



Tecnológico
de Monterrey

Tecnológico de Monterrey
Campus Monterrey
Profesional

TC2004B. Análisis de Ciencia de Datos

Grupo 201

E2: Reporte final del reto

Samantha Brito Ozuna	A01611690
Natalia Sofía Guevara Hernández	A00835884
Gael Arnulfo Ordaz Zamora	A01734114
José David Banda Rodríguez	A01285601
Mario Alberto Landa Flores	A00836172
Alejandro Adriaensens Martínez	A01285665

03 de Mayo de 2024

1. Descripción del negocio

Hoy en día, la cantidad de datos que se extraen de diversos ámbitos va creciendo exponencialmente, lo cual resulta complicado su manejo, entendimiento y análisis. Es por ello que la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es *una sobresaliente opción para abordar proyectos de análisis de datos* (Luis Miguel Arias, 2023). Esta metodología mayormente se trata de un proceso de 6 pasos a seguir, con la finalidad de conseguir el proceso más adecuado y eficiente para después aplicarlo en la toma de decisiones.

La principal razón por la que destaca es que, gracias a su comprensible estructura de 6 pasos, facilita la administración y el seguimiento del progreso del proyecto; estos pasos son: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Al mismo tiempo, ayuda a reforzar el enfoque del negocio al poner un énfasis en los objetivos y necesidades del mismo. De igual forma, su adaptabilidad para aplicarse en diversos contextos y que sea aplicable en cualquier ámbito empresarial, lo hace atractivo para las organizaciones. No obstante, no necesariamente se requiere de seguir las 6 fases en el orden establecido, ya que cualquier negocio puede ser un blanco de retrocesos si no se están obteniendo los resultados esperados y se debe de dar prioridad a la etapa con la que se está batallando.

1.1 Objetivos del Negocio

El Instituto Nacional de Estadística y Geografía (INEGI), organismo público autónomo que se encarga de crear y captar datos del país para almacenarlos e informar sobre temas relacionados a diferentes características de México, cómo información sobre el territorio, los recursos, la población y la economía (Instituto Nacional de Estadística y Geografía, 2024).

El INEGI, cuenta con grandes fuentes de datos, los cuales pueden ser benéficos para varios sectores del estado. Derivado de esto, un manejo distinto a estas bases de datos nos permitirá realizar un modelo predictivo, para poder identificar insights en los datos que no llegan a ser visibles en una primera instancia, lo cual llegaría a impactar de manera positiva en el análisis de los mismos empleando herramientas como data mining para el uso de estas grandes bases de datos de una forma distinta e innovativa.

Como criterios de éxito, se adentra a ellos la generalización del modelo, es decir, que no sea únicamente aplicable a un caso específico y sea útil para varios sectores del país; también, se espera tener

un impacto social y económico, dado que al tener acceso a una gran cantidad de datos, se espera contribuir positivamente a la comunidad mexicana a través del desarrollo de un proyecto.

1.2 Problema Actual

El problema general que se tiene es que se cuenta con cantidades masivas de datos que no se les están dando un uso, por lo que, se busca encontrar un área de oportunidad en estos recursos para resolver una problemática o proponer algo de valor para la sociedad. El problema específico que se decidió abordar es la falta de sistemas, modelos e información en general que ayude a tomar decisiones sobre dónde abrir un negocio. Al indagar por los recursos del INEGI, es notable que los datos de sus diversas encuestas y censos pueden ser usados para ofrecer apoyo al momento de tomar decisiones empresariales. Esto tendría un gran impacto en la sociedad mexicana al influir en el aumento de empleos y ayudaría a los empresarios y gobiernos tomar mejores decisiones.

El Instituto Nacional de Estadística y Geografía (INEGI) cuenta con diversos recursos para consulta que servirán como herramienta para el desarrollo del proyecto. Uno de ellos es el Marco Geoestadístico Nacional (MGN) el cual presenta la división geoestadística del territorio continental e insular en diferentes niveles de desagregación, para referir geográficamente la información estadística de los censos y encuestas institucionales y de las Unidades del Estado (INEGI, 2024). A su vez, se tiene accesibilidad al Censo de Población y Vivienda, en donde este presenta las principales características demográficas, socioeconómicas y culturales de la población mexicana; de ella misma, cabe la posibilidad de consultar esta información por manzana urbana si se desea trabajar con áreas específicas del país. Finalmente, toda la inmensa información que el INEGI posee se encuentra dentro del Directorio Estadístico Nacional de Unidades Económicas (DENU), en donde se puede descargar toda la información de un estado en específico.

Con todo lo expuesto anteriormente, el organismo público busca encontrar un valor de impacto social basado en algún aspecto de interés del equipo, del cual no se tiene restricción alguna dado que la información es de acceso público para el usuario, e inclusive está la opción de recabar información de otras fuentes de datos.

1.3 Marco teórico - Modelos de Clasificación

De acuerdo con Espinosa-Zúñiga (2020) “en julio de 2010 el Instituto Nacional de Estadística y Geografía (INEGI) puso a disposición del público el Directorio Estadístico Nacional de Unidades Económicas (DENU). Esta primera versión del DENU (generada a partir de la información recabada por los Censos Económicos 2009) proporcionaba los datos de identificación y ubicación de poco más de

cuatro millones de unidades económicas (entendiendo como tales establecimientos y empresas) de todos los sectores de la actividad económica (excepto las actividades agropecuarias y forestales) que se hallaron activos durante los censos en el territorio nacional.”

Esto ofreció un cúmulo de información útil para lo siguiente:

- a. Estudios e investigaciones académicas.
- b. Toma de decisiones para inversión y optimización de recursos.
- c. Investigaciones de mercado y desarrollo de negocios.
- d. Política de protección civil y desarrollo urbano.

Con toda esta información pública lista para el análisis se realizaron varios trabajos similares al propuesto en este proyecto, Espinosa-Zúñiga (2020) buscó obtener un modelo de segmentación geográfica de las unidades económicas del DENUE .

Mediante técnicas de segmentación (no supervisadas) se hizo la búsqueda de patrones (segmentos o clusters) en un conjunto de observaciones y donde cada cluster es homogéneo en el sentido de que agrupa observaciones con características comunes

A través de la técnica PAM (Partitioning Around Medoids) que es más robusta que k-medias. Se obtuvo un modelo de segmentación geográfica con la base pública del DENUE, aplicando la metodología CRISP-DM

Este modelo partía al país en 4 grandes clusters, de los cuales el cluster 1 abarca áreas con alto desarrollo, excepto Michoacán. El cluster 2 incluye entidades del norte con desarrollo medio-alto. Nuevo León podría pertenecer a este cluster según la validación del modelo. El cluster 3 agrupa entidades con desarrollo medio-bajo en el sur y norte, y podría subdividirse. El cluster 4 contiene entidades del centro, mayormente de desarrollo bajo, excepto el Estado de México. El modelo se basa en datos del DENUE, pero para una segmentación más precisa socioeconómica, se necesitan variables adicionales como el PIB estatal o el índice de marginación.

Este trabajo es una muestra de los resultados que se pueden obtener a partir de un conjunto de datos del INEGI por lo que proporciona de igual manera más ideas para implementar el proyecto de constituir en la efectiva evaluación de mercado desde el panorama geográfico para la instalación de un negocio físico, ya que al segmentar el país en 4 regiones dependiendo de su crecimiento económico, se puede intentar dividir a un estado/municipio/región de la misma manera para tener un mejor indicador de que zonas tienen mejor crecimiento que otras.

Un trabajo realizado por Verdejo (2016) explora la arquitectura de integración para un sistema de recomendación geográfica. Este sistema tiene como objetivo sugerir ubicaciones óptimas para establecer puntos de venta basándose en la geolocalización y los intereses de los usuarios. El trabajo propone una arquitectura escalable y mantenible, distribuyendo funcionalidades e interrelaciones entre componentes clave. Es de relevancia la manera en que resuelven el problema ya que el sistema propuesto combina sistemas de recomendación con información geográfica, como los son los datos del censo del INEGI y el DENU, además es interesante la manera en la que estructuran su modelo con una estructura modular, lo que lo hace un proyecto escalable.

Para evaluar un proyecto similar, se puede seguir un enfoque que combine análisis de datos y aprendizaje automático. Para esto es primordial que se definan claramente los objetivos, recopilar datos relevantes como demográficos y geográficos, y realizar un análisis exploratorio para comprender mejor las relaciones entre las variables. Además, de acuerdo con Milenka Linneth , se considera que la información de proyecciones de población se utiliza para la toma de decisiones , debido a la incertidumbre en los escenarios futuros, el análisis de sensibilidad es un tema trascendental en la Demografía actual. Por lo que realizar un análisis de sensibilidad es fundamental para comprender cómo pequeñas variaciones en los parámetros clave afectan las predicciones.

Para Caswell (2008) la importancia de los análisis de sensibilidad radica en su aplicación en política pública y en teoría del muestreo, ya que los parámetros más sensibles son los que deberían ser estimados de forma más precisa. Este mismo autor ha venido profundizando en el modelado de los sistemas estocásticos, en los cuales considera la "individualidad estocástica" en el ámbito demográfico. Uno de los hallazgos más importantes, ha sido llamado el "efecto mariposa" con el cual se explica el hecho de que a pequeñas variaciones en los parámetros de un modelo de predicción se observan resultados muy diferentes. Este concepto ha dado origen a la teoría del caos que trata con sistemas que se comportan de manera estable e inestable en determinados períodos.

De acuerdo de igual manera con Linneth (2012) el actual desarrollo computacional y sus aplicaciones permiten crear escenarios. De esta forma el modelaje y la simulación brindan la oportunidad de manejar la incertidumbre en el ámbito demográfico. Lo que brinda una herramienta que permite un acercamiento entre la teoría y la práctica, para acceder al estudio de estos sistemas estocásticos.

1.4 Proyecto de Análisis Predictivo

1.4.1 Objetivo del proyecto

Los objetivos fundamentales del proyecto se constituyen en la efectiva evaluación de mercado desde el panorama geográfico para la instalación de un negocio físico. Ubicar un negocio con instalaciones físicas requiere de una minuciosa investigación, especialmente para seleccionar el área que mejor favorezca una actividad económica acorde al giro del negocio, tal y como Fernandez, C. y Aqueveque, C. (2001) lo realizaron en su estudio, pues visualizaron las relaciones existentes entre las características demográficas y las creencias y motivaciones de los consumidores.

1.4.2 Hipótesis

Así, la hipótesis que estudia esta investigación es la siguiente: *¿En qué medida es posible identificar la óptima ubicación geográfica para colocar un negocio físico específico dentro de un área delimitada, como un estado nacional, para beneficio de la actividad económica del negocio utilizando análisis de ciencia de datos?*

Con el propósito de facilitar esta tarea de evaluación geográfica, el proyecto busca integrar habilidades avanzadas de análisis de datos y algoritmos de aprendizaje automático para proporcionar información de valor respecto al mejor punto de ubicación para un negocio específico. Como señala Heckman, L (2024), algunas cosas que no sabemos perfectamente, y herramientas como el machine learning o AI, no ayudan a comprender los datos, sin embargo, realmente para tomar decisiones y conocer los riesgos es necesario comenzar a hacer análisis.

Para ello, se analizarán y evaluarán bases de datos proporcionadas por la INEGI así como por otras fuentes de acceso libre. En otras palabras, el objetivo del proyecto es identificar la mejor posición geográfica para ubicar un negocio dentro de un área delimitada, como pudiera ser un estado del país.

1.4.3 Metodología CRISP-DM

- Entendimiento del negocio:**

Como ya fue presentado, se plantea trabajar junto con el Instituto Nacional de Estadística y Geografía, del cual cuya función rige en proporcionar datos confiables y relevantes de distintos ámbitos sociales, económicos y estadísticos del país, para así apoyar a la toma de decisiones de diversos ámbitos.

- Entendimiento de los datos:**

Para manejar el proyecto de manera eficiente, se piensa trabajar con el DENUE (Directorio Estadístico Nacional de Unidades Económicas) y el Censo del INEGI (Instituto Nacional de

Estadística y Geografía). Para dar uso de su contenido, primero es importante entender qué es lo que se trata de comunicar: para el primer caso se tiene que brinda información de los negocios privados y públicos del país, abarcando aspectos como su ubicación, información general y la fecha de registro, permitiendo así entender el contexto empresarial y económico en el que se encuentran las unidades económicas; por otro lado, el CENSO trata de entender el contexto social, económico y demográfico de cada manzana registrada del país, proporcionando información como la población total o características como viviendas particulares, si se cuenta con ciertos electrodomésticos, etcétera.

- **Preparación de los datos:**

Dado que estas bases de datos manejan grandes dimensiones de registros y columnas, es importante realizar una depuración precisa para no perder información relevante que ayudará a acercarse al objetivo del proyecto. En primera instancia se piensa en identificar datos faltantes y cómo manejarlos: si su presencia es o no necesaria o si se requiere de reemplazarlos con algún otro valor. Luego, priorizando el enfoque del proyecto, hay que decidir cuáles son las variables relevantes del DENU que establecen si un negocio es exitoso o no, mientras que del CENSO es importante tomar en cuenta ciertas variables demográficas.

- **Modelado:**

Como se construirá un sistema de recomendación de cierto giro de negocio en una zona donde posiblemente resulte exitoso, difícilmente se va a predecir mediante algún problema de regresión o clasificación pues no se cuenta con una columna que determine dicho apartado. Es por ello que se optó por realizar un algoritmo de aprendizaje no supervisado, específicamente de clustering, para así lograr interpretar y extraer información de los agrupamientos formados y sacar como conclusiones, la recomendación del negocio.

- **Evaluación:**

Existen diversos métodos para realizar clustering, como por ejemplo K-Means, DBSCAN, Gaussian Mixture Model. Estos, pueden compartir o no medidas de evaluación que determinan el número óptimo de clusters, basándose en diversos factores como la cohesión entre los clusters, la homogeneidad de los datos, entre otros. Por mencionar algunos métodos, existe el método del codo, el coeficiente de silueta, el BIC y más.

- **Despliegue:**

El INEGI da acceso a formatos .shp que facilitan la visualización geográfica de cómo se visualizarán estos clusters por manzanas, esto con el propósito de darle, primero que nada, una interpretación más rápida y sencilla para después comunicar al socio formador los hallazgos del proyecto.

1.4.4 Resultados esperados

Se espera obtener un modelo estadístico o sistema que indique a empresarios dónde pueden colocar sus instalaciones o locales, según las necesidades de cada empresario y de sus emprendimientos, ya que en la actualidad, el análisis predictivo ya no es una opción sino un requisito crucial para las empresas, debido a que estos pueden hacer o deshacer el éxito de una empresa. (Business Wire, 2020).

1.4.5 Beneficios esperados e impacto social

De esta forma, se espera obtener un beneficio económico en la venta de este sistema/modelo. Por otro lado, también se espera un impacto social, ya que el adecuado uso de los espacios urbanos no solo implica un incremento en la actividad comercial e impulso económico para el área empresarial, sino también implica un espacio de mayor provecho para los habitantes del área de trabajo, ya que los negocios estarán pensados en satisfacer las necesidades específicas de la zona, y así, como indica Andrews, J. (2023) desempeñar un papel aún más importante en la configuración del futuro.

1.5 Definir el plan del proyecto

Etapa	Tiempo	Responsable
Conocimiento del Negocio <ul style="list-style-type: none">● Comprender el problema● Investigar al socio formador● Investigar casos similares	2 días	Todo el equipo
Comprensión de los datos <ul style="list-style-type: none">● Recopilar y leer datos● EDA● Buscar fuentes de datos externas	2 días	Todo el equipo

Preparación de los datos	1.5 semana	Todo el equipo
<ul style="list-style-type: none"> ● Seleccionar variables de interés ● Limpieza de datos ● Transformación de datos 		
Modelación	1.5 semanas	Todo el equipo
<ul style="list-style-type: none"> ● Elegir modelo ● Separar datos de entrenamiento y test ● Hacer ajustes al modelo 		
Evaluación	2 días	Todo el equipo
<ul style="list-style-type: none"> ● Evaluación de métricas ● Revisar si es necesario regresar a pasos anteriores y hacer cambios 		
Implantación	2 días	Todo el equipo

2) Entendimiento de los datos

2.1 Descripción del set de datos

El desarrollo de la exploración de datos se llevó a cabo en un notebook de Python. El primer paso en la realización de esta exploración fue la integración de las bases de datos, las cuales fueron el censo de población y vivienda de la INEGI así como el DENU. El censo de población y vivienda del INEGI incluye más de 200 columnas y contiene registros de cada manzana de Nuevo León. Para este análisis se utilizarán únicamente algunas columnas relacionadas con la demografía de la población en cuanto a edad, escolaridad y acceso a salud, así como columnas relacionadas con el tipo de viviendas y sus características. Estas fueron seleccionadas basándose en los indicadores que utiliza la CONEVAL en ciertas bases de datos para evaluar el rezago social. Así mismo se filtró para tener únicamente registros de los municipios del Área Metropolitana de Monterrey.

Posteriormente, se limpió la base de datos del censo de INEGI, se comprobó la inexistencia de elementos nulos, se eliminaron datos inválidos, se redujo la cantidad de columnas, se filtró por clave de municipios y se agregó una columna de identificación. De esta forma, la base de datos terminó con 17 columnas y 52565 filas. Además, se modificó el tipo de dato que algunas columnas manejaban para facilitar el análisis. Así, las variables o columnas que contiene la base de datos y sus características son las siguientes:

1. MUN (int64): Cualitativa. Código numérico del municipio.
2. NOM_MUN (object): Cualitativa. Nombre del municipio.
3. AGEB (object): Cualitativa. Código de la Área Geoestadística Básica.
4. POBTOT (int64): Cuantitativa. Población total.
5. VIVPAR_HAB (int64): Cuantitativa. Número de viviendas particulares habitadas.
6. P18YM_PB (int64): Cuantitativa. Población de 18 años o más con primaria completa.
7. PSINDER (int64): Cuantitativa. Población sin derechohabiencia a servicios de salud.
8. VPH_EXCSA (int64): Cuantitativa. Viviendas particulares con excusado o sanitario.
9. VPH_LAVAD (int64): Cuantitativa. Viviendas particulares con lavadora.
10. VPH_REFRI (int64): Cuantitativa. Viviendas particulares con refrigerador.
11. VPH_TELEF (int64): Cuantitativa. Viviendas particulares con teléfono fijo.
12. P15YM_AN (int64): Cuantitativa. Población de 15 años o más analfabeta.
13. P6A11_NOAM (int64): Cuantitativa. Población de 6 a 11 años sin asistencia escolar.
14. P12A14NOA (int64): Cuantitativa. Población de 12 a 14 años que no asiste a la escuela.

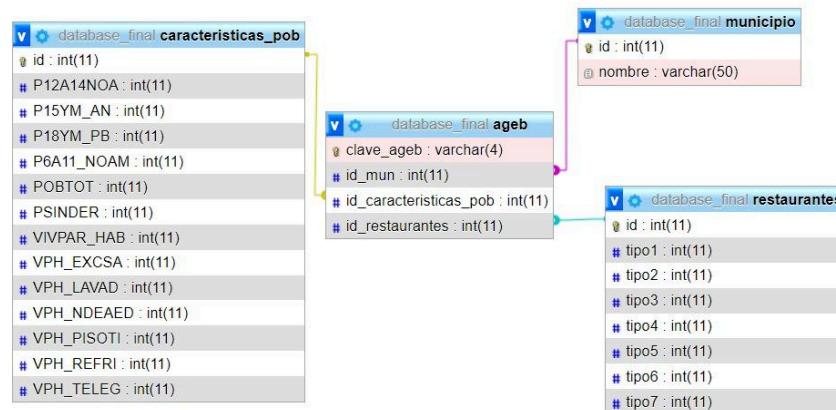
15. VPH_PISOTI (int64): Cuantitativa. Viviendas particulares con piso de tierra.
16. VPH_NDEAED (int64): Cuantitativa. Viviendas particulares habitadas que disponen de agua entubada en el ámbito de la vivienda.
17. CVEMA (object): Cualitativa. Clave compuesta por Municipio y AGEBS.

2.2 Integración y preparación de datos

Para este proyecto se busca segmentar por AGEBS, por lo tanto se agruparon las manzanas de un mismo AGEBS y se sumaron sus valores, creando un nuevo data frame que contiene registros de cada AGEBS de Nuevo León en vez de Manzanas. Con el proceso de limpieza completamente finalizado para el censo de INEGI, se obtuvo la base de datos final sobre el censo, que contiene 17 columnas y 1645 filas.

Además del censo de INEGI, se utilizó la base de datos DENUE. El data frame cargado del DENUE contiene los registros de que cada uno de los negocios del estado de Nuevo León, contiene 186,665 negocios con 43 columnas con respecto a la localización, tipo, tamaño e información general del negocio. Para este proyecto el DENUE se usó para contabilizar el número de restaurantes de cierto tipo en cada AGEBS. Para esto se filtró para tener únicamente registros de restaurantes que forman parte del Área Metropolitana de Monterrey, incluyendo en estos 8 municipios: Apodaca, Escobedo, Guadalupe, Juárez. Monterrey, San Nicolás, San Pedro y Santa Catarina. Otros procesos que se llevaron a cabo en la limpieza de este data frame fueron filtrar por clave de municipio y añadir una columna de identificación. Finalmente, se creó otro data frame que expone la frecuencia de restaurantes según su tipo en cada AGEBS, para una mejor comprensión de los datos.

2.3 Modelo Conceptual de Datos



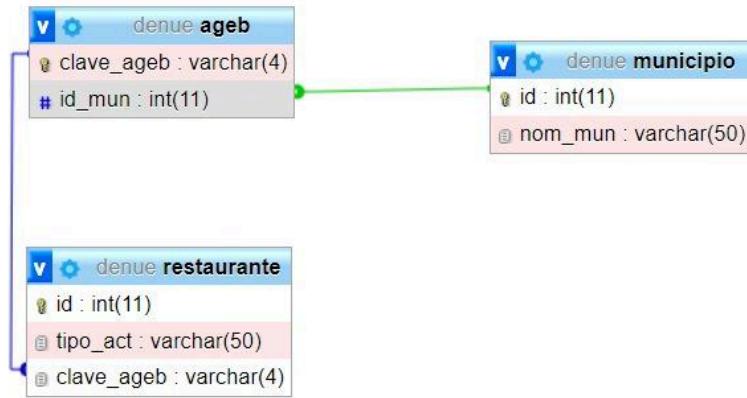


Figura 1. Diagrama Entidad Relación

2.4 Exploración de datos, correlación entre variables y principales hallazgos

Con el trabajo mencionado anteriormente, fue posible concretar un análisis más claro y conciso sobre las variables que se analizarán en el desarrollo de este proyecto, permitiendo una mejor comprensión de los datos que se trabajan. La exploración de datos permitió definir que el proyecto se centrará en predecir la mejor ubicación para colocar un nuevo restaurante según el tipo de restaurante, el enfoque en este giro favorece la precisión de los resultados y facilita el manejo de las bases de datos.

Ahora bien, se procedió a un análisis de las variables cualitativas y cuantitativas del censo ya limpio, es decir, con la información a usar. En primer lugar, se arroja un análisis descriptivo de las variables numéricas del censo.

	P12A14NOA	P15YM_AN	P18YM_PB	P6A11_NOAM	POBTOT	PSINDER
count	1645.000000	1645.000000	1.645000e+03	1645.000000	1.645000e+03	1645.000000
mean	22.421884	89.478419	4.261238e+03	15.054711	1.027080e+04	1918.936778
std	235.178502	884.356736	3.543405e+04	163.411336	8.316953e+04	15939.169666
min	0.000000	0.000000	0.000000e+00	0.000000	0.000000e+00	0.000000
25%	0.000000	3.000000	9.250000e+02	0.000000	2.430000e+03	310.000000
50%	5.000000	14.000000	1.840000e+03	3.000000	5.096000e+03	813.000000
75%	11.000000	46.000000	3.094000e+03	7.000000	7.414000e+03	1419.000000
max	6578.000000	28134.000000	1.019099e+06	4884.000000	2.285946e+06	450013.000000

VIVPAR_HAB	VPH_EXCSA	VPH_LAVAD	VPH_NDEAED	VPH_PISOTI	VPH_REFRI	VPH_TELEF
1645.000000	1645.000000	1645.000000	1645.000000	1645.000000	1645.000000	1645.000000
2696.091185	2885.543465	2600.084498	0.152584	15.993921	2827.564134	1862.483283
21776.202445	23464.475690	21168.920173	2.708541	171.573500	22997.950012	15579.403293
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
639.000000	704.000000	636.000000	0.000000	0.000000	689.000000	390.000000
1348.000000	1461.000000	1318.000000	0.000000	0.000000	1436.000000	896.000000
1908.000000	2046.000000	1834.000000	0.000000	6.000000	2010.000000	1343.000000
579832.000000	650341.000000	586635.000000	96.000000	5226.000000	637308.000000	454930.000000

Figura 2. Análisis descriptivo de variables numéricas del censo

En ello, de manera general se interpreta que, debido a que la mayoría de los cuartiles se mantienen en números bajos, la cantidad de población en cada AGEB por cada variable es baja, infiriendo en primera instancia gran parte de AGEB's con poco rezago social. De igual manera, se espera que la gran parte de las variables estén sesgadas a la izquierda y con la intervención de valores atípicos.

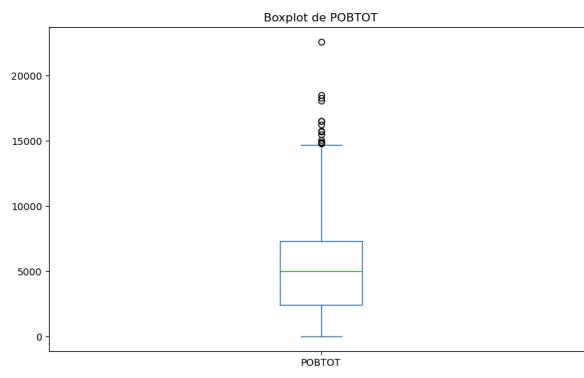
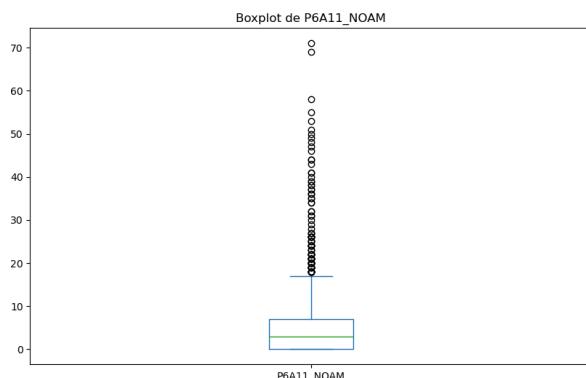
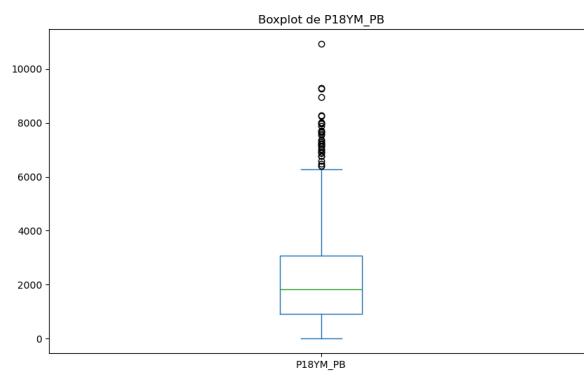
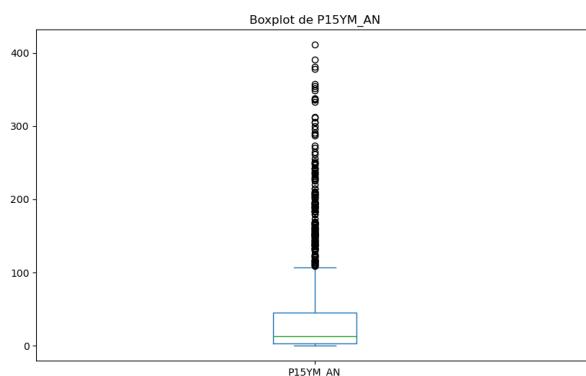
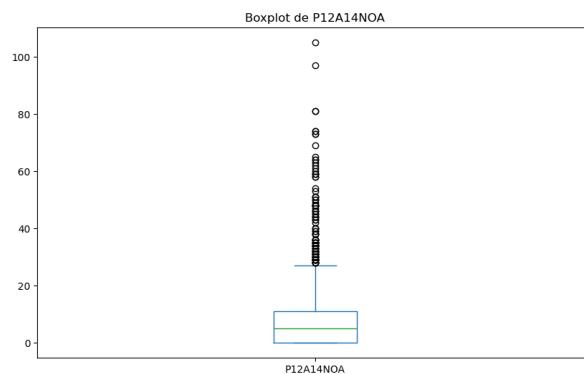
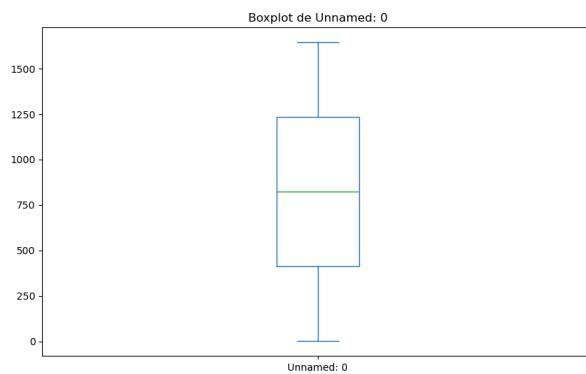
Para las variables cualitativas se llevó a cabo el mismo proceso, dando como salida la siguiente tabla:

	CVEMA	MUN	NOM_MUN	AGEB
count	1645	1645	1645	1645
unique	1645	8	8	1395
top	0060000	39	Monterrey	0000
freq	1	492	492	8

Figura 3. Análisis variables cualitativas del censo

De ella se observa que la clave creada *CVEMA*, que consiste en juntar AGEB + manzana, es única para cada registro. De igual manera, se ve que Monterrey es el municipio que más AGEB's tiene.

Así mismo, se realizó un análisis visual de las variables tanto cualitativas como cuantitativas para una mejor interpretación y análisis de los datos de las mismas obteniendo así los siguientes boxplot, diagramas de dispersión e igualmente un heatmap para poder identificar la correlación existente entre entre las variables del dataset.



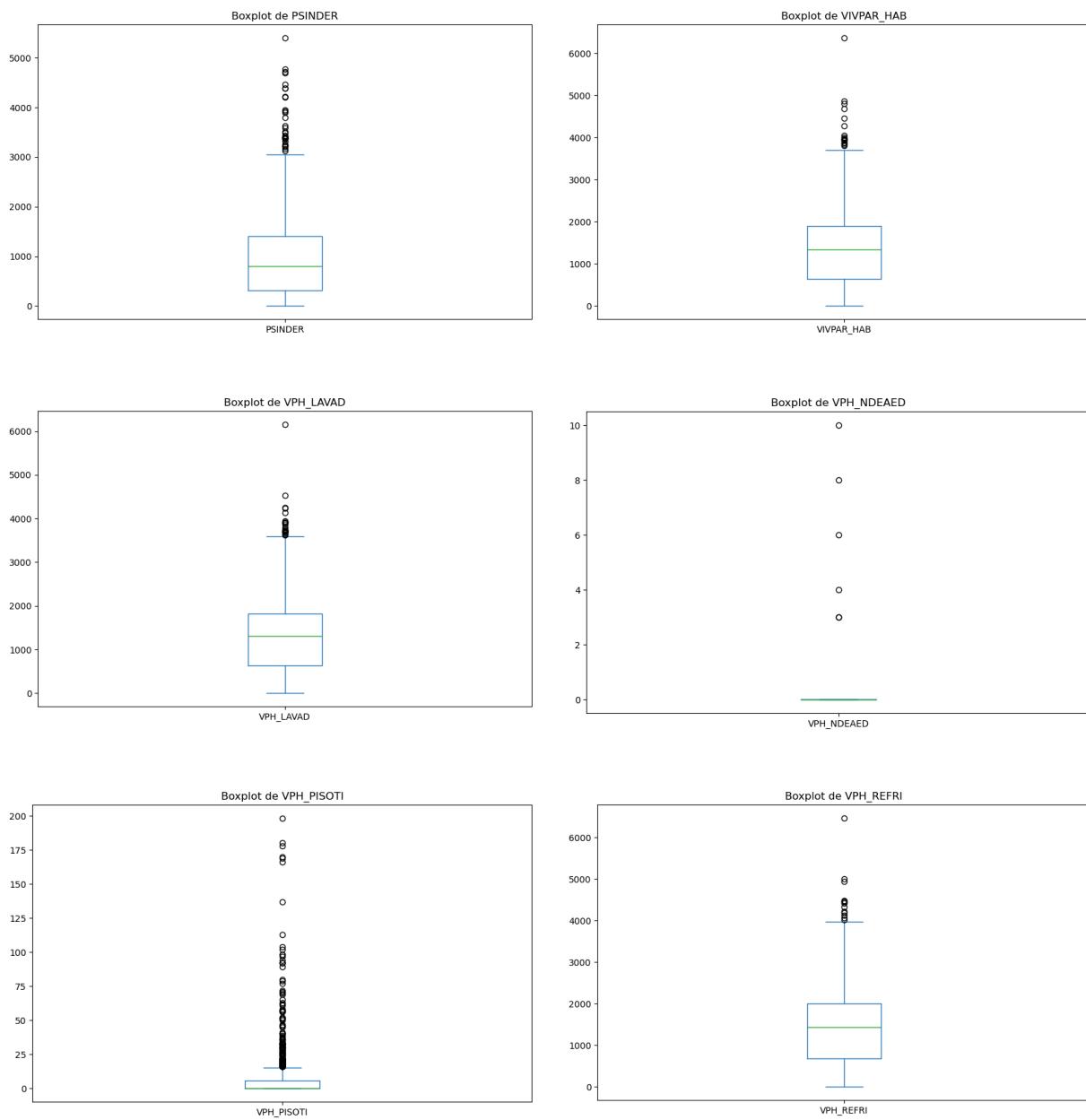
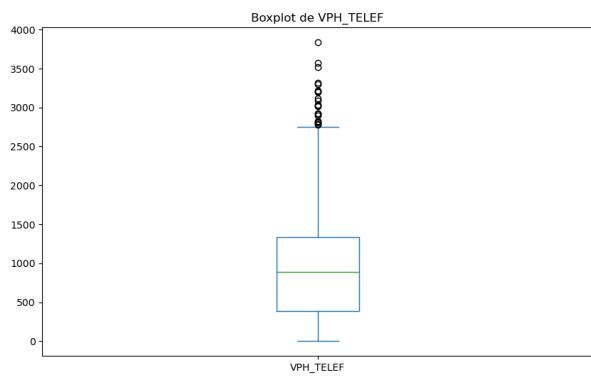


Figura 4. Boxplot variables numéricas del censo



Histogramas de las Variables Cuantitativas

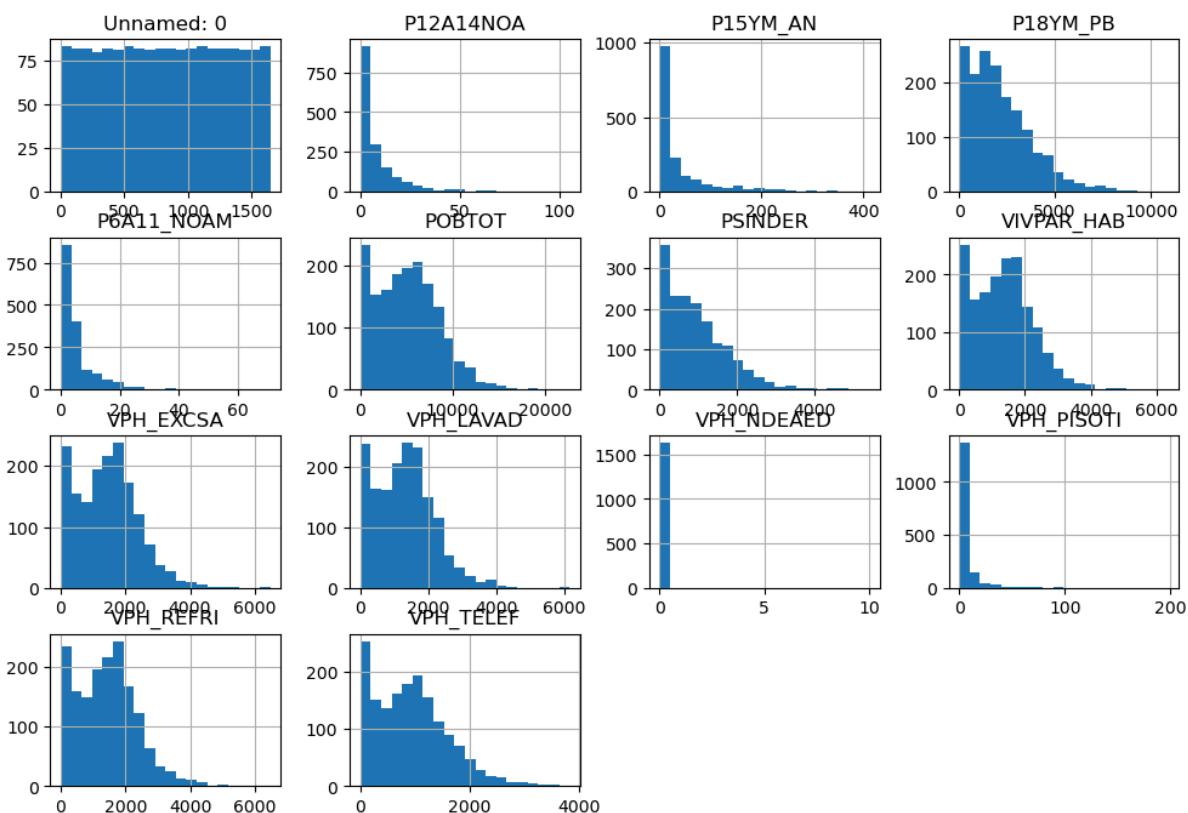


Figura 5. Histogramas variables cuantitativas del censo

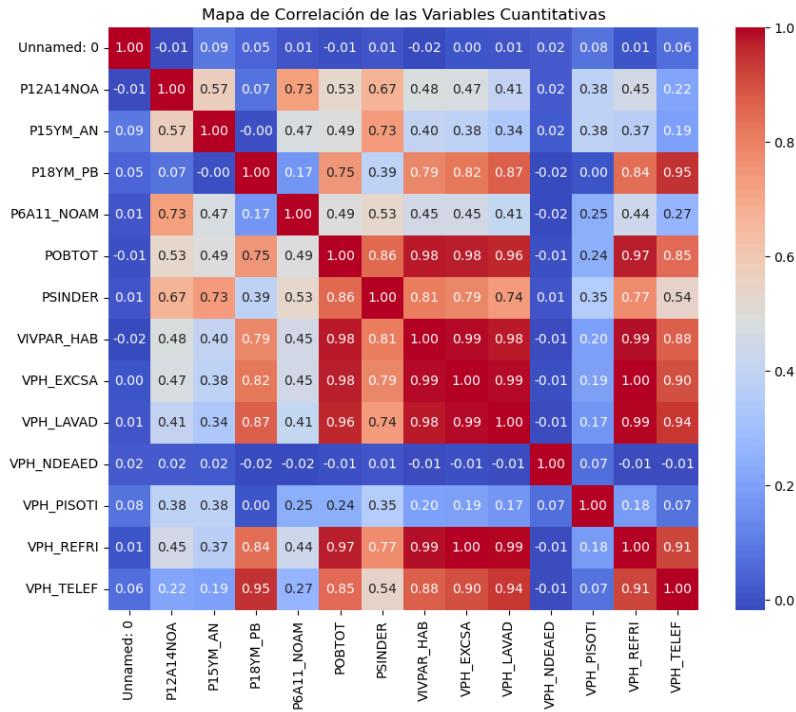


Figura 6. Heatmap variables del censo

Por otro lado, ya que se está empleando información de 8 municipios del área metropolitana de Monterrey, realizó un histograma para conocer la distribución de los datos dados los municipios, teniendo así una distribución semejante a una exponencial, donde Monterrey es quién tiene mayor cantidad de negocios y San Pedro menor. Así mismo, la presencia de negocios entre Escobedo, San Nicolás y Santa Catarina, llega a ser muy parecida. Por lo que se puede esto igualmente atribuir a la densidad de población que llega a existir en la zona metropolitana.

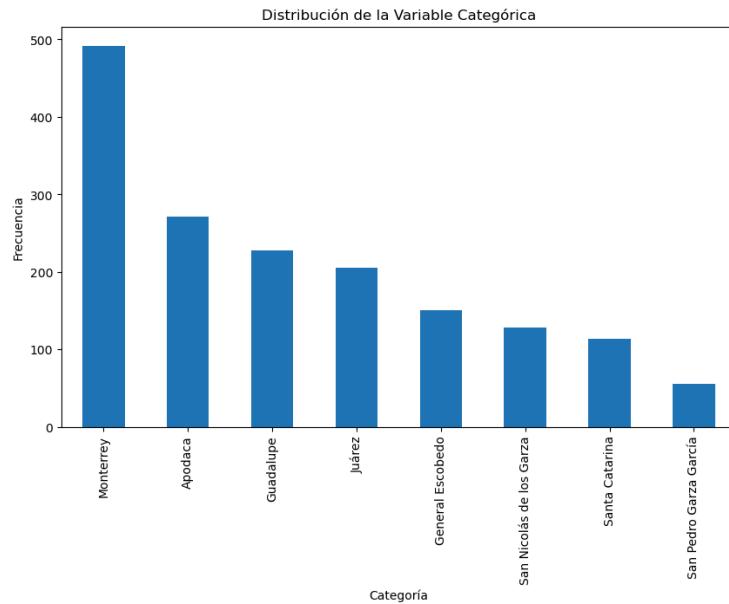


Figura 7. Diagrama barras municipios del censo

Al igual que el censo, es de suma importancia observar la estadística descriptiva perteneciente al DENUE, que en este caso, consiste en el conteo de cada tipo de restaurante por AGEB.

	Restaurante_Restaurantes con servicio de preparación de alimentos a la carta o de comida corrida	Restaurante_Restaurantes con servicio de preparación de antojitos	Restaurante_Restaurantes con servicio de preparación de pescados y mariscos
count	1293.000000	1293.000000	1293.000000
mean	1.712297	1.486466	0.438515
std	4.141739	1.683330	0.879928
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	0.000000	1.000000	0.000000
75%	2.000000	2.000000	1.000000
max	62.000000	14.000000	12.000000

Restaurante_Restaurantes con servicio de preparación de pizzas, hamburguesas, hot dogs y pollos rostizados para llevar	Restaurante_Restaurantes con servicio de preparación de tacos y tortas	Restaurante_Restaurantes de autoservicio	Restaurante_Restaurantes que preparan otro tipo de alimentos para llevar
1293.000000	1293.000000	1293.000000	1293.000000
1.978345	3.846868	0.543697	0.574633
2.004135	3.756695	1.407845	1.279164
0.000000	0.000000	0.000000	0.000000
0.000000	1.000000	0.000000	0.000000
1.000000	3.000000	0.000000	0.000000
3.000000	5.000000	1.000000	1.000000
14.000000	35.000000	15.000000	30.000000

Figura 8. Estadística descriptiva DENU

Trabajando con un análisis de la media de cada una de las variables, es posible inferir que para 4 tipos de restaurantes de los 7 que hay, al menos se encuentra 1 restaurante de ese tipo en cada AGEB, mientras que de los otros 3 es difícil encontrar uno en cada AGEB. También es interesante observar los máximos que hay en cada variable, dado que dentro de esos AGEB's se podría tratar de una zona altamente con alto consumo. Sin embargo, esta interpretación está sujeta a cambiar dado al comportamiento que tengan los datos y la media no sea la mejor herramienta para interpretar.

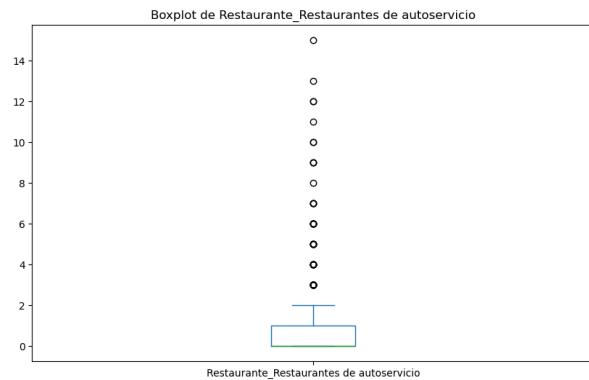
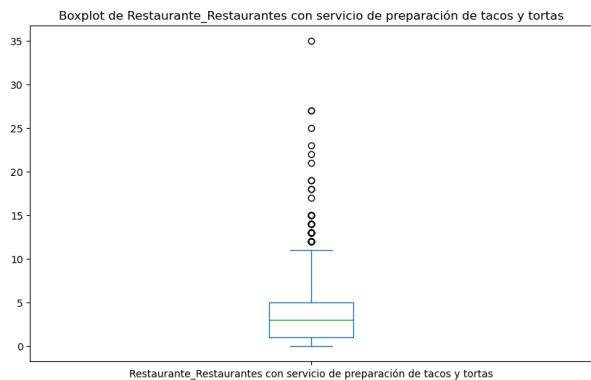
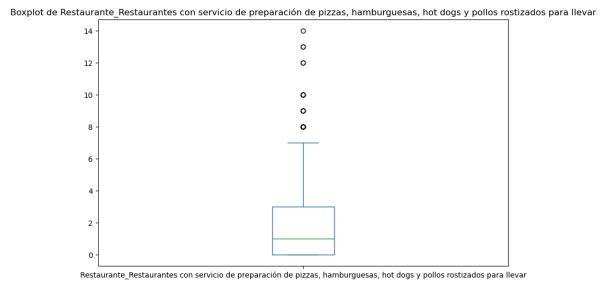
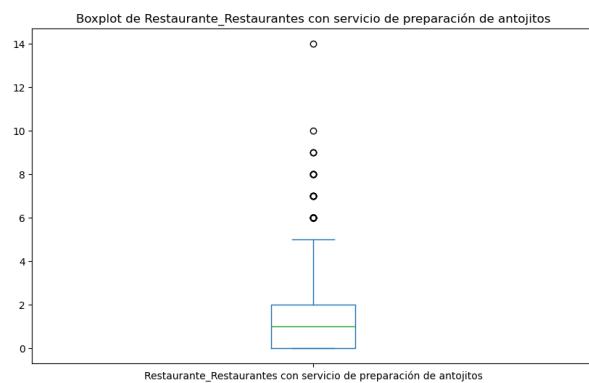
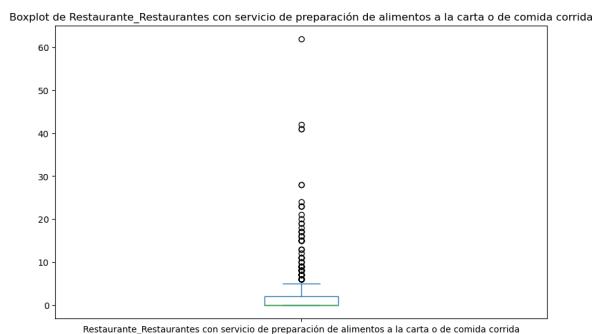
Mientras tanto, para la única variable cualitativa que se maneja dentro de esta base de datos se tiene lo siguiente:

CVEMA	
count	1293
unique	1293
top	0060051
freq	1

Figura 9. Análisis variables cualitativas DENU

Esta se trata únicamente del código creado para distinguir cada AGEB, en donde se ve que es único.

Como complemento visual se realizaron boxplots, histogramas y un heatmap para ver el comportamiento y relación de las variables.



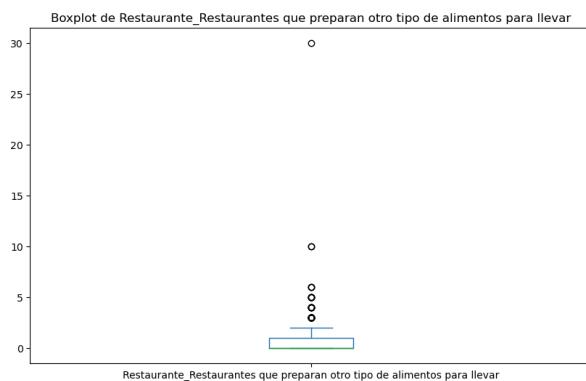


Figura 10. Boxplot variables numéricas DENUE

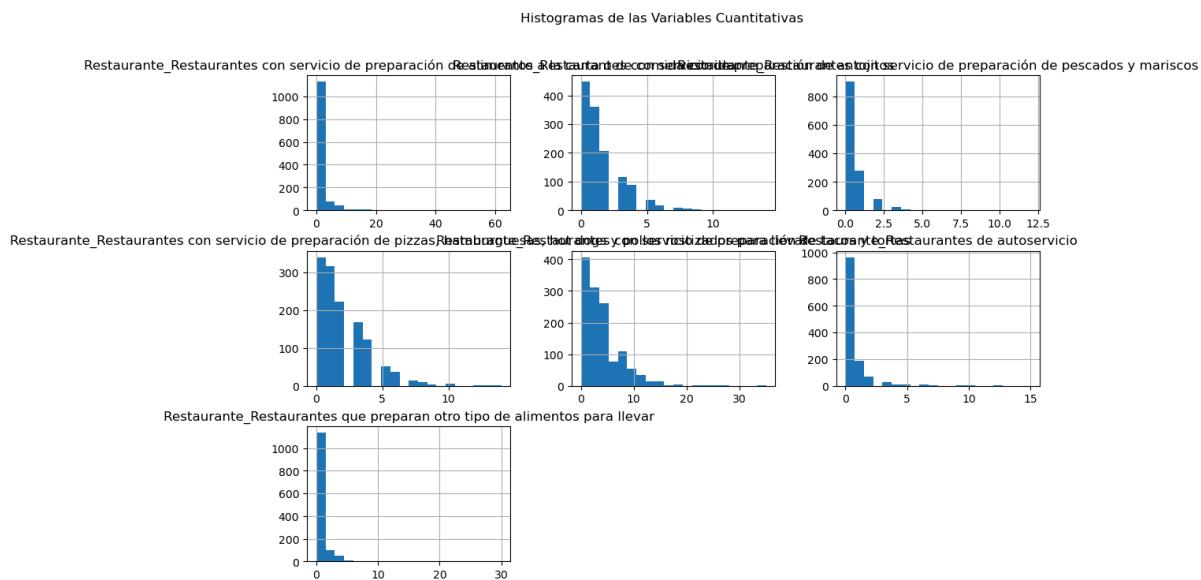


Figura 11. Histogramas variables cuantitativas DENUE

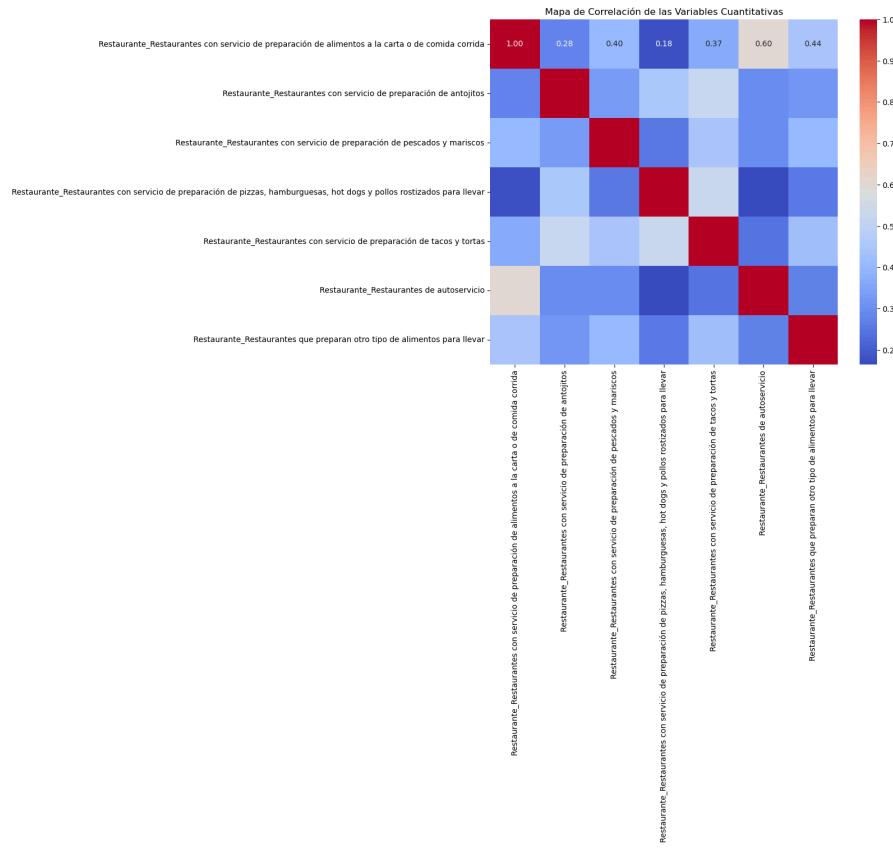
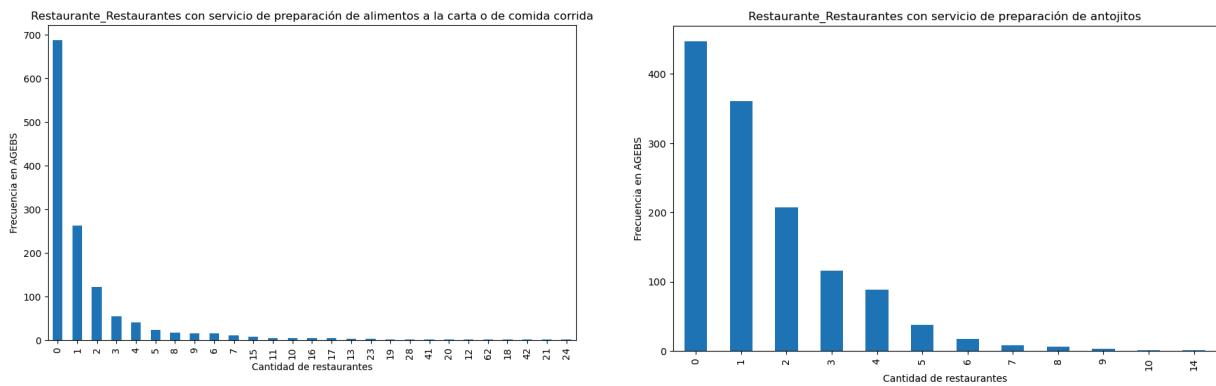


Figura 12. Heatmap variables DENUE

En el mismo contexto, se realizaron gráficas de barras para cada tipo de restaurante y la frecuencia que se tiene en cada AGEB.



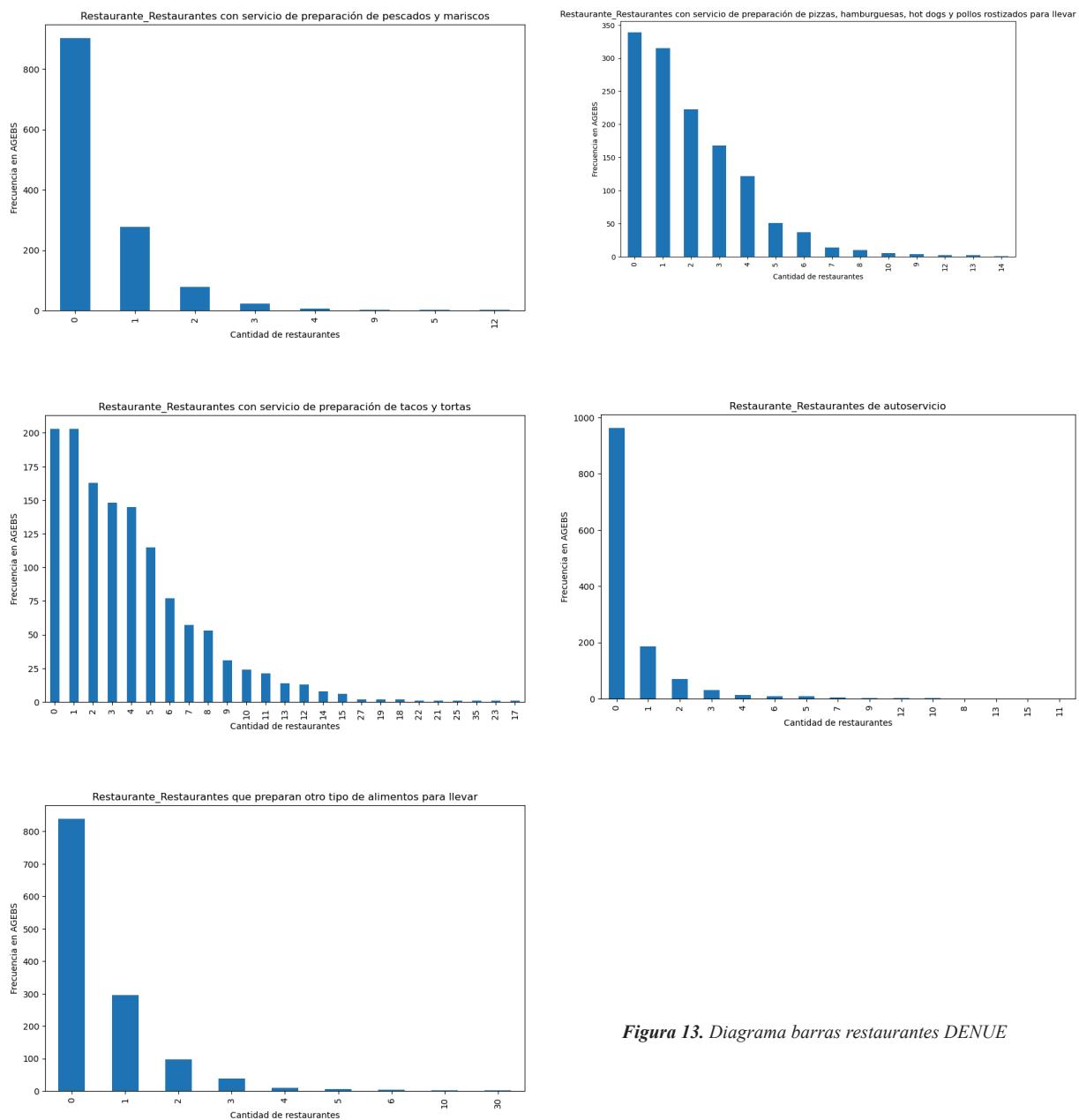


Figura 13. Diagrama barras restaurantes DENUE

En general, en cada tipo de restaurante predomina el conteo de 0, esto se debe a que de todos los AGEBS a usar, la mayoría de ellos no contienen todos los tipos de restaurantes, e incluso puede que haya AGEB's sin ningún tipo de restaurante. A su vez, analizando los histogramas se comprende que el análisis del número de restaurantes que hay en cada AGEBS no debería de realizarse con la media dado que están sesgados.

2.5 Preparación de los datos

La preparación de los datos se realizó a partir de diferentes bases de datos, se utilizaron los data frames de INEGI final con 17 columnas y 1645 filas así como el de frecuencias de tipo de restaurantes de DENUE con 8 columnas y 1293 filas que se obtuvieron en la etapa de exploración. Con el objetivo de agregar una columna que contenga información geográfica de los agebs y que permita una visualización en mapa se leyó el archivo shapefile de los agebs del estado de Nuevo León.

Enseguida, se realizaron diferentes merge para integrar las bases de datos. A partir de esto es posible graficar en un mapa los AGEBS y pueden ser clasificados a partir de las columnas del censo. Por ejemplo a continuación se muestra la separación de los AGEBs por municipio:

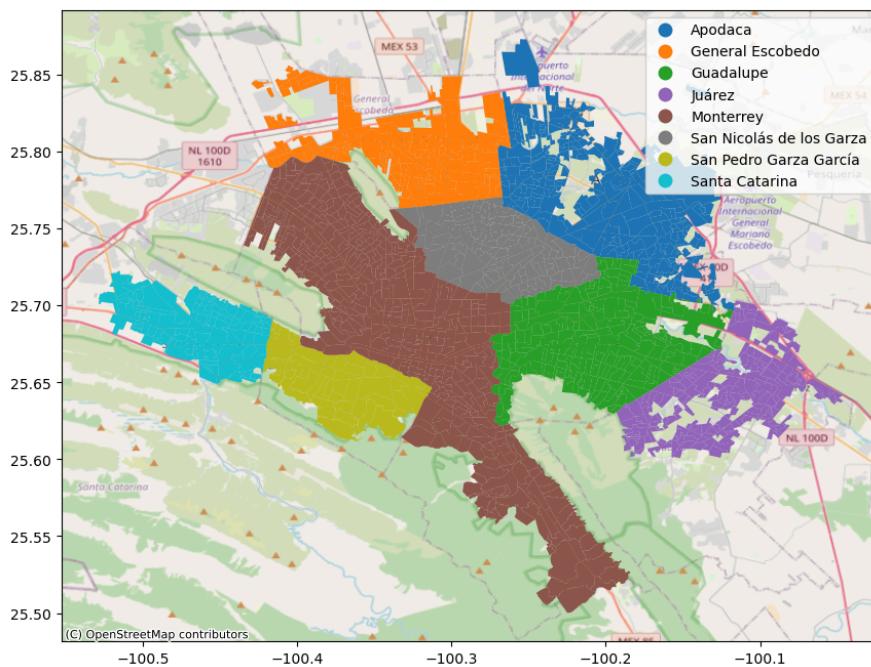


Figura 14. Mapa de separación de AGEBS por municipio

Los datos finales del data frame incluyen información demográfica y de frecuencia de restaurantes para diferentes AGEBS en Monterrey. Estos datos son relevantes para el análisis demográfico y de restaurantes en Monterrey, ya que proporcionan información sobre la población en diferentes áreas, así como sobre la disponibilidad y la frecuencia de diferentes tipos de restaurantes. En este caso se excluyeron más de cien columnas y registros que fueron descartadas en la etapa de exploración.

3. Generación de Modelos de Aprendizaje

3.1 Variables relevantes para el análisis

Nombre de la variable	Descripción	Tipo de dato	Tipo de variable
Rest_Carta	Define el número de restaurantes que hay de este tipo en el AGEB	Entero	Numérica
Rest_Antojitos	Define el número de restaurantes que hay de este tipo en el AGEB	Entero	Numérica
Rest_Mariscos	Define el número de restaurantes que hay de este tipo en el AGEB	Entero	Numérica
Rest_Rapida	Define el número de restaurantes que hay de este tipo en el AGEB	Entero	Numérica
Rest_TacosTortas	Define el número de restaurantes que hay de este tipo en el AGEB	Entero	Numérica
Rest_Autoserv	Define el número de restaurantes que hay de este tipo en el AGEB	Entero	Numérica
Rest_OtrosLlevar	Define el número de restaurantes que hay de este tipo en el AGEB	Entero	Numérica
POBTOT	Define el número total de personas que hay en el AGEB	Entero	Numérica

Ya que se trabajaron diversos métodos de aprendizaje no supervisado no hay variables predictoras/objetivo. Todas las variables de la tabla se usarán para hacer clustering.

3.2 Datos disponibles

	Rest_Carta	Rest_Antojitos	Rest_Mariscos	Rest_Rapida	Rest_TacosTortas	Rest_Autoserv	Rest_OtrosLlevar	POBTOT
0	0	2	0	3	3	1	3	10712
1	0	0	0	1	1	3	0	9224
2	0	0	0	2	1	0	0	4650
3	8	2	1	3	3	6	0	0
4	0	1	0	1	0	0	0	8268
5	0	2	1	4	4	1	1	13304
6	0	3	1	2	6	0	0	11738
7	0	0	0	2	0	0	0	3844
8	0	0	0	0	1	0	0	2426
9	0	0	0	2	0	0	0	2016

3.3 Modelos de Aprendizaje No Supervisado:

- **K-Means:**

El algoritmo K-means es una técnica de clustering que agrupa datos en k grupos basados en sus características. En nuestro caso, estamos utilizando datos de frecuencia de restaurantes y la población total de cada AGEB como características para el clustering. Para garantizar que todas las características contribuyan equitativamente al cálculo de la distancia entre puntos, normalizamos los datos utilizando el MinMaxScaler de scikit-learn, lo que nos permite escalar todas las características al mismo rango ([0, 1]).

Exploramos diferentes combinaciones de hiperparámetros, incluyendo el número de clusters (num_clusters) y el estado aleatorio (random_state). El número de clusters es un factor crítico que afecta la estructura de los clusters, mientras que el estado aleatorio controla la inicialización de los centroides. Se probaron varias opciones para estos hiperparámetros, incluyendo num_clusters_options = [3, 4, 5] y random_state_options = [0, 42, 100].

Utilizamos dos métricas para evaluar los modelos de clustering: la inercia y el coeficiente de silueta. La inercia, que representa la suma de las distancias al cuadrado de cada punto a su centroide más cercano, disminuye a medida que aumenta el número de clusters, como era de esperarse. Por otro lado, el coeficiente de silueta mide la cohesión dentro de los clusters y la separación entre ellos.

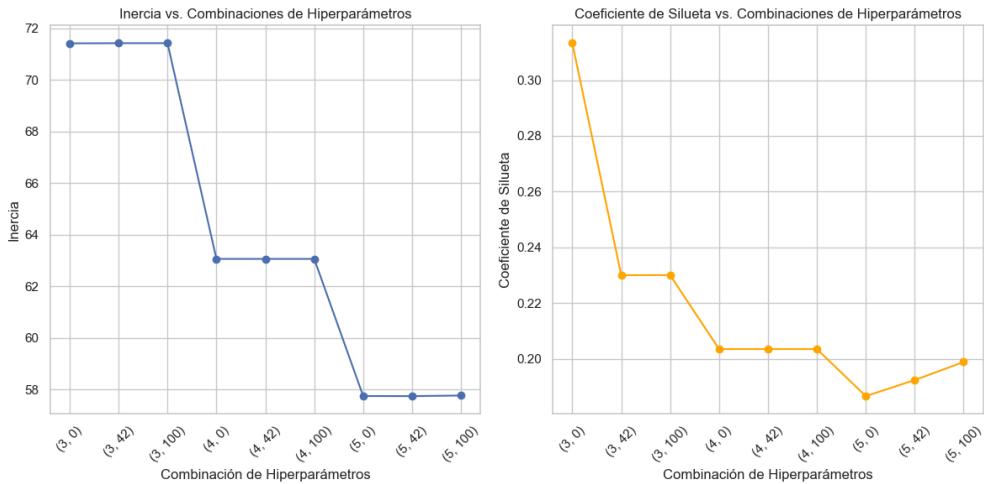


Figura 15. Gráficos de Inercia y Coeficiente de Silueta para diferentes combinaciones de hiperparámetros (número de clusters y estado aleatorio) en el modelo de clustering K-means.

Basándonos en las métricas de inercia y coeficiente de silueta, las combinaciones de hiperparámetros con `num_clusters=3` y `random_state=0` parecen ser las más adecuadas para este conjunto de datos. Estas configuraciones producen clusters más compactos y bien definidos en comparación con otras configuraciones.

● Gaussian Mixture Models:

Este modelo de clustering se basa en asumir que los datos son parte de una combinación de distribuciones normales (gaussianas), y cada registro de la base de datos tiene una probabilidad de pertenecer a cada una de estas figuras. Cada distribución representa un cluster en los datos, y para el ajuste del modelo se requiere de estimar los mejores parámetros como la media, la covarianza y la proporción de cada distribución.

Para experimentar con este modelo se usa la librería `sklearn.mixture` importando `GaussianMixture`; de ahí, para la búsqueda del mejor modelo se prueba con distintos hiperparámetros usando `GridSearchCV`, en donde con la creación de un diccionario se varía el número de componentes (distribuciones normales o clusters) y el tipo de covarianza a usar.

En el número de componentes se decidió explorar desde 2 hasta 7 componentes, mientras que en los tipos de covarianza se maneja `full`, donde cada componente tiene su propia matriz de covarianza completa; `tied` establece que todos los componentes comparten la misma matriz de covarianza; con `diag` cada componente tiene su propia matriz de covarianza diagonal; y con `esférica`, cada componente tiene su propia varianza única y sus características también.

Lo que hace `GridSearchCV` es realizar todas las combinaciones posibles de los hiperparámetros y son calificadas con el `BIC` (Bayesian Information Criterion). Entre menor puntaje para el `BIC`, mejor

modelo es. Esta métrica se basa en encontrar el equilibrio entre la capacidad de ajuste a los datos y la complejidad del modelo.

Una vez entrenado el modelo y hacer la búsqueda de los mejores hiperparámetros, se estableció que el mejor número de componentes es 7 con una covarianza diag, el cual arroja un BIC score de 1491.6. Si bien estos son los mejores hiperparámetros, la métrica BIC decide que entre más cercano sea su valor 0, mejor es el modelo, lo cual, no es bien representado con Gaussian Mixture Model.

Number of components	Type of covariance	BIC score
17	7	diag 1491.603805
16	6	diag 1840.195207
14	4	diag 2120.877108
3	5	full 2172.272724
5	7	full 2186.840292

Figura 16. Hiperparametros

Asimismo, se presenta un gráfico que compara todos los BIC score, demostrando una vez más cuáles son los mejores hiperparámetros.

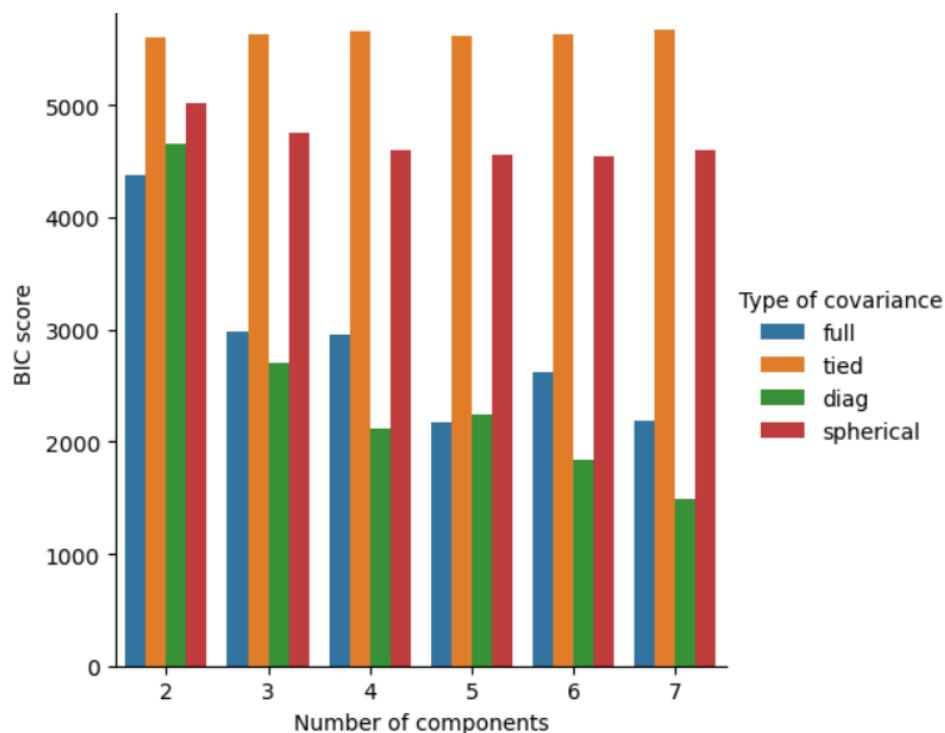


Figura 17. Distribución de los hiper parámetros

- Self-Organizing Maps:

Los Self Organizing Maps, también conocidos como SOM, es una red neuronal no supervisada, la cual es entrenada empleando técnicas de aprendizaje no supervisado para la producción de una representación de baja dimensión a partir de un espacio de entrada, también conocido como mapa, por ende ayuda para la reducción de las dimensiones de los datos. A diferencia de otras redes neuronales, esta emplea funciones de vecindad para la preservación topológica de las variables de entrada.

Las unidades neuronales, forman un espacio bidimensional, creando un mapeo desde un espacio de alta dimensión. Se conserva la distancia relativa entre los puntos y ayuda al análisis de conglomerados de datos de alta dimensión, además tienen la capacidad de generalización reconociendo entradas que no habían tenido antes. (Dey, V, 2024)

En el caso del empleo de este modelo, se usó la librería de *minisom* importando *MiniSom*. Por otro lado, para la determinación del número de clusters a emplear, se definió un función, en la cual se realizaba variación del espacio muestral y del número de clusters dentro de un rango de (1-10), en el cual se empleaba el *Quantization error*; este contrasta la información de los datos agregando o removiendo, con esto mientras este se encuentre dentro de un parámetro de [0.5-1], se considera que es un buen modelo, sin embargo se llegó a obtener un error en promedio de 1.16, que si bien no es tan bueno llega a predecir de buena manera y este es afectado por el número de clusters, sin embargo a medida que estos se aumentan, llegan a dispersar más los datos bajando el quantization error. Con esto se obtuvo que el mejor número de clusters a emplear es de 3, con un quantization error de 1.15 con un random state de: 377, en el cual las frecuencias de los restaurantes dadas las características de los mismos se encuentran centrados en la distribución del mismo a forma de perímetro del mapa en el que se llegan a encontrar y teniendo una mayor carga en el extremo inferior derecho del mapa los valores del cluster.

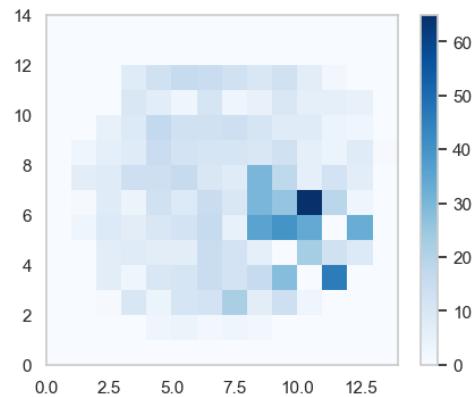




Figura 18. Distribución de clusters en el mapa dada la clasificación de los restaurantes.

- **Hierarchical Clustering:**

El método de clustering jerárquico es una técnica que agrupa datos en una jerarquía de clusters, formando un árbol que representa la relación de agrupación entre los datos. En esta ocasión, se exploró el uso del clustering jerárquico en un conjunto de datos que contiene información sobre la frecuencia de diferentes tipos de restaurantes en áreas geográficas específicas, así como la población total de cada área.

Antes de aplicar el clustering jerárquico, se realizó un preprocesamiento de los datos que incluyó la normalización de las características utilizando StandardScaler de scikit-learn. Esto fue importante para garantizar que todas las características contribuyen equitativamente al cálculo de la distancia entre puntos.

Se utilizó el algoritmo de clustering jerárquico aglomerativo, que es un enfoque ascendente donde cada observación comienza en su propio clúster y se van fusionando en función de la distancia entre ellos. Utilizamos la función AgglomerativeClustering de scikit-learn para realizar el clustering jerárquico.

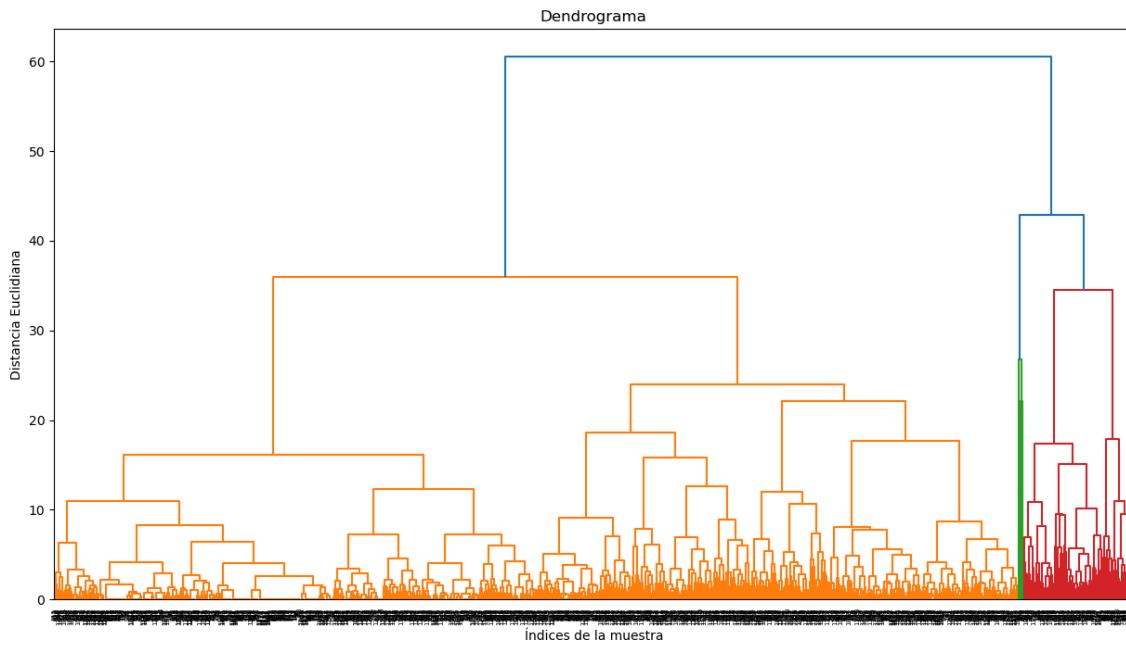


Figura 19. Aglomeración de datos en clusters visualizado en dendrograma

Se exploraron diferentes combinaciones de hiperparámetros para el clustering jerárquico, incluyendo el número de clusters y el método de enlace. Se probó con un rango de 2 a 10 clusters y se utilizaron los métodos de enlace 'ward', 'complete', 'average', y 'single'.

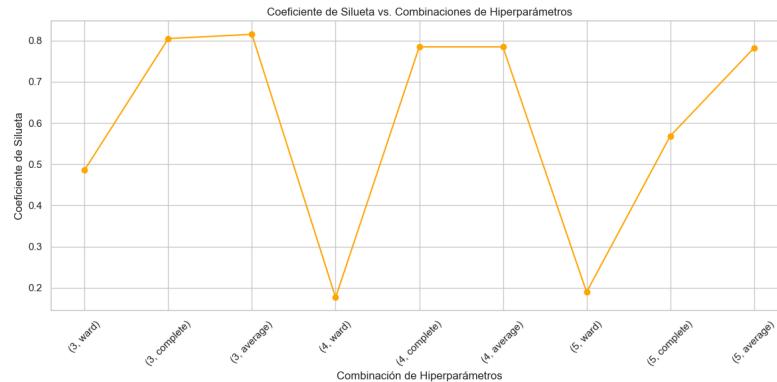


Figura 20. Gráfico de coeficiente de silueta de combinaciones de hiperparámetros.

Para evaluar el modelo de clustering jerárquico, calculamos el índice de Davies-Bouldin para cada partición obtenida. El índice de Davies-Bouldin es una medida de la "bondad" de una partición basada en la dispersión intra-cluster y la separación inter-cluster. Cuanto menor sea el valor del índice de Davies-Bouldin, mejor será la partición. Visualizamos los resultados utilizando un gráfico que compara el índice de Davies-Bouldin para diferentes números de clusters. Este gráfico nos permite identificar el

número óptimo de clusters que minimiza el índice de Davies-Bouldin y proporciona la mejor partición de los datos.

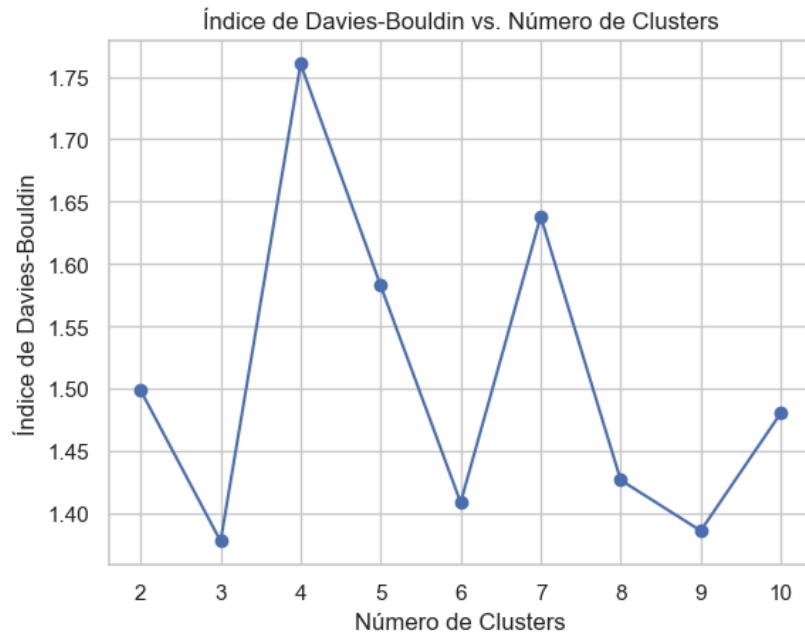


Figura 21. Gráfico de indice de Davies-Bouldin según el número de clusters creado

Tomando como fundamento el análisis del índice de Davies-Bouldin, se identificó el número óptimo de clusters y el método de enlace que produce la partición más coherente de los datos. Esto permite segmentar las áreas geográficas en clusters que comparten características similares en términos de la frecuencia de diferentes tipos de restaurantes y la población total. Estos resultados pueden ser útiles para la toma de decisiones en áreas como la ubicación de nuevos restaurantes o la segmentación de mercado.

- **DBSCAN:**

El algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es una técnica de aprendizaje no supervisado utilizada para agrupar puntos de datos en conjuntos basados en su densidad en el espacio. A diferencia de otros algoritmos de clustering, como K-means, DBSCAN no requiere que el usuario especifique el número de clusters de antemano, lo que lo hace útil en situaciones donde el número de clusters no es conocido de antemano o puede variar.

El DBSCAN es un algoritmo sencillo que define los clústeres mediante la estimación de la densidad local. Se puede dividir en 4 etapas:

1. Para cada observación se mira el número de puntos a una distancia máxima ϵ de ella. Esta zona se denomina ϵ -vecindad de la observación.
2. Si una observación tiene al menos un cierto número de vecinos, incluida ella misma, se considera una observación central. En este caso, se ha detectado una observación de alta densidad.
3. Todas las observaciones en la vecindad de una observación central pertenecen al mismo clúster. Puede haber observaciones centrales cercanas entre sí. Por lo tanto, de un paso a otro, se obtiene una larga secuencia de observaciones centrales que constituyen un único clúster.
4. Cualquier observación que no sea una observación central y que no tenga ninguna observación central en su vecindad se considera una anomalía.

Por tanto, es necesario definir dos datos antes de utilizar DBSCAN, los cuales se podrían considerar como los hiperparametros más importantes de este modelo:

epsfloat, predeterminado = 0,5

La distancia máxima entre dos muestras para que una se considere cercana a la otra. Este no es un límite máximo para las distancias de los puntos dentro de un grupo. Este es el parámetro DBSCAN más importante para elegir adecuadamente para su conjunto de datos y función de distancia.

min_samplesint, predeterminado = 5

El número de muestras (o peso total) en una vecindad para que un punto se considere central. Esto incluye el punto en sí. Si min_samples se establece en un valor más alto, DBSCAN encontrará grupos más densos, mientras que si se establece en un valor más bajo, los grupos encontrados serán más dispersos.

Para definir el mejor epsilon posible se hizo lo siguiente , se realizó un análisis de vecinos más cercanos (Nearest Neighbors) en un conjunto de datos previamente normalizado. Este análisis tiene como objetivo principal visualizar la estructura y la densidad de los datos en función de la distancia entre los puntos.

Se creó un objeto llamado `neigh` configurado para buscar los dos vecinos más cercanos a cada punto en los datos normalizados. Luego, se ajusta este objeto `NearestNeighbors` al conjunto de datos normalizados usando `fit()`, calculando así la estructura de vecindad basada en la distancia entre los

puntos. Se utiliza `kneighbors()` para calcular las distancias y los índices de los vecinos más cercanos para cada punto.. Las distancias se ordenan y se selecciona la columna para obtener las distancias al segundo vecino más cercano.

Finalmente, se crea un gráfico de líneas con `matplotlib`, donde las distancias se trazan en el eje y y los puntos de datos ordenados por distancia en el eje x. Este gráfico ayuda a determinar el valor óptimo de `epsilon` para algoritmos de clustering como DBSCAN.

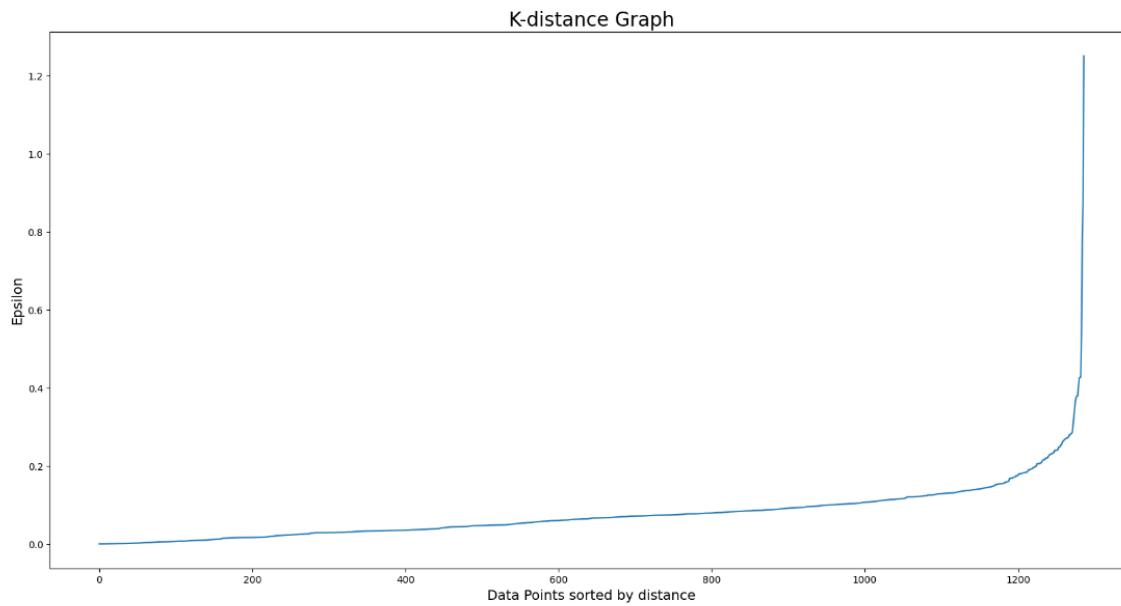
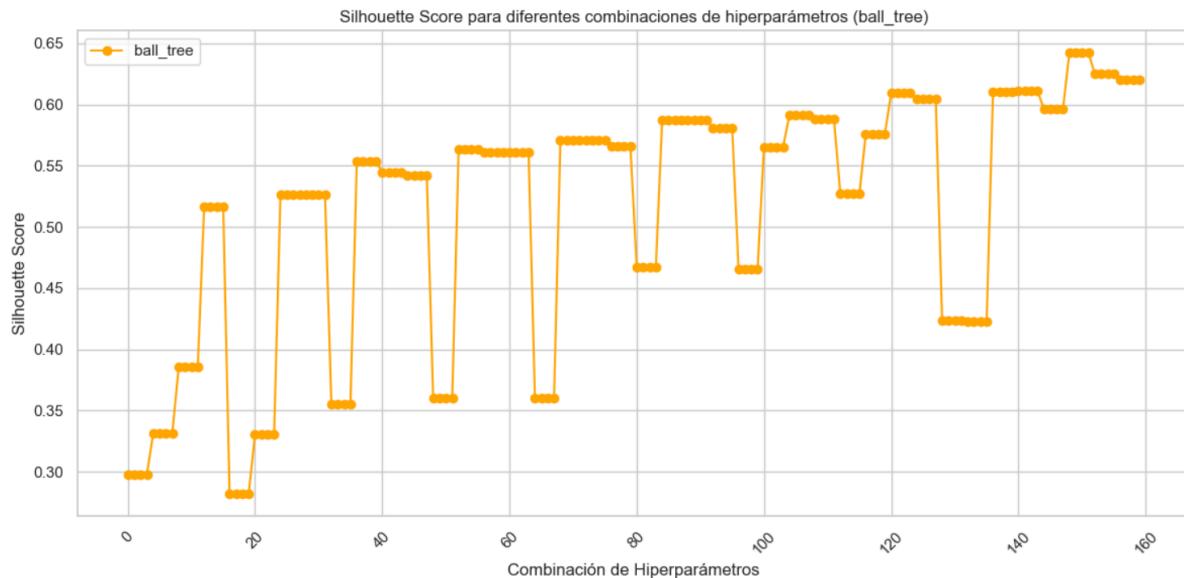


Figura 22. K-distance

Luego se hizo una búsqueda de hiperparametros para definir cuál era la mejor combinación de estos para conseguir la mejor silhouette score.



Eps	Min_Samples	Algorithm	Silhouette_Score	
149	0.300000	3	ball_tree	0.642179

Figura 23. Silhouette

Como se puede observar el mejor eps, es de 0.3 como se había calculado anteriormente. Ahora para evaluar el modelo de clustering por dbscan, calculamos el índice de la silueta para cada partición obtenida. El índice de la silueta es una medida de cuán similar es un objeto a su propio cluster(cohesión) en comparación con otros clusters (separación). La silueta va de -1 a +1, donde un valor alto indica que el objeto está bien emparejado con su propio cúmulo y mal emparejado con los clusters vecinos. Por lo que con este valor se puede decir que los clusters generados por este algoritmo no son tan malos.

- **Mean Shift:**

El algoritmo de Mean Shift es un método de aprendizaje no supervisado de clustering basado en densidad. Su funcionamiento consiste en una búsqueda iterativa con el objetivo de encontrar la región con mayor densidad. Sus usos más comunes se encuentran en el procesamiento de imágenes, específicamente en la segmentación de imágenes, y agrupamiento por visión computacional.

Para aplicar este algoritmo se usó la librería `sklearn.cluster` importando `MeanShift`. Este método permitió realizar los clusters en nuestro DataFrame previamente limpiado y ordenado. El parámetro más importante de este algoritmo es bandwidth, que define la distancia máxima permitida entre los puntos, es decir, se encarga de establecer la ventana de búsqueda para calcular los centroides de los clusters. Este parámetro fue algo que se estuvo variando para ver cómo se comportan los clusters.

Primero, se decidió utilizar otra herramienta de la librería de *sklearn* para calcular el mejor tamaño de *bandwidth*. Al no indicar ningún valor de ancho de banda, el código automáticamente utiliza *estimate_bandwidth* para calcular este parámetro. Su valor con esta herramienta fue de 0.2813464798026591. Al revisar cuántos clusters se formaron y la cantidad de datos que había en cada uno se obtuvo que el método de MeanShift creó 30 clusters teniendo los primeros tres 1094, 54 y 40 datos respectivamente. Además, aproximadamente la mitad de estos clusters solo incluyen un dato. Esto nos indica rápidamente que este método no es muy bueno para nuestro tipo de problema. Otro motivo por el cual descartar este método para nuestro proyecto es que su coeficiente de silueta fue de 0.32170785528285, lo que indica que no fue muy bueno al acomodar los datos respecto a sus similitudes a pesar de que no los acomodo erróneamente pues el coeficiente es positivo. Al tener tantos clusters con un solo dato esto se vuelve poco útil para resolver el objetivo del proyecto.

Por otra parte, al hacer una pequeña búsqueda cambiando los valores de los parámetros se llegó a que la combinación con el coeficiente de silueta más alto (0.57) fue el de un bandwidth de 0.2813464798026591 y seeds generadas con un bin_size de 0.8. Esta combinación arrojó dos clusters, uno con 1259 datos y otro con 28. Eliminando los outliers, quedan 1217 y 28 respectivamente. Al visualizar en un mapa los clusters no se puede apreciar ninguna relación entre los AGEBS dentro de cada cluster. Además, al revisar las frecuencias de cada tipo de restaurante en cada cluster tampoco se puede visualizar ninguna relación útil. Por lo tanto, este método no nos serviría en nuestro trabajo.

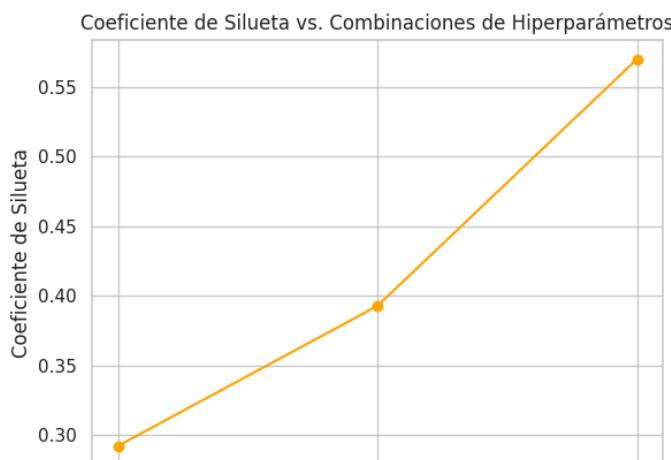


Figura 24. Coeficiente de silueta de cada combinación

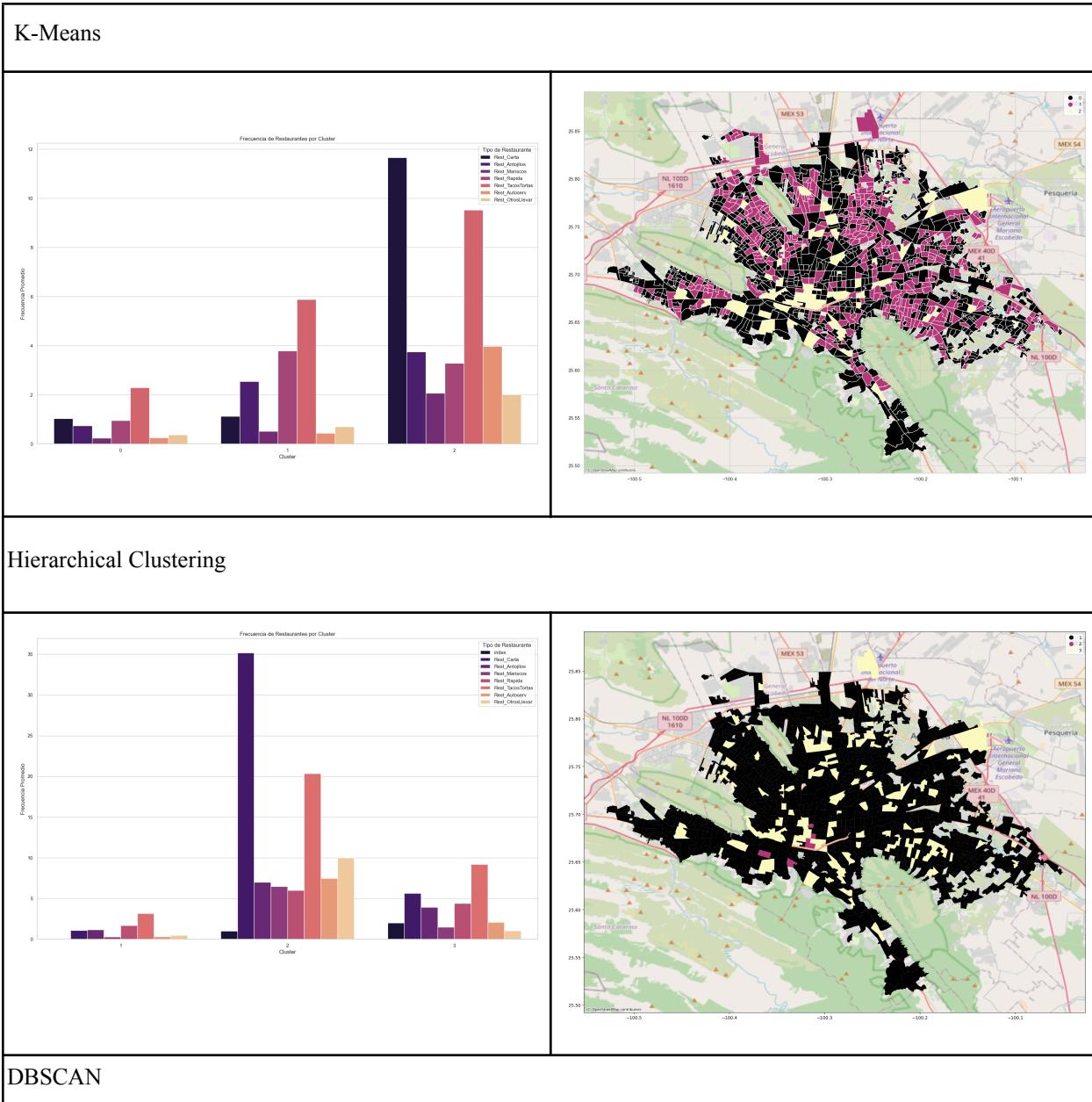
3.4 Generación y Evaluación de los Modelos de Aprendizaje

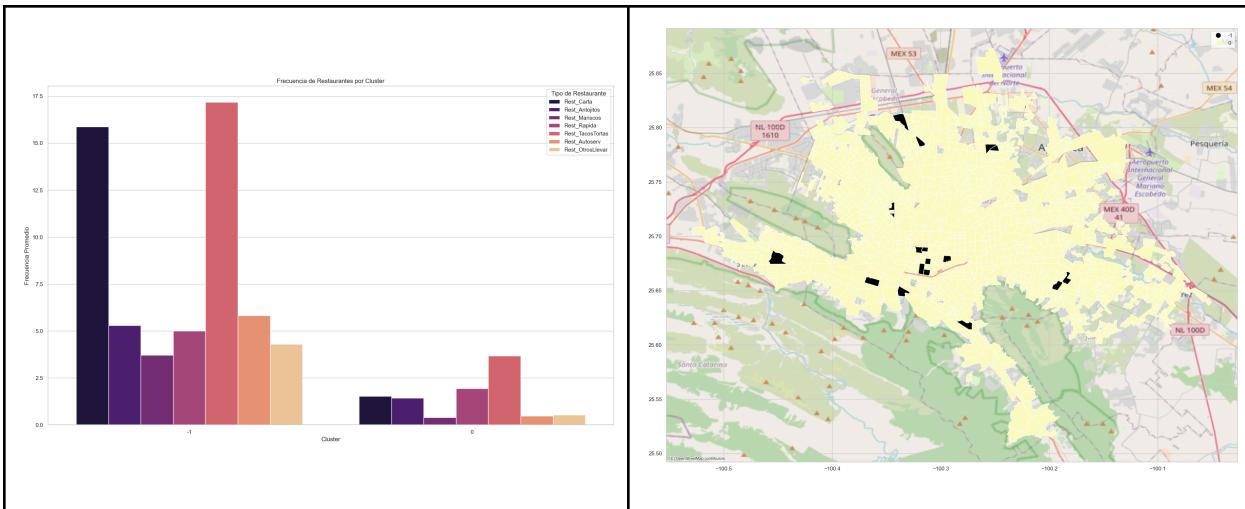
Alumno	Tipo de Clustering	Parámetros 1	Parámetros 2	Parámetros 3	Parámetros 4

Alejandro	K-Means	<i>n_clusters: 3</i> <i>random_state: 0</i> <i>Inercia: 71</i> <i>Silueta: 0.32</i>	<i>n_clusters: 3</i> <i>random_state: 42</i> <i>Inercia: 71</i> <i>Silueta: 0.23</i>	<i>n_clusters: 4</i> <i>random_state: 0</i> <i>Inercia: 63</i> <i>Silueta: 0.21</i>	<i>n_clusters: 5</i> <i>random_state: 0</i> <i>Inercia: 57</i> <i>Silueta: 0.10</i>
Mario Landa	Hierarchical Clustering	<i>n_clusters: 3</i> <i>método de enlace: average</i> <i>Silueta:~ 0.8</i> <i>Índice de Davies-Bouldin: 1.37</i>	<i>n_clusters: 4</i> <i>método de enlace: complete</i> <i>Silueta:~ 0.8</i> <i>Índice de Davies-Bouldin: 1.76</i>	<i>n_clusters: 4</i> <i>método de enlace: average</i> <i>Silueta:~ 0.8</i> <i>Índice de Davies-Bouldin: 1.76</i>	<i>n_clusters: 5</i> <i>método de enlace: average</i> <i>Silueta:~ 0.8</i> <i>Índice de Davies-Bouldin: 1.58</i>
Gael Ordaz	DBSCAN	<i>eps: 0.3</i> <i>min samples: 3</i> <i>algorithm: ball_tree</i> <i>Silueta:0.642179</i> <i>Índice de Davies-Bouldin: 1.272243</i>	<i>eps: 0.288889</i> <i>min samples: 5</i> <i>algorithm: kd_tree</i> <i>Silueta:0.610968</i> <i>Índice de Davies-Bouldin: 1.414430</i>	<i>eps: 0.277778</i> <i>min samples= 4</i> <i>algorithm:brute</i> <i>Silueta:0.609429</i> <i>Índice de Davies-Bouldin: 1.424418</i>	<i>eps: 0.211111</i> <i>min samples: 2</i> <i>algorithm: brute</i> <i>Silueta:0.281739</i> <i>Índice de Davies-Bouldin: 2.027583</i>
Natalia Guevara	Mean Shift	<i>bandwidth: 0.225963955500</i> <i>seeds: calculadas con un bin_size: 8</i> <i>silueta: 0.570050</i>	<i>bandwidth: 0.2259639555009</i> <i>seeds: calculadas automáticamente</i> <i>silueta: 0.2892839893787</i>	<i>bandwidth: 0.225963955500</i> <i>seeds: calculadas con un bin_size: 0.28134</i> <i>silueta: 0.64798026591</i>	<i>bandwidth: 0.225963955500</i> <i>seeds: calculadas con un bin_size: 0.5</i> <i>silueta: 0.392939</i>

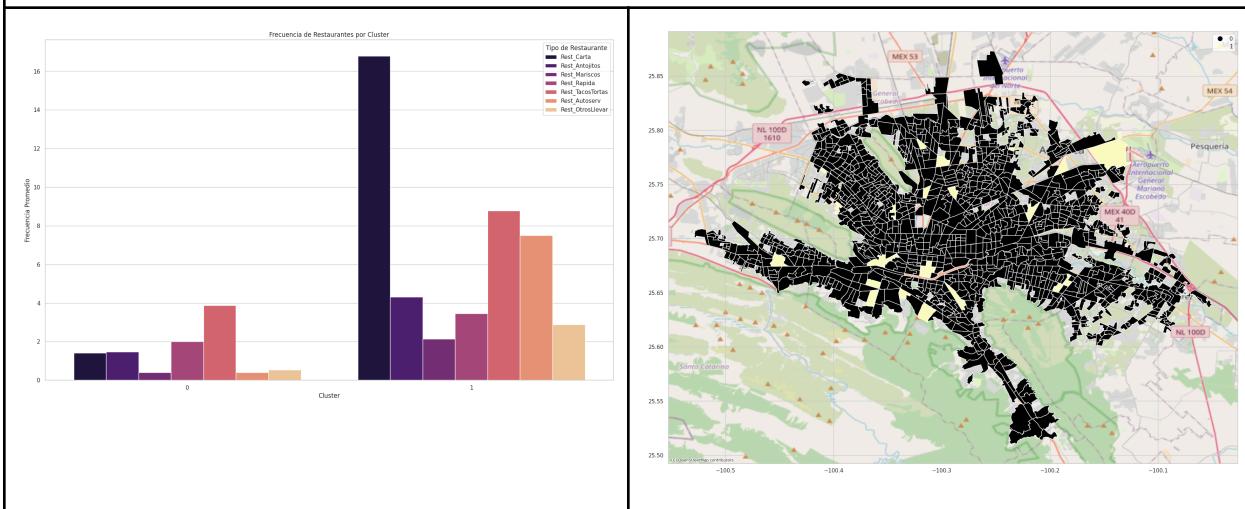
José David Banda	Gaussian Mixture Model	<i>Número de componentes: 7</i> <i>Tipo de covarianza: diag</i> <i>BIC score: 1491.60</i>	<i>Número de componentes: 6</i> <i>Tipo de covarianza: diag</i> <i>BIC score: 1840.19</i>	<i>Número de componentes: 4</i> <i>Tipo de covarianza: diag</i> <i>BIC score: 2120.87</i>	<i>Número de componentes: 5</i> <i>Tipo de covarianza: full</i> <i>BIC score: 2172.27</i>
Samantha Brito	Self-Organizing Maps	<i>n_clusters:3</i> <i>quan_err:1.152</i> <i>rand_sta:377</i>	<i>n_clusters:6</i> <i>quan_err: 1.156</i> <i>rand_sta: 877</i>	<i>n_clusters: 9</i> <i>quan_err: 1.158</i> <i>rand_sta: 3588</i>	<i>n_clusters: 3</i> <i>quan_err: 1.155</i> <i>rand_sta: 3964</i>

Clusters realizados por cada modelo representación

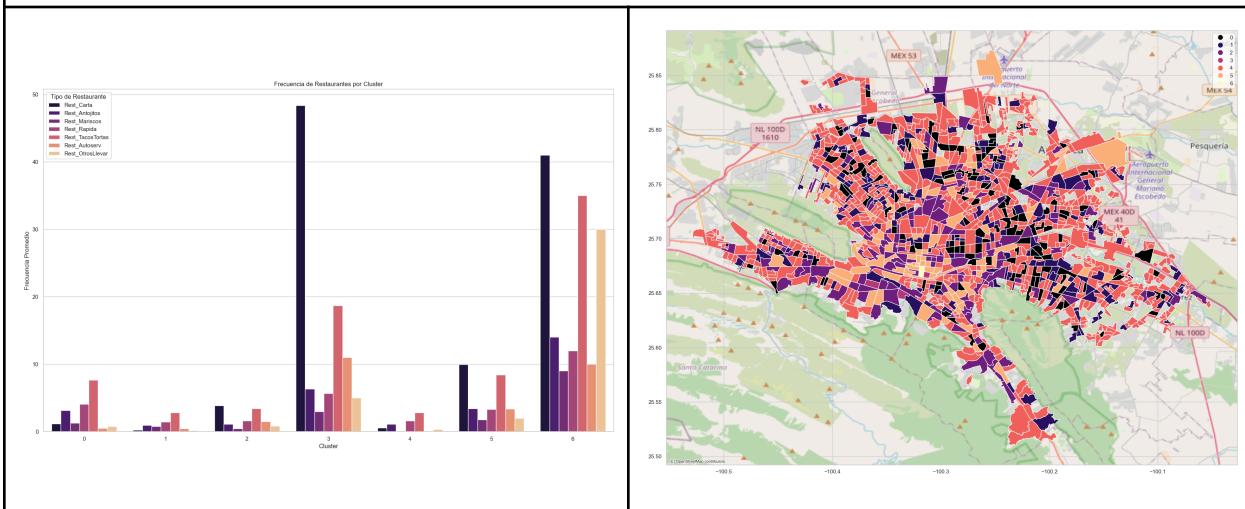




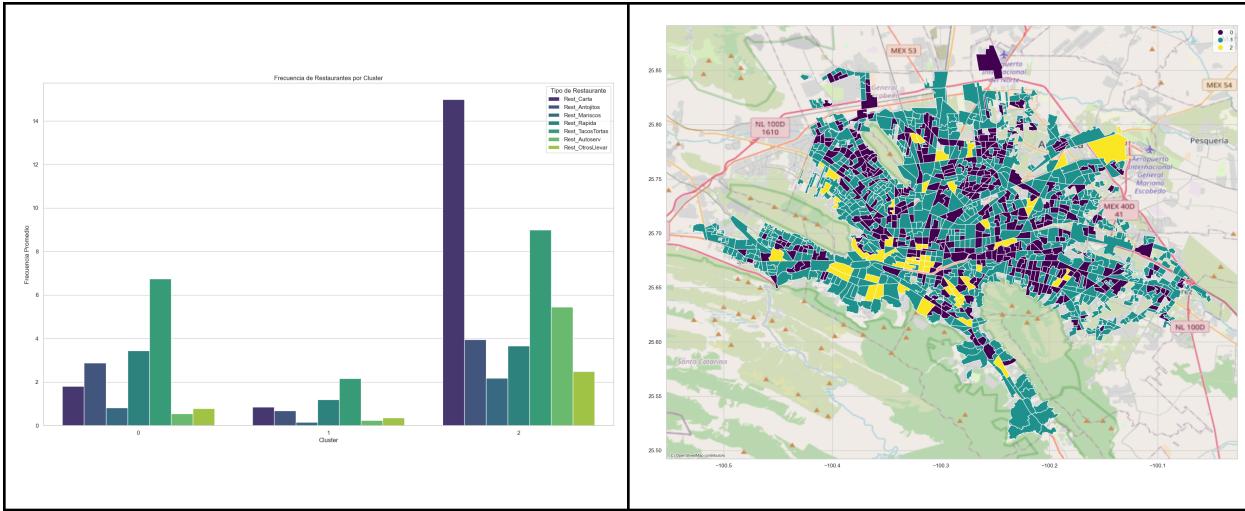
Mean Shift



Gaussian Mixture Model



Self-Organizing Maps



El modelo que se llegó a seleccionar para la predicción fué el de K-means, esto debido a que nuestro objetivo es poder predecir dado la categoría de restaurantes, en donde se puede establecer uno nuevo por la identificación de menor densidad de restaurantes dado un ageb. Por esta razón, al emplear la clusterización gracias al algoritmo de k-means, debido a que el funcionamiento del mismo se basa en la sumatoria de las distancias al cuadrado dentro de los clusters sea minimizado, por ende este llega a ser el mejor algoritmo dado al objetivo planteado. Ya que, como se busca conocer la región para el establecimiento de un local, al basarse en distancia y la minimización del mismo, es el que mejor llega a adaptarse al modelo planteado.

4 Selección y Despliegue

4.1 Ejemplo de aplicación del modelo

Se aplicó el modelo K-means debido a su gran capacidad para clusterizar los datos, junto con su gran rendimiento. De esta manera se agruparon en tres grandes grupos, con el fin de poder predecir dado la categoría de restaurantes, en donde se puede establecer uno nuevo por la identificación de densidad de restaurantes dado un ageb. Con el conjunto de datos que teníamos se aplicó el modelo a la zona Metropolitana de Monterrey , lo que nos da el siguiente resultado:

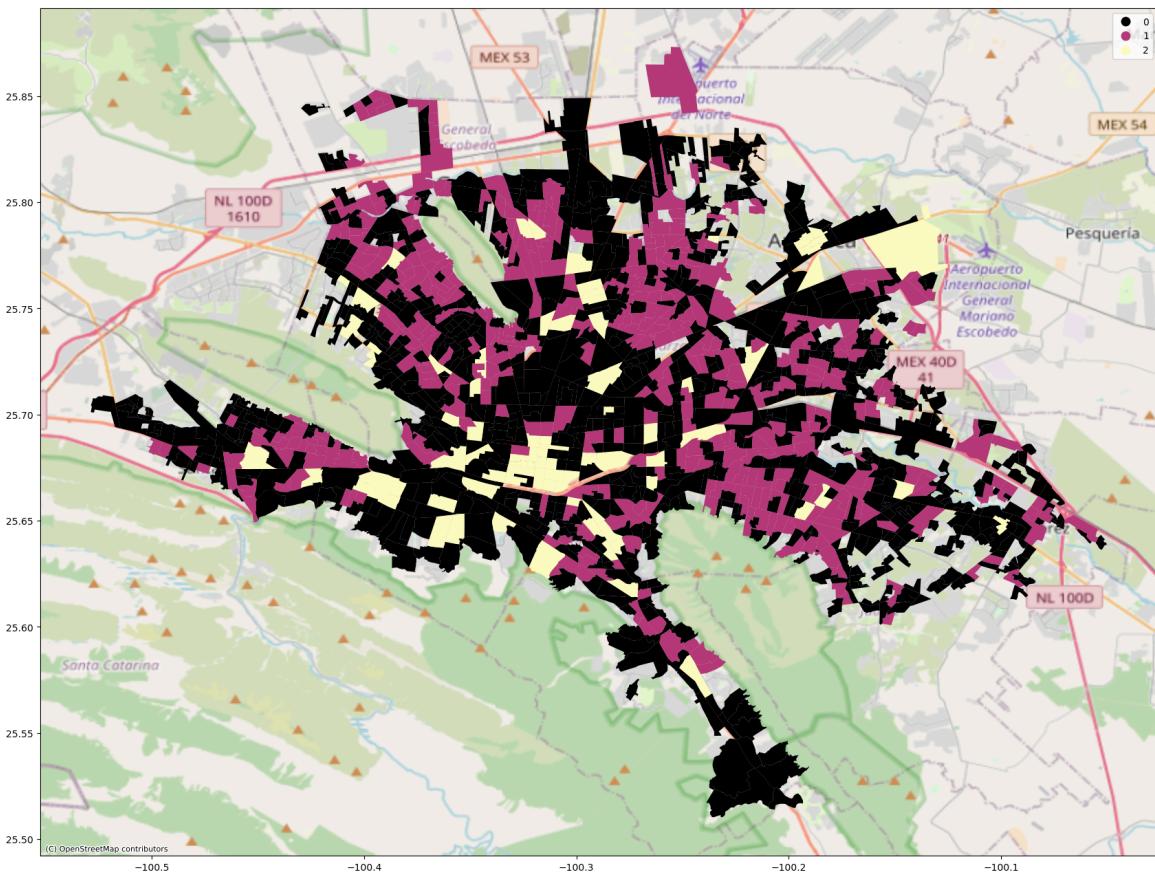


Figura 25.Creación de clusters con el mejor modelo

En el mapa se identifican claramente los clusters obtenidos, donde el más oscuro pertenece a la agrupación número 0 con 806 AGEB's, y la más clara a la formación número 2 con 78 AGEB's, mientras que la agrupación 1 contiene 403 AGEB's, de esta forma logramos categorizar la zona Metropolitana de Monterrey en restaurantes.

4.2 Interpretación de los resultados obtenidos, su impacto y utilidad.

A partir de nuestro modelo y los clusters generados , decidimos verificar la frecuencia de los restaurantes para cada tipo de cluster con el propósito de poder tener un mejor entendimiento de los clusters y la agrupación de que se hizo,

Cluster 0: Frecuencia relativamente alta en todos los tipos de restaurantes. Esto sugiere una zona con una variedad amplia de opciones de comida, que podría ser un área comercial o residencial densamente poblada.

Cluster 1: Frecuencias más bajas en general. Esto sugiere un área con menos opciones de comida y posiblemente menos densamente poblada, como áreas suburbanas o rurales.

Cluster 2: Frecuencias moderadas en la mayoría de los tipos de restaurantes. Esto podría indicar un área más equilibrada en términos de opciones de comida, posiblemente una combinación de áreas residenciales y comerciales.

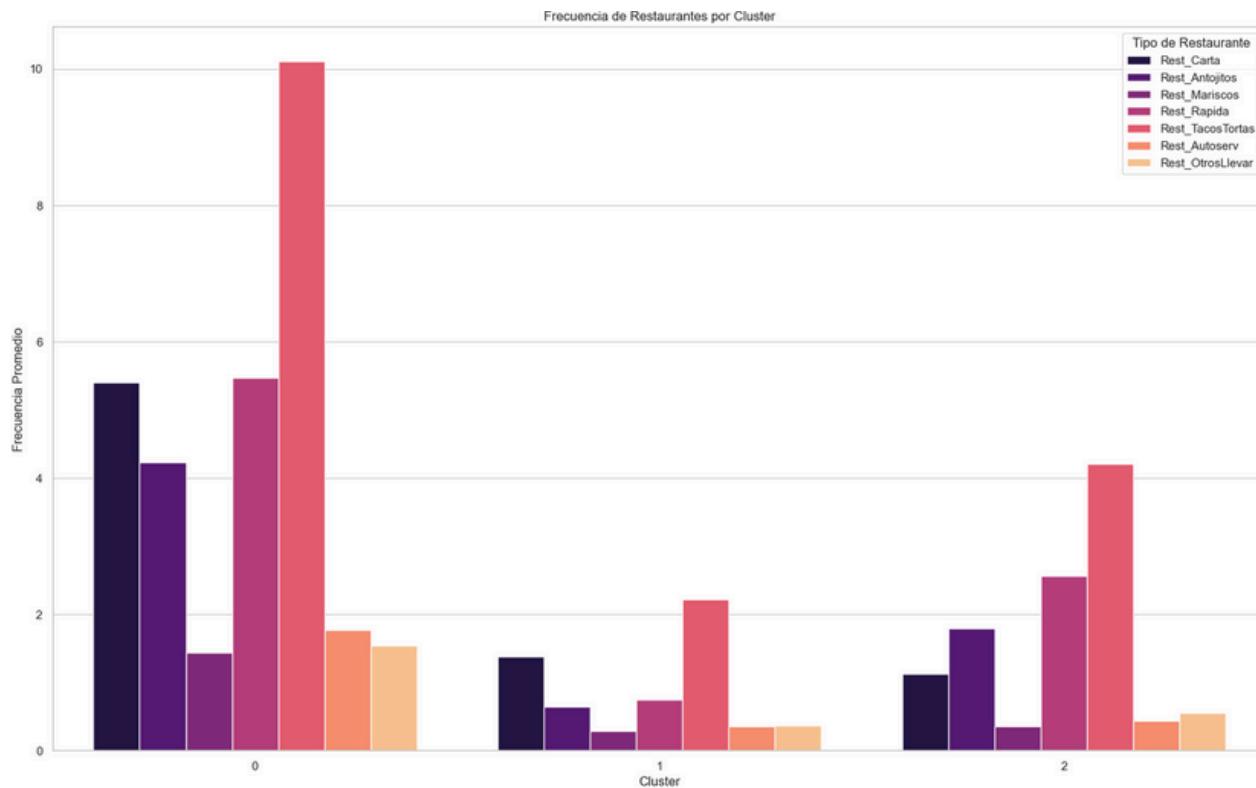


Figura 26. Frecuencia de restaurantes en cada cluster

De las imágenes que obtuvimos podemos apreciar que un restaurante del tipo de mariscos puede tener buena oportunidad, debido a que se encuentra en una zona residencial altamente poblada pero con poca competencia .

4.3. Descripción del prototipo y aplicación del modelo

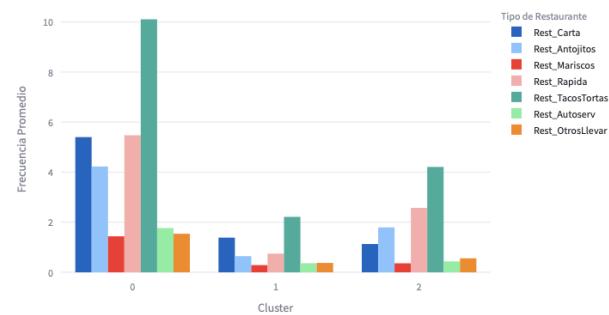
Una vez que se seleccionó qué modelo era el más adecuado para el modelado de los datos. Para tener una mejor interfaz con la cual el usuario fuera capaz de acceder a estos modelos desarrollados.

Se empleó Streamlit para el desarrollo de una página web, en la cual se muestran los histogramas de frecuencias con los datos de los restaurantes, donde el usuario es capaz de poder interactuar con el mismo. Por otro lado, se muestra el mapa de los clusters sobre un mapa, donde igualmente al emplear el área metropolitana de Monterrey, este es mostrado y el usuario es capaz de seleccionar y explorar sobre los clusters para poder ubicar sobre qué áreas y calles se llegan a encontrar.

Así mismo, se muestra una descripción de qué se encuentra en cada cluster y finalmente, se muestra un display dada la densidad de población por restaurante y los agebs, se encontraron los 3 clusters con mayor oportunidad de negocio, los cuales son mostrados al final con pinos, en donde muestra los restaurantes existentes y permite analizar para que el usuario pueda observar en donde puede poner un negocio del sector restaurantes.

Análisis de Clusters de AGEBS basados en Restaurantes

Frecuencia de Restaurantes por Cluster



Descripción de Clusters

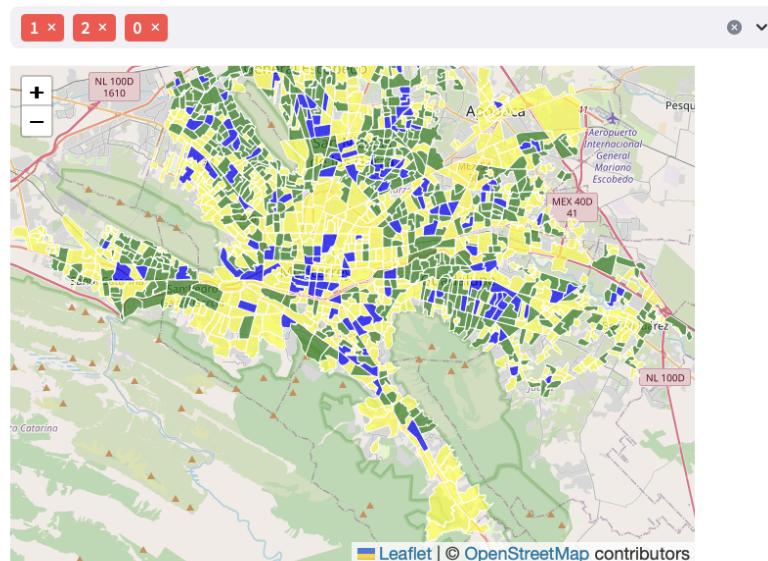
Selecciona un cluster:

0 1 2

Cluster 0: Tiene una frecuencia relativamente alta en todos los tipos de restaurantes, especialmente en los de alimentos a la carta o de comida corrida, tacos y tortas, y de autoservicio. Esto sugiere una zona con una variedad amplia de opciones de comida, que podría ser un área comercial o residencial densamente poblada.

Mapa de Clusters

Selecciona los clusters:



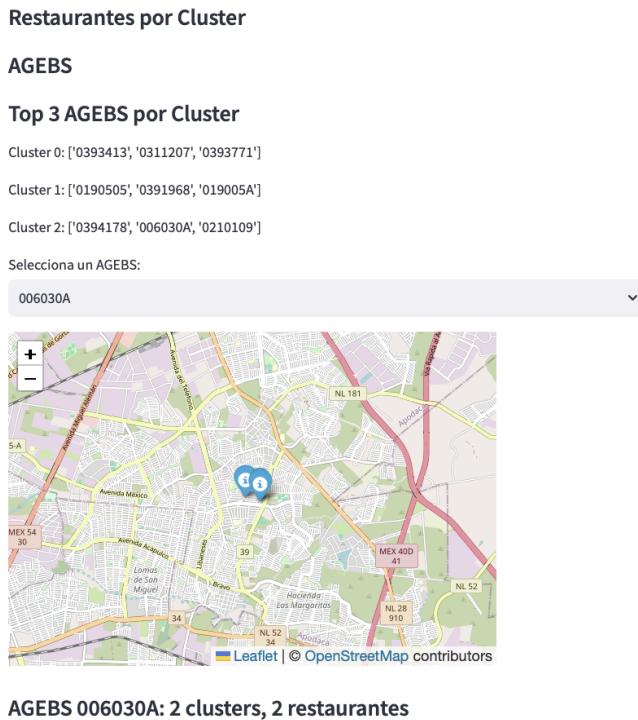


Figura 27. Deploy de la página web (prototipo final)

4.4. Interpretación de los resultados obtenidos, su impacto y utilidad

Finalmente, gracias a los resultados obtenidos, estos permiten derivado del uso y limpieza de las bases de datos del INEGI del DENUE y CENSO, poder la población interactuar de una manera interactiva y con una interfaz visual y gráfica, los datos sobre los restaurantes en el área metropolitana de la ciudad de Monterrey, lo cual tiene un impacto positivo ya que este tipo de bases de datos si bien son fuentes de información abiertas, no todas las personas llegan a tener la factibilidad de poder interpretar las mismas.

Por ende, al nosotros detectar como área de oportunidad poder realizar un modelado con este tipo de datos, esto permite el acceso a más personas para poder hacer uso de los mismos y tener la factibilidad de que llegue a ser muy visual y cualquier persona pueda interpretar este tipo de datos. Siendo estos útiles para los comerciantes en el área metropolitana de Monterrey que busquen poner un negocio tipo restaurante, y tengan información de antemano sobre la densidad del sector restaurantero, las ubicaciones y en donde les beneficiaría poner uno, teniendo así un poco de certeza sobre la ubicación y una pequeña probabilidad de que el restaurante puede ser revolucionario en dicha zona.

5 Conclusiones y recomendaciones

5.1 Conclusiones

La elaboración de este proyecto incluyó la descarga de los datos de la plataforma del INEGI, la limpieza e integración de estos en un solo archivo para su posterior uso. Se realizó un análisis exploratorio de datos para conocer sus características. Después, un método de aprendizaje no supervisado de clustering permitió segmentar el Área Metropolitana de Monterrey con base en las similitudes que tienen los AGEBS en relación al tipo de restaurantes. Esto, junto con la creación de la variable Personas por Restaurante, proporcionan información valiosa para encontrar oportunidades de negocio, conociendo la prevalencia de cada tipo de restaurantes para cada cluster e identificando aquellos agebs donde existe mayor demanda sin satisfacer. Finalmente, el despliegue de esta información en forma de página web como prototipo final, permitió la implementación y summarización de todo el proceso llevado a cabo en forma de una página, con la cual los usuarios son capaces de interactuar con la misma y poder obtener información de una forma más fácil e interactiva sin la necesidad de acceder directamente a las bases de datos del INEGI.

5.2 Recomendaciones

- Estandarizar el layout entre las diferentes bases de datos y a través de los años.
- Abrir la exploración y análisis de los datos a distintas áreas de negocios no solamente restaurantes
- Incluir nuevas columnas de información sobre la demografía de la población

Referencias:

- [1] Andrews, J. (2023, October 2). *Revolutionizing Supply Chains: The Power of Predictive Analytics for Inventory Optimization*. LinkedIn.com.
<https://www.linkedin.com/pulse/revolutionizing-supply-chains-power-predictive-justin-andrews>
- [2] Argote Cusi, Milenka Linneth. (2015). Análisis de sensibilidad de proyecciones de población. *Papeles de población*, 21(84), 45-67. Recuperado en 15 de abril de 2024, de
http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-74252015000200003&lng=es&tlang=es.
- [3] Arias, L. (2023, November 8). *Metodología CRISP-DM: La guía definitiva para la Minería de Datos*. LinkedIn.com.
<https://www.linkedin.com/pulse/metodolog%C3%ADA-crisp-dm-la-gu%C3%ADa-definitiva-para-miner%C3%ADA-de-arias-xyusf>
- [4] Business Wire. (2020, February 7). How Predictive Analytics Is Transforming the Transport and Logistics Industry.
<https://0-link-gale-com.biblioteca-ils.tec.mx/apps/doc/A613346317/PPES?u=itesmgic&sid=ebsco&xid=368899ee>
- [5] Espinosa-Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería, Investigación y Tecnología*, 21(1), 1-13. <https://doi.org/10.22201/fi.25940732e.2020.21n1.008>
- [6] Fernández, C., & Aqueveque, C. (2001). Segmentación de mercados: buscando la correlación entre variables sociológicas y demográficas. *Revista Colombiana de Marketing*, 2(2).
- [7] Heckmann, L. (2024). Crystal Ball: Pentagon Harnessing Data for Predictive Logistics Planning. *National Defense*, 108(844), 26–27.
- [8] IBM Documentation. (2021, August 17). Ibm.com.
<https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview>

- [9] Instituto Nacional de Estadística y Geografía (INEGI). (2020). *Censo de Población y Vivienda 2020*. Inegi.org.mx.
<https://www.inegi.org.mx/programas/ccpv/2020/#herramientas>
- [10] Instituto Nacional de Estadística y Geografía (INEGI) (2020). *SCITEL*. Inegi.org.mx.
<https://www.inegi.org.mx/app/scitel/Default?ev=10>
- [11] Instituto Nacional de Estadística y Geografía (INEGI). (2024). *Marco Geoestadístico*. Inegi.org.mx. <https://www.inegi.org.mx/temas/mg/>
- [12] Instituto Nacional de Estadística y Geografía (INEGI). (s/f). *Quiénes somos*. Recuperado el 10 de abril de 2024, de https://www.inegi.org.mx/inegi/quienes_somos.html
- [13] Dey, V. (2024, 18 marzo). Beginners Guide to Self-Organizing Maps. Analytics India Magazine. <https://analyticsindiamag.com/beginners-guide-to-self-organizing-maps/>
- [14] Perry, P. (2009, septiembre 16). Cross-Validation for Unsupervised Learning. Cornell University. <https://arxiv.org/abs/0909.3052>
- [15] Verdejo, E. (2014). Arquitectura de integración para el desarrollo de un sistema de geo-recomendación para establecer puntos de venta. Instituto Tecnológico de Orizaba. https://www.rcs.cic.ipn.mx/2016_113/Arquitectura%20de%20integracion%20para%20el%20desarrollo%20de%20un%20sistema%20de%20geo-recomendacion.pdf