MCDI
MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

“Segmentación de Incidencia Delictiva en Municipios de la República Mexicana”

(TIPO DE PROYECTO)
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Pablo Gómez García

Asesor:

(Nombre y Apellidos)

Ciudad de México, Mayo, 2023.

CONACYT

INFOTEC POSGRADOS

Autorización de impresión

Agradecimientos

Tabla de contenido

Tabla de contenido 5

Introducción 1

Capítulo 1. Planteamiento de la Problemática 3

1.1 Objetivo 4

1.2 Alcance de la solución 4

Capítulo 2. Técnicas de modelación no supervisadas 2

2.1 K-Means 3

2.2 Agrupamiento Jerárquico 3

2.3 DBSCAN 4

2.4 Modelos de mezclas gaussianas 5

Capítulo 3. Metodología Aplicada al Problema 2

3.1 Procesamiento de los Datos 2

3.2 Análisis Exploratorio 2

3.3 Reducción de la Dimensionalidad 3

3.4 Modelado de los datos 5

Capítulo 4. Resultados y Conclusiones 2

Conclusiones 13

Bibliografía 15

ANEXO 1 15

Índice de términos 16

Índice de figuras

Índice de gráficos

Índice de cuadros

Siglas y abreviaturas

SIGLAS	Descripción
SIGLAS	Descripción
SIGLAS	Descripción
SIGLAS	Descripción
SIGLAS	Descripción
SIGLAS	Descripción

Glosario

“A”

Aenean: Sollicitudin sem lorem, nec tristique lacus auctor in. Etiam luctus purus non dui fringilla tempor. Suspendisse euismod urna id nisl cursus, nec tincidunt lacus sagittis. Sed mollis sem mattis ligula rutrum scelerisque. Ut mattis condimentum blandit. Curabitur ipsum mauris, ullamcorper non accumsan id, eleifend id velit. Nunc at semper massa, sit amet pellentesque metus. Pellentesque pulvinar eget mauris sit amet dapibus. Integer vel lorem ut sem pretium semper vel at est. Aenean vitae varius libero. Sed accumsan nisl eu nulla consectetur fringill.

“B”

Blandit: Aenean laoreet ligula massa, ut varius lorem convallis ut. Integer at libero neque. Donec vestibulum neque in efficitur iaculis. Donec eros eros, porta suscipit auctor et, rutrum ut tortor.

“C”

Congue: Quam nibh convallis mauris, et tincidunt massa dolor maximus mauris. Fusce pretium lectus vitae aliquet aliquet. Fusce elit ligula, cursus eu velit eu, maximus tempus lorem.

Introducción

En la actualidad existen en México diversos estudios a nivel municipio, la gran mayoría realizados por el Instituto Nacional de Estadística y Geografía (INEGI), dichos estudios muestran características poblacionales de cada uno de ellos, también muestran estados de opinión, preferencias electorales, índices de Pobreza y Marginación etc.

Toda la información que se tiene acerca de los municipios es valiosa para distintos fines, sin embargo a la fecha hay muy pocos estudios que describen a los municipios por medio de varias de sus características, si es que existen tales estudios, lo hacen a nivel descriptivo simple.

Dentro de este proyecto planteamos la posibilidad de aplicar una técnica no supervisada para encontrar una segmentación muy característica de los municipios del País. Para este proyecto delimitaremos el tema únicamente considerando variables referentes a la incidencia delictiva en los municipios.

Tomando como punto de partida la información pública acerca de la incidencia delictiva a nivel municipio, plateamos la siguiente Hipótesis:
"*Existen variables entre los datos públicos de los municipios referentes a incidencia delictiva que permiten diferenciarlos en segmentos cuyos elementos que los conforman poseen características homogéneas entre sí pero heterogéneas con los elementos de otros segmentos*".
Bajo el supuesto anterior la propuesta de este trabajo consiste en tomar la información pública de fuentes como el INEGI y la PGR referente a la incidencia delictiva a nivel municipio para construir un Modelo no supervisado con el que se pueda encontrar a los segmentos planteados en la hipótesis.

El importante enfatizar que el modelo propuesto no tiene como objetivo clasificar municipios que sean creados en el futuro, debido a que la creación de municipios no es algo que sucede a menudo, sin embargo busca mostrar de manera descriptiva el valor y aporte de cada segmento encontrado por el modelo.

Finalmente este trabajo muestra una serie de sugerencias sobre la aplicación de los resultados obtenidos.

Conceptos clave

- Definición de segmentación: La segmentación, en el ámbito general, podría definirse como el proceso de dividir un conjunto de elementos o datos en grupos o segmentos más pequeños y homogéneos, basados en características o propiedades comunes. El objetivo de la segmentación es identificar y comprender las similitudes y diferencias entre los elementos, lo

que permite una mejor comprensión de la estructura y la variabilidad de los datos.

- Incidencia delictiva: La incidencia delictiva se refiere a la medida de la frecuencia y el tipo de delitos que ocurren en un determinado período de tiempo y en una determinada área geográfica. Generalmente, se utiliza para describir la cantidad y la variedad de delitos cometidos en una comunidad, ciudad, región o país.
- Factores de riesgo: Los factores de riesgo son variables o condiciones que aumentan la probabilidad de que ocurra un evento adverso o problemático. En el contexto de la delincuencia o la seguridad, los factores de riesgo son aquellos elementos o características que están asociados con un mayor riesgo de que se produzcan actos delictivos o comportamientos antisociales.

Casos de estudio y aplicaciones

Realizando una investigación al respecto sobre temas relacionados encontramos algunos estudios que se han hecho a nivel municipio.

El primer artículo encontrado fue "***Identificación de clústeres en la Zona Metropolitana de Guadalajara: restaurantes [1]".***

Este artículo se basa en identificar la metodología de clusterización más apropiada para aplicarse en el sector restaurantero de la Zona Metropolitana de Guadalajara. Lo que se hizo en esta investigación fue llevar a cabo un recuento de las distintas técnicas de clusterización espacial, para después identificar que la más conveniente es la técnica de Kulldorff, la cual fue utilizada para mapear los clústeres de los restaurantes existentes en la metrópoli. Los resultados muestran diez clústeres de restaurantes en la Zona Metropolitana de Guadalajara, siete de ellos con alta concentración de unidades económicas. El presente estudio es innovador respecto a la detección de clústeres en la industria restaurantera de la Zona Metropolitana de Guadalajara.

Otra investigación encontrada es "***Distribución espacial de un índice de creatividad a nivel municipal en México [2]***".

La investigación antes mencionada aborda una discusión sobre la creatividad en México mediante la estimación de un índice a nivel municipal que integra variables relacionadas con el Talento, la Tolerancia y la Tecnología, elementos que de acuerdo con [3], las regiones deben poseer para que sean más competitivas. El análisis empírico combina técnicas de componentes principales junto con un análisis exploratorio de datos espaciales con el fin de obtener un ranking del Índice de Creatividad, conocer la distribución espacial del mismo y mostrar los municipios mejor posicionados. Los resultados señalan que el Índice de Creatividad se

concentra en los grandes centros urbanos de México (Ciudad de México, Guadalajara y Monterrey) y que existe una fuerte relación entre el capital humano y la tecnología, mas no entre la clase bohemia y la clase creativa, por lo que se llegan a resultados parciales en comparación con los trabajos realizados por [3] en otros contextos.

La tercer investigación encontrada es "***Histéresis y asimetría en delitos: un análisis de los robos a nivel colonia en la Zona Metropolitana de Guadalajara [4]***".

La investigación plantea que una de las características de la delincuencia que ha sido subestimada en estudios empíricos es su persistencia temporal y, por consiguiente, su respuesta asimétrica a cambios en sus variables explicativas. La importancia del efecto histéresis ha sido notada en diferentes estudios, en los que se argumenta que éste puede reducir de manera significativa la efectividad de las políticas de lucha contra el crimen. Utilizando las denuncias de diferentes tipos de robos a nivel colonia en la Zona Metropolitana de Guadalajara, encontramos que dichos delitos presentan un fuerte componente de histéresis, y muestran un comportamiento asimétrico ante cambios en las condiciones económicas.

Capítulo 1

Planteamiento de la Problemática

Capítulo 1. Planteamiento de la Problemática

Como podemos ver en nuestro día a día, la información se genera en cantidades desproporcionadas por todos lados, se genera a través de las redes sociales, los sistemas de información tales como los sistemas bancarios, aeropuertos, los sistemas de cobro e inventarios de los centros comerciales por mencionar algunos. El gran reto de la ciencia de datos es dotar de utilidad para la toma de decisiones a dicha información.

Debe tenerse claro que la información por sí sola, no aporta más que un panorama muy general de lo que existe en todos esos sistemas. Por ello y para poder dar una propuesta de valor basada en el conocimiento se debe traducir dicha información por medio de la inteligencia a aplicaciones direccionadas a dar solución a un problema de negocio.

Derivado de lo anterior y siguiendo la misma directriz vamos a asumir que nuestro sistema de información es el Instituto Nacional de Estadística y Geografía (INEGI), que de acuerdo con lo citado en su página[1] se trata de "*Un organismo público autónomo responsable de normar y coordinar el Sistema Nacional de Información Estadística y Geográfica, así como de captar y difundir información de México en cuanto al territorio, los recursos, la población y economía, que permita dar a conocer las características de nuestro país y ayudar a la toma de decisiones*". Hasta este punto sabemos que el INEGI provee de mucha información referente a temas del País México, particularmente a nivel de municipios. Ahora bien, esta información por si sola carece de inteligencia (definiremos el termino inteligencia como la capacidad de predecir eventos futuros a partir de información observada) por lo que no podría dar una respuesta concreta a una pregunta de negocio con enfoque predictivo.
Finalmente podemos apreciar que el problema se centra en poder tomar la información que existe y transformarla, combinarla y modelarla de una manera que pueda dar respuesta a la pregunta de negocio que planteamos en la introducción.

De una Manera más especifica la nuestra pregunta de negocio es la siguiente:

¿Existen variables entre los datos públicos sobre incidencia delictiva de los municipios que permiten diferenciarlos en segmentos cuyos elementos que los conforman poseen características homogéneas entre sí pero heterogéneas con los elementos de otros segmentos?

¿Por qué esta pregunta de negocio? La respuesta es porque se busca llevar más allá esta información, aportando un análisis más profundo que combine diversos

[1] https://www.inegi.org.mx/inegi/quienes_somos.html

tipos de datos acerca de la incidencia delictiva de los municipios de la República Mexicana con la finalidad de aportar conocimiento e inteligencia para la toma de decisiones.

Generalmente la información proporcionada por el INEGI y de más organismos que proporcionan información referente al país y a nivel regional, se usan para el lanzamiento de propuestas en beneficio de la población, tales como programas de alimentación, programas en beneficio del campo entre otras acciones más. Justo por ello la propuesta que se plantea en este trabajo busca ir más allá y por ello plantemos el objetivo del proyecto como se plantea en el siguiente apartado.

1.1 Objetivo

Plantear una solución desde un enfoque Analítico basada en un Modelo no supervisado que logre segmentar a los municipios de la República Mexicana. Se pretende que dicha segmentación resulte de utilidad para estrategias de implementación de Políticas Públicas de Seguridad.

1.2 Alcance de la solución

El alcance de esta solución depende de la calidad de la información, sin embargo depende de otros factores como la cultura de denuncia y la Histéresis y asimetría en delitos mencionada en [4]. Aun con lo anterior se pueden obtener los siguientes beneficios:

- Identificación de patrones: La segmentación permite identificar patrones y tendencias específicas en la incidencia delictiva. Al agrupar los datos en segmentos más pequeños y homogéneos, es posible detectar áreas geográficas, períodos de tiempo o tipos de delitos que presentan características similares y pueden requerir enfoques de prevención y respuesta específicos.

- Enfoque en áreas de alto riesgo: La segmentación ayuda a identificar áreas geográficas que presentan una incidencia delictiva más alta o concentrada. Esto permite una asignación más eficiente de recursos y esfuerzos de seguridad para abordar los problemas en áreas específicas y reducir el riesgo de delitos en esas zonas.

- Personalización de estrategias de prevención: Al comprender mejor los diferentes segmentos de la incidencia delictiva, es posible adaptar las estrategias de prevención y respuesta de acuerdo con las características y necesidades específicas de cada grupo. Esto puede incluir programas de intervención dirigidos a poblaciones en riesgo, campañas de concientización focalizadas o medidas de seguridad específicas para ciertos tipos de delitos.

- Evaluación de políticas y programas: La segmentación permite una evaluación más precisa y detallada de la efectividad de las políticas y programas de prevención y respuesta delictiva. Al analizar los resultados en función de los diferentes segmentos, es posible determinar qué enfoques son más exitosos en la reducción de la incidencia delictiva en cada grupo y

- Mejor comprensión de las causas subyacentes: La segmentación puede ayudar a identificar factores de riesgo específicos asociados con cada segmento de la incidencia delictiva. Esto proporciona información más detallada sobre las causas subyacentes de los delitos y puede contribuir a la comprensión de los factores sociales, económicos, culturales o individuales que contribuyen a la delincuencia en cada grupo. ajustar las intervenciones en consecuencia.

Capítulo 2

Técnicas de modelación no supervisadas

Capítulo 2. Técnicas de modelación no supervisadas

De acuerdo con [9], [10] y [11] las técnicas de modelación no supervisadas, son métodos utilizados en el campo del aprendizaje automático y la minería de datos para descubrir patrones, estructuras y relaciones ocultas en conjuntos de datos sin la necesidad de etiquetas o categorías predefinidas. A diferencia de las técnicas de modelación supervisadas, en las que se cuenta con datos de entrada y salidas esperadas para entrenar un modelo, las técnicas de modelación no supervisadas se aplican a conjuntos de datos no etiquetados, donde el objetivo principal es explorar y extraer información valiosa sin tener una guía externa. Algunas de las técnicas de modelación no supervisadas más comunes incluyen:

- Clustering (agrupamiento): Es el proceso de agrupar datos similares en grupos o clústeres basados en la similitud de sus características. El objetivo es encontrar estructuras o patrones intrínsecos en los datos sin conocer de antemano las categorías a las que pertenecen.
- Reducción de dimensionalidad: Estas técnicas se utilizan para reducir la cantidad de variables o dimensiones en un conjunto de datos, preservando al mismo tiempo la información relevante. Esto ayuda a visualizar y comprender mejor los datos al reducir la complejidad y el ruido.
- Reglas de asociación: Se utilizan para descubrir relaciones y asociaciones interesantes entre diferentes variables o elementos en un conjunto de datos. Estas técnicas encuentran patrones comunes o secuencias frecuentes que pueden ser útiles en áreas como el análisis de mercado o la recomendación de productos.
- Detección de anomalías: Se enfoca en identificar observaciones o eventos inusuales o atípicos en un conjunto de datos. Estas técnicas buscan identificar patrones que difieren significativamente del comportamiento normal, lo que puede ser útil en la detección de fraudes o problemas de seguridad.

Las técnicas de modelación no supervisadas son útiles para explorar y analizar grandes volúmenes de datos sin la necesidad de etiquetas o categorías predefinidas. Proporcionan una forma de descubrir patrones ocultos, estructuras subyacentes y relaciones interesantes en los datos, lo que puede conducir a una mejor comprensión y toma de decisiones en diversas áreas de aplicación.

2.1 K-Means

De acuerdo con lo visto en [10] y [11] el método de K-Means es un algoritmo de clustering ampliamente utilizado que agrupa un conjunto de datos en K grupos o clústeres, donde K es un número predefinido. El objetivo del algoritmo es asignar cada dato al clúster más cercano en función de la distancia euclidiana entre ellos, minimizando la varianza intra-cluster.

El algoritmo de K-Means sigue los siguientes pasos:

1. Inicialización: Se seleccionan aleatoriamente K centroides, que representan los centros iniciales de los clústeres.
2. Asignación: Cada dato se asigna al clúster cuyo centroide está más cerca, utilizando la distancia euclidiana como medida de proximidad.
3. Actualización del centroide: Los centroides de los clústeres se recalculan como el centroide de todos los datos asignados a ese clúster.
4. Iteración: Los pasos de asignación y actualización del centroide se repiten hasta que los centroides convergen y no se producen cambios significativos.

El algoritmo de K-Means puede variar en función de la estrategia de inicialización y del criterio de convergencia utilizado. También se pueden aplicar técnicas para mejorar la convergencia y evitar los mínimos locales, como ejecutar el algoritmo varias veces con diferentes inicializaciones y seleccionar el mejor resultado.

2.2 Agrupamiento Jerárquico

Basado en el trabajo de [8], [10] y [13] podemos definir que el método de agrupamiento jerárquico es una técnica de clustering que organiza los datos en una estructura jerárquica de clústeres, formando un árbol o dendrograma. El objetivo es agrupar los datos en función de su similitud o distancia, de manera que los elementos más similares se agrupen en clústeres cercanos y los elementos menos similares se encuentren en clústeres más distantes.

Existen dos enfoques principales para el agrupamiento jerárquico:

1. Agrupamiento jerárquico aglomerativo (bottom-up): Comienza considerando cada punto de datos como un clúster individual y luego fusiona iterativamente los clústeres más cercanos hasta que todos los puntos de datos estén agrupados en un único clúster.
2. Agrupamiento jerárquico divisivo (top-down): Comienza considerando todos los puntos de datos como un único clúster y luego divide iterativamente los clústeres en subclústeres más pequeños hasta que cada punto de datos forme su propio clúster individual.

El resultado final del agrupamiento jerárquico es un dendrograma que muestra la estructura de los clústeres y las relaciones de similitud entre ellos. Este dendrograma se puede utilizar para determinar el número óptimo de clústeres y para visualizar los diferentes niveles de agrupamiento.

2.3 DBSCAN

De acuerdo con [8] y [14] el método DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de clustering que se basa en la densidad de los puntos de datos para agruparlos. A diferencia de otros algoritmos de clustering, como K-Means, DBSCAN no requiere que el número de clústeres sea especificado de antemano. En cambio, puede descubrir automáticamente el número de clústeres en función de la distribución de densidad de los datos.

El algoritmo DBSCAN define los siguientes elementos clave:

1. Puntos centrales (core points): Son aquellos puntos que tienen un número mínimo de puntos vecinos dentro de un radio especificado (eps).
2. Vecinos directos (directly reachable points): Son puntos que se encuentran dentro del radio especificado (eps) de otro punto central.
3. Puntos de borde (border points): Son puntos que tienen menos vecinos dentro del radio especificado (eps) y no son considerados centrales, pero están dentro del radio de vecindad de un punto central.
4. Puntos de ruido (noise points): Son aquellos puntos que no son centrales, no tienen vecinos directos y están fuera del radio especificado (eps) de cualquier punto central.

El algoritmo DBSCAN se ejecuta de la siguiente manera:

- Se selecciona un punto de datos aleatorio que no ha sido visitado.
- Si el punto seleccionado es un punto central, se forma un nuevo clúster y se expande visitando todos los puntos directamente alcanzables desde ese punto.
- Si el punto seleccionado es un punto de borde, se asigna al mismo clúster que su punto central correspondiente.
- Se repiten los pasos anteriores hasta que todos los puntos hayan sido visitados.

El resultado final del algoritmo DBSCAN es un conjunto de clústeres, donde cada clúster consiste en puntos de datos densamente conectados, y puntos que no están en ningún clúster se consideran ruido.

2.4 Modelos de mezclas gaussianas

Tomando en consideración lo planteado en [5] en el Capítulo 9: Mixture Models and EM y lo mencionado en [15] y [16] el método de Modelos de Mezclas Gaussianas (Gaussian Mixture Models, GMM) es una técnica de aprendizaje automático no supervisado utilizada para la estimación de densidad y clustering.

En el contexto de clustering, los GMMs se utilizan para modelar la distribución de los datos y asignar puntos a diferentes clústeres.

En resumen, un GMM asume que los datos se generan a partir de una combinación de componentes Gaussianos (distribuciones Gaussianas). Cada componente representa un clúster en el conjunto de datos y se caracteriza por su media y matriz de covarianza. La probabilidad de que un punto de datos pertenezca a cada componente se determina mediante la estimación de la probabilidad a posteriori utilizando el algoritmo Expectation-Maximization (EM). El algoritmo EM se utiliza para ajustar los parámetros del GMM de manera iterativa, maximizando la verosimilitud de los datos observados.

Capítulo 3
Metodología Aplicada al Problema

Capítulo 3. Metodología Aplicada al Problema

A continuación se presenta la metodología aplicada al proyecto. Los datos utilizados pueden descargarse directamente de la página de la PGR[2].

3.1 Procesamiento de los Datos

El procesamiento de los datos consistió en la transposición de los datos pues la versión original viene con la información hacía abajo. Se realizo un procesamiento para obtener datos acumulados de delitos. Tal como se muestra en la imagen 3.1.

id	Aborto	Abuso_de_confianza	Abuso_sexual	Acoso_sexual	Allanamiento_de_morada	Amenazas	Contra_el_medio_ambiente	Corrupcion_de_menores
1010001	0.000000	0.291667	0.177083	0.031250	0.010417	0.479167	0.000000	0.000000
1010002	0.000000	0.000000	0.010417	0.000000	0.000000	0.062500	0.000000	0.000000
1010003	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333	0.000000	0.000000
1010004	0.000000	0.468750	0.333333	0.041667	0.062500	0.947917	0.000000	0.000000
1010005	0.125000	34.020833	17.354167	1.895833	8.583333	66.093750	0.041667	0.989583
...	...	...	...	...	...	...	...	...
99014	0.281250	34.770833	10.406250	3.895833	4.781250	61.072917	3.989583	0.760417
99015	0.552083	66.395833	34.760417	11.427083	6.125000	131.354167	6.343750	2.583333
99016	0.364583	22.166667	12.062500	4.020833	3.500000	52.333333	3.312500	0.739583
99017	0.854167	19.531250	14.541667	3.239583	3.187500	63.093750	2.291667	0.833333
99998	0.072917	0.187500	1.447917	0.604167	0.000000	0.656250	0.156250	0.239583

Imagen 3.1 Muestra el layout final después de realizar el procesamiento y transposición de los datos.

3.2 Análisis Exploratorio

Para el análisis exploratorio no se detectaron anomalías ni valores faltantes y todas las variables en su totalidad son del tipo intervalo, esto se puede observar en la imagen 3.2.

[2] https://www.gob.mx/sesnsp/acciones-y-programas/datos-abiertos-de-incidencia-delictiva

	count	mean	std	min	25%	50%	75%	max
Aborto	2479.0	0.020641	0.103366	0.0	0.00	0.00	0.010	1.97
Abuso_de_confianza	2479.0	0.824712	3.707413	0.0	0.00	0.04	0.240	66.40
Abuso_sexual	2479.0	0.676958	2.973642	0.0	0.01	0.05	0.220	50.75
Acoso_sexual	2479.0	0.134433	0.677515	0.0	0.00	0.00	0.030	15.64
Allanamiento_de_morada	2479.0	0.434885	3.573997	0.0	0.00	0.03	0.150	154.16
Amenazas	2479.0	3.179000	14.026048	0.0	0.04	0.21	0.930	217.26
Contra_el_medio_ambiente	2479.0	0.060803	0.398351	0.0	0.00	0.00	0.010	9.18
Corrupcion_de_menores	2479.0	0.071154	0.512703	0.0	0.00	0.00	0.020	16.99
Dano_a_la_propiedad	2479.0	4.244078	18.110651	0.0	0.06	0.29	1.445	308.27
por_servidores_publicos	2479.0	0.588495	3.766842	0.0	0.01	0.04	0.130	104.36
Despojo	2479.0	0.874296	3.068362	0.0	0.03	0.14	0.440	45.21
Electorales	2479.0	0.043667	0.163427	0.0	0.00	0.01	0.030	2.70
Evasion_de_presos	2479.0	0.003780	0.020601	0.0	0.00	0.00	0.000	0.41
Extorsion	2479.0	0.243852	1.074441	0.0	0.00	0.01	0.080	21.84
Falsedad	2479.0	0.104558	0.753174	0.0	0.00	0.00	0.020	27.36
Falsificacion	2479.0	0.543647	3.503101	0.0	0.00	0.01	0.070	83.49
Feminicidio	2479.0	0.006523	0.020495	0.0	0.00	0.00	0.010	0.36
Fraude	2479.0	2.399137	12.807845	0.0	0.02	0.10	0.570	294.99
Homicidio	2479.0	0.148471	0.548769	0.0	0.01	0.03	0.090	15.85
Hostigamiento_sexual	2479.0	0.049484	0.304511	0.0	0.00	0.00	0.020	6.95
Incesto	2479.0	0.000710	0.010123	0.0	0.00	0.00	0.000	0.45
Incump_obligaciones_asistencia_fam	2479.0	0.764712	3.491686	0.0	0.00	0.03	0.230	73.43
Lesiones	2479.0	0.732283	2.920760	0.0	0.01	0.06	0.280	55.86
Narcomenudeo	2479.0	1.937608	19.184218	0.0	0.00	0.04	0.250	731.50

Imagen 3.2 Muestra el análisis descriptivo de algunas de las variables referentes a tipo de delitos.

3.3 Reducción de la Dimensionalidad

Para la reducción de la dimensionalidad se utilizó el método de clustering de variables, este consiste en realizar agrupaciones de variables y elegir una representante de cada una de ellas para ser utilizada en la modelación dejando fuera la colinealidad que existe entre las variables pertenecientes a cada cluster.

	Cluster	Variable	RS_Own	RS_NC	RS_Ratio
0	0	Abuso_de_confianza	0.830808	0.769288	0.733349
1	0	por_servidores_publicos	0.824154	0.621512	0.464602
2	0	Falsedad	0.760237	0.409843	0.406269
3	0	Falsificacion	0.822116	0.567534	0.411324
4	0	Fraude	0.909022	0.681342	0.285505
5	0	Trata_de_personas	0.770087	0.522953	0.481950
6	1	Abuso_sexual	0.872288	0.757028	0.525624
7	1	Amenazas	0.838310	0.567483	0.373835
8	1	Dano_a_la_propiedad	0.895934	0.750435	0.416989
9	1	contra_el_patrimonio	0.586850	0.414141	0.705204
10	1	libertad_seguridad_sexual	0.599838	0.477581	0.765978
11	1	Violacion_simple	0.874484	0.742551	0.487539
12	1	Violencia_familiar	0.934205	0.729318	0.243070
13	2	Otros_contra_la_familia	0.919739	0.276414	0.110921
14	2	Rapto	0.919739	0.137501	0.093056
15	3	Feminicidio	0.775939	0.627296	0.601178
16	3	libertad_personal	0.798321	0.570432	0.469494
17	3	Secuestro	0.636053	0.453582	0.666061
18	3	Aborto	0.690397	0.459194	0.572485
19	3	Despojo	0.843609	0.798878	0.777591
20	3	Evasion_de_presos	0.668282	0.515286	0.684358
21	4	Trafico_de_menores	1.000000	0.105126	0.000000
22	5	Otros_Fuero_Comun	0.889798	0.500419	0.220589

Imagen 3.3 Muestra el análisis de clustering de variables para algunos clusters y su valor de RS_ratio.

Tras el Análisis de Clustering de variables podemos tomar la variable con la distancia menor al centroide dentro de cada cluster, esto es con el valor del RS_Ratio y así elegiríamos una variable representante de cada grupo de acuerdo a su nivel de correlación. Por otro lado aquellos cluster cuyo número de variables es uno significa que son variables que no se correlacionan con ninguna otra.

El criterio tomado es que la iteración del algoritmo pare cuando se tenga el 70% de la Varianza explicada.

En este caso de 40 variables que entraron al análisis podemos explicar el 70% de la varianza del fenómeno con solo con solo 11 variables.

3.4 Modelado de los datos

Antes de proceder al modelado de los datos fue necesario realizar un proceso de estandarización a las variables. Posteriormente se probó con el método de K-means para obtener los primeros resultados de este ejercicio. Es importante mencionar que estos no son definitivos pues se seguirá probando con otros métodos.

En la imagen 3.4 se pueden observar hasta los primeros 12 clusters formados en este ejercicio.

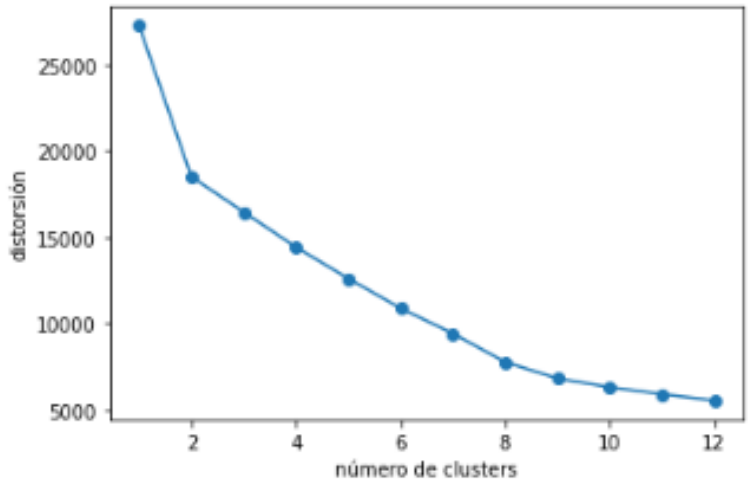

Imagen 3.4 Muestra los primeros 12 clusters formados por el método K-means.

Capítulo 4
Resultados y Conclusiones

Capítulo 4. Resultados y Conclusiones

Como un Primer ejercicio se obtienen 7 segmentos de municipios donde el cluster 1 de color azul fuerte concentra el 95% de los municipios y el otro 5% se logra diferenciar del resto.

Imagen 4.1 Muestra la distribución del promedio de delitos denunciados por cluster en la República Mexicana

En la imagen se puede apreciar que el tamaño del círculo representa el número promedio de delitos denunciados por mes, es decir, entre más grande tiene muchos más delitos.

Conclusiones

Conclusiones

Observando la caracterización de los clusters obtenidos en la siguiente imagen podemos obtener una conclusión.

cluster	Promedio de Fraude	Promedio de Violencia_familiar	Promedio de Rapto	Promedio de libertad_personal	Promedio de Trafico_de_menores	Promedio de Otros_Fuero_Comun	Promedio de Incesto
0	19.54	59.89	0.03	5.45	0.02	59.41	0.00
1	0.59	1.95	0.00	0.18	0.00	1.92	0.00
2	69.39	393.58	0.00	21.32	0.06	227.35	0.05
3	92.88	152.86	0.02	12.41	0.02	166.45	0.01
4	5.18	38.81	0.00	10.98	0.00	16.04	0.45
5	35.07	145.16	0.81	21.48	0.02	25.36	0.01
6	30.50	132.40	0.04	8.41	2.09	61.98	0.01
Total	**2.40**	**6.42**	**0.01**	**0.58**	**0.00**	**5.90**	**0.00**

cluster	Promedio de Homicidio	Promedio de Contra_el_medio_ambiente	Promedio de Corrupcion_de_menores	Otros_contra_la_sociedad	Recuento de Municipio
0	1.26	0.40	0.61	116.59	78
1	0.07	0.02	0.02	154.10	2361
2	8.88	0.53	11.50	2.33	3
3	1.97	2.42	1.52	185.60	27
4	0.13	0.02	0.19	0.10	1
5	1.71	0.01	1.51	4.35	8
6	2.25	0.03	2.18	0.79	1
Total	**0.15**	**0.06**	**0.07**	**463.88**	**2479**

Imagen 4.1 Muestra el promedio de delitos denunciados por cluster en la República Mexicana

De los 7 segmentos obtenidos se observa que hay uno con 78 municipios que se distingue por altos índices promedio de delitos por fraude, violencia familiar, fuero común y contra la sociedad.

Por otro lado hay 2 segmentos con un solo municipio, uno solo con 3, uno con 8 y uno con 27 municipios.

Como primer acercamiento a los resultados buscados de nuestro proyecto, concluimos lo siguiente:

- 7 segmentos no son suficientes para discriminar la totalidad de los municipios. Se logra diferenciar el 5% de los municipios del resto como aquellos con mayor incidencia delictiva en fraudes, secuestros, trata de personas y homicidios.
- Es necesario incorporar más variables que ayuden a discriminar mejor a los municipios.
- Detectamos un sesgo en cuanto a la cultura de la denuncia, es decir municipios que se perciben con altos índices de delitos (por ejemplo los

mencionados constantemente en las noticias azotados por el narcotráfico) no reflejan dicha característica en los datos oficiales.

Bibliografía

[1] Luquín-García, D. . ., & Fong Reynoso, C. (2022). Identificación de clústeres en la Zona Metropolitana de Guadalajara: restaurantes. Estudios Demográficos y Urbanos, 37(3), 1063–1104.

[2] Villarreal González, A., Flores Segovia, M. A., & Gasca Sánchez, F. M. (2018). Distribución espacial de un índice de creatividad a nivel municipal en México. Estudios Demográficos y Urbanos, 33(1), 149–186.

[3] Florida, Richard (2002), The rise of the creative class, Nueva York, Basic Books.

[4] Cortez Yactayo, W. W. (2017). Histéresis y asimetría en delitos: un análisis de los robos a nivel colonia en la Zona Metropolitana de Guadalajara. Estudios Demográficos y Urbanos, 32(3), 593–629.

[5] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[6] Everitt, B. S., Landau, S., & Leese, M. (2011). Cluster Analysis. Wiley.

[7] Aggarwal, C. C. (2015). Data Clustering: Algorithms and Applications. CRC Press.

[8] Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice Hall.

[9] Tan, P.-N., Steinbach, M., & Kumar, V. (2013). Introduction to Data Mining. Pearson (Capítulo 8: Cluster Analysis).

[10] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. (Capítulo 14: Unsupervised Learning)

[11] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1(14), 281-297.

[12] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[13] Kaufman, L., & Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons. (Capítulo 5: Hierarchical Clustering).

[14] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226-231.

[15] McLachlan, G., & Peel, D. (2004). Finite mixture models. John Wiley & Sons.

[16] Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, 3(1), 72-83.

ANEXOS

ANEXO 1

Índice de términos

“A”

Aliquam……12

“B”

Blandit……...3

“C”

Consectetur……7

“D”

Donec……12