****Inteligencia de negocios  
5A. Componentes de BI para analítica predictiva  
Por David Aarón Ramírez Olmeda  
Programa: Maestría en Ciencia de Datos e Información**

**Introducción General**

El presente análisis busca abordar la problemática de los accidentes viales en México, utilizando técnicas avanzadas de análisis de datos para identificar patrones y tendencias que puedan ser útiles en la toma de decisiones. Dado que la información disponible abarca un periodo mensual y está desagregada por municipio, nuestra finalidad es segmentar los datos de manera que podamos obtener una visión más clara de los factores que influyen en la ocurrencia de accidentes en distintas zonas geográficas y momentos del tiempo. Este enfoque nos permitirá entender si existe una segmentación clara y recurrente en la forma en que se presentan los accidentes viales.

***Definición del análisis y técnica utilizada***

El objetivo principal de este estudio es realizar una segmentación de los datos de accidentes viales para identificar patrones que nos permitan categorizar municipios de México de acuerdo con las características asociadas a la incidencia de accidentes. La información utilizada en el análisis incluye variables como la **densidad poblacional, número de accidentes, heridos, muertos, condiciones climáticas, etc.** Estas variables están desglosadas a nivel mensual y por cada municipio del país, proporcionando un amplio panorama que nos ayudará a identificar comportamientos específicos en cada región.

Para lograr este objetivo, utilizaremos el algoritmo de **K-means**, una técnica de **aprendizaje no supervisado** que permite agrupar los datos en clusters o segmentos según su similitud. Optamos por esta técnica porque resulta adecuada cuando no se tiene una categorización previa de los datos, y queremos descubrir patrones ocultos o estructuras en la información. En el caso de nuestro proyecto, que tiene un enfoque social, nos interesa entender si existen agrupaciones naturales en la forma en que ocurren los accidentes viales. Identificar estos clusters podría ser de gran utilidad para desarrollar estrategias de prevención más específicas y focalizadas, así como para asignar recursos de manera más eficiente.

La segmentación nos permitirá, entre otras cosas, observar si los accidentes siguen patrones estacionales, si ciertas regiones presentan características comunes en los meses de mayor incidencia, o si hay municipios que destacan por factores climáticos o demográficos en relación con los accidentes. Estos insights pueden traducirse en políticas públicas más efectivas.

***Definición de Casos de Uso***

A partir de la información disponible, podemos derivar diversos casos de uso que serían de gran utilidad para el proyecto. Algunos de los más relevantes incluyen:

1. **Identificación de zonas de riesgo**: Con la segmentación de los municipios, podemos identificar aquellos que presentan una alta concentración de accidentes en ciertos periodos del año. Esto permitirá a las autoridades y organizaciones civiles implementar campañas preventivas o mejoras en la infraestructura de tránsito en dichas regiones.
2. **Análisis de factores climáticos**: Al incluir variables como la temperatura media y la precipitación, podemos analizar si existen relaciones entre las condiciones climáticas y la frecuencia de accidentes. Este análisis puede ser crucial para el diseño de políticas de seguridad vial adaptadas a diferentes condiciones meteorológicas.
3. **Optimización de recursos**: Los municipios con un alto índice de accidentes podrían requerir mayor inversión en infraestructura, señalización, o campañas de concienciación. La segmentación nos permitirá focalizar estos esfuerzos en las zonas más vulnerables.
4. **Detección de patrones temporales**: Al trabajar con datos mensuales, es posible identificar si ciertos meses presentan más accidentes debido a factores estacionales, como temporadas vacacionales o eventos específicos. Esto facilitará la creación de campañas preventivas que se alineen con los picos de accidentes.

***Abstracción de datos analizados***

Los datos que serán analizados incluyen una combinación de variables demográficas, geográficas y climáticas que nos permitirán observar patrones a nivel municipal. Estas variables incluyen:

* **Densidad poblacional** por municipio
* **Número de accidentes** registrados por mes
* **Total de muertos y heridos**
* **Condiciones del pavimento y urbanización**
* **Condiciones climáticas** como temperatura media y precipitación media

Al segmentar estos datos utilizando el algoritmo de K-means, esperamos identificar grupos de municipios que compartan características similares en términos de accidentes viales, lo que permitirá una mejor comprensión de los factores que inciden en la siniestralidad.

***Tipología de aprendizaje automático***

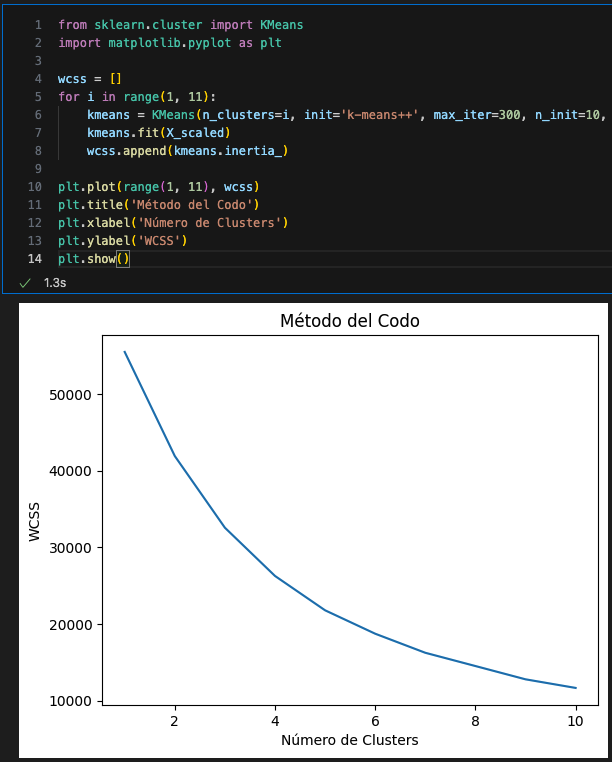
La información con la que estamos trabajando se corresponde con un escenario de **aprendizaje no supervisado**. Esto es porque no contamos con etiquetas predefinidas o categorías claras en las que clasificar los datos desde el inicio, y nuestro objetivo es descubrir patrones ocultos o agrupaciones naturales en los datos. Utilizamos el algoritmo de **K-means** específicamente por su capacidad para agrupar datos en función de sus características y encontrar segmentaciones que no habrían sido evidentes con otro tipo de técnicas.

**Desarrollo**

### Definición de Clusters

Para determinar la cantidad óptima de clusters, implementamos el método del codo. Este análisis sugirió que la mejor opción era dividir los datos en 4 clusters. Por otro lado, el coeficiente de silueta nos proporcionó las siguientes métricas:

* **2 clusters**: 0.834
* **3 clusters**: 0.393
* **4 clusters**: 0.401 (elección final)
* Otros valores de clusters no ofrecieron mejoras significativas en el coeficiente.



El puntaje promedio de silueta final fue **0.401** para los 4 clusters seleccionados, lo cual indica una moderada cohesión y separación entre los grupos formados.

### Características Analizadas

Se seleccionaron varias características clave para realizar la segmentación, entre las cuales destacan:

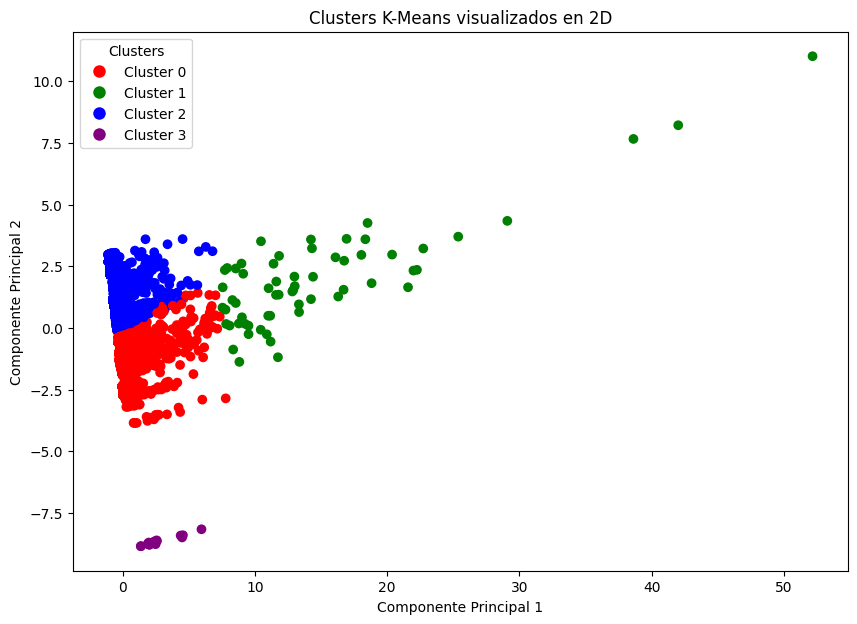
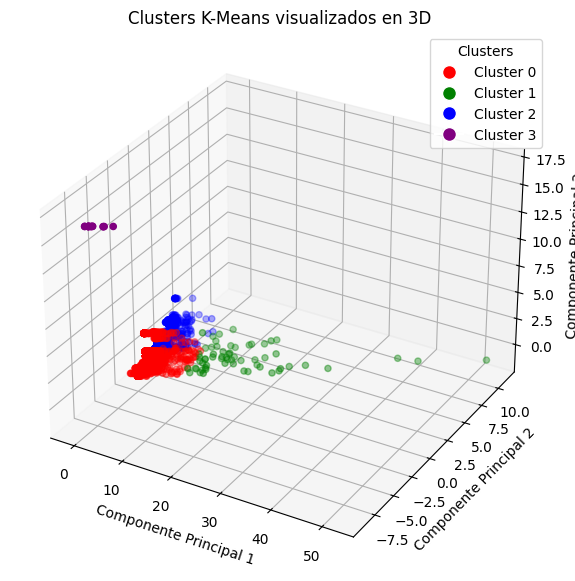
* **Densidad poblacional**
* **Incidencia**
* **Total de muertos**
* **Total de heridos**
* **Sumatoria de pavimento**
* **Sumatoria urbana**
* **Temperatura media**
* **Precipitación media**

### Distribución de Clusters

Después de realizar el análisis, los tamaños de los clusters resultaron en lo siguiente:

* Cluster 0: 2,916 instancias
* Cluster 1: 61 instancias
* Cluster 2: 3,944 instancias
* Cluster 3: 16 instancias

Los clusters más grandes, el 0 y el 2, mostraron una clara diferenciación en las visualizaciones tanto en **2D** como en **3D**. Aunque estos dos clusters están cercanos, logramos segmentarlos utilizando **PCA** para la reducción de dimensionalidad, donde los clusters 1 y 3 se separan de forma más clara.

### Valores Promedio por Cluster

Aquí mostramos los valores promedio de cada característica en su forma original para los cuatro clusters:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Característica** | **Cluster 0** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
| Densidad Poblacional | 82.58 | 53.41 | 35.04 | 6163.30 |
| Incidencia | 0.318 | 0.296 | 0.211 | 0.533 |
| Total de muertos | 2.14 | 45.14 | 0.81 | 9.38 |
| Total de heridos | 31.29 | 1222.00 | 19.98 | 100.19 |
| Sumatoria pavimentada | 136.72 | 6057.44 | 55.11 | 409.19 |
| Sumatoria urbana | 124.25 | 5988.79 | 47.60 | 406.94 |
| Temperatura media | 20.37 | 21.61 | 24.12 | 18.30 |
| Precipitación media | 784.76 | 652.47 | 1265.47 | 469.70 |

Para tener una visión más clara de la comparación entre clusters, también normalizamos los valores. Las siguientes son las medias escaladas por cluster:

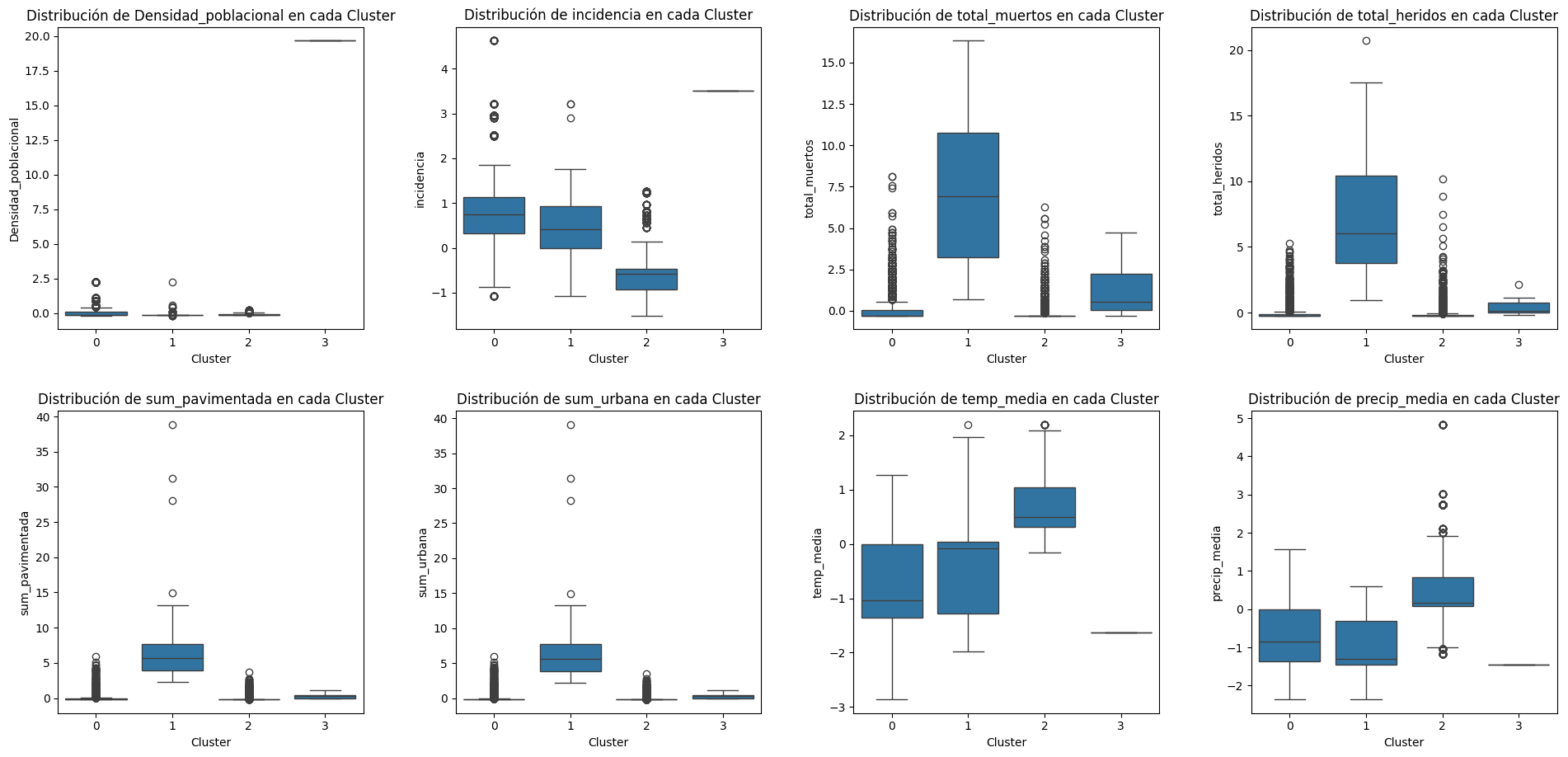
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Característica** | **Cluster 0** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
| Densidad Poblacional | 0.043 | -0.051 | -0.111 | 19.662 |
| Incidencia | 0.773 | 0.492 | -0.593 | 3.516 |
| Total de muertos | 0.060 | 7.279 | -0.162 | 1.275 |
| Total de heridos | -0.027 | 7.695 | -0.101 | 0.420 |
| Sumatoria pavimentada | -0.008 | 7.326 | -0.109 | 0.330 |
| Sumatoria urbana | -0.011 | 7.296 | -0.106 | 0.341 |
| Temperatura media | -0.825 | -0.344 | 0.622 | -1.623 |
| Precipitación media | -0.669 | -0.995 | 0.516 | -1.445 |

### Los resultados muestran que el Cluster 3, con alta densidad poblacional, presenta la mayor incidencia de accidentes, destacando la relación entre urbanización y accidentalidad. Por otro lado, el Cluster 1 sorprende al tener accidentes más graves, pese a su baja densidad, lo que sugiere problemas relacionados con la infraestructura vial. Las condiciones climáticas también parecen influir, ya que el Cluster 2, con mayor precipitación, muestra menor incidencia, posiblemente debido a una menor actividad vial en esos municipios.

### Este análisis destaca la posibilidad de ajustar las políticas de prevención y recursos según las características específicas de cada cluster.

### Visualización de Resultados

En la siguiente imagen, se observan las distribuciones de las variables seleccionadas dentro de cada cluster. Podemos notar que los clusters 1 y 3 presentan características más extremas, mientras que los clusters 0 y 2, aunque cercanos, están claramente diferenciados en algunas de las métricas.



La segmentación realizada muestra cómo los diferentes clusters representan grupos con características bien definidas, especialmente en términos de densidad poblacional, incidencia y variables climáticas. El uso de PCA permitió visualizar de manera clara la separación de los grupos, destacando que los clusters más grandes (0 y 2) requieren un análisis más detallado debido a sus similitudes.\

### *Acercamientos vanguardistas a situaciones similares (Estado del arte)*

**Diversos autores han abordado la predicción de accidentes y la severidad de lesiones utilizando técnicas avanzadas de Machine Learning y Deep Learning, aplicando enfoques novedosos para mejorar la precisión y segmentación de los datos. A continuación, se presentan algunos estudios relevantes:**

**Pourroostaei Ardakani (2023)**: En su artículo "Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis", los autores utilizaron técnicas de ML como Random Forest y Gradient Boosting para predecir accidentes automovilísticos. Concluyeron que la combinación de datos demográficos, condiciones climáticas y características de las vías mejora significativamente la precisión de los modelos de predicción de accidentes, lo que resalta la importancia de segmentar datos geográficos y temporales.

**Zhengjing Ma (2021)**: En su estudio "An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors", los autores desarrollaron un marco basado en **redes neuronales profundas (DNN)** para predecir la severidad de las lesiones en accidentes de tráfico. El uso de factores como la velocidad y las condiciones climáticas mejoró la precisión del modelo en comparación con otros métodos más tradicionales.

**Shangguan (2020)**: En el artículo "Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique", los autores aplicaron técnicas de ML en combinación con **clustering** para predecir la severidad de las lesiones en accidentes de tráfico. Este enfoque híbrido permitió una mejor identificación de grupos de accidentes similares, optimizando la predicción de resultados severos. El estudio muestra cómo la combinación de algoritmos de aprendizaje automático con técnicas de agrupamiento mejora la precisión al abordar problemas complejos.

En comparación con estos estudios, nuestro enfoque utiliza K-Means para segmentar los datos de accidentes a nivel mensual y por municipio en México, identificando patrones y comportamientos que permiten inferir relaciones entre la incidencia de accidentes y las características del entorno. Aunque empleamos clustering de manera similar a Shangguan, la diferencia clave radica en el contexto y la granularidad de los datos, donde nuestro interés principal es detectar segmentaciones que puedan tener implicaciones sociales y de políticas públicas a nivel local.

<https://colab.research.google.com/drive/1JcdhuqQ8m0rJjcRWGAhk63Az3kY1HtZW?usp=sharing>

Bibliografía

Pourroostaei Ardakani S, Liang X, Mengistu KT, So RS, Wei X, He B, Cheshmehzangi A. Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. Sustainability. 2023; 15(7):5939. doi:10.3390/su15075939.

Ma Z, Mei G, Cuomo S. An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. Accident Analysis & Prevention. 2021; 156:106178. doi:10.1016/j.aap.2021.106178.

Shangguan Z, Wu C, Xie K, Zhu H, Zheng J, Gan Z. Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol. Int J Environ Res Public Health. 2020; 17(15):5497. doi:10.3390/ijerph17155497. PMCID: PMC7432564.