UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Aprendizaje automático para el análisis del comportamiento del comercio al por menor en México durante los últimos años

ARTÍCULO ACADÉMICO

Que para obtener el título de

Ingeniera en Computación

P R E S E N T A

Patricia Soto Vázquez

ASESOR DE ARTÍCULO ACADÉMICO

Dr. Guillermo Gilberto Molero Castillo

Ciudad Universitaria, Cd. Mx., 2022

Resumen

En años recientes, los métodos de aprendizaje automático han ganado protagonismo al ser utilizados como herramienta para el análisis de datos en diferentes áreas como la economía. El presente documento, a manera de tesina, expone información relevante que resulta del análisis de datos del comercio al por menor en México y las ramas industriales que lo conforman. Dicho análisis es resultado del desarrollo del proyecto de investigación, realizado bajo la Modalidad de Titulación por Actividad de Investigación, aprobado por el Comité de Titulación de la División de Ingeniería Eléctrica de la Facultad de Ingeniería. **Objetivo.** Emplear un algoritmo de aprendizaje automático, para el análisis del comportamiento del comercio al por menor en México durante los últimos años. **Método.** Para el análisis de datos, se utilizó el aprendizaje no supervisado, específicamente el agrupamiento basado en K-means. Este método no supervisado se realizó aplicando un algoritmo, cuyo proceso de entrenamiento se basó en un conjunto de datos proporcionados de manera libre por el INEGI (Instituto Nacional de Estadística, Geografía e Informática). Así, mediante K-means fue posible organizar clústeres con información sobre las características de las diferentes ramas industriales del comercio minorista. **Resultados.** Como resultado de la aplicación del algoritmo K-means, se realizó una categorización de las ramas industriales analizadas en cinco grupos, de los cuales se destacan sus características más sobresalientes. **Conclusiones.** Fue posible identificar que el comercio al por menor, fuente de una gran cantidad de empleos en México, posee características que permiten comprender su comportamiento y reconocer tendencias a lo largo de los años. Además, la información obtenida establece las bases para la realización de análisis que permitan impulsar el desarrollo de este sector, proponiendo soluciones que le permitan crecer dentro del mercado agresivo y exigente en el que subsiste.

Índice

Capítulo 1
Introducción

1.1. Contexto de la investigación

El mundo actual se desarrolla en un contexto que demanda el uso de la información como fuente de conocimiento, que no solo se limite a comprender aquello que los datos representan, sino que permita generar valor en cualquier área que sea aplicado. Así, a raíz de esta búsqueda constante y necesaria de hacer uso pleno del conocimiento, la Inteligencia Artificial (IA) aplicada en la Industria 4.0, entendida como ciencia, beneficia a la industria en el proceso de la transformación digital, con el propósito de generar conocimiento impulsado principalmente por el aprendizaje automático y aprendizaje profundo (Molero-Castillo *et al*, 2018).

El conocimiento, en materia de economía, se construye con base en el análisis de la información que deriva de la actividad económica de un país (Montuschi, 2001). De la categorización de la actividad económica, surgen los sectores económicos, pilares en el crecimiento y desarrollo económico (Piedras, 2006; Lee y Shin, 2020). Por lo tanto, dentro del contexto económico en el que México se desarrolla, es necesaria una economía basada en conocimiento, siendo el aprendizaje automático un pilar importante en el empuje hacia la cuarta revolución industrial, que ha llevado de la mano la aplicación de algoritmos para el análisis de datos con el propósito de construir un sistema económico eficiente y de calidad.

En este sentido, ante esta necesidad creciente del análisis de datos en el ámbito económico, el uso del aprendizaje automático, como una herramienta de apoyo para la analítica avanzada de datos, es fundamental, puesto que ofrece una amplia variedad de algoritmos (Hansen, 2018; Kumar *et al*, 2021), dentro de los que se encuentran los supervisados, no supervisados, por reforzamiento, profundo y mixtos; logrando tener en la

actualidad un importante posicionamiento como respuesta a la amplia digitalización y almacenamiento de los datos.

Por otro lado, el aprendizaje automático actual no solo está cambiando la forma en la que se produce, comercializa y vende un producto, sino también forma parte del estudio del crecimiento económico, determinado por el aumento de la productividad y los ingresos de un país (Quiroga Persivale, 2018; Mathur, 2019). Dicho estudio establece las bases para analizar el desarrollo económico, medido con base en las mejoras de las condiciones de vida de la población. No cabe duda que el desarrollo de nuevos productos y empresas, con niveles de automatización y robotización sin precedentes, pueden transformar de forma transversal la economía y el mercado laboral.

1.2. Problema de investigación

La economía mexicana se sostiene gracias a la participación de diferentes actividades y unidades económicas. Estas últimas representan el lugar o la entidad donde se realizan estas actividades, siendo importante no solo las fábricas, cadenas de tiendas, oficinas, escuelas u hospitales a lo largo del territorio mexicano, sino también un espacio de vivienda que funge como establecimiento comercial, o incluso un trabajador por cuenta propia sin establecimiento (Instituto Nacional de Estadística y Geografía [INEGI], 2021b). Todas estas unidades económicas son imprescindibles para la economía del país, gracias a su enorme participación en el mercado.

En este sentido, para entender como el comercio está directamente relacionado con la calidad de vida de las personas, es meritorio enfocar esfuerzos en el análisis de uno de los sectores económicos que tiene un importante impacto en la económica mexicana, este es, el comercio al por menor, el cual es la actividad económica definida por la venta individual de bienes y servicios directamente a consumidores finales (INEGI, 2004). Dicha actividad, por su naturaleza, forma parte de la cadena de suministro gracias a su modelo enfocado en la venta entre empresa y consumidor.

Este tipo de comercio es un sector fundamental en México, puesto que, en términos del producto interno bruto (PIB), las actividades terciarias tuvieron una estructura porcentual anual de 60%, correspondiente al 2020, dentro de la cual 9.2% corresponde al comercio al por menor (INEGI, 2021a; INEGI 2021b). Por otro lado, este tipo de comercio destaca por su importancia, puesto que además de su considerable participación porcentual respecto al PIB, concentra también una amplia población que encuentra, en este sector, una fuente de empleo.

Aunado a lo anterior, la desinformación sobre el comportamiento de la actividad económica no solo lleva a las personas y empresas a tener un mal manejo de sus negocios, sino que además fomenta el desinterés para establecer medidas o legislaciones que beneficien al comercio al por menor (Arana, 2018). Por lo que, el análisis del comercio al por menor es fundamental en pro del entendimiento del desarrollo y crecimiento económico de una determinada región a lo largo de los años. Este tipo de análisis se puede lograr con base en la observación de similitudes, tendencias y comportamientos, para los cuales los algoritmos de aprendizaje automático son útiles. El propósito es emplear este tipo de algoritmos para identificar evidencia en forma de patrones a partir de los datos, con los cuales se puede hacer un análisis informado y reflexivo sobre la situación actual del comercio al por menor y su impacto en la economía nacional.

1.3. Objetivos

1.3.1. Objetivo general

- Emplear algoritmos de aprendizaje automático para el análisis del comportamiento del comercio al por menor en México durante los últimos años.

1.3.2. Objetivos específicos

- Adquirir y hacer un análisis exploratorio de datos sobre el comercio al por menor en México.

- Establecer el número adecuado de grupos mediante la aplicación de un método para segmentación de datos.
- Identificar características, tendencias y similitudes en el conjunto de vectores de datos con base en un tipo de aprendizaje no supervisado.
- Interpretar los resultados obtenidos, de manera que se permita mostrar una visión sobre el comercio al por menor en México, su economía y población dedicada a este.

1.4. Justificación

La observación de datos en el sector financiero se ha vuelto clave para el análisis del crecimiento económico en una determinada región, el cual está determinado por el aumento de la productividad y los ingresos dentro de su territorio (Diferenciador, 2020). Así, en virtud de los datos que indican un importante posicionamiento del comercio al por menor en la economía de México, es oportuno analizar la importancia de dicho comercio en la población mexicana, puesto que es visible el tamaño de estas unidades económicas en la población ocupada.

En la actualidad, el sector de las pequeñas y medianas empresas (PYMES) es uno de los más vulnerables, puesto que como cualquier negocio necesitan de una correcta administración de sus ingresos financieros, sin embargo, la mayoría carecen de ésta (Pavón, 2016). Estos ingresos dependen, en gran medida, de una adecuada gestión financiera, de la que muchas PYMES carecen. Por lo que, de acuerdo con los datos del Centro de Desarrollo para la Competitividad Empresarial, el 75% de las PYMES cierran sus operaciones apenas dos años después de haber sido creadas. Además, el Instituto Nacional de Geografía y Estadística (INEGI) señala que las empresas de nueva creación en México solo viven en promedio 7.7 años (INEGI, 2016).

Por lo tanto, es notable el impacto que tiene el comercio al por menor y en su conjunto las pequeñas y medianas empresas en la economía del país. En consecuencia, enfocar un análisis de datos en este sector, a través de algoritmos de aprendizaje automático, se vuelve importante en momentos no únicamente de decrecimiento económico, sino también de crisis,

como el que se está viviendo en los últimos años a consecuencia de la pandemia por COVID-19.

1.5. Organización del documento

El documento está organizado de la siguiente manera, el Capítulo 2 presenta los antecedentes de la economía como Ciencia Social, se mencionan algunas de las aportaciones más significativas del aprendizaje automático en la economía, se discuten sus aplicaciones, el uso de algoritmos basados en segmentación (clustering) y se presentan los trabajos relacionados. El Capítulo 3 describe el método definido como propuesta de solución. El Capítulo 4 presenta los resultados obtenidos, basado en datos del comercio minorista en México, y el Capítulo 5 resume las principales conclusiones y el trabajo futuro.

Se presenta además tres anexos, en los que se incluye información relacionada sobre el trabajo de investigación efectuado. En el Anexo A se presenta la carta de aceptación de la publicación del artículo de investigación en la revista Research in Computing Science. El Anexo B muestra el artículo de investigación aceptado para su publicación en la revista mencionada (www.rcs.cic.ipn.mx), cuyo título es ‘Machine learning for the retail trade behavior analysis in Mexico’. En el Anexo C se presenta el código en Python de los métodos de aprendizaje automático utilizados para el análisis de comportamiento del comercio minorista en México.

Capítulo 2
Marco teórico y estado del arte

2.1. Comercio al por menor

La economía estudia la forma en la que participan las personas y las organizaciones de la sociedad, ya sea en la producción, distribución y consumo de bienes y servicios (Ávila-Lugo, 2007). Esta participación económica se entiende como un conjunto de actividades de compra-venta que realiza el ser humano con el objetivo de satisfacer sus necesidades. A partir de la concepción de dicha participación, es posible abstraer el concepto de comercio, el cual hace referencia al intercambio de bienes, generalmente a cambio de dinero.

El comercio al por menor está definido por unidades económicas, dentro de las cuales se encuentran diversos establecimientos que están bajo el control de una entidad propietaria, asentada de manera permanente y delimitada por instalaciones fijas (INEGI, 2021c). Además, estas unidades económicas, están situadas en diversos niveles geográficos, por ejemplo, país, estado, municipio y localidad, donde cumplen la función de permitir actividades de compra-venta de mercancías, o prestación de servicios, independientemente de si tienen o no fines mercantiles (Rajesh Kumar *et al.*, 2021; Gabel *et al.*, 2019). En este grupo participan los micronegocios y las pequeñas y medianas empresas (INEGI, 2021c).

Convencionalmente, el estudio del crecimiento económico se basa en el análisis de indicadores como el PIB, gracias al cual es notable la importante participación del comercio al por menor en la economía de México. Si bien el PIB no es un indicador suficiente para determinar el crecimiento económico del país, es uno de los más importantes, gracias a que un crecimiento en este indicador fácilmente podría traducirse en el alza de empleos. Sin embargo, el PIB no es el único indicador que expone la importancia del comercio al por

menor en la economía mexicana, sino que también los indicadores de ocupación y empleo proporcionan información relevante en este rubro.

2.2. Indicadores económicos

Con respecto a la población ocupada por tamaño en la unidad económica, los indicadores de ocupación y empleo establecen la existencia de 20.1 millones de personas ocupadas en micronegocios, 7.5 millones en pequeños establecimientos, y 5.2 millones en medianos establecimientos (INEGI, 2020). Esta información es representativa, puesto que tiene una cobertura del 63.9% (84556) de las viviendas de la Encuesta Nacional de Ocupación y Empleo. Entre los organismos que ofrecen información sobre las actividades, indicadores económicos y mercado laboral destacan:

- La Encuesta Nacional de Ocupación y Empleo (ENOE) es la principal fuente de información sobre el mercado laboral. Constituye el proyecto estadístico más grande del país, debido a que ofrece datos mensuales y trimestrales sobre la fuerza de trabajo, ocupación, informalidad laboral, subocupación y desocupación (INEGI, 2021d). El 2020 difundieron las características ocupacionales de la población de 15 años a más, así como de variables demográficas y económicas para el análisis de la fuerza de trabajo.

- La Encuesta Anual del Comercio (EAC) proporciona información sobre las actividades comerciales y brinda un panorama estadístico frecuente que coadyuven en la toma de decisiones de los diferentes sectores productivos del país (INEGI, 2019a). La EAC tiene como base la Encuesta Mensual sobre Empresas Comerciales (EMEC), cuyo propósito principal es la generación de información estadística.

- La Organización Mundial del Comercio (OMC) es la única organización internacional, de la que México forma parte, que se ocupa de las normas que rigen el comercio entre los países. Su objetivo es garantizar que los intercambios comerciales se realicen de forma fluida, previsible y libre (Organización Mundial del Comercio, 2021). En sus publicaciones contemplan el comercio al por mayor y menor, siendo una de las más

recientes: Helping MSMES Navigate The Covid-19 Crisis, donde se explica cómo se han visto afectadas las PYMES ante la crisis por la pandemia COVID-19 y cómo prevalecen en los sectores económicos más afectados por el choque derivado de la demanda de bienes y servicios.

En México, el comercio minorista no ha permanecido inmutable, ya que ha cambiado su estructura y dinámica a través del tiempo (Bocanegra-Gastelum, 2008), es por ello que entender en que aspectos dicho sector permanece constante o cambia, se vuelve fundamental.

Frecuentemente, se analiza el dinamismo del comercio al por menor desde la perspectiva de las grandes empresas y su posicionamiento en el mercado, ya que dicho estrato proporciona grandes cantidades de información debido a la capacidad de los mismos de tener fácil acceso a medios tecnológicos que les permiten la recopilación masiva de datos, no obstante, gracias a que diferentes organizaciones enfocan esfuerzos en recabar datos de los micros, pequeños y medianos negocios cada determinado periodo de tiempo, es que es posible dirigir esfuerzos en analizar el comercio minorista desde otra perspectiva, sin embargo, los datos sin procesar no aportan información significativa.

Bajo esta idea, el uso de Inteligencia Artificial como herramienta de mejora ante la necesidad de impulsar el desarrollo tecnológico, ha ganado terreno en diferentes ramas de la economía, del mismo modo el aprendizaje automático, siendo una rama de la Inteligencia Artificial, utiliza métodos basados en algoritmos que permiten dar sentido a datos reconociendo patrones y prediciendo comportamientos, lo cual es de gran ayuda para comprender el comportamiento de sectores económicos.

2.3. Aprendizaje automático basado en clustering

El aprendizaje automático es uno de los principales dominios de la Inteligencia Artificial, consiste en un conjunto de algoritmos para el análisis impulsado por datos, que permiten establecer modelos, a partir de los datos de ejemplos o experiencias, para entrenar a las máquinas (computadoras) y aprender a partir de estos (Palma-Méndez y Merín-Morales, 2008; Mathur, 2019). Dentro del aprendizaje automático, los métodos no supervisados son

algoritmos que basan su proceso de entrenamiento en un conjunto de datos sin etiquetas o clases, previamente definidas. Es decir, no se conoce ningún valor objetivo o de clase, ya sea categórico o numérico. Por lo tanto, estos métodos no requieren de la intervención humana (Rouhiainen, 2018).

En este sentido, las principales aplicaciones del aprendizaje no supervisado están relacionadas con el agrupamiento de datos, donde el objetivo es encontrar grupos con elementos similares, de tal forma que los objetos internos de un grupo tengan una alta similitud entre ellos, y sean diferentes (disimiles) con objetos de otros grupos (García Cambronero y Gómez Moreno, 2006). Existen dos grupos principales de algoritmos de agrupamiento (Pla *et al.*, 2007): i) los métodos jerárquicos, que producen una organización jerárquica de los elementos que forman el conjunto de datos, posibilitando de esta manera distintos niveles de agrupación; y ii) los métodos particionales, que generan grupos de elementos que no responden a ningún tipo de organización jerárquica. Estos algoritmos se basan en la distancia entre elementos.

Un algoritmo de agrupamiento jerárquico, de tipo aglomerativo, comienza la agrupación a partir de cada elemento individual, considerándolo como un grupo unitario; y de manera iterativa se unen los dos grupos más cercanos hasta obtener un único grupo general. Los elementos similares con menor distancia son los primeros en unirse, y continúan uniéndose en forma progresiva de acuerdo a sus similitudes. En contrapartida, los métodos jerárquicos divisivos trabajan en el sentido opuesto. Un único grupo inicial de elementos se divide en dos subgrupos, de manera que los elementos de un subgrupo sean disímiles con los elementos del otro. Estos subgrupos son divididos sucesivamente hasta que queden tantos subgrupos como cantidad de elementos (Pla *et al*., 2007).

Los algoritmos particionales asumen un conocimiento a priori del número de grupos en el que deben ser divididos el conjunto de datos, esto es, llegan a una división que optimiza un criterio predefinido (Soto *et al.*, 2006). Entre los algoritmos que emplean este tipo de agrupamientos destaca K-means, cuya idea principal es definir k centroides (uno para cada grupo) y luego tomar cada elemento (registro) de la base de datos y situarlo en el grupo del centroide más cercano.

2.4. K-means

El algoritmo K-means, pertenece al conjunto de algoritmos particionales dentro del aprendizaje no supervisado. Este algoritmo fue propuesto por J. MacQueen en 1967 (MacQueen, 1967), su objetivo principal es calcular una partición óptima de elementos. Consiste en un proceso de segmentación en el cual un conjunto de elementos se divide en un determinado número de grupos. Tiene su fundamento en la idea de agrupar los elementos de acuerdo a su media, dicha media es llamada centroide.

El centroide es un punto que ocupa la posición media en un grupo. El algoritmo se realiza de manera iterativa con base en la cantidad de centroides que se establezcan, es decir, tiene la característica de inicializar el número de grupos en el que se dividirán los elementos, sin embargo, para dar solución a este inconveniente, en K-means se hace uso del método del codo (Elbow Method), con diferente número de clústeres (configuraciones de k), para tener una aproximación sobre el número adecuado de grupos (Martín-Pérez, 2018).

Esta aproximación del número adecuado de grupos se logra calculando la suma del error al cuadrado (SSE, por sus siglas en inglés) entre los elementos de cada grupo y su centroide. El objetivo es mostrar de manera gráfica los diferentes resultados de SSE para cada configuración de k (Nainggolan *et al*., 2019). Las etapas del algoritmo del método de codo, para determinar el valor adecuado de grupos en K-means, son las siguientes:

1. Inicializar el valor inicial de k.
2. Aumentar el valor de k.
3. Calcular la suma de los resultados de SSE de cada valor de k, y con dichos resultados trazar la curva que permita su identificación.
4. El análisis de SSE resulta del valor de k en el que la varianza intraclúster disminuye drásticamente.
5. Ubicar el valor k en forma de codo.

De manera general, en el algoritmo K-means, inicialmente los centroides se calculan de manera aleatoria y se asignan los elementos a su centroide más cercano. Posteriormente, en cada iteración, se recalcula el centroide de cada grupo y con base en los nuevos centroides se distribuyen todos los elementos según el centroide más cercano. El proceso se repite hasta que ya no existan cambios en los grupos formados. En este sentido, la aplicación del método del codo es una herramienta clave, para que los grupos en los que se basará el análisis proporcionen información relevante, evitando sesgo en los resultados.

2.5. Trabajos relacionados

En los últimos años, el comercio en México es objeto de estudio debido a su importancia, no solo en la economía nacional e internacional, sino también por su impacto en la población mexicana. En este sentido, el análisis del comercio al por menor es fundamental en beneficio del entendimiento del desarrollo y crecimiento económico de una determinada región a lo largo de los años. Este tipo de análisis se puede lograr con base en la observación de similitudes, tendencias y comportamientos, para los cuales los algoritmos de aprendizaje automático son útiles. Algunos trabajos relacionados con la agrupación de datos y algoritmos de particionamiento en materia de economía son:

- *Data mining and machine learning in retail business: developing efficiencies for better customer retention* (Rajesh Kumar *et al.*, 2021), presenta un análisis del marketing minorista y discute la aplicación de técnicas de minería de datos y aprendizaje automático. Destaca dentro de la metodología, el uso de K-means como una herramienta para la identificación de puntos de datos incompletos, dicho algoritmo, en conjunto con un método de predicción de intereses y generación de patrones, permitió identificar patrones de compra de los registros de un usuario.

- *P2V-MAP: Mapping Market Structures for Large Retail Assortments* (Gabel *et al.*, 2019), analiza estructuras del mercado, a través de avances en el procesamiento del lenguaje natural y el aprendizaje automático. Además, el enfoque utilizado permite comparar técnicas de reducción de la dimensionalidad que muestran una aportación relevante para el análisis del mercado, ya que impulsa la aplicación de técnicas de

aprendizaje automático para proponer soluciones a problemas relacionados con las estructuras del mercado para minoristas.

- *Machine learning for enterprises: Applications, algorithm selection, and challenges* (Lee y Shin, 2020), expone la importancia del aprendizaje automático aplicado a las empresas, con el objetivo de impulsar su desarrollo tecnológico y con ello, reducir costos de productos y servicios; promoviendo de esta forma, la aceleración de los procesos comerciales. Se discute el uso de métodos de agrupamiento, clasificación y predicción, además de los desafíos al aplicar algoritmos de aprendizaje automático. No obstante, también se analizó el incremento en la implementación de herramientas y técnicas de aprendizaje automático por las empresas para impulsar su potencial.

- *Sustainability of SMEs in the Competition: A Systemic Review on Technological Challenges and SME Performance* (Prasanna *et al*, 2019), expone la importancia de las pequeñas y medianas empresas como un motor de desarrollo económico. Además, expone los retos a los que se enfrentan y se revisa la necesidad del progreso tecnológico para impulsar la innovación en la economía, revelando los efectos positivos que tiene sobre los niveles de producción y crecimiento económico. Por otra parte, se plantea la adopción de tecnologías de la información como un medio para enfrentar los retos competitivos a lo que las PYMES se enfrentan.

El trabajo de análisis realizado por diversos investigadores, enfocado en el comercio que tiene como base el aprendizaje automático y donde se emplean diversos algoritmos especializados, como los de segmentación de datos, no es suficiente en materia de economía mexicana, por lo tanto, surge la necesidad de utilizar estos métodos y algoritmos para identificar evidencia en forma de patrones a partir de los datos, gracias a los cuales es posible hacer un análisis informado y reflexivo sobre la situación actual del comercio al por menor y su impacto en la economía nacional.

Dada la falta de enfoque al comercio minorista aplicando algoritmos de aprendizaje no supervisado en aspectos que vayan más allá de la estructura del mercado y la mercadotecnia, es que se vuelve primordial hacer uso de métodos innovadores en el análisis de datos

correspondientes al comercio al por menor en México, es por ello que la investigación actual está enfocada a analizar elementos que constituyen diferentes actividades inherentes del comercio minorista.

Capítulo 3
Método de solución

El método de trabajo definido para el análisis del comportamiento del comercio al por menor en México durante los últimos años, a través de aprendizaje automático, fue dividido en cuatro etapas: *i*) adquisición de datos, *ii*) análisis exploratorio de datos, *iii*) selección de variables, y *iv*) aplicación del algoritmo.

3.1. Adquisición de datos

Los datos analizados se obtuvieron de la Encuesta Anual de Comercio. Esta encuesta parte de las unidades económicas provenientes del Marco Estadístico Nacional de Unidades Económicas (MENUE), la cual se alimenta del Registro Estadístico de Negocios de México (RENEM) con variables de diseño referenciadas (INEGI, 2020b). Esta fuente de datos fue adquirida a través del sitio web oficial del Instituto Nacional de Estadística, Geografía e Informática, www.inegi.org.mx/app/descarga/ficha.html?tit=110334&ag=0&f=csv, la cual está estructurada en matrices de datos sobre los principales indicadores económicos de la actividad comercial por sector, subsector y rama de actividad a nivel nacional. Estos indicadores son útiles para comparar y analizar las tendencias y factores que influyen en el comportamiento de la actividad comercial en México.

La información global está conformada por 40 ramas de actividad económica y está integrada por 2134549 empresas del sector comercial. De estas, 18 ramas pertenecen al comercio al por mayor (integrado por 126933 empresas) y 22 al comercio al por menor (integrado por 2007616 empresas), siendo este último el objeto de estudio en este trabajo de investigación. Además, las variables utilizadas están determinadas por el Sistema de

Clasificación Industrial de América del Norte, el cual permite crear agrupaciones de manera sistemática, siempre bajo una misma lógica, lo que ayuda a evitar controversias y errores de interpretación (Sistema de Clasificación Industrial de América del Norte, 2021). Asimismo, dentro del comercio al por menor, la estratificación por número de trabajadores de cada empresa según el INEGI está definido como (INEGI, 2019b): i) micro (de hasta 10 personas), ii) pequeña (11 a 30 personas), y iii) mediana (31 a 100 personas).

3.2. Análisis exploratorio de datos

El periodo de análisis comprende de 2016 a 2019, puesto que en 2020 se publicaron las cifras definitivas como parte del levantamiento de los Censos Económicos 2019. En este sentido, sobre el conjunto de datos se realizó inicialmente un análisis exploratorio, el cual fue útil para conocer los datos y comprender sus principales características. Para ejecutar este análisis, se utilizó como herramienta el lenguaje de programación Python.

Así, con base en la exploración de datos se observó que la estructura de los datos está conformada por 61 variables que representan las actividades económicas. Estas actividades están conformadas en tres niveles de agregación: i) sector, ii) subsector, y iii) rama. Los datos registrados son números no negativos, incluyendo el cero, que se agruparon en número de establecimientos, número de personas y miles de pesos (moneda nacional). Además, no se tuvieron valores nulos. Se observó también que no se tienen valores atípicos o fuera de rango. La Tabla 1 muestra los niveles de agregación, donde los primeros dos dígitos corresponden al sector (código 46), los primeros tres dígitos al subsector, y los cuatro dígitos en su conjunto a la rama de la actividad económica.

Tabla 1. Niveles de agregación de las actividades económicas por sector, subsector y rama.

Código	Niveles de agregación
4611	Comercio al por menor de abarrotes y alimentos
4612	Comercio al por menor de bebidas, hielo y tabaco
4621	Comercio al por menor en tiendas de autoservicio
4622	Comercio al por menor en tiendas departamentales
4631	Comercio al por menor de productos textiles, excepto ropa

4632	Comercio al por menor de ropa, bisutería y accesorios de vestir
4633	Comercio al por menor de calzado
4641	Comercio al por menor de artículos para el cuidado de la salud
4651	Comercio al por menor de artículos de perfumería y joyería
4652	Comercio al por menor de artículos para el esparcimiento
4653	Comercio al por menor de artículos de papelería, libros, revistas y periódicos
4659	Comercio al por menor de mascotas, regalos, artículos religiosos, desechables, artesanías y otros artículos de uso personal
4661	Comercio al por menor de muebles para el hogar y otros enseres domésticos
4662	Comercio al por menor de mobiliario, equipo y accesorios de cómputo, teléfonos y otros aparatos de comunicación
4663	Comercio al por menor de artículos para la decoración de interiores
4664	Comercio al por menor de artículos usados
4671	Comercio al por menor de artículos de ferretería, tlapalería y vidrios
4681	Comercio al por menor de automóviles y camionetas
4682	Comercio al por menor de partes y refacciones para automóviles, camionetas y camiones
4683	Comercio al por menor de motocicletas y otros vehículos de motor
4684	Comercio al por menor de combustibles, aceites y grasas lubricantes
4691	Comercio al por menor exclusivamente a través de Internet, y catálogos impresos, televisión y similares

Por otro lado, se midió también el grado de relación lineal entre pares de variables, el cual varía entre -1 y 1. Esta información fue estructurada en una matriz de correlaciones, encontrándose mayormente una relación débil entre las variables. No obstante, aquellas variables que presentaron cierta relación fueron las pertenecientes a una misma actividad económica. Además, con el objetivo de realizar un análisis que genere valor en el agrupamiento, fue necesario realizar una selección de variables. Dicha selección permitió enfocar el análisis sobre las variables significativas que representan diferentes actividades económicas del comercio al por menor.

3.3. Selección de variables

De las 22 ramas del comercio al por menor, mostradas en la Tabla 1, se obtuvo un conjunto de 58 variables y un total de 88 registros por cada una de estas. Dicho conjunto de variables se obtuvo como resultado de la depuración de variables categóricas, por ejemplo, la descripción de la actividad y el estatus de las cifras proporcionadas. Además, se descartó el

año debido a que inherentemente representa ya un agrupamiento de datos, el cual se busca evitar, puesto que el objetivo es obtener una segmentación que conjunte todas las variables a través de la medición de similitudes entre los elementos disponibles.

En este sentido, las variables seleccionadas ofrecen información de los siguientes rubros: a) estrato de la empresa: micro, pequeña o mediana; b) tipos de establecimientos: auxiliares o comerciales (número de establecimientos); y c) personal dependiente y no dependiente: mujeres y hombres (número de personas). Además, las otras variables seleccionadas representan las diferentes actividades del comercio al por menor, las cuales se agruparon como:

- Consumo de bienes y servicios para uso propio o reventa: mercancías, materiales, materias primas y auxiliares, consumo de combustibles y lubricantes, energía eléctrica, envases y empaques, alquiler de bienes muebles e inmuebles, pagos por personal no dependiente de la razón social, publicidad y servicios de comunicación (miles de pesos).

- Impuestos que gravan la actividad y específicos a los productos, gastos fiscales, financieros y donaciones (miles de pesos).

- Ventas netas e ingresos por suministro de bienes y servicios: ventas de mercancías adquiridas para su reventa, productos elaborados, ingresos por consignación y comisión, por prestación de servicios y por alquiler de bienes muebles e inmuebles (miles de pesos).

- Activos fijos: compra y venta de maquinaria y equipo de producción, bienes inmuebles, unidades y equipo de transporte, equipo de cómputo y periféricos, mobiliario, equipo de oficina y otros activos físicos (miles de pesos).

Estas actividades financieras representan al conjunto de operaciones que se ejecutan en el mercado de ofertantes y demandantes, cuya vía es la adquisición de ingresos y la realización de gastos.

3.4. Aplicación del algoritmo

El análisis de los datos se basó en la aplicación del algoritmo K-means por su funcionalidad y características de eficiencia, donde a través de un método como el del codo (Elbow Method) fue posible definir el número adecuado de grupos, en los cuales se asignaron los vectores de datos (elementos) que comprende el objeto de estudio. Además, a través de este algoritmo se optimiza la solución de problemas en los que los elementos se distribuyen en k clústeres, de forma que la suma de las varianzas internas de todos estos sea la más baja posible.

Adicionalmente, debido a que los algoritmos particionales asumen un conocimiento a priori del número de grupos en el que deben ser divididos el conjunto de datos y que, a diferencia de los algoritmos de agrupamiento jerárquico, donde es posible definir cualquier número de grupos basados en las necesidades del observador, para el presente análisis resultó eficiente el conocimiento anticipado del número adecuado de clústeres. Además, el uso de los centroides permitió representar, de manera general, a cada uno de los grupos obtenidos, para así describir sus comportamientos.

En este sentido, como algoritmo se utilizó K-means, el cual fue implementado en Python con el propósito de encontrar similitudes entre los registros de las diferentes ramas del comercio al por menor, de tal forma que se generaron grupos basados en el siguiente proceso de asignación de elementos y actualización de los centroides:

1. Inicio: se establecieron centroides aleatorios para la formación de grupos.
2. Asignación: se asignó a cada elemento (vector de datos) a su centroide más cercano.
3. Actualización: se calculó la media de todos los puntos asignados en el clúster para establecer el nuevo centroide.
4. Repetir: se repitieron los pasos 2 y 3 de manera iterativa hasta que los centroides no cambiaron más.

Para la implementación del algoritmo de K-means se utilizó el módulo *sklearn.cluster* que reúne algoritmos de segmenatción de datos no supervisados, teniendo como pseudocódigo:

K-MEANS (P, k)
Entradas:
elementos de un conjunto de datos $P = \{p_1, \dots, p_n\}$
número de clústeres k
Salida:
centroides $\{c_1, \dots, c_k\}$ *que implícitamente dividen al conjunto de datos P en k clústeres*

1. *elegir k centroides iniciales* $C = \{c_1, \dots, c_k\}$
2. ***mientras*** *que el criterio de parada no se ha cumplido*
3. ***hacer*** → *paso de asignación:*
4. ***para*** $i = 1, \dots, N$
5. ***hacer*** → *encuentra el centroide más cercano* $c_k \in C$ *a la instancia* p_1
6. *asignar instancia* p_1 *para establecer* C_k
7. → *paso de actualización*
8. ***para*** $i = 1, \dots, k$
9. ***hacer*** → *establecer* c_i *como el centroide de todos los elementos en* C_i

Dentro del método K-means en Python, es posible definir la métrica de distancia a utilizar. Una buena métrica de distancia ayuda a mejorar significativamente el rendimiento del proceso de segmentación. Estas medidas de distancia, conocidas también como búsqueda de similitud vectorial, juegan un papel importante en el aprendizaje automático, las cuales se emplean en función de la situación en la que esté trabajando. Por ejemplo, en algunas áreas, la distancia *Euclidiana* (por Euclides) puede ser óptima y útil para calcular distancias entre elementos, donde la distancia viene a ser la longitud de la hipotenusa; en otras, como en información geoespacial, la distancia de Manhattan, o geometría del taxista, puede resultar útil si se necesita calcular la distancia entre dos puntos en una ruta similar a una cuadrícula.

Saber qué medida de distancia usar es útil para obtener modelos más precisos. Por ejemplo, si se quiere analizar búsquedas similares de un grupo de usuarios, estas pueden ser imprecisas y variadas. Unos pueden buscar algo genérico como ‘zapatos negros’ o algo más preciso como ‘Nike AF1 LV8’. Matemáticamente, una distancia es una función, que asigna un valor positivo a cada par de elementos (puntos) de un espacio n-dimensional. Esta tiene las siguientes propiedades: a) no negativa, el valor puede ser mayor o igual a cero; b)

simétrica, la distancia entre a y b es la misma que entre b y a; y c) la distancia de dos objetos en un mismo punto es cero.

En este sentido, para este trabajo de investigación se utilizó, como métrica, la distancia E*uclidiana*, debido a su flexibilidad en casos de uso de propósito general para calcular la distancia entre elementos, conocida también como espacio euclidiano. Una de sus bases se encuentra en la aplicación del teorema de Pitágoras, y otro debido al enfoque refinado de K-means para asignar cada elemento al clúster más cercano, entre el elemento y el centroide del clúster, aplicando principalmente la distancia euclidiana en un espacio n-dimensional. Por lo que, se buscó asignar los elementos a través de las distancias mínimas, entre cada elemento y los centroides, logrando tener así una alta similitud intraclúster y baja semejanza interclúster. La ecuación de la distancia euclidiana es la siguiente:

$$\mathrm{dist}(p, c) = \sqrt{\sum_{i=1}^{n} (p_i - c_i)^2}$$

Donde: $P = \{p_1, \dots, p_n\}$ son los elementos del conjunto de datos y $C = \{c_1, \dots, c_k\}$ corresponde a los centroides. En virtud de la necesidad de un conocimiento a priori sobre el número adecuado de grupos, para la implementación del algoritmo se estableció un rango de configuraciones de k. Este rango permitió ejecutar el algoritmo de manera iterativa para obtener los grupos.

Por lo tanto, dado que en el algoritmo K-means se necesita especificar el número de clústeres (grupos) en los cuales segmentar los datos, se utilizó, como se mencionó previamente, el método Elbow Method, que es una heurística que se utiliza para determinar ese número adecuado de grupos. Este método consistió en calcular la suma de las distancias al cuadro de cada elemento del clúster a su centroide correspondiente (SSE, por sus siglas en inglés), esto para cada configuración de k, basada en la siguiente ecuación:

$$\mathrm{SSE} = \sum_{k=1}^{k} \mathrm{dist}(p_i, c_k) = \sum_{k=1}^{k} \sum_{p_i \in C_k} (p_i - c_k)^2$$

Donde: k es el número de clústeres formados y representa el índice de la sumatoria para cada configuración de k; p_i son los elementos presentes en cada clúster; y c_k son los centroides. Posteriormente, con base en las estimaciones, se generó una gráfica para identificar donde la distorsión (efecto del codo) cambia de manera significativa. El algoritmo de este método el siguiente:

1. Calcular el agrupamiento para diferentes valores de k. Por ejemplo, k de 2 a 12 grupos.
2. Para cada *k*, calcular la suma total de las distancias al cuadrado dentro de los grupos (SSE), esto es, entre los centroides (c_k) y sus respectivos elementos (p_i)
3. Trazar la curva de SSE según el número de k grupos.
4. Identificar la ubicación del punto de la línea con máxima curvatura (efecto del codo) dentro de la curva trazada. Se considera a ese punto como un indicador de distorsión, que sugiere dejar de dividir los datos en grupos adicionales; obteniéndose así, de manera representativa, ese número adecuado de grupos.

Bajo la idea del objetivo de K-means, que es la minimización de la varianza intraclúster y la maximización de la varianza interclúster, la gráfica resultante permite observar que mientras el número de clústeres es mayor, la varianza explicada intraclúster tiende a disminuir y viceversa, fenómeno mediante el cual es posible la identificación del efecto del codo, que permite establecer la cantidad de grupos adecuada. La Figura 1 muestra el trazado de dicha curva, en la que se observó que el efecto del codo está en k igual a 5, donde la distorsión cambia de manera significativa en ese punto.

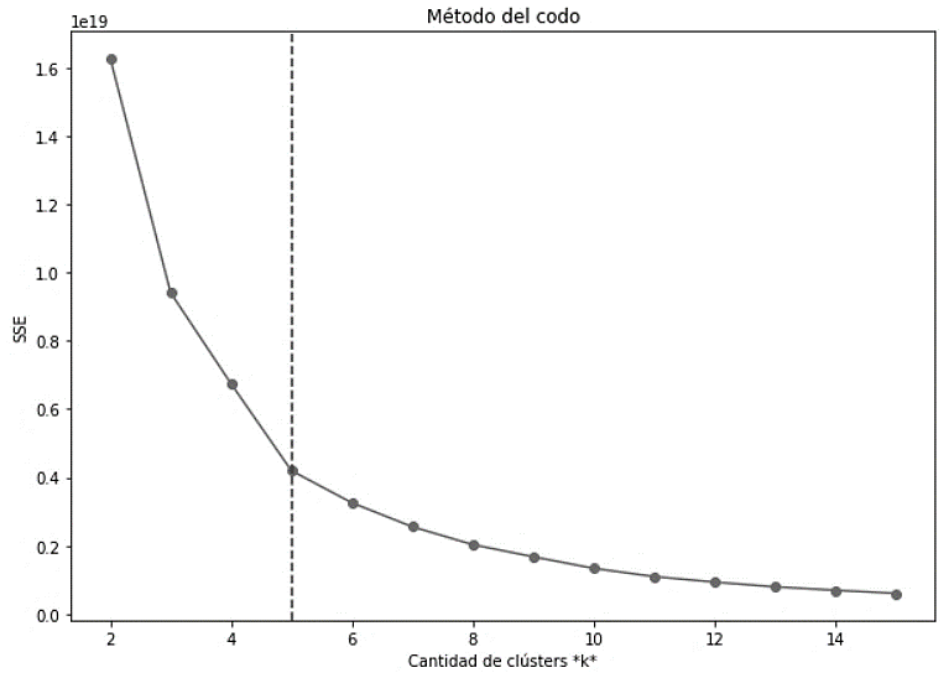

Fig. 1. Método del codo para la identificación del número adecuado de grupos.

En la práctica, puede que no exista un codo afilado y, como método heurístico, ese 'codo' no siempre puede identificarse sin ambigüedades. Por lo que, como método de validación en la definición de grupos, se utilizó *Kneed* de Python, la cual es una API (interfaz de programación de aplicaciones), denominada *KneeLocator*, que una vez instanciada, identifica el punto de inflexión máximo (cambio) en la trayectoria de la línea ajustada a los datos de entrada. Este cambio en la trayectoria, se define como el punto de la línea con máxima curvatura.

Kneed integra un algoritmo para encontrar el codo, similar a lo mostrado previamente en la Figura 1, pero de manera automática. La identificación de esta ubicación puede ser útil en varios casos, sin embargo, en el aprendizaje automático se puede emplear para ayudar con la selección de un valor apropiado de k en la clusterización de datos a través de K-means. Así, con base en esta función, se confirmó que la cantidad adecuada de clústeres fue cinco.

Capítulo 4

Resultados

4.1. Resultados alcanzados

De los resultados obtenidos, derivados de la aplicación del algoritmo K-means, se observó una segmentación diferenciada para el periodo de evaluación (2016-2019). Con base en las similitudes internas de los grupos y los valores máximos y mínimos de sus centroides, la Tabla 2 muestra un resumen de la conformación de éstos, a partir de la cual se determinaron aspectos que resaltan en los grupos, como mayor o menor participación en los rubros analizados; estrato (micro, pequeña y mediana empresa), tipos de establecimientos y personal (dependiente y no dependiente); y consumo, impuestos, ventas y activos fijos.

Tabla 2. Resumen de los grupos obtenidos y sus características.

Clúster	Características
1	**Rama industrial** – Grupo que concentra la mayor cantidad de ramas industriales, como: Bebidas, hielo y tabaco | Productos textiles, excepto ropa | Ropa, bisutería y accesorios de vestir | Calzado | Artículos de perfumería y joyería | Artículos para el esparcimiento | Artículos de papelería, libros, revistas y periódicos | Mascotas, regalos, artículos religiosos, desechables, artesanías y otros artículos de uso personal | Muebles para el hogar y otros enseres domésticos | Mobiliario, equipo y accesorios de cómputo, teléfonos y otros aparatos de comunicación (2016 y 2017) | Artículos para la decoración de interiores | Artículos usados | Partes y refacciones para automóviles, camionetas y camiones (2016 y 2017) | Motocicletas y otros vehículos de motor | Comercio al por menor exclusivamente a través de internet, y catálogos impresos, televisión y similares. **Estrato, tipos de establecimientos y personal (dependiente y no dependiente)** – Menor número de empresas pequeñas y medianas. – Menor número de establecimientos auxiliares. – Menor número de personal dependiente (hombres) y no dependiente (hombres y mujeres). **Consumo, impuestos, ventas y activos fijos**

	– Menor consumo de mercancías, materiales, materias primas y auxiliares. – Menor cantidad de impuestos que gravan la actividad y específicos a los productos. – Menor cantidad de ventas netas de mercancías, productos elaborados, prestación de servicios, alquiler de bienes muebles e inmuebles. – Menor cantidad de compra y venta de maquinaria y equipo de producción, bienes inmuebles, unidades y equipo de transporte, equipo de cómputo y periféricos, mobiliario, equipo de oficina y otros activos fijos.
2	**Rama industrial** – Comercio al por menor en tiendas de autoservicio. **Estrato, tipos de establecimientos y personal (dependiente y no dependiente)** – Mayor participación de personal no dependiente (hombres y mujeres). **Consumo, impuestos, ventas y activos fijos** – Mayor cantidad de dinero por consumo de mercancías, combustibles y lubricantes, energía eléctrica, envases y empaques; pagos por alquiler de bienes muebles e inmuebles, personal no dependiente de la razón social, publicidad y servicios de comunicación. – Mayor cantidad de impuestos específicos a los productos. – Mayor cantidad de ventas netas de mercancías, productos elaborados, ingresos por prestación de servicios, alquiler de bienes muebles e inmuebles. – Mayor cantidad de compra y venta de maquinaria y equipo de producción, bienes inmuebles, equipo de cómputo y periféricos, mobiliario, equipo de oficina y otros activos fijos.
3	**Rama industrial** – Abarrotes y alimentos | Artículos de ferretería, tlapalería y vidrios (2018 y 2019). **Estrato, tipos de establecimientos y personal (dependiente y no dependiente)** – Mayor cantidad de micro-empresas. – Mayor cantidad de establecimientos auxiliares y comerciales. – Mayor cantidad de personal dependiente (hombres y mujeres). **Consumo, impuestos, ventas y activos fijos** – Mayor cantidad de materiales consumidos para la prestación de servicios y materias primas y auxiliares. – Mayor cantidad de impuestos que gravan la actividad. – Menor cantidad de ingresos por consignación y comisión.
4	**Rama industrial** – Comercio al por menor en tiendas departamentales | Artículos para el cuidado de la salud | Artículos de ferretería, tlapalería y vidrios (2016 y 2017) | Automóviles y camionetas | Mobiliario, equipo y accesorios de cómputo, teléfonos y otros aparatos de comunicación (2018 y 2019) | Partes y refacciones para automóviles, camionetas y camiones (2018 y 2019). **Estrato, tipos de establecimientos y personal (dependiente y no dependiente)** – Menor cantidad de micro-empresas. **Consumo, impuestos, ventas y activos fijos** – Mayor cantidad de ingresos por consignación y comisión.

5	**Rama industrial** – Combustibles, aceites y grasas lubricantes. **Estrato, tipos de establecimientos y personal (dependiente y no dependiente)** – Mayor participación de empresas pequeñas y medianas. – Menor cantidad de establecimientos comerciales. – Menor cantidad de personal dependiente (mujeres). **Consumo, impuestos, ventas y activos fijos** – Mayor cantidad de compra y venta de unidades y equipo de transporte.

El *Clúster 1* se caracteriza por tener la menor cantidad de dinero dirigido al pago de impuestos que gravan la actividad comercial, e impuestos específicos a los productos que se venden. Se caracteriza además por la falta de control y gravamen fiscal a las empresas. Un ejemplo de esto son aquellas empresas que venden a través de Internet, donde a pesar del aumento del consumo de productos y servicios a través de plataformas digitales, el pago de impuestos sigue siendo bajo. De acuerdo con el Estudio económico realizado por la OCDE (Organización para la Cooperación y el Desarrollo Económicos), los ingresos fiscales en esta rama continúan siendo bajos, y la política fiscal tiene un bajo impacto redistributivo. Esto ocasiona una afectación en la recaudación tributaria, generando debilidad administrativa y de control para el pago de impuestos.

De manera contraria al grupo anterior, el *Clúster 2* se caracteriza por destinar la mayor cantidad de recursos al consumo, pago de impuestos, ventas y activos fijos. Dicho comportamiento se debe a que el comercio al por menor, en tiendas de autoservicio, sobresale de otras ramas por tener una mayor productividad y una distribución más eficiente de los productos.

El *Clúster 3* resalta por estar conformado por el comercio al por menor de abarrotes y alimentos, el cual es una de las ramas más comunes en la sociedad mexicana, donde la competencia es local. Además, este tipo de comercio está relacionado con el tamaño de las microempresas, que poseen menos barreras de entrada gracias al comportamiento de los consumidores, quienes acuden a las tiendas más cercanas a su domicilio o van a la siguiente para obtener mejores precios, mayor variedad de productos o algún otro beneficio.

El *Clúster 4* tiene diferencias entre las ramas industriales que la conforman, puesto que estas no siguen una misma tendencia que les permita permanecer unidas a lo largo del periodo analizado. Esto puede ser a consecuencia de que estas ramas han atravesado un proceso de trasformación en los últimos años, por ejemplo, el comercio al por menor de artículos de ferretería, tlapalería y vidrios ha presentado importantes bajas con respecto a la inversión extranjera.

Por su parte, el *Clúster 5* se caracteriza por tener la mayor participación de pequeñas y medianas empresas. Lo anterior es relevante debido a que durante el 2020 continuó esta misma tendencia, de acuerdo con información proporcionada por DataMexico, puesto que se evidenció un alza del 15.1% de empresas pequeñas y de 10.6% de empresas medianas, esto con respecto al año anterior (2019). Mientras que las empresas del estrato micro continuaron a la baja con una caída del 2.45%.

Como resultado, con respecto a la participación en los diferentes rubros y actividades financieras, se observó que el Clúster 1, a diferencia del Clúster 2, es el que más ramas industriales reúne, y que poseen una menor participación en todas las actividades financieras. Por lo tanto, es posible notar que no se encuentren bien posicionadas. Se observó además que los clústeres 3 y 5 cuentan con una mayor participación en los tres estratos: micro, pequeñas y medianas empresas, esto en comparación con los clústeres 1, 2 y 4.

4.2. Discusión

En la conformación de los cinco grupos, las ramas industriales que tuvieron participación en más de un clúster fueron: i) Comercio al por menor de mobiliario, equipo y accesorios de cómputo, teléfonos y otros aparatos de comunicación; ii) Comercio al por menor de partes y refacciones para automóviles, camionetas y camiones; y iii) Comercio al por menor de artículos de ferretería, tlapalería y vidrios. Este comportamiento se debe a los años de análisis, de 2016 a 2019, asociados con las 22 ramas industriales del comercio al por menor. Por lo que, en su conjunto, conforman el total de variables analizadas. En este sentido, una misma rama puede tener características que la hagan pertenecer a uno o más grupos. Lo que

demuestra que las ramas de comercio al por menor no siempre se comportan de la misma manera y que pueden verse afectadas por factores externos.

Por consiguiente, es posible observar que las diferentes tendencias que prevalecieron en las diferentes ramas industriales, para el periodo de análisis, son un reflejo de la actividad económica del país y de sus sectores industriales, actividades como la importación afectan directamente a los proveedores de los comerciantes. Además, la inversión extranjera en los diferentes sectores industriales también repercute directamente en los comercios minoristas.

Sectores como las industrias manufactureras, la preparación de alimentos y bebidas, el comercio al por mayor, servicios inmobiliarios, construcción, minería, agricultura, cría y explotación de animales, aprovechamiento forestal, pesca y caza, así como la generación, transmisión, distribución y comercialización de energía eléctrica, suministro de agua y de gas natural por ductos al consumidor final (en este caso para su uso en las unidades económicas dedicadas al comercio al por menor), tienen un gran impacto en los rubros analizados para cada rama industrial.

Por otro lado, es importante analizar el mercado en el que el comercio al por menor se desenvuelve desde el punto de vista del consumidor y cómo sus decisiones afectan al comerciante minorista. Los resultados obtenidos brindan las bases para comprender el poder que tiene el consumidor dentro del comercio en México. En la actualidad, las cadenas de supermercados han desplazado a los comercios minoristas, provocando que los consumidores prefieran a estos supermercados sobre unidades económicas pequeñas. Según la OCDE, esto se debe a la capacidad que las grandes empresas tienen para ofrecer precios de oferta más bajos y para conjuntar una enorme variedad de productos en una misma unidad económica; estas características han llevado a algunos minoristas a la quiebra de sus negocios.

Para evitar el cierre de los negocios se debe promover la participación de los comercios minoristas en regímenes fiscales. No obstante, dicha participación, a la fecha, ha sido poco favorable, y se ve reflejada en los resultados obtenidos. Un ejemplo de esto es que la Secretaría de Hacienda y Crédito Público, en conjunto con el Servicio de Administración Tributaria, brindan la opción de la incorporación de las PYMES al Régimen de Incorporación Fiscal (RIF), para obtener beneficios como descuentos sobre el Impuesto Sobre la Renta

(ISR), deducción de pagos, emisión de facturación electrónica, seguridad social, financiamiento o créditos. Sin embargo, las políticas a favor del comercio minorista para muchos emprendedores son soluciones parciales e insuficientes.

Lo anterior confirma las características mencionadas con respecto a las PYMES y cómo éstas subsisten en medio de un mercado agresivo y demandante. Es relevante considerar que algunas de las características destacadas no contribuyen de manera beneficiosa al desarrollo del comercio al por menor en México, ya que las unidades económicas no cuentan con innovaciones logísticas ni tecnológicas.

Capítulo 5
Conclusiones y trabajo futuro

5.1. Conclusiones

El comercio al por menor ha atravesado a lo largo de los años riesgos y dificultades que impiden su desarrollo y crecimiento. Dado que este tipo de comercio forma parte importante en la estabilidad económica y social del país, es fundamental comprender su comportamiento y de la mano con algoritmos especializados de aprendizaje automático, es posible hacerlo.

El uso de algoritmos de aprendizaje automático permite identificar patrones de datos para impulsar un mejor entendimiento del comportamiento del comercio minorista y que, en conjunto con el uso de tecnologías de la información, se busca que los micronegocios y las PYMES tengan crecimiento, rentabilidad y puedan afrontar los retos en materia de tecnología, impulsada en la actualidad por la Industria 4.0.

La importancia de los datos abiertos en materia de economía, para este tipo de análisis, es útil para el desarrollo de trabajos de impacto social que brindan información relevante. Bajo esta idea, trabajar con datos abiertos del comercio al por menor representó un reto significativo, puesto que se identificaron variables claves para la obtención de resultados útiles sobre la población dedicada al comercio minorista.

De la mano con lo anterior, es fundamental resaltar la importancia del comercio al por menor dentro de la economía mexicana, puesto que es el sector al que muchas PYMES se dedican y, en términos del PIB, este es fuente de una gran cantidad de empleos. Sin embargo, también atraviesa dificultades para subsistir en el mercado.

Estas PYMES subsisten en medio de la vulnerabilidad de sus negocios, ya que carecen de gestión financiera, que no les permite desarrollarse dentro de un mercado demandante. Además, no cuentan con los recursos para actualizar su tecnología. De esto, se deriva la necesidad de establecer políticas que beneficien a los trabajadores dedicados al comercio minorista, sin requerir de trámites administrativos complejos.

La incorporación de los comercios minoristas a regímenes fiscales es una solución que busca la protección y el bienestar de los trabajadores ante cualquier eventualidad. Sin embargo, el informe de operaciones es una actividad compleja de gestionar para la mayoría de los trabajadores. Por lo que, enfocar esfuerzos para promover programas de apoyo para los comercios minoristas, micros, pequeños y medianos es fundamental en pro de su crecimiento y desarrollo económico.

Por otra parte, con el objetivo de identificar fortalezas, debilidades, oportunidades, amenazas e incluso riesgos que posee y afronta el comercio al por menor, recurrir al uso de algoritmos de aprendizaje automático toma relevancia, gracias a que estos proporcionan una vía mediante la cual se innove la resolución de problemas y se concrete una propuesta de solución. Por esa razón, la aplicación del algoritmo de agrupamiento K-means permitió un mejor entendimiento del comportamiento del comercio al por menor.

Sin duda, hacer uso de algoritmos de agrupamiento, como K-means resultó ser una poderosa herramienta para observar el dinamismo a través de los años del sector comercial. Por lo que, se ratifica la importancia de la aplicación del aprendizaje automático como una herramienta de apoyo para la analítica avanzada de datos en materia de economía y, con ello, conocer el crecimiento y desarrollo económico de México.

En este sentido, el valor de esta investigación es la contribución de la implementación de K-means aplicado al mercado minorista, debido a que brinda información significativa y se comprueba que los diferentes sectores industriales afectan la manera en la que fluctúa la actividad económica de las ramas que conforman el comercio al por menor. Se observó además la existencia de factores externos que influyen en las características de pertenencia a un determinado clúster y sus similitudes con otras ramas industriales.

5.2. Trabajo futuro

Como trabajo futuro, y ante la disponibilidad de nuevos datos del sector comercial, que son puestos a disposición cada año, se pretende hacer un nuevo análisis con información actualizada para enriquecer los resultados obtenidos. Esto puede ser importante debido al comportamiento de la pandemia por COVID-19 y su impacto en el comercio al por menor, que según fuentes consultadas, éste registró una importante caída, en comparación con meses previos a la pandemia.

Sería relevante también enfocar esfuerzos en la visualización de resultados a través de recursos visuales como gráficos, mapas o una interfaz gráfica dirigida a usuarios interesados y contribuir así a la toma de decisiones o al análisis de la información en esta área.

Por oto lado, es indudable que el gobierno y las empresas en general deben poner énfasis en establecer medidas de apoyo al comercio minorista y a las pequeñas y medianas empresas. Además, se debe establecer medidas que garanticen planes de apoyo, sin importar el tamaño de estas; orientar sobre la escasez de mano de obra calificada; asesorar sobre la subsistencia en el sector después de la crisis por la pandemia ocasionada por COVID-19; y diversificar los canales de venta, en especial ayudando a los pequeños minoristas físicos a vender en línea.

Anexo A

Carta de aceptación

En este apartado se presenta la carta de aceptación de la publicación del artículo de investigación en la revista Research in Computing Science, emitidada por el editor.

RESEARCH IN COMPUTING SCIENCE

ISSN 1870-4069

Centro de Investigación en Computación, Instituto Politécnico Nacional,
Av. Juan de Dios Bátiz, s/n, Col. La Escalera, CP 07320, DF, México
Tel.: +52-55-5729 6000, ext. 56518, 56653
http://www.rcs.cic.ipn.mx

Mexico City, October 3rd, 2021

To whom it may concern:

Hereby I confirm that the paper

"Machine learning for the retail trade behavior analysis in Mexico"

by Patricia Soto Vázquez and Guillermo Gilberto Molero Castillo

after thorough reviewing process is accepted for publication in our journal.

It is scheduled for the volume 150(11), 2021, which is now in the process of technical production.

With best regards,

……..
Dr. Grigori Sidorov
Editor-in-Chief

Anexo B

Artículo publicado

En este apartado se presenta el artículo de investigación aceptado para su publicación en la revista Research in Computing Science, indizada en DBLP, LatIndex y Periodica.

Research in Computing Science

eISSN (applied for)
Indexing: DBLP, LatIndex, Periodica

Research in Computing Science, eISSN (applied for), is an internationally refereed open access scientific research journal published by the National Polytechnic Institute, a government-owned PhD-granting university subordinated to the Ministry of Public Education of Mexico. All papers submitted for publication are subject to rigorous international review process. Publication in this journal is free of charge. For the moment the journal is *not* indexed in WoS or EI. Contact: Prof. Grigori Sidorov, Editor-in-Chief. See the Editorial Board.

The topics of interest, number of pages per paper, submission procedure, deadlines, and contact for submissions are specified in the Call for Papers of a respective special issue or conference.

The format of papers is identical to Springer LNCS series format (though the journal is *not* published by Springer). You can find useful these formatting tips. Papers that do not follow these format requirements may be rejected without review or may be not included in the journal even if they have been accepted for publication. For the moment we do not require a copyright form, so please do not send it to us. We may contact you later for a copyright form.

Machine learning for the retail trade behavior analysis in Mexico

Patricia Soto-Vázquez, Guillermo Molero-Castillo,
Everardo Bárcenas, Rocío Aldeco-Pérez

Universidad Nacional Autónoma de México,
Facultad de Ingeniería,
Mexico

soto.holden@gmail.com, gmoleroca@fi-b.unam.mx,
{ebarcenas, raldeco}@unam.mx

Abstract. In recent years, machine learning methods have gained prominence when used as a tool for data analysis in different areas such as the economy. This article presents the result of the data analysis of the retail trade in Mexico and the industrial branches that comprise it. For the data analysis, unsupervised learning was used, specifically clustering based on the K-means, through which it was possible to organize clusters with information on the characteristics of the different industrial branches of the retail trade. As a result, it was possible to identify that this sector, the source of many jobs, subsists in the midst of an aggressive, demanding market, with insufficient access to update its technology and complex administrative procedures.

Keywords: Machine learning, clustering, k-means, economics, retail trade.

1 Introduction

The use of information as a source of knowledge is not only limited to understand what the data represents, but also generates value in any area it is applied, which is essential in the context of the actual, globalized world. Thus, as a result of this constant and necessary search to make total utilization of knowledge, Artificial Intelligence, applied in Industry 4.0, benefits the society in the process of digital transformation, mainly driven by machine learning and deep learning [1].

Knowledge in economics is built on the analysis of information derived from the economic activity of a country [2]. From the categorization of economic activity, economic sectors emerge as pillars of growth and development [3] [4]. Therefore, within the economic context in which Mexico develops, a knowledge-based economy is necessary, with machine learning being an important pillar in the push towards the fourth industrial revolution, which has led to the application of data analysis algorithms in order to build an efficient and quality economic system.

In this sense, given this growing need for data analysis in the economic field, the use of machine learning, as a support tool for advanced data analytics, is important [5] [6],

since it offers a wide variety of algorithms, among which are supervised, unsupervised, deep, reinforcement, and mixed, currently achieving an important position in response to the extensive digitization and storage of data.

On the other side, current machine learning is not only changing the way a product is produced, marketed and sold, but is also part of the study of economic growth, determined by the increase in productivity and income of a country [7] [8]. This study establishes the basis for analyzing economic development, measured based on improvements in the living conditions of the population. There is no doubt that the development of new products and companies, with unprecedented levels of automation and robotization, can transversally transform the economy and the labor market.

Thus, to understand how trade is directly associated with people quality of life, it is worthwhile to focus efforts on the analysis of one of the economic sectors that have a predominant impact on the Mexican economy, this is, retail trade; which is the economic activity defined by the individual sale of goods and services directly to final consumers [9]. This activity (by its nature) is a component of the supply chain in view to its model focused on the sale between the company and the consumer. This type of trade is a fundamental sector in Mexico, since, in terms of gross domestic product (GDP), tertiary activities had an annual percentage structure of 60%, corresponding to 2020, within which 9.2% corresponds to trade retail [10] [11].

In addition to the above, misinformation on the demeanor of economic activity not only leads people and companies to mismanage their business but encourages disinterest in establishing measures or laws that benefit the retail trade [12]. Consequently, this paper aims to show insight about retail trade in Mexico, which represents a field of opportunity, through an analysis, based on machine learning, since, moreover to its considerable percentage share with respect to GDP, it also concentrates a large population that finds, in this sector, a source of employment.

The document is organized as follows, Section 2 presents the antecedents of economics as social science, some of the contributions of machine learning in the economy, discuss its applications, the use of algorithms and related work are also presented. Section 3 describes the method established as a proposed solution. Section 4 presents the results obtained, based on an example of application, and Section 5 summarizes some conclusions and future work.

2 Background

2.1 Retail trade

Retail trade is defined by economic units, within which are several establishments that are under the control of a proprietary entity, permanently established and delimited by fixed facilities [10] [13]. Furthermore, these economic units are located at different geographical levels. For example, country, state, municipality, and locality, where they perform the task of enabling activities of buying and selling merchandise, or providing services, regardless of whether they have mercantile purposes [6] [15]. This group includes micro-businesses and small and medium-sized enterprises (SMEs) [14].

At present, the SME sector is one of the most vulnerable since, like any business, they require correct management of their financial income. These incomes depend, to a large extent, on proper financial management, which many SMEs lack. According to the Development Center for Business Competitiveness, 75% of SMEs close their operations just two years after being created. Moreover, the National Institute of Geography and Statistics (INEGI, by its acronym in Spanish) denotes that the new businesses in Mexico only live on average 7.7 years [16].

Conventionally, the study of economic growth is based on the analysis of indicators such as GDP, thanks to which the significant share of retail trade in the Mexican economy is notable. Although GDP is not a sufficient indicator to determine the economic growth of the country, it is one of the most important, since a rise in this indicator could easily translate into an increase in employment. Also, the occupation and employment indicators provide relevant information in this area.

2.2 Economic indicators

Concerning the employed population by size in the economic unit, the employment and occupation indicators establish that 20.1 million people are employed in micro-businesses, 7.5 million in small businesses, and 5.2 million in medium sized businesses. This information is representative since it has a coverage of 63.9% (84556) of the dwellings in the National Survey of Occupation and Employment (ENOE, by its acronym in Spanish) [17]. Among the organizations that provide information on the activities, economic indicators, and the labor market include:

- The National Survey of Occupation and Employment, that is the primary source of information on the labor market. It provides monthly and quarterly data on the labor force, occupation, labor informality, underemployment, and unemployment [18]. In 2020, they disseminated the occupational characteristics of the population aged 15 years and over, along with demographic and economic variables for the analysis of the labor force.
- The Annual Trade Survey (EAC, by its acronym in Spanish), which provides information on commercial activities and provides a frequent statistical overview that contributes to the decision-making of the different productive sectors of the country [19]. The Annual Trade Survey is based on the Monthly Survey on Commercial Companies (EMEC, by its acronym in Spanish), whose main purpose is to generate statistical information.
- The World Trade Organization (WTO) is the only international organization, of which Mexico is a member, that deals with the rules that govern trade between countries. Its aim is to ensure that commercial exchanges take place in a fluid, predictable and free manner [20]. One of his recent publications was 'Helping MSMES Navigate The Covid-19 Crisis', which explains how SMEs have been affected by the COVID-19 pandemic.

2.3 Clustering based machine learning

Machine learning consists of a set of algorithms for data-driven analysis, which allow establishing models, from the data of examples or experiences, to train machines (computers) and learn from them [8] [21]. Within machine learning, unsupervised methods are algorithms that base their training process on a previously defined, labelless data set. That is, no target or class value is known, either categorical or numeric. Therefore, these methods do not require human intervention [22].

The main applications of unsupervised learning are related to data clustering, where the objective is to find clusters with similar elements, in such a way that the internal elements of a cluster have a high similarity, and are different (dissimilar) with elements of others clusters [23]. There are two main types of clustering algorithms [24]: i) hierarchical, which produce a hierarchical organization of the elements that make up the data set, thus enabling different levels of clustering; and ii) partitional, which generate clusters of elements that do not correspond to any type of hierarchical organization. These algorithms are based on the distance between elements.

Partitional algorithms assume a priori knowledge of the number of clusters into which the data set must be divided, that is, they arrive at a division that optimizes a predefined criterion [25]. Among the algorithms that use this type of clustering highlights K-means, whose main idea is to define k centroids (one for each cluster) and then take each element (a record) from the database and place it in the nearest centroid cluster. The centroid is a point that occupies the middle position in a cluster. The next step is to recalculate the centroid of each cluster and redistribute all elements according to the nearest centroid. The process is repeated until there are no longer changes in the clusters formed [24]. In addition, in K-means the elbow method is used, with different configurations of k, to obtain an approximation to the adequate number of clusters [26].

2.4 Related work

In recent years, in Mexico, trade is the object of study due to its importance, not only in the national and international economy but also because of its impact on the Mexican population. In this sense, the analysis of retail trade is essential for the benefit of understanding the development and economic growth of a certain region over the years. This type of analysis can be achieved based on the observation of similarities, trends, and behaviors, for which machine learning algorithms are useful, some works related to data grouping and partitioning algorithms in economics are:

- Data mining and machine learning in the retail business: developing efficiencies for better customer retention [6], presented an analysis of retail marketing and discussed the application of data mining and machine learning techniques. Within the methodology, the use of K-means as a tool for the identification of incomplete data points stands out. This algorithm, together with another one for predicting customer interest and pattern mining techniques, made it possible to identify purchase patterns from user records.
- P2V-MAP: Mapping of market structures for large retail assortments [15], where market structures were analyzed through advances in natural language processing

and machine learning. The approach used made it possible to compare data dimensionality reduction techniques, which show a contribution to the market analysis. In addition, the use of machine learning algorithms is proposed to propose solutions in problems related to the structures of the retail market.

- Machine learning for enterprises: Applications, algorithm selection, and challenges [4], in this work the importance of machine learning applied to companies, is exposed, with the aim of promoting their technological development, and thus reducing costs of products and services. In addition, the use of methods and challenges in the application of machine learning algorithms for grouping, classification, and forecasting were analyzed. As well as the increase in the implementation of machine learning tools and algorithms in companies to increase their potential.
- Sustainability of SMEs in the Competition: A Systemic Review on Technological Challenges and SME Performance [27], in this paper the importance of SMEs as an engine of economic development was discussed, described the challenges they face, and reviewed the need for technological progress to drive innovation in the economy and the positive effects it has on production levels and economic growth. In addition, the adoption of information technologies as a means to face competitive challenges in SMEs was exposed.

Due to the growing need to increase research focused on the Mexican economy, based on machine learning, it is important to include, in the solutions, algorithms and varied approaches for understanding trade retail. The purpose is to identify evidence in the form of patterns from the data, with which various informed and thoughtful analyzes can be carried out on the current situation of retail trade and its impact on the national economy.

3 Method

The method defined for the analysis of the behavior of retail trade in Mexico was divided into four stages: a) data acquisition, b) exploratory data analysis, c) selection of variables, and d) algorithm application.

3.1 Data acquisition

The analyzed data were obtained from the Annual Trade Survey. This survey is based upon the economic units from the National Statistical Framework of Economic Units (MENUE, by its acronym in Spanish), supplied by the Mexican Business Statistical Registry (RENEM, by its acronym in Spanish) with referenced design variables [28]. The data source corresponds to open data available through the official website of the INEGI (www.inegi.org.mx/app/descarga/ficha.html?tit=110334&ag=0&f=csv), which is made up of data matrices on prime economic indicators of commercial activity by sector, subsector, and branch of economic activity at the national level.

The global information is made up of 40 branches of economic activity and is made up of 2134549 businesses in the commercial sector. Of these, 18 branches belong to the

wholesale trade (comprising 126933 business) and 22 to the retail trade (comprising 2007616 business), being, the latter, the object of study in this research work. Furthermore, the variables used are determined by the North American Industrial Classification System, which allows us to create clusters systematically, always under the same logic, which helps to avoid controversies and errors of interpretation [29]. Further, within the retail trade, the stratification by a number of workers in each company according to INEGI is defined as [30]: i) micro (up to 10 people), ii) small (11 to 30 people), and iii) medium (31 to 100 people).

3.2 Exploratory data analysis

The analysis period was from 2016 to 2019, since in 2020 the final figures were published as part of the 2019 Economic Censuses. In this sense, an exploratory analysis was initially carried out on the data set, which was useful to know the data and understand its main characteristics.

Thus, based on data exploration, it was observed that the data structure is made up of 61 variables that represent economic activities. The data recorded are non-negative numbers, which were grouped into number of establishments, number of people, and thousands of pesos (national currency). Also, there are no null values. It was also observed that there are no out-of-range values. Table 1 shows the levels of aggregation, where the first two digits correspond to the sector (46), the first three digits to the subsector, and the four digits as a whole to the branch of economic activity.

Table 1. Aggregation levels of economic activities by sector, subsector, and industrial branch.

Code	Levels of aggregation
4611	Retail trade of groceries and food products
4612	Retail trade of beverages, ice, and tobacco
4621	Retail trade in self-service stores
4622	Retail trade in department stores
4631	Retail trade of textile products, except apparel
4632	Retail trade of clothing, costume jewelry, and clothing accessories
4633	Retail trade of footwear
4641	Retail trade of health care items
4651	Retail trade of perfumery and jewelry
4652	Retail trade of entertainment articles
4653	Retail trade of stationery, books, magazines, and newspapers
4659	Retail trade of pets, gifts, religious articles, disposables, handicrafts, and other articles for personal use
4661	Retail trade of household furniture and other household goods
4662	Retail trade of furniture, computer equipment and accessories, telephones, and others communication devices
4663	Retail sale of articles for interior decoration
4664	Retail trade of used goods
4671	Retail trade of hardware, plumbing, and glassware
4681	Retail trade of cars and trucks
4682	Retail trade of parts and spare parts for automobiles, vans, and trucks
4683	Retail trade of motorcycles and other motor vehicles
4684	Retail trade of fuels, oils, and lubricating grease
4691	Retail trade exclusively through the Internet, and printed catalogs, television, and similar

The degree of linear relationship between pairs of variables was also measured. This information was structured in a correlation matrix, finding mostly a weak relationship between the variables. However, those variables that presented a certain relationship were those belonging to the same economic activity. In addition, it was necessary to carry out a selection of variables. This selection allowed to focus the analysis on the significant variables of different economic activities of the retail trade.

3.3 Feature selection

From each of the 22 retail trade branches, shown in Table 1, a set of 58 variables was obtained, from which categorical variables were filtered, and the year was discarded because it inherently already represents a clustering of data, which is to be avoided, since the objective is to obtain a segmentation that combines all the variables through the measurement of similarities between the available elements. In this sense, the selected variables offer information on the following categories: i) stratum of the business: micro, small or medium; ii) types of establishments: auxiliary or commercial (number of establishments); and iii) dependent and non-dependent personnel: women and men (number of people).

The other selected variables represent different significant activities of the retail trade, such as: a) consumption of goods and services; b) taxes on the activity; c) net sales; and d) fixed assets. These financial activities represent the set of operations that are executed in the supply and demand market, whose path is the acquisition of income and the realization of expenses.

3.4 Algorithm application

For the cluster analysis, the K-means algorithm was used due to its functionality and efficiency characteristics, where through a method, such as the Elbow, it is possible to define the appropriate number of clusters into which the data vectors should be divided (elements), which make up the data set. Furthermore, through this algorithm optimization problems are dealt with, in which the elements are distributed in K clusters so that the sum of the internal variances of all of them is as low as possible.

Thus, in order to find similarities in the different branches of retail trade, the algorithm was implemented in Python, in such a way that clusters were generated based on the following process of assigning elements and updating centroids:

1. Start: centroids chosen during each iteration were established randomly for the formation of clusters.
2. Assignment: each data point (vector) was assigned to its nearest centroid.
3. Update: the average of all assigned points in the cluster was calculated to set the new centroid.
4. Repeat: steps 2 and 3 were repeated iteratively until the centroids no longer changed.

The elements were assigned employing the minimum distances, measured through the Euclidean distance, between each element and the centroids, achieving thus, a high intra-cluster similarity and low inter-cluster similarity. The equation of the Euclidean distance is as follows:

$$\text{dist}(p, c) = \sqrt{\sum_{i=1}^{n} (p_i - c_i)^2}$$

Where, $p = \{p_1, \ldots, p_n\}$ are the elements of the data set and $c = \{c_1, \ldots, c_k\}$ corresponds to the centroids. Therefore, by virtue of the need for a priori knowledge about the adequate number of clusters, a range of k configurations was established for the implementation of the algorithm. This range allowed the algorithm to be run iteratively to obtain the clusters. Subsequently, based on the resulting categorization, the sum of the squared error (SSE) between each element of the formed cluster and its closest centroid was calculated. This SSE estimate was for each configuration of k based on the following equation:

$$\text{SSE} = \sum_{k=1}^{k} \text{dist}(p_i, c_i) = \sum_{k=1}^{k} \sum_{p_i \in C_k} (p_i - c_i)^2$$

Since this is a measure of error, the goal of K-means is to try to minimize this value. This measurement of error is used to carry out the elbow method, in which a curve is drawn with the values obtained from SSE to find an inflection point (elbow), through which the optimal number of clusters to be analyzed is established. Figure 1 shows the layout of this curve, in which it is observed that the elbow effect suddenly changes its orientation in k equal to 5.

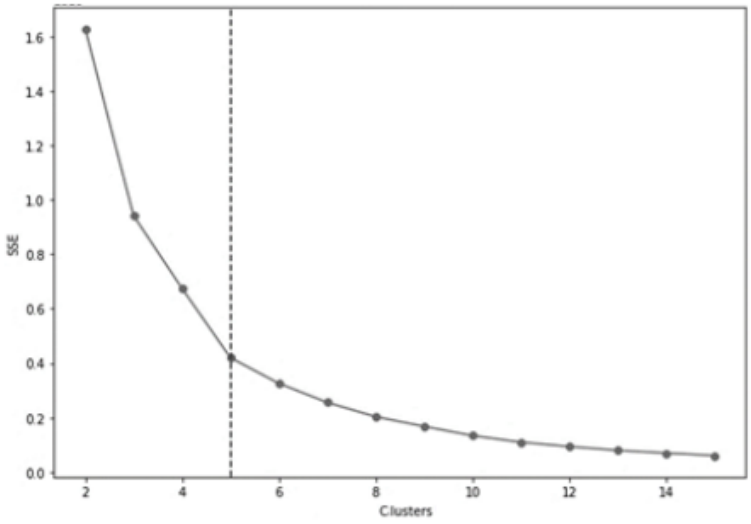

Fig. 1. Elbow method for identifying the adequate number of clusters.

4 Results

From the results obtained, a differentiated segmentation was observed for the evaluation period (2016-2019). Thus, based on the internal similarities of the clusters and the values of their centroids, Table 2 shows a summary of their most significant characteristics, determined by the aspects in which each cluster stands out for having greater or lesser participation in the retail trade activities; stratum (micro, small and medium-sized business), types of establishments and personnel (dependent and non-dependent); and consumption, taxes, sales and fixed assets.

Table 2. Summary of the clusters obtained.

Cluster	Characteristics
1	**Industrial branch** — Beverages, ice, and tobacco | Textile products, except apparel | Clothing, costume jewelry and clothing accessories | Footwear | Perfumery and jewelry articles | Entertainment articles | Stationery, books, magazines, and newspapers | Pets, gifts, religious articles, disposables, handicrafts and other articles for personal use | Household furniture and other household goods | Furniture, computer equipment and accessories, telephones, and other communication devices (2016 and 2017) | Interior decorating articles | Used goods | Parts and spare parts for automobiles, vans, and trucks (2016 and 2017) | Motorcycles and other motor vehicles | Retail trade exclusively through the Internet, and printed catalogs, television and similar. **Stratum, types of establishments and personnel (dependent and non-dependent)** — The lowest number of small and medium-sized companies. — The lowest number of auxiliary establishments. — The lowest number of dependent personnel (men) and non-dependent personnel (men and women). **Consumption, taxes, sales, and fixed assets** — The lowest consumption of merchandise, materials, raw and auxiliary materials. — The lowest number of taxes levied on the activity and specific to the products. — The lowest net sales of merchandise, manufactured products, services rendered, rental of movable and immovable property. — The lowest purchase and sale of machinery and production equipment, real estate, transportation units and equipment, computer and peripheral equipment, furniture, office equipment, and other fixed assets.
2	**Industrial branch** — Retail trade in department stores. **Stratum, types of establishments and personnel (dependent and non-dependent)** — The greatest participation of non-dependent personnel (men and women). **Consumption, taxes, sales, and fixed assets** — The biggest amount of money for consumption of merchandise, fuels and lubricants, electrical energy, containers, and packaging; payments for the rental of movable and immovable property, personnel non-dependent, advertising, and communication services. — The highest number of specific taxes on products. — The highest net sale of merchandise, manufactured products, income from services rendered, rental of movable and immovable property. — The greatest purchase and sale of machinery and production equipment, real estate, computer and peripheral equipment, furniture, office equipment, and other fixed assets.
3	**Industrial branch** — Groceries and food products | Hardware, plumbing, and glassware products (2018 and 2019). **Stratum, types of establishments and personnel (dependent and non-dependent)** — The greatest number of micro-companies.

	— The greatest number of auxiliary and commercial establishments. — The greatest number of dependent personnel (men and women). **Consumption, taxes, sales, and fixed assets** — The greatest number of materials consumed for services rendered, raw and auxiliary materials. — The greatest number of taxes levied on the activity. — The lowest consignment and commission income.
4	**Industrial branch** — Retail trade in department stores | Health care items | Hardware, plumbing and glassware products (2016 and 2017) | Cars and trucks | Furniture, computer equipment and accessories, telephones and other communication devices (2018 and 2019) | Parts and spare parts for automobiles, vans, and trucks (2018 and 2019). **Stratum, types of establishments and personnel (dependent and non-dependent)** — The lowest number of small companies. **Consumption, taxes, sales, and fixed assets** — The greatest consignment and commission income.
5	**Industrial branch** — Retail trade of fuels, oils and lubricating grease. **Stratum, types of establishments and personnel (dependent and non-dependent)** — The greatest participation of small and medium-sized companies. — The lowest number of commercial establishments. — The lowest number of dependent personnel (women). **Consumption, taxes, sales, and fixed assets** — The greatest purchase and sale of transportation units and equipment.

Cluster 1 is distinguished by having the least amount of money directed to the tax payment levied on commercial activity, and specific taxes on the products sold. Also characterized by the lack of control and tax burden. A case is the businesses that sell through the Internet, which, despite the increase in the consumption of products and services through digital platforms, the payment of their taxes is still low. According to the economic study carried out by the OECD, tax revenues in this branch continue to be low and fiscal policy has a lower redistributive impact. This causes an impact on tax collection, generating administrative and control weaknesses for the payment of taxes.

Contrary to the previous cluster, Cluster 2 is characterized by allocating the largest amount of resources to consumption, tax payments, sales, and fixed assets. This behavior is because the retail trade, for example, in self-service stores, it stands out for having a higher productivity and more efficient distribution of products.

Cluster 3 stands out for being made up of the grocery and food retail trade, as one of the most common industrial branches in Mexican society, where competition is local. In addition, this type of commerce consists mainly of micro-businesses, since they have fewer barriers thanks to the behavior of consumers, who go to the stores closest to their home to obtain better prices, greater variety, or another benefit.

Cluster 4 has differences between the industrial branches that comprise it, since they do not follow the same trend throughout the period analyzed. This may be due to the fact that these branches have undergone a process of transformation in recent years, for example, the retail trade of hardware stores and glass items has presented significant declines with respect to foreign investment.

For its part, Cluster 5 is characterized by having increased participation of small and medium-sized businesses. This is important because, during 2020, the same trend

continued according to the information provided by DataMexico, there was an increase of 15.1% in small businesses and 10.6% in medium businesses, this compared to the previous year (2019). Meanwhile, the micro businesses continued with a fall of 2.45%.

Regarding the participation in the different financial items and activities, it was observed that Cluster 1, unlike Cluster 2, is the one with the most industrial branches and that they have lower participation in all financial activities. This confirms the characteristics mentioned in the background section regarding SMEs and how they survive in the midst of an aggressive and demanding market. It was also observed that clusters 3 and 5 have greater participation in the three strata (micro, small and medium business), compared to clusters 1, 2, and 4.

On the other side, in the conformation of the five clusters, there are industrial branches that have participated in more than one cluster, such as: i) Retail trade of the furniture, computer equipment, and accessories, telephones, and other communication devices; ii) Retail trade of parts and spare parts for automobiles, vans, and trucks; and iii) Retail trade of hardware and glass items. This behavior is due to the years of analysis, from 2016 to 2019, associated with the 22 branches of the retail trade.

Consequently, it is possible to observe that the different trends that prevailed over the years, in the different industrial branches, are a reflection of the economic activity of the country and its industrial sectors. Activities such as imports directly affect merchants' suppliers. On the other hand, foreign investment in different industrial sectors also has a direct impact on retail businesses.

Sectors such as manufacturing industries, food and beverage preparation, wholesale trade, real estate services, construction, mining, agriculture, animal husbandry and exploitation, forestry, fishing, and hunting; as well as generation, transmission, distribution, and commercialization of electrical energy, supply of water and natural gas through pipelines to the final consumer (in this case for use in economic units dedicated to retail trade), have a great impact on the items analyzed for each industrial branch.

Certainly, it is important to analyze the market in which the retail trade operates from the point of view of the consumer and how its decisions affect the retailer. The results obtained provide the basis for understanding the power that the consumer has within commerce in Mexico. For example, today supermarkets have displaced retailers, causing consumers to prefer supermarkets instead of small businesses. This is due to the ability of large companies to offer lower prices and offer a huge variety of products. These characteristics have led some retailers out of business.

To avoid business closures, the participation of retail businesses in tax regimes should be promoted. For example, the Ministry of Finance and Public Credit, together with the Tax Administration Service, must continue with the incorporation of SMEs to the Tax Incorporation Regime (RIF), to obtain benefits, such as discounts on income tax (ISR), deduction of payments, issuance of electronic invoicing, social security, financing or credits. However, pro-retail policies for many entrepreneurs are partial solutions, causing them to fail or survive in the midst of an aggressive and demanding market.

5 Conclusions

The retail trade has experienced risks and difficulties over the years that impede its development and growth. Since this type of trade is an important part of the economic and social stability of a country, it is essential to understand its behavior, which, hand in hand with specialized machine learning algorithms, is possible to do so.

The importance of data in economics is useful for the development of social impact studies that provide relevant information. Under this idea, working with open data from the retail trade represented a significant challenge, since key variables were identified for obtaining useful results on the population dedicated to the retail trade.

In line with the above, it is essential to highlight the importance of retail trade within the Mexican economy, considering that it is the sector to which many SMEs are dedicated and, in terms of GDP, it is the source of numerous jobs. However, it is also struggling to survive in the market.

These SMEs subsist in the midst of the vulnerability of their businesses since they lack financial management, which does not allow them to develop within a demanding market. They also do not have the resources to update their technology. From the foregoing, the need for policies to benefit workers dedicated to the retail trade without requiring complex administrative procedures is derived.

The incorporation of retail businesses into tax regimes is a solution that seeks the protection and well-being of workers in any eventuality. However, it is a complex activity to manage for most workers, so focusing efforts to promote support programs for micro, small and medium-sized businesses is essential for their growth and economic development.

On the other side, in order to identify strengths, weaknesses, opportunities, threats and even risks that the retail trade faces, the use of machine learning algorithms becomes essential, since they provide a way to the resolution of specific problems. For this reason, the application of the K-means clustering algorithm allowed a better understanding of the behavior of the retail trade.

Undoubtedly, making use of clustering algorithms proved to be a powerful tool for observing the dynamism over the years of the trade sector, thus confirming the importance of applying machine learning as a support tool in data analytics in the field of economics and, with this, to know the growth economic of Mexico.

It was observed that the retail trade has experienced risks and difficulties that impede its development and growth, therefore, measures must be established to guarantee support plans, regardless of their size; provide guidance on the shortage of skilled labor; advise on subsistence in the sector after the crisis due to the COVID-19 pandemic; and diversify sales channels, especially by helping small physical retailers sell online.

As future work, it is intended to make a new analysis with updated information to enrich the results obtained. This may be important due to the behavior of the COVID-19 pandemic and its impact on the retail trade, which, according to sources consulted, registered a significant drop, compared to months prior to said pandemic. In addition, it would be relevant to focus efforts on the visualization of results through visual resources such as graphics, maps or a graphical interface aimed at interested users and thus contribute to decision-making, or information analysis in this area.

References

1. G. Molero-Castillo, G. Maldonado-Hernández, C. Mezura-Godoy, E. Benítez-Guerrero: Interactive system for the analysis of academic achievement at the upper-middle education in Mexico. Computación y Sistemas, 22(1), 223-233 (2018).
2. L. Montuschi: Datos, información y conocimiento. De la sociedad de la información a la sociedad del conocimiento. Universidad del CEMA, 192(6), 2-32 (2001).
3. E. Piedras: Industrias y patrimonio cultural en el desarrollo económico de México. Cuicuilco, 13(38), 29-46 (2006).
4. I. Lee, Y. J. Shin: Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons, 63(2), 157-170 (2020).
5. S. Hansen: Aplicación del aprendizaje automático al análisis económico y la formulación de políticas. Papeles de economía española, 157, 216-234 (2018).
6. M. R. Kumar, J. Venkatesh, A. M. Z. Rahman: Data mining and machine learning in retail business: developing efficiencies for better customer retention. Journal of Ambient Intelligence and Humanized Computing, 1-13 (2021).
7. G. Quiroga Persivale: ¿Qué es la inteligencia artificial y cómo se aplica en los negocios? https://repositorioacademico.upc.edu.pe/handle/10757/624220, last accessed 2021/06/07.
8. P. Mathur: Overview of Machine Learning in Retail. In Machine Learning Applications Using Python. Apress, Berkeley, CA, 147-157 (2019).
9. Comercio al por menor. http://centro.paot.org.mx/documentos/inegi/comercio_menor.pdf, last accessed 2021/06/07.
10. INEGI: Producto Interno Bruto Trimestral: Por actividad económica, www.inegi.org.mx/temas/pib, last accessed 2021/30/03.
11. INEGI: Glosario, www3.inegi.org.mx/contenidos/temas/economia/empresas/glosario.pdf, last accessed 2021/30/03.
12. D. Arana: Pymes mexicanas, un panorama para 2018, www.forbes.com.mx/pymes-mexicanas-un-panorama-para-2018/, last accessed 2021/30/03.
13. J. Ávila-Lugo: Introducción a la economía. Ed. Plaza y Valdez, México, pág. 390, ISBN: 970-722-256-5 (2007).
14. INEGI: Clasificación para Actividades Económicas. Encuesta Nacional de Ocupación y Empleo (ENOE), www.inegi.org.mx, last accessed 2021/30/03.
15. S. Gabel, D. Guhl, D. Klapper: P2V-MAP: Mapping market structures for large retail assortments. Journal of Marketing Research, 56(4), 557-580 (2019).
16. INADEM: Conflictos en el emprendimiento, www.inadem.gob.mx/conflictos-en-el-emprendimiento, last accessed 2021/30/03.
17. ENOE: Resultados del tercer trimestre de 2020, www.inegi.org.mx/contenidos/programas/enoe/15ymas/doc/enoe_n_presentacion_ejecutiva_trim3.pdf, last accessed 2021/30/03.
18. INEGI: Encuesta Nacional de Ocupación y Empleo (ENOE), población de 15 años y más de edad, www.inegi.org.mx/programas/enoe/15ymas/ , last accessed 2021/05/05.
19. INEGI: Encuesta Anual del Comercio 2019, www.inegi.org.mx/programas/eac/2013/, last accessed 2021/05/05.
20. OMC: México y la OMC. Available in: www.wto.org/spanish/thewto_s/countries_s/mexico_s.htm, last accessed 2021/05/05.
21. J. T. Palma Méndez, R. Marín Morales: Inteligencia artificial: métodos, técnicas y aplicaciones. Madrid: MacGraw-Hill, 1022 (2008).

22. L. Rouhiainen: Inteligencia Artificial. Madrid: Alienta Editorial, https://static0planetadelibroscom.cdnstatics.com/libros_contenido_extra/40/39308_Inteligencia_artificial.pdf, last accessed 2021/20/05.
23. C. G. Cambronero, I. G. Moreno: Algoritmos de aprendizaje: knn & kmeans, www.it.uc3m.es/~jvillena/irc/practicas/08-09/06.pdf, last accessed 2021/20/05.
24. D. Pla, F. Pascual, S. Sánchez: Algoritmos de agrupamiento. Método Informáticos Avanzados, 164-174 (2007).
25. A. J. Soto, I. Ponzoni, G. E. Vazquez: Análisis Numérico De Diferentes Criterios De Similitud En Algoritmos De Clustering. Mecánica Computacional, 993-1012 (2006).
26. M. Pérez: Aplicación de K-means y SOM, oa.upm.es/53779/1/TFG_MARTA_MARTIN_PEREZ.pdf, last accessed 2021/20/05.
27. R. Prasanna, J. Jayasundara, S. Naradda Gamage, E. Ekanayake, P. Rajapakshe, G. Abeyrathne: Sustainability of SMEs in the Competition: A Systemic Review on Technological Challenges and SME Performance. Journal of Open Innovation: Technology, Market, and Complexity, 5(4), 1-18 (2019).
28. EAC: Síntesis metodológica: Encuestas Económicas Nacionales, www.inegi.org.mx/contenido/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825194260.pdf, last accessed 2021/05/05.
29. SCIAN: NAICS – SCIAN, https://naics-scian.inegi.org.mx/naics_scian/default_e.aspx, last accessed 2021/20/05.
30. INEGI: Censos Económicos 2019, Micro, pequeña, mediana y gran empresa, www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825198657.pdf, last accessed 2021/09/28.

Anexo C

Código fuente

En este apartado se muestra el código fuente implementado en Python para el análisis de los datos del comercio al por menor en México mediante el algoritmo K-means y método del codo.

```
#Provee estructuras de datos, genera gráficos de alta calidad con
matplotlib y se integra con NumPy
Import pandas as pd

#Da soporte para crear vectores y matrices grandes multidimensionales,
permite el uso de funciones matemáticas
Import numpy as np

#Trazado y visualización de figuras en 2D
Import matplotlib.pyplot as plt

#Biblioteca para visualización de datos basado en matplotlib
import seaborn as sns

#Visualización dentro del notebook
%matplotlib inline

eac = pd.read_csv('/content/drive/MyDrive/ProyectoInvestigación/EAC_comer
cio_al_por_menor_2016_2019.csv')

#Se eliminan 4 filas quecorresponden a la suma de de toda la columna
Correspondiente a cada año entre dos
Variables = eac.drop([0, 32, 64, 96], axis=0)

#Se seleccionan todas las filas y se descartan las variables (columnas)
categóricas
VariablesSel = Variables.drop(['DESCRIPCION_ACTIVIDAD', 'ANIO', 'ESTATUS'
], axis=1)

#Corresponde al conjunto de datos numéricos sobre los cuales se realizará
el análisis exploratorio de datos
VariablesSel

from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min
```

```
from kneed import KneeLocator

#Corresponde a la selección de las ramas industriales para su análisis
VariablesSelDer = VariablesSel.drop(VariablesSel[VariablesSel['CODIGO_ACT
IVIDAD']<4000].index)

#Inicialización de la lista que contendrá las sumas de las distancias SEE
para cada k
SSEDer = []
#Al menos dos clusters hasta 15 clusters
for i in range(2, 16):
#Se generan centros aleatorios que cambian, por lo que para evitarlo es
necesario definir la posicisión 0 para la generación de centros
aleatorios
kmDer = KMeans(n_clusters=i, random_state=0)

#init: controla la técnica de inicialización. Default "random".
#n_clusters: establece k para el paso de clustering.
#n_init: establece el número de inicializaciones a realizar.
#El comportamiento predeterminado del algoritmo scikit-learn es realizar
diez ejecuciones de K-means y devolver los resultados del que tiene el
SSE más bajo
#max_iter: establece el número máximo de iteraciones para cada
inicialización del algoritmo k-means. Default 300 iteraciones
#random_state: inicializa el generador interno de numeros random a 0.
#Se usa para inicializar un nuevo objeto RandomState
#Esto realizará diez ejecuciones del algoritmo k-means en sus datos con
un máximo de 300 iteraciones por ejecución:
kmDer.fit(VariablesSelDer)
#Llenado de la lista SEE a partir de la funcion inertia
#La funcion inertia contiene el valor de SSE para cada configuración de k
SSEDer.append(kmDer.inertia_)

Se grafica SSEDer en función de k
plt.figure(figsize=(10, 7))
plt.plot(range(2, 16), SSEDer, marker='o')
plt.axvline(x=5, color='green', linestyle='--')
plt.xlabel('Cantidad de clústers *k*')
plt.ylabel('SSEDer')
plt.title('Método del codo')
plt.show()

klDer = KneeLocator(range(2, 16), SSEDer,
curve="convex", direction="decreasing")

#El punto Knee es el punto de máxima curvatura
klDer.elbow

#Etiquetado de cada elemento respecto al clúster que pertenece
```

```
MParticionalDer =
KMeans(n_clusters=5, random_state=0).fit(VariablesSelDer)

#El metodo predict determina el cluster más cercano al que pertenece cada
muestra de x (en este caso, x 0 VariablesSelDer)
MParticionalDer.predict(VariablesSelDer)

#Se muestra el cluster al que cada rama industrial pertenece
MParticionalDer.labels_

#Se muestra la cantidad de ramas industriales que pertenecen a cada
grupo.
VariablesSelDer.groupby(['clusterP'])['clusterP'].count()

#Se muestra los centros o centroides de cada clúster
CentroidesPDer = MParticionalDer.cluster_centers_
pd.DataFrame(CentroidesPDer.round(4))
```

Referencias bibliográficas

Arana, D. (2018). Pymes mexicanas, un panorama para 2018. Disponible en: www.forbes.com.mx/pymes-mexicanas-un-panorama-para-2018

Ávila-Lugo, J. (2007). Introducción a la economía. Ed. Plaza y Valdez, México, pág. 390, ISBN: 970-722-256-5

Bocanegra Gastelum, C. O. (2008). Para entender el comercio minorista en México a partir de los noventa. Revista Nicolaita de Estudios Económicos, 3(2), 89-104

Diferenciador (2020). Crecimiento y desarrollo económico. Disponible en: www.diferenciador.com/diferencia-entre-crecimiento-y-desarrollo-economico

Gabel, S., Guhl, D., Klapper, D. (2019). P2V-MAP: Mapping market structures for large retail assortments. Journal of Marketing Research, 56(4), 557-580

García Cambronero, G., y Gómez Moreno, I. (2006). Algoritmos de aprendizaje: knn & kmeans. Inteligencia en Redes de Comunicación, Universidad Carlos III de Madrid, 23

Hansen, S. (2018). Aplicación del aprendizaje automático al análisis económico y la formulación de políticas. Papeles de economía española, 157, 216-234

Instituto Nacional de Estadística, Geografía e Informática. (2004). Comercio al por menor. http://centro.paot.org.mx/documentos/inegi/comercio_menor.pdf

Instituto Nacional de Estadística, Geografía e Informática. (2019a). Encuesta Anual del Comercio 2019. Disponible en: www.inegi.org.mx/programas/eac/2013

Instituto Nacional de Estadística, Geografía e Informática. (2019b). Censos Económicos 2019, Micro, pequeña, mediana y gran empresa. Disponible en: www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825198657.pdf

Instituto Nacional de Estadística, Geografía e Informática. (2020a). Encuesta Nacional de Ocupación y Empleo. Resultados del tercer trimestre de 2020. Disponible en: www.inegi.org.mx/contenidos/programas/enoe/15ymas/doc/enoe_n_presentacion_ejecutiva_trim3.pdf

Instituto Nacional de Estadística, Geografía e Informática. (2020b). Encuesta Anual del Comercio. Síntesis metodológica: Encuestas Económicas Nacionales. Disponible en:

www.inegi.org.mx/contenido/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825194260.pdf

Instituto Nacional de Estadística, Geografía e Informática. (2021a). Producto Interno Bruto Trimestral: Por actividad económica. Disponible en: www.inegi.org.mx/temas/pib

Instituto Nacional de Estadística, Geografía e Informática. (2021b). Glosario. Disponible en: www3.inegi.org.mx/contenidos/temas/economia/empresas/glosario.pdf

Instituto Nacional de Estadística, Geografía e Informática. (2021c). Clasificación para Actividades Económicas. Encuesta Nacional de Ocupación y Empleo (ENOE). Disponible en: www.inegi.org.mx/rnm/index.php/catalog/209/download/6081

Instituto Nacional de Estadística, Geografía e Informática. (2021d). Encuesta Nacional de Ocupación y Empleo (ENOE), población de 15 años y más de edad. Disponible en: www.inegi.org.mx/programas/enoe/15ymas

Instituto Nacional del Emprendedor. (2016). Conflictos en el emprendimiento. Disponible en: www.inadem.gob.mx/conflictos-en-el-emprendimiento

Lee, I., Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons, 63(2), 157-170

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In 5-th Berkeley Symposium on Mathematical Statistics and Probability, 281-297

Martín Pérez, M. (2018). Aplicación de k-means y som (self-organizing maps) al análisis micro de accidentes de tráfico.

Mathur, P. (2019). Overview of Machine Learning in Retail. In Machine Learning Applications Using Python. Apress, Berkeley, CA, 147-157

Molero-Castillo, G., Maldonado-Hernández, G., Mezura-Godoy, C., Benítez-Guerrero, E. (2018). Interactive system for the analysis of academic achievement at the upper-middle education in Mexico. Computación y Sistemas, 22(1), 223-233

Montuschi, L. (2001). Datos, información y conocimiento. De la sociedad de la información a la sociedad del conocimiento. Universidad del CEMA, 192(6), 2-32

Nainggolan, R., Perangin-angin, R., Simarmata, E., Tarigan, A. F. (2019). Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. Journal of Physics: Conference Series, 1361(1), 012015. https://doi.org/10.1088/1742-6596/1361/1/012015

Organización Mundial del Comercio. (2021). México y la OMC. Disponible en: www.wto.org/spanish/thewto_s/countries_s/mexico_s.htm

Palma-Méndez, J. T., Marín Morales, R. (2008). Inteligencia artificial: métodos, técnicas y aplicaciones. Madrid: MacGraw-Hill, 1022

Pavón, L. (2016). Inclusión financiera de las pymes en el Ecuador y México. Serie Financiamiento para el Desarrollo No. 263, Comisión Económica para América Latina y el Caribe (Cepal), https://doi.org/http://hdl.handle.net/11362/40848

Piedras, E. (2006). Industrias y patrimonio cultural en el desarrollo económico de México. Cuicuilco, 13(38), 29-46

Pla, D., Pascual, F., Sánchez, S. (2007). Algoritmos de agrupamiento. Método Informáticos Avanzados, 164-174

Prasanna, R., Jayasundara, J., Naradda Gamage, S., Ekanayake, E., Rajapakshe, P., Abeyrathne, G. (2019). Sustainability of SMEs in the Competition: A Systemic Review on Technological Challenges and SME Performance. Journal of Open Innovation: Technology, Market, and Complexity, 5(4), 1-18

Quiroga Persivale, G. (2018). ¿Qué es la inteligencia artificial y cómo se aplica en los negocios? Gestión. Disponible en: http://hdl.handle.net/10757/624220

Rajesh Kumar, M., Venkatesh, J., Md Zubair Rahman, A. M. J. (2021). Data mining and machine learning in retail business: developing efficiencies for better customer retention. Journal of Ambient Intelligence and Humanized Computing, 1-13

Rouhiainen, L. (2018). Inteligencia Artificial. Madrid: Alienta Editorial. Disponible en: https://static0planetadelibroscom.cdnstatics.com/libros_contenido_extra/40/39308_Inteligencia_artificial.pdf

Sistema de Clasificación Industrial de América del Norte. (2021). Sistema de Clasificación Industrial de América del Norte, NAICS – SCIAN. Disponible en: https://naics-scian.inegi.org.mx/naics_scian/default_e.aspx

Soto, A. J., Ponzoni, I., Vazquez, G. E. (2006). Análisis Numérico De Diferentes Criterios De Similitud En Algoritmos De Clustering. Mecánica Computacional, 993-1012