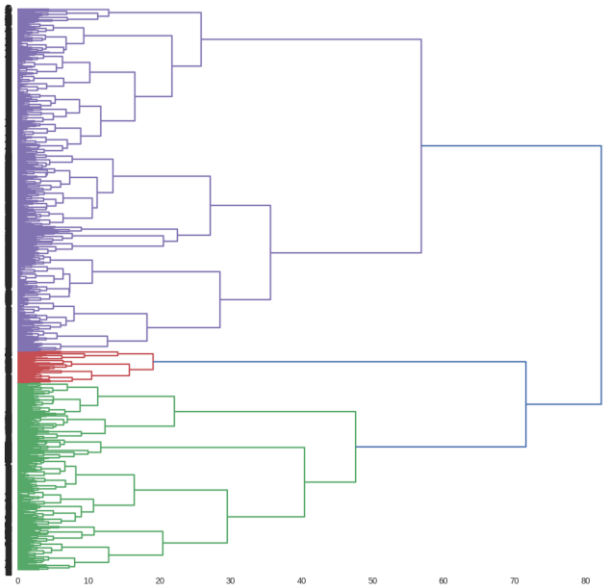


Ilustración 23. Dendrograma de la variable Probable manera de muerte

Finalmente se calculan las métricas para este algoritmo en la base reducida y se obtiene lo siguiente:

```
Silhouette Score: 0.3989032208823436
Calinski-Harabasz Score: 286.1722122965244
```

Ilustración 24. Métricas Hierarchical dataset reducido

Para este algoritmo, el *Silhouette score* es el mejor de todos, sin embargo el índice *Calinski-Harabasz* para este caso es muy bajo.

l. Analizar cuál es el algoritmo óptimo:

El índice *Calinski-Harabasz* se basa en la comparación de la relación ponderada entre la suma de los cuadrados (la medida de la separación del clúster) y la suma de los cuadrados dentro del clúster (la medida de cómo se empaquetan estrechamente los puntos dentro de un clúster), como se mencionó, aunque el *Hierarchical* tuvo mejor puntaje de silueta, el puntaje *Calinski* fue muy bajo, por lo tanto se elige el algoritmo *K-means* con *dataset* reducido como el algoritmo óptimo, ya que tiene mucho mejor *Calinski-Harabasz Score* y la diferencia en el *Silhouette score* no es mucha.

Tabla 1. Comparación entre algoritmos aplicados

Algoritmo	*Silhouette Score*	*Calinski-Harabasz Score*
K-means D.E.	0.2509019854780971	1302.310275870395
K-means D.R.	0.3335556172144477	1960.036799241166
DBSCAN D.O.	0.27401473659011294	31.06441990973646
DBSCAN D.R.	0.14986081158806078	105.03330618500573
HIERARCHICAL D.R.	0.3989032208823436	286.1722122965244

Elección del Algoritmo K-Means:

Después de un análisis exhaustivo de las diferentes técnicas de *clustering* disponibles, se optó por aplicar el algoritmo *K-Means* en la base reducida para agrupar las funciones en clústeres. Esta elección se basó en su capacidad para manejar grandes conjuntos de datos y su eficacia en la detección de patrones ocultos. Además, *K-Means* es un algoritmo altamente escalable y se ajusta bien a los datos numéricos presentes en nuestro conjunto de datos.

K-Means* vs. *Hierarchical: A pesar de que el *clustering* jerárquico puede ser efectivo, *K-Means* se eligió debido a su capacidad para manejar grandes volúmenes de datos de manera más eficiente. Además, *K-Means* permite una asignación definitiva de puntos a clústeres, lo que facilita la interpretación.

***K-Means* vs. DBSCAN:** Aunque DBSCAN es excelente para detectar clústeres con formas arbitrarias, su rendimiento puede ser sensible a la elección de hiperparámetros y no funcionar óptimamente en conjuntos de datos de alta dimensionalidad. *K-Means*, en este caso, proporcionó una solución más robusta y escalable.

m. Interpretar clústers con el algoritmo óptimo:

Una vez que obtuvimos nuestros clústers, llevamos a cabo un análisis detallado de cada uno. Esto implicó identificar perfiles característicos para cada clúster, considerando las variables más influyentes y las tendencias clave.

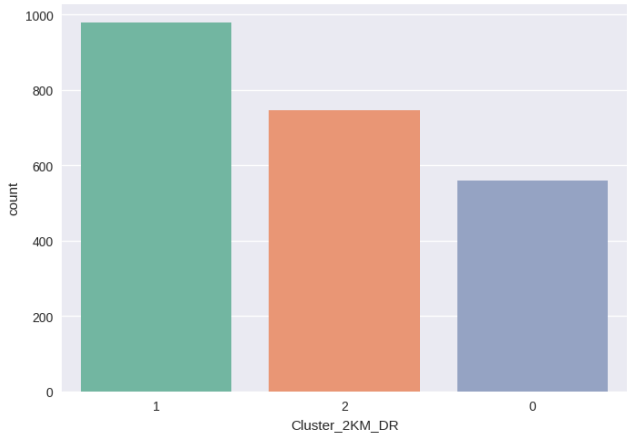


Ilustración 25. Distribución de las características en los clusters

Se tiene una distribución de datos que no es muy pareja, pero se ordena descendentemente así: el grupo 1 tiene más datos, le sigue el grupo 2 con una mediana diferencia y está en último lugar el grupo 0 con una diferencia similar de observaciones por debajo en comparación con la que hay entre el grupo 1 y el grupo 2.

Finalmente, graficamos las variables que pasaron de categóricas a numéricas a través de la metodología de codificación para ver su comportamiento en los clústers definidos.

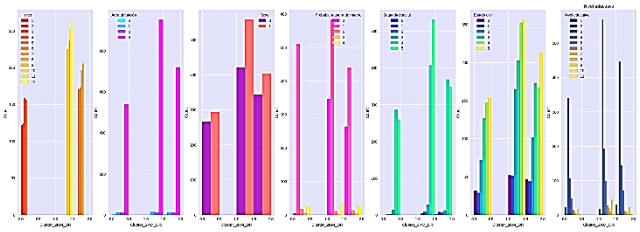


Ilustración 26. Distribución de las variables categóricas por clúster

Para la variable mes, puede verse que cada grupo quedó con una asignación de las características de cuatro meses, teniendo una similar distribución de características entre si. En el grupo 0 hay menor cantidad de características correspondiente a los meses 1, 2, 3 y 4 que corresponden respectivamente a Noviembre, septiembre, diciembre y octubre, siendo diciembre el de mayor cantidad de características. En el grupo 1 está la mayor cantidad de características y agrupa los meses 9, 10, 11 y 12 que corresponden a mayo, junio, julio y enero, siendo enero el de mayor cantidad de características. Finalmente está el grupo 2 que reúne características de los meses 5, 6, 7 y 8 que representan a Marzo, agosto, abril y febrero respectivamente, siendo febrero el de mayor cantidad de características.

En cuanto al área de defunción, el grupo 1 también tiene mayor cantidad de características, la distribución en los tres grupos está dominada por el área 3 que corresponde a cabecera municipal, para los grupos 1 y 2 le sigue el área 1 que corresponde a rural disperso y en el grupo 0, se tiene igual proporción entre el área 1 y 2 que corresponde a rural disperso y centro poblado respectivamente. El área 0 representa la categoría sin información que fue tratada previamente para reducir su cantidad.

Para la variable sexo, dominan el sexo Masculino (1) en cuanto a cantidad de características para los tres grupos.

Para la variable de probable manera de muerte, se tiene mayor cantidad de características en el tipo enfermedad (1), para los tres grupos, en segundo lugar están las características clasificadas sin información (0) para los grupos 1 y 2, mientras que en el grupo 0, es el homicidio (4), el que está en segundo lugar.

Se tiene diferentes distribuciones en los grupos para la variable Seguridad social, en el grupo 0 y 2 domina el tipo subsidiado (4), seguido por poca diferencia del tipo contributivo (5). En el grupo 1 domina el tipo contributivo (5) seguido por poca diferencia del tipo subsidiado (4).

Para el estado civil se presenta dominancia del estado civil 5 en todos los grupos, correspondiente a VIUDO (A), en los grupos 0 y 1, se tiene similar distribución de las categorías proporcionalmente hablando, ya que el grupo 1 cuenta con mayor cantidad, en segundo lugar tienen el estado civil 4 que corresponde a CASADO (A). Para el grupo 2, el segundo lugar está ocupado por ele stado civil 3 que corresponde a SOLTERO (A).

Finalmente, en cuanto a nivel educativo se tiene una distribución similar de categorías en los tres grupos, proporcionalmente hablando. El grupo 1 mantiene la mayor cantidad de características seguido del grupo 2 y el grupo 0 con menor cantidad, la categoría con más características en todos los grupos es básica primaria, seguido de básica secundaria.

Se grafica también las variables edad y año como variables numéricas y su distribución con respecto a los clústers definidos:

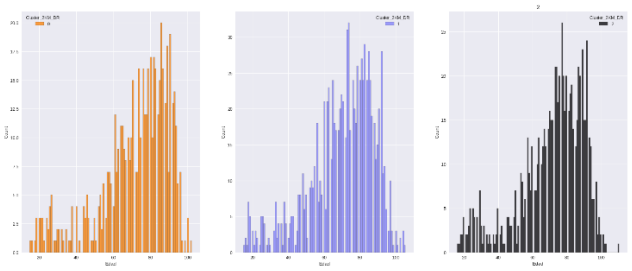


Ilustración 27. Distribución por clúster de la variable numérica Edad

Se tiene una distribución muy similar entre los tres grupos, los cuales incluyen características de todas las edades, pareciéndose en gran medida a la distribución normal. Los rangos de edad con mayor ocurrencia de muertes por grupo son: grupo 0: edades entre los 70 y los 95 años. Grupo 1: edades entre los 60 y los 95 años. Grupo 2: edades entre los 75 y los 95 años.

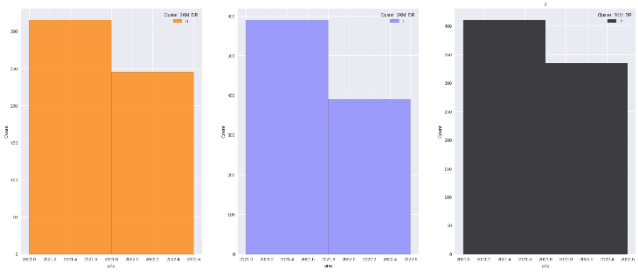


Ilustración 28. Distribución por clúster de la variable numérica Año

Para la otra variable numérica Año, al haber solo dos años para los datos, se observa que se comparten datos en similar proporción en todos los grupos, presentándose una diferencia mayor en el grupo 1 y presentándose mayor ocurrencia en el año 2021.

n. Aplicar análisis de factores: En un intento de realizar un análisis factorial exploratorio, realizamos pruebas de adecuación de datos, incluyendo la Prueba de Bartlett y el KMO (Kaiser-Meyer-Olkin). Sin embargo, los resultados indicaron que nuestros datos no eran adecuados para un análisis factorial en esta etapa, ya que el KMO general de nuestros datos es 0.54 (<0.6), por lo tanto es más adecuado utilizar PCA.

IV. CONCLUSIONES

- Clúster 0: Presenta una mayor ocurrencia de defunciones en los meses de diciembre y octubre, con una distribución geográfica significativa en la cabecera municipal, así como una proporción similar entre la zona rural dispersa y el centro poblado. Las defunciones son predominante de individuos de sexo masculino y se deben principalmente a enfermedades, seguidas por homicidios y suicidios. En cuanto a la seguridad social, las personas en este clúster suelen pertenecer a los tipos subsidiado y contributivo, en ese orden. El estado civil viudo es más común, seguido por el estado casado. Además, el nivel educativo predominante es Básica Primaria, seguido de Básica Secundaria. Por último, la edad de las personas

fallecidas en este clúster se encuentra principalmente en el rango de 70 a 95 años mayormente en el 2021.

- Clúster 1: Muestra una mayor ocurrencia de defunciones en los meses de julio y enero. La distribución geográfica es más pronunciada en la cabecera municipal, seguida de la zona rural dispersa. Las defunciones son principalmente de individuos de sexo masculino y se deben a enfermedades, seguidas de casos sin información y homicidios. En términos de seguridad social, las personas de este clúster suelen pertenecer a los tipos contributivo y subsidiado, con una inclinación hacia el primero. El estado civil viudo es predominante, seguido de casado. Similar al Clúster 0, el nivel educativo más común es Básica Primaria, seguido de Básica Secundaria. Las edades de las personas fallecidas en este clúster se encuentran principalmente en el rango de 60 a 95 años mayormente en el 2021.

- Finalmente, el Clúster 2 presenta una mayor ocurrencia de defunciones en los meses de abril y febrero. La distribución geográfica se concentra en la cabecera municipal y la zona rural dispersa. Al igual que en los clústers anteriores, las defunciones son mayoritariamente de individuos de sexo masculino y se deben a enfermedades, seguidas de casos sin información y homicidios. En lo que respecta a la seguridad social, las personas de este clúster suelen pertenecer a los tipos subsidiado y contributivo en ese orden. El estado civil viudo es más común, seguido de soltero. El nivel educativo predominante es Básica Primaria, seguido de Básica Secundaria, y las edades de las personas fallecidas en este clúster se encuentran principalmente en el rango de 75 a 90 años mayormente en el 2021.

Estas conclusiones indican las características distintivas de cada clúster y son fundamentales para comprender las tendencias de mortalidad en la población estudiada.

V. RECOMENDACIONES

- Después de realizar el estudio recomendamos analizar con mayor profundidad los Factores Socioeconómicos, recomendamos realizar una investigación exhaustiva sobre las relaciones entre las variables de seguridad social, área de defunción y nivel educativo, ya que estas conexiones pueden arrojar luz sobre cómo influyen en las tasas de mortalidad.

- Además, consideramos relevante tener en cuenta análisis como estos en la realización de políticas públicas, dado que estos datos son esenciales para la formulación de políticas de salud y prevención basadas en resultados concretos, en particular en lo que respecta a las posibles causas de defunciones, como enfermedades y eventos violentos.

- Sumado a eso, recomendamos hacer énfasis en la atención diferenciada por género, teniendo en cuenta que, reconociendo las diferencias en la distribución de defunciones por sexo, se pueden adaptar los servicios de salud y prevención para satisfacer las necesidades específicas de hombres y mujeres. Esto garantiza una atención diferenciada y efectiva, abordando de manera más precisa los factores de riesgo y las estrategias de prevención.

- A pesar del desafío presentado por el resultado de la prueba KMO, se recomienda continuar explorando un análisis factorial. Este enfoque puede ayudar a identificar posibles factores subyacentes que podrían estar influyendo en las defunciones, proporcionando una comprensión más profunda de los determinantes de la mortalidad y permitiendo una mejor toma de decisiones en políticas de salud y prevención.

VI. REFERENCIAS

[1] GOV.CO DATOS ABIERTOS, «CANTIDAD DEFUNCIONES POR NIVEL EDUCATIVO 2018 - 2022,» 8 Mayo 2023. [En línea]. Available: https://www.datos.gov.co/Salud-y-Protecci-n-Social/CANTIDAD-DEFUNCIONES -POR-NIVEL-EDUCATIVO-2018-2022/vq8d-mtti. [Último acceso: 29 Septiembre 2023].