Ingeniería Investigación y Tecnología
volumen XXI (número 1), enero-marzo 2020 1-17
ISSN 2594-0732 FI-UNAM artículo arbitrado
Información del artículo: Recibido: 5 de febrero de 2019, reevaluado: 26 de julio y
7 de noviembre de 2019, aceptado: 7 de noviembre de 2019

http://dx.doi.org/10.22201/fi.25940732e.2020.21n1.008

Ingeniería
Investigación y
Tecnología

Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública

Implementation of the CRISP-DM methodology for geographical segmentation using a public database

Espinosa-Zúñiga Javier Jesús
Grupo Financiero Ve por Más S.A. de C.V. Gerencia CRM, México
Correo: jjespinoza@vepormas.com
https://orcid.org/0000-0001-6828-2145

Resumen

El avance tecnológico ha permitido a las organizaciones en todos los niveles almacenar grandes volúmenes de datos. Sin embargo, un problema al cual se están enfrentando actualmente es el análisis de dichos datos a fin de extraer conocimiento útil para toma de decisiones en problemas reales. Actualmente existen varias metodologías que facilitan el análisis de datos para extraer información que se pueda convertir en conocimiento: una de ellas es la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) que a pesar de ser la metodología más utilizada para proyectos de minería de datos, y de tener más de veinte años desde su creación, no es muy conocida en el ámbito laboral de muchas organizaciones de todo tipo en México. El presente artículo tiene como objetivo aplicar la metodología CRISP-DM en la obtención de un modelo de segmentación geográfica sobre la base pública de unidades económicas del Directorio Nacional de Unidades Económicas (DENUE). Para ello, se aplicaron los seis pasos de la metodología (comprensión del problema, comprensión de datos, preparación de datos, modelado, evaluación del modelo e implementación del mismo) para obtener un modelo de segmentación geográfica que clasificó las entidades de la República Mexicana de acuerdo con sus unidades económicas. Aunque se trata de un modelo sobre la base del DENUE susceptible de mejora, muestra el beneficio de aplicar la metodología CRISP-DM, lo cual sin duda es de utilidad para las organizaciones que aplican actualmente dichos proyectos en México, y también muestra la potencialidad de explotar una base pública con información valiosa como la base del DENUE en muchos sentidos (comercial, académico, etcétera) mediante minería de datos.
Descriptores: Segmentación, metodología, DENUE, CRISP-DM, minería de datos.

Abstract

Technological progress has allowed to the organizations to store big amounts of data. However, organizations are facing to the challenge of analyzing such data for getting useful knowledge for decision making in real situations. Nowadays there are several methodologies that allow organizations to analyze big amounts of data in order to get information and knowledge. One of them is CRISP-DM (Cross Industry Standard Process for Data Mining) that despite the fact of be the most widely used methodology for Data Mining projects and to have more than twenty years old, it is yet not well known for many organizations in Mexico. This article aims to illustrate how to apply CRISP-DM for getting a geographical segmentation model for a public database called DENUE which contains a directory of business units in Mexico. The six steps of the methodology (understanding problem, understanding data, preparation of data, modeling, evaluation and implementation) has been applied in order to get a geographical segmentation model that divides Mexican geographical entities according to their business units. Albit some observations were classified not properly (according to the evaluation that was applied to the model) in general the clusters are acceptable considering the variables used for getting them, and in order to improve the model we suggest to consider additional variables that are no disposable in DENUE database nowadays. Although it is a segmentation model over DENUE database which is susceptible of improvement, it shows the potential of applying CRISP-DM for Data Mining projects and also shows the potential of exploiting public databases in order to get knowledge useful for many purposes (business, scholars, etc.).
Keywords: Segmentation, methodology, DENUE, CRISP-DM, datamining.

Introducción

En julio de 2010 el Instituto Nacional de Estadística y Geografía (INEGI) puso a disposición del público el Directorio Estadístico Nacional de Unidades Económicas (DENUE). Esta primera versión del DENUE (generada a partir de la información recabada por los Censos Económicos 2009) proporcionaba los datos de identificación y ubicación de poco más de cuatro millones de unidades económicas (entendiendo como tales establecimientos y empresas) de todos los sectores de la actividad económica (excepto las actividades agropecuarias y forestales) que se hallaron activos durante los censos en el territorio nacional (INEGI, 2017).

La segunda versión del DENUE fue publicada en marzo 2011 y proporcionaba información de las unidades económicas ubicadas en las cabeceras municipales, en las localidades urbanas de 2,500 habitantes y más, así como en las localidades rurales importantes de todo el territorio nacional (corredores, ciudades y parques industriales y localidades de 2,500 habitantes que por su importancia dentro de alguna actividad económica son cubiertas por los censos económicos, así como negocios ubicados en zonas rurales que por su tamaño o importancia económica son captados por los censos económicos y registrados en el DENUE).

En 2013 se publicó el DENUE interactivo que conservó el mismo número de unidades económicas que las versiones anteriores, pero incorporó la posibilidad de que los informantes registraran o actualizaran en línea los datos de sus negocios.

La versión actual del DENUE (novena) contiene información de más de cinco millones de unidades económicas y no solo permite la actualización de datos en línea sino que también las despliega mediante un Sistema de Información Geográfica (GIS) disponible gratuitamente en Internet y que ofrece datos de identificación, tamaño (en términos de número de personas que emplea), ubicación geográfica y datos de contacto de cada unidad económica. Adicionalmente, cada unidad económica se asocia a su correspondiente código de actividad económica de acuerdo con el Sistema de Clasificación Industrial de América del Norte (INEGI, 2011). La información del DENUE se valida previamente por el INEGI. Otras fuentes de actualización del DENUE son: instituciones de gobierno, encuestas económicas, censos y estudios sobre demografía de las unidades económicas (DENUE-INEGI, 2018).

Lo anterior ofrece un cúmulo de información útil para lo siguiente:

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

En el presente estudio se aplicará la metodología CRISP-DM (que es un estándar empleado a nivel mundial tanto en la industria como en la academia para proyectos de minería de datos) para obtener un modelo de segmentación geográfica de las unidades económicas del DENUE. Para ello, el trabajo se divide en el marco teórico donde se verán brevemente las características básicas de la metodología CRISP-DM, en la aplicación de la metodología CRISP-DM que aplicará cada una de las seis etapas que conforman esta metodología, así como proponer un modelo de segmentación de la base de unidades económicas del DENUE. Finalmente las conclusiones donde se comentarán los resultados derivados del estudio, así como los siguientes pasos que se proponen a fin de darle continuidad.

Para la realización de este estudio se utilizó el lenguaje de programación R, versión 3.3.3, que es de uso común para análisis estadístico y es de distribución gratuita.

Marco teórico

La metodología CRISP-DM es una de las más empleadas actualmente para el desarrollo de proyectos de minería de datos. En 1997, se puso en marcha bajo el financiamiento del Programa de Investigación y Desarrollo en Tecnologías de Información de la Unión Europea (ESPRIT) y se dirigió inicialmente por cinco empresas: SPSS (empresa enfocada a software estadístico que posteriormente sería adquirida por IBM), Teradata (empresa encargada a Inteligencia de Negocios), Daimler AG (empresa automotriz que contaba con un equipo de Minería de Datos relevante), NCR (una de las mayores empresas en informática en aquel entonces que posteriormente produjo su propio software de Minería de Datos) y Ohra (compañía aseguradora que en aquellos años empezaba a explorar el uso potencial de la minería de datos). La primera versión de la metodología se presentó en Bruselas en 1999 y se publicó ese mismo año como una guía paso a paso de minería de datos (Gallardo, 2018).

Actualmente IBM es la principal empresa que promueve el uso de esta metodología (aunque se usa por muchos profesionales de minería de datos que no pertenecen a esta empresa y también en medios académicos para proyectos de minería de datos). Esta empresa

incorporó la metodología CRISP-DM a uno de sus productos (SPSS Modeler) (Wikipedia, 2018).

De acuerdo con encuestas realizadas (Piatetsky (2002) What main methodology are you using for data mining?; Piatetsky, 2004 y Piatetsky (2007) la metodología CRISP-DM es la más utilizada para proyectos de minería de datos. El único otro estándar de minería de datos mencionado en estas encuestas fue SEMMA (Sample, Explore, Modify, Model and Assess, que es una metodología desarrollada por el instituto SAS). Una revisión y crítica de los modelos de minería de datos en 2009 llamó a CRISP-DM "el estándar de facto para el desarrollo de la minería de datos y los proyectos de descubrimiento de conocimiento" (Marbán, 2018).

La metodología CRISP-DM consta de seis etapas:

1. *Comprensión del problema o negocio:* Esta es la etapa más importante, ya que si no se tiene una correcta comprensión del problema, o negocio, de nada servirán las etapas siguientes. Las actividades principales de esta etapa son:

   - *Identificación del problema*: Consiste en entender y delimitar la problemática, así como identificar los requisitos, supuestos, restricciones y beneficios del proyecto.

   - *Determinación de objetivos:* Puntualiza las metas a lograr al proponer una solución basada en un modelo de minería de datos. En el presente estudio el objetivo es obtener un patrón de comportamiento de las entidades de la República Mexicana, de acuerdo con las unidades económicas ubicadas en cada entidad.

   - *Evaluación de la situación actual:* Especifica el estado actual antes de implementar la solución de minería de datos propuesta, a fin de tener un punto de comparación que permita medir el grado de éxito del proyecto.

2. *Comprensión de datos:* Las actividades principales de esta etapa son:

   - *Recolección de datos:* Consiste en obtener los datos a utilizar en el proyecto identificando las fuentes, las técnicas empleadas en su recolección, los problemas encontrados en su obtención y la forma como se resolvieron los mismos.

   - *Descripción de datos:* Identifica el tipo, formato, volumetría y significado de cada dato.

   - *Exploración de datos:* Radica en aplicar pruebas estadísticas básicas que permitan conocer las propiedades de los datos a fin de entenderlos lo mejor posible.

3. *Preparación de datos:* Generalmente esta es la etapa que consume más tiempo en el proyecto, y es donde se seleccionan los datos que se transforman de acuerdo con los resultados de la etapa anterior a fin de utilizarlos en la etapa de modelado. Las actividades principales de esta etapa son:

   - *Limpieza de datos:* Aplicación de diferentes técnicas, por ejemplo, normalización de datos, discretización de campos numéricos, tratamiento de valores ausentes, tratamiento de duplicados e imputación de datos.

   - *Creación de indicadores:* Genera indicadores que potencien la capacidad predictiva de los datos a partir de los datos existentes y ayuden a detectar comportamientos interesantes para modelar.

   - *Transformación de datos:* Cambia el formato o estructura de ciertos datos sin modificar su significado, a fin de aplicarles alguna técnica particular en la etapa de modelado.

4. *Modelado:* En esta etapa se obtiene propiamente el modelo de minería de datos. Sus actividades principales son:

   - *Selección de técnica de modelado*: Elige la técnica apropiada de acuerdo con el problema a resolver, los datos disponibles, las herramientas de minería de datos disponibles, así como el dominio de la técnica elegida.

   - *Selección de datos de prueba:* En algunos tipos de modelos se requiere dividir la muestra en datos de entrenamiento y de validación.

   - *Obtención del modelo:* Genera el mejor modelo mediante un proceso iterativo de modificación de parámetros del mismo.

5. *Evaluación del modelo:* En esta etapa se determina la calidad del modelo con base en el análisis de ciertas métricas estadísticas del mismo, comparando los resultados con resultados previos, o bien, analizando los resultados con apoyo de expertos en el dominio del problema. De acuerdo con los resultados de

esta etapa se determina seguir con la última fase de la metodología, regresar a alguna de las etapas anteriores o incluso partir de cero con un nuevo proyecto.

6. *Implementación del modelo:* Esta etapa explota, mediante acciones concretas, el conocimiento adquirido mediante el modelo. Aquí también es importante documentar los resultados de manera clara para el usuario final y asegurarse de que todas las etapas de la metodología se documenten debidamente para hacer una revisión del proyecto a fin de obtener lecciones aprendidas durante el proceso. Asimismo, monitorear las acciones para detectar áreas de oportunidad o incluso nuevos problemas.

Comprensión del problema

En la base del DENUE se identifica la necesidad de explotar la información que ofrece para obtener conocimiento útil. Una forma clara es detectar posibles patrones de comportamiento de las unidades económicas que la conforman, lo cual sería de utilidad para los siguientes propósitos:

a) *Comerciales:* Conocer los patrones de comportamiento de las empresas en México permite detectar mercados potenciales, diseñar estrategias de negocio específicas por patrón de comportamiento y dar pauta a estudios mercadológicos a fin de obtener mayor conocimiento sobre dichos patrones.

b) *Gubernamentales:* Conocer los patrones de comportamiento de las empresas en México reconoce necesidades específicas por patrón de comportamiento, lo que puede definir las políticas de diversa índole como económica, legal, etcétera, para las empresas del país.

c) *Académicos:* Conocer los patrones de comportamiento de las empresas en México da pauta a investigaciones de diferente tipo, económico, geográfico, estadístico, etcétera, que generaran nuevo conocimiento sobre las mismas.

Existen estudios previos sobre comportamiento y distribución geográfica de las unidades económicas del país, por ejemplo el análisis del INEGI, que comprende el ciclo de vida de las pequeñas y medianas empresas (INEGI, 2012)

Comprensión de datos

En el Anexo I se detalla la lista completa de datos que se pueden descargar desde el sitio del DENUE. Para efectos del presente estudio se considerarán solo los siguientes datos:

1. *Código de actividad económica:* Contiene el código de actividad económica SCIAN (Sistema de Clasificación Industrial de América del Norte) que es un código desarrollado por los gobiernos de Canadá, Estados Unidos y México, a fin de generar estadísticas comparables entre los tres países. En el caso de México se toma como base por el INEGI para proporcionar un marco único, consistente y actualizado de recopilación, análisis y presentación de estadísticas económicas (Secretaría de Comunicaciones y Transportes, 2016). Dicho código se compone de seis dígitos que se interpretan como se indica en la Tabla 1.

El SCIAN contiene un total de 22 sectores económicos, 94 subsectores económicos, 306 ramas económicas, 615 subramas económicas y 1,084 clases económicas. En el presente estudio se considerará solo el sector económico SCIAN. En el Anexo II se detalla la lista de dichos sectores.

2. *Tamaño del establecimiento:* Contiene el número de empleados de cada unidad económica agrupados en rangos, los valores posibles se indican en la Tabla 2.

Tabla 1. Significado de códigos SCIAN

Código	Significado
123456	Los dos primeros dígitos indican el sector económico
123456	Los tres primeros dígitos indican el subsector económico
123456	Los cuatro primeros dígitos indican la rama económica
123456	Los cinco primeros dígitos indican la subrama económica
123456	Los seis dígitos indican la clase económica

Tabla 2. Tamaños de empresa

Tamaño de empresa	Descripción
0 a 5 personas	La unidad económica tiene menos de 5 empleados
6 a 10 personas	La unidad económica tiene entre 6 y 10 empleados
11 a 30 personas	La unidad económica tiene entre 11 y 30 empleados
31 a 50 personas	La unidad económica tiene entre 31 y 50 empleados
51 a 100 personas	La unidad económica tiene entre 51 y 100 empleados
101 a 250 personas	La unidad económica tiene entre 101 y 250 empleados
251 y más personas	La unidad económica tiene más de 250 empleados

3. *Nombre de entidad:* Integra el nombre de la entidad federativa donde se ubica la unidad económica y los valores posibles corresponden a las 32 entidades federativas del país.

Se descargan un total de 5,078,730 registros desde el sitio DENUE (descarga realizada el viernes 19 de octubre 2018 en archivos csv). Al cargar estos registros en R (considerando solo nombre de establecimiento, código de actividad, tamaño del establecimiento y entidad) se realiza un análisis exploratorio básico donde se obtiene lo siguiente:

a) Se detectaron un total de 2,416,546 registros distintos por nombre de establecimiento, lo que significa que hay registros duplicados. Adicionalmente se descubren establecimientos con nombre genérico y caracteres raros (la Figura 1 muestra los 10 establecimientos con mayor número de duplicados y la Figura 2 la proporción de registros duplicados *vs.* no duplicados).

b) No se detectan nulos los campos código de actividad SCIAN, tamaño de establecimiento y entidad (Figura 3).

c) La Figura 4 muestra la gráfica de barras para sector económico SCIAN (extraído del campo código de actividad) donde se observan valores en blanco y códigos no válidos.

d) En la Figura 5 se observa la gráfica de barras para tamaño de empresa donde se observan valores en blanco y códigos no válidos.

e) La Figura 6 muestra la gráfica de barras para entidad donde se observan valores en blanco, nombres de entidad no válidos y nombres de entidad repetidos.

f) En la Tabla 3 se observan los resultados del análisis de correlaciones entre pares de variables para sector económico SCIAN, tamaño de empresa y entidad. Se ve claramente que hay relación estadística significativa entre estas variables debido a que en todos los casos se tiene p-value <= 0.05 y la fuerza de asociación es mayor a 0.3 (Amat, 2016).

	nombre_establecimiento	conteo
1	TIENDA DE ABARROTES SIN NOMBRE	72459
2	ABARROTES SIN NOMBRE	29086
3	CONSULTORIO DENTAL SIN NOMBRE	20948
4	PAPELERÍA SIN NOMBRE	20491
5	TALLER MECÁNICO SIN NOMBRE	19753
6	MISCELÁNEA SIN NOMBRE	18936
7	CARPINTERÃA SIN NOMBRE	18567
8	ESTÉTICA SIN NOMBRE	16773
9	TIENDA DE ROPA SIN NOMBRE	13583
10	PESCA Y CAPTURA DE OTROS PECES CRUSTÃCEOS MOLUSC>	10458

Figura 1. Registros con mayor número de duplicados por nombre de establecimiento, con base en unidades económicas DENUE

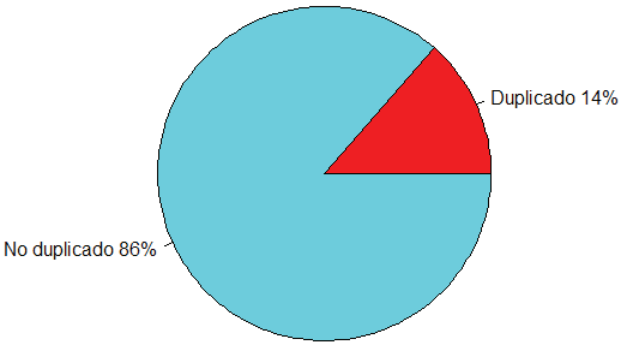

Figura 2. Proporción de establecimientos no duplicados *vs.* duplicados, con base en unidades económicas DENUE

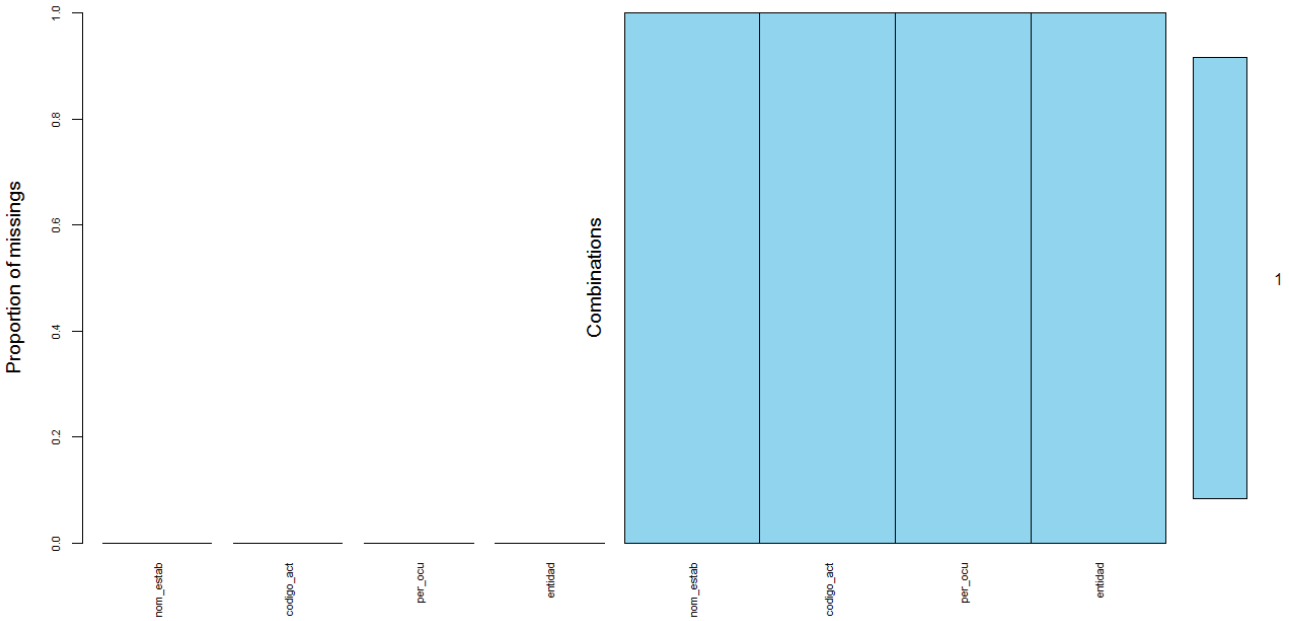

Figura 3. Código de actividad económica DENUE

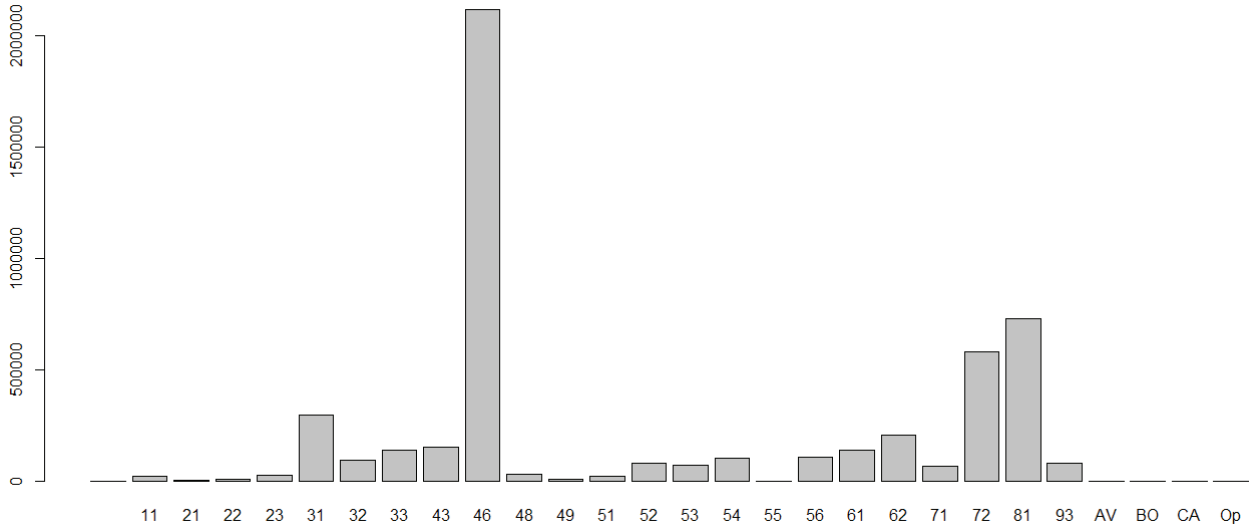

Figura 4. Sector económico SCIAN

http://dx.doi.org/10.22201/fi.25940732e.2020.21n1.008

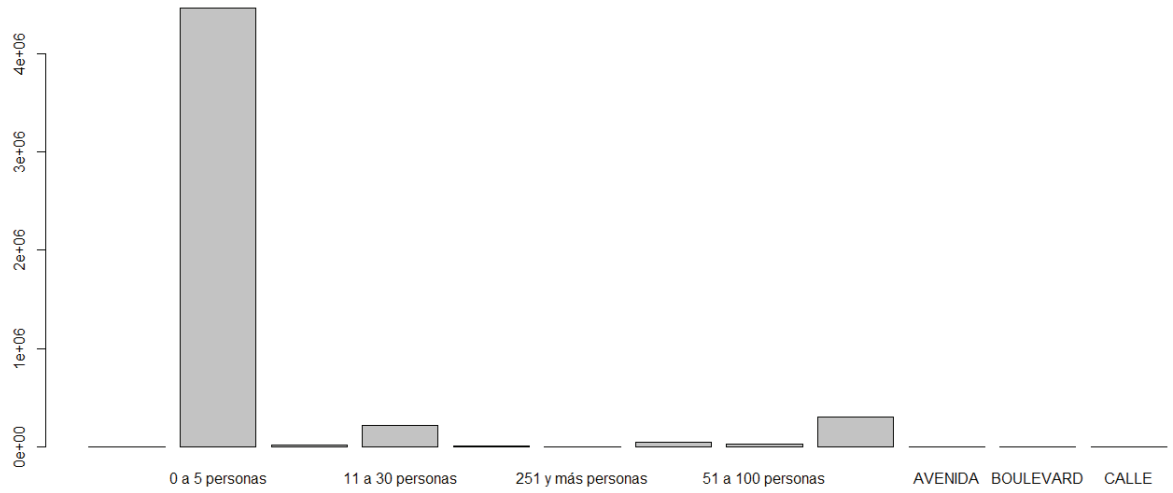

Figura 5. Tamaño de empresa

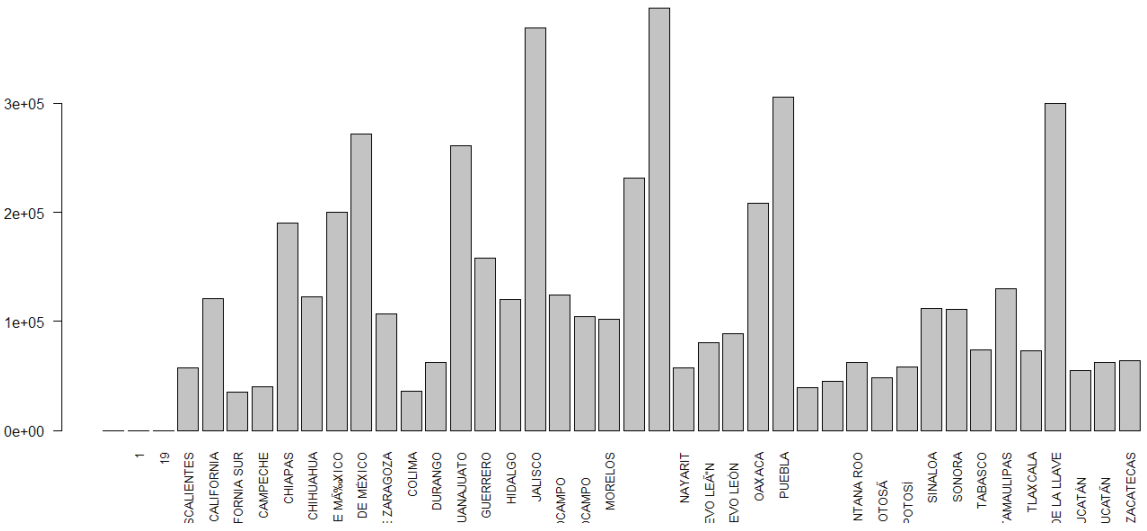

Figura 6. Entidades

Tabla 3. Análisis de correlaciones

Variable 1	Variable 2	p-value (1)	Fuerza de asociación (2)
Sector SCIAN	Entidad	< 2.2 e-16	0.336
Sector SCIAN	Tamaño de empresa	< 2.2. e-16	0.559
Entidad	Tamaño de empresa	< 2.2 e-16	0.396

1) H0= las variables son independientes.
   H1= las variables no son independientes.

2) La fuerza de asociación es medida en este caso mediante el coeficiente Cramer V, que es una mejora al coeficiente Chi cuadrado y permite medir la fuerza de asociación entre dos variables mediante un valor entre 0 (no hay asociación entre las variables) y 1 (asociación totalmente fuerte entre las variables).

Preparación de datos

De acuerdo con los resultados de la etapa anterior se realizó lo siguiente: (R-CRAN, 2019) (Alba, 2011):

a) Se consideran solo registros no duplicados (2,088,713 registros que representan 86 % de la base original). Esta base se denomina en lo sucesivo *muestra limpia*.

b) Se descartan los registros con sector económico en blanco o no válido. La Figura 7 descubre la volumetría de unidades económicas DENUE por sector económico SCIAN para la muestra limpia después de aplicar esta limpieza.

c) Se descartan los registros con tamaño de empresa en blanco o no válido. La Figura 8 revela la volumetría de unidades económicas SCIAN por tamaño de empresa para la muestra limpia después de aplicar esta limpieza.

d) Se eliminan los registros con nombre de entidad en blanco o no válido, asimismo se homologan los nombres de las entidades. La Figura 9 deja ver la volumetría de unidades económicas DENUE por entidad de la muestra limpia después de aplicar esta limpieza.

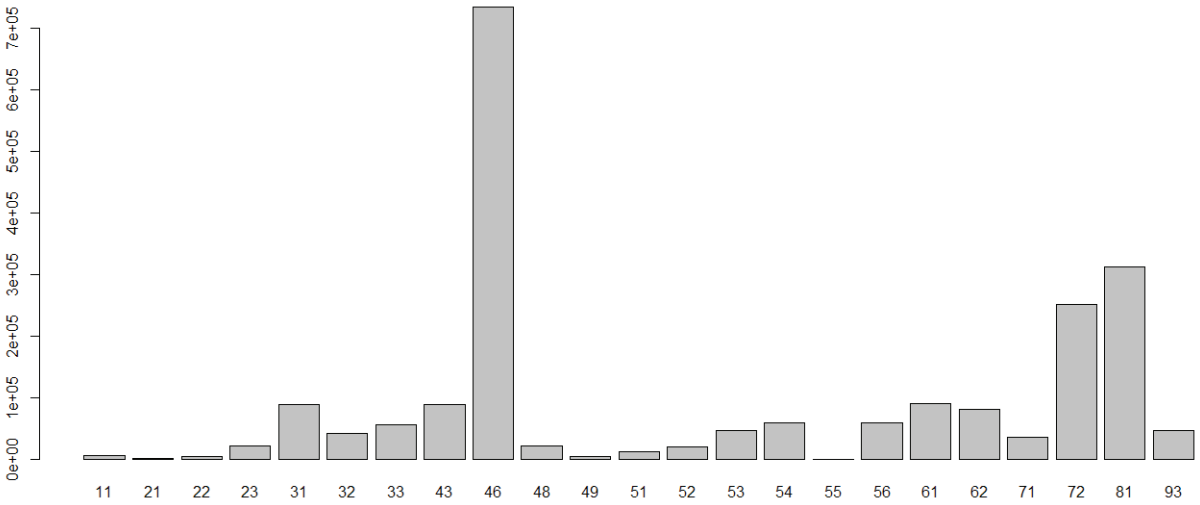

Figura 7. Sector económico SCIAN (muestra limpia)

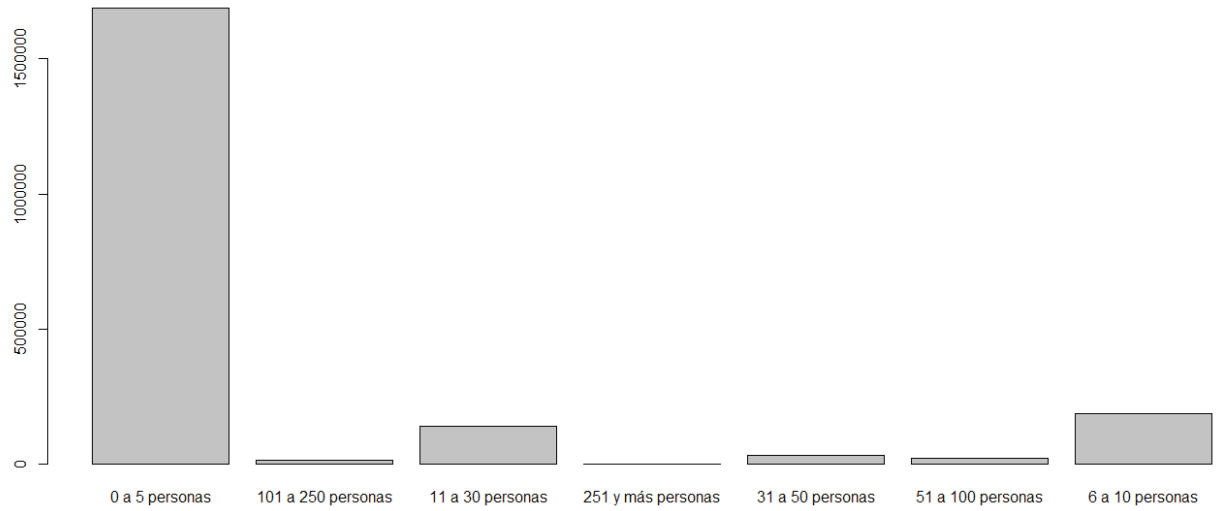

Figura 8. Tamaño de empresa (muestra limpia)

http://dx.doi.org/10.22201/fi.25940732e.2020.21n1.008

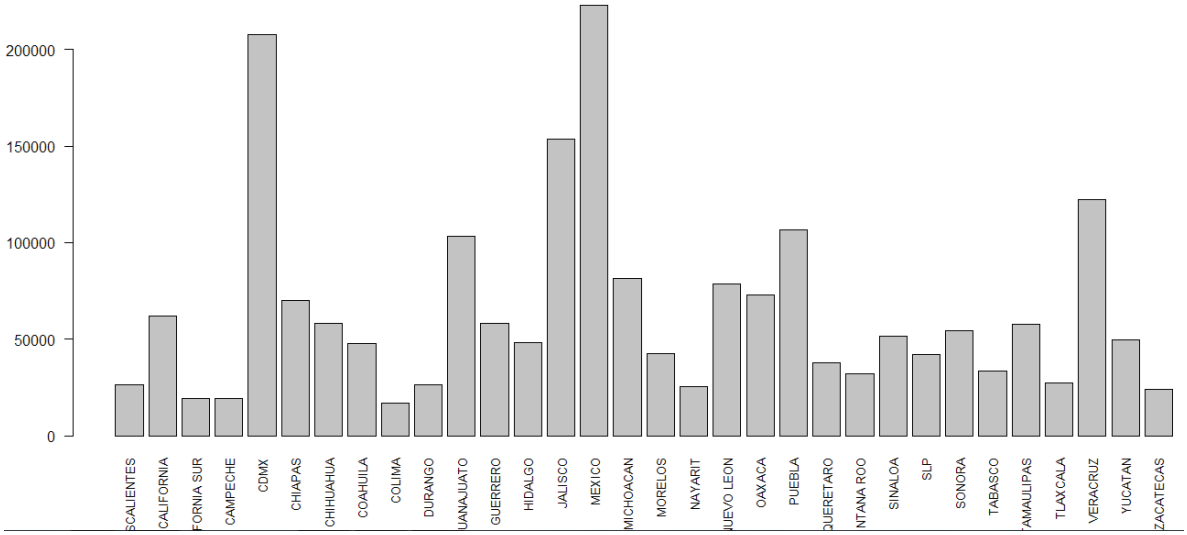

Figura 9. Entidad (muestra limpia)

	entidad	sectorSCIAN11_normalizado	sectorSCIAN21_normalizado	sectorSCIAN22_normalizado	sectorSCIAN23_normalizado	sectorSCIAN31_33_normalizado
1	AGUASCALIENTES	0.0006062440	0.0004167930	0.001250379	0.014625644	0.10226584
2	BAJA CALIFORNIA	0.0028152250	0.0008043500	0.000949133	0.011872205	0.07319584
3	BAJA CALIFORNIA SUR	0.0184965130	0.0010333250	0.002686644	0.016016533	0.06866443
4	CAMPECHE	0.0185500960	0.0011884460	0.002376892	0.017620007	0.07451041
5	CDMX	0.0000674452	0.0001878830	0.001382626	0.009355610	0.07684896
6	CHIAPAS	0.0032667620	0.0002853070	0.001669044	0.011854494	0.08243937
7	CHIHUAHUA	0.0010335740	0.0013091940	0.001498682	0.010387418	0.07553703
8	COAHUILA	0.0006507270	0.0023720060	0.001742270	0.015344571	0.08705053
9	COLIMA	0.0058204480	0.0035863370	0.001822565	0.019107531	0.06843436
10	DURANGO	0.0009009690	0.0022899620	0.002853067	0.013777311	0.08371499
11	GUANAJUATO	0.0006198970	0.0006489540	0.001249479	0.010218610	0.12161599
12	GUERRERO	0.0082180970	0.0007205850	0.002848025	0.005421542	0.10083039
13	HIDALGO	0.0023099450	0.0019387040	0.002866807	0.008022934	0.08445737
14	JALISCO	0.0015488340	0.0004685550	0.001119325	0.010125989	0.09956138
15	MEXICO	0.0007049460	0.0004849310	0.002303423	0.004368871	0.08466540
16	MICHOACAN	0.0062591460	0.0014264460	0.002877486	0.007476544	0.09961757
17	MORELOS	0.0056208840	0.0006820320	0.004680151	0.006208843	0.08085607
18	NAYARIT	0.0068450530	0.0008945240	0.002916926	0.010384256	0.07580118
19	NUEVO LEON	0.0002038030	0.0006750990	0.000662361	0.018023870	0.09481957
20	OAXACA	0.0044913880	0.0002197620	0.004216685	0.010424965	0.10275252
21	PUEBLA	0.0013527730	0.0012870130	0.002733729	0.008839997	0.10453931
22	QUERETARO	0.0004217190	0.0010279390	0.001054296	0.014232999	0.09064312
23	QUINTANA ROO	0.0013280620	0.0000617703	0.002316388	0.009852369	0.05080610
24	SINALOA	0.0147556280	0.0009720440	0.002624519	0.013725261	0.08404293
25	SLP	0.0009484290	0.0011618260	0.003509188	0.014226437	0.08901008
26	SONORA	0.0090885680	0.0025573090	0.003090849	0.016006182	0.08946904
27	TABASCO	0.0083577100	0.0010447140	0.005581756	0.015551310	0.06524984
28	TAMAULIPAS	0.0019866630	0.0006737380	0.002211243	0.013958470	0.06770203
29	TLAXCALA	0.0050196020	0.0009159860	0.003260909	0.006228703	0.11475470
30	VERACRUZ	0.0041895680	0.0006125100	0.001829363	0.008779309	0.07795209

Figura 10. Base final normalizada para modelado (fragmento)

Posteriormente se obtiene base con muestra limpia de 2,081,573 registros. Como siguiente paso, se normalizó esta base dividiendo para cada entidad de la República Mexicana el total de unidades económicas por sector entre el total de unidades económicas en la entidad. La Figura 10 muestra un fragmento de los valores obtenidos al normalizar.

Modelado

El término *segmentación* hace referencia a una serie de técnicas no supervisadas (es decir, que no requieren una variable respuesta a diferencia de las técnicas supervisadas donde sí se necesita) cuyo fin es encontrar patrones (segmentos o clusters) en un conjunto de observaciones y donde cada cluster es homogéneo en el sentido de que agrupa observaciones con características comunes (Berzal, 2018).

Las técnicas de segmentación tienen múltiples aplicaciones, por ejemplo, mercadotecnia (donde se divide un mercado de clientes potenciales en segmentos más pequeños con diferentes necesidades, características y comportamientos, que por tanto, son alcanzados por las empresas con estrategias diferenciadas), procesamiento de imágenes (donde se divide una imagen digital en varias partes a fin de simplificar o cambiar la representación de la imagen en otra más significativa y fácil de analizar), Sociología y Demografía (donde se divide un grupo social en subgrupos con comportamiento similar a fin de analizarlos).

Las técnicas de segmentación se pueden clasificar en tres grupos (Mathur, 2014) (Amat, 2017):

1. *Particionales:* Requieren que el usuario especifique de antemano el número de clusters a crear. Ejemplo de estas técnicas son: k-medias, k-medoides (PAM) y Clara. Estas técnicas (generalmente la k-medias) son las más empleadas por su facilidad de implementación. Para estas técnicas se requiere conocer de antemano el número de clusters (denotado como k) existiendo a su vez diversas maneras para ello.
2. *Jerárquicos:* No requieren que el usuario especifique con anterioridad el número de clusters. Los clusters se representan gráficamente en una estructura de árbol llamado dendograma. Estas técnicas pueden ser de dos tipos: aglomerativo (el agrupamiento se inicia en la base del árbol donde cada observación forma un cluster individual y se van combinando hasta converger en una única rama central) y divisivo (inicia con todas las observaciones agrupadas en un mismo cluster, el cual se divide sucesivamente hasta que cada observación forma un cluster individual).
3. *Híbridos:* Combinan características de los dos anteriores. Ejemplos de estas técnicas son: k-medias jerárquico (que combina las técnicas k-medias y jerárquicos), segmentación fuzzy (que es semejante a k-medias, pero con dos diferencias: los centroides son calculados de forma distinta y por cada observación se obtiene una probabilidad de pertenecer a cada cluster) y segmentación basada en densidad (DBSCAN, que trata de identificar clusters mediante redes neuronales) (Kassambara, 2018).

Este estudio utilizó la técnica PAM (*Partitioning Around Medoids*) que es más robusta que k-medias. Está basada en el concepto de *medoide*, que se define como la observación más cercana al centro dentro de un conjunto de datos.

Consiste en lo siguiente:

- Seleccionar k observaciones aleatorias como clusters iniciales.
- Calcular matriz de distancias entre todas las observaciones.
- Asignar cada observación a su medoide más cercano.
- Para cada cluster, se valida si seleccionando otra observación como medoide se reduce la distancia promedio del cluster, en caso afirmativo, se selecciona dicha observación como nuevo medoide, si al menos un medoide ha cambiado en el paso 4 se repite el paso 3, en caso contrario, se concluye el proceso.

La Figura 11 muestra el resultado de determinar el número óptimo de clusters a utilizar (existen diferentes técnicas para ello, en este caso se utiliza la técnica WSS-Total Within Sum of Square) que sugieren como número óptimo de clusters el que minimize la suma de cuadrados de las observaciones. Aplicando esta técnica a la base normalizada se tiene que a mayor número de clusters disminuye la suma de cuadrados (Figura 11), por lo que se probó con un número de clusters entre 4 y 12 y se seleccionó aquel donde los clusters se distinguen mejor entre sí.

Después de múltiples pruebas se consideró que el modelo de segmentación con 4 clusters (obtenido mediante técnica PAM) es donde mejor se marca la diferencia de los clusters entre sí (Figura 12).

Evaluación del modelo

En esta etapa se evaluó la calidad de los clusters obtenidos en la etapa de modelado. Existen tres formas para ello (Amat, 2017):

1. *Validación interna:* Dos métricas comunes son el coeficiente Silhouette (su valor puede estar entre –1 y 1, cuando vale 0 significa que la observación se encuentra en la frontera entre dos clusters, cuando es cercano a –1 significa que es mayor la evidencia de que la observación se ha asignado al cluster correcto y cuando es cercano a Symbol1 significa que es mayor la evidencia de que el cluster se ha clasificado en un cluster incorrecto) y el índice Dunn (mientras mayor sea este valor significa que los clusters están más compactos y separados, este indicador debe usarse con cuidado, ya que si todos los clusters tienen calidad aceptable excepto uno, el índice estará influenciado por dicho cluster).

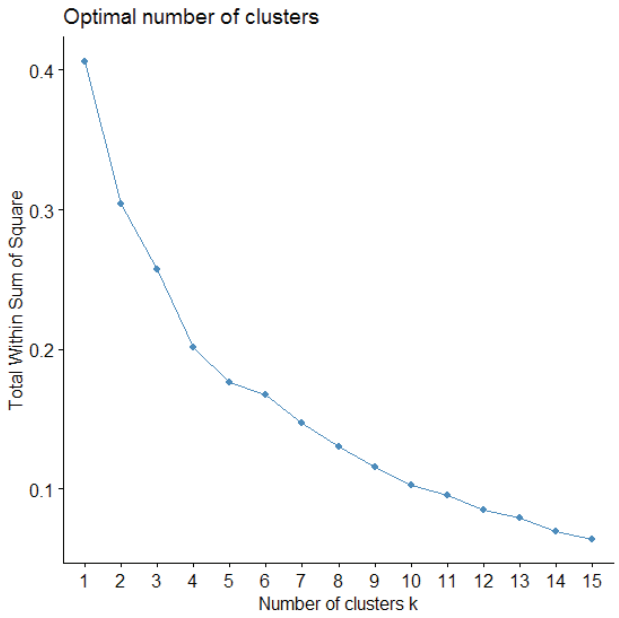

Figura 11. Número óptimo de clusters

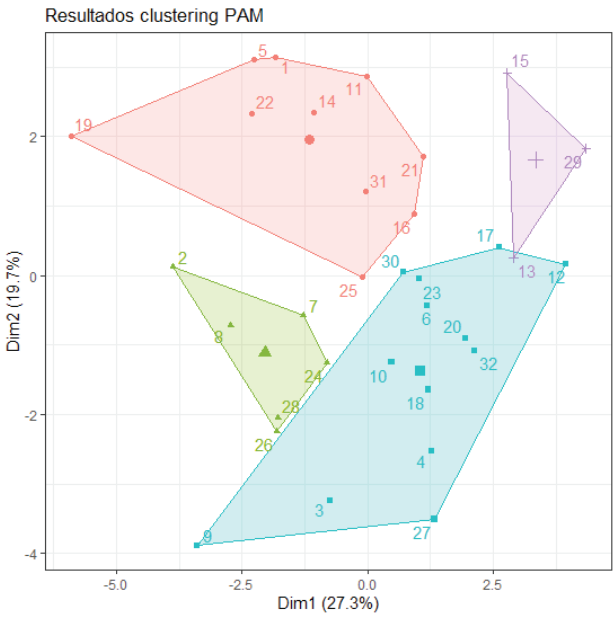

Figura 12. Modelo para 4 clusters

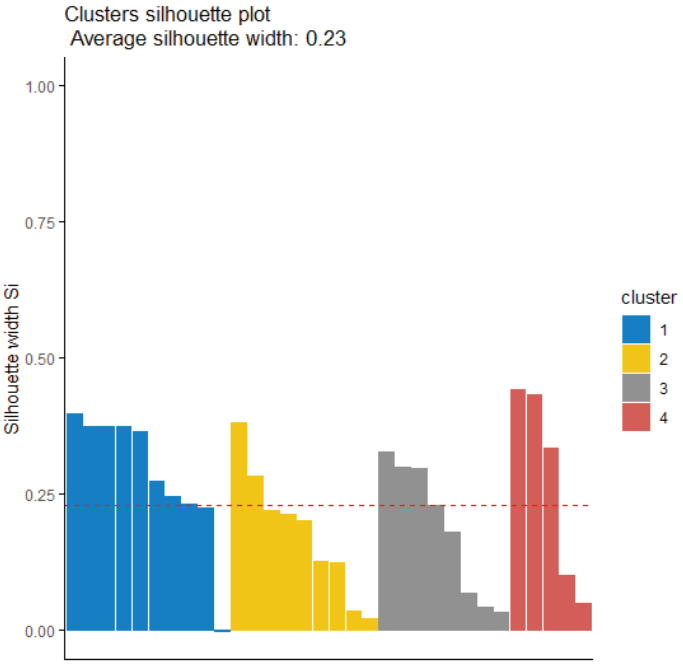

Figura 13. Coeficiente Silhouette para modelo de 4 clusters

2. *Validación externa:* Compara los resultados del modelo *vs.* un conjunto de datos del que se conoce el cluster al que pertenece cada observación.

3. *Significancia:* Se evalúa la probabilidad (p-value) de que los clusters obtenidos sean al azar.

Para el presente estudio se utilizaron métricas de validación interna. La Figura 13 muestra el resultado de calcular el coeficiente Silhouette para el modelo obtenido. El promedio de dicho coeficiente es 0.23 (mientras más cercano sea este valor a 1, mayor evidencia de que las observaciones se clasificaron en el cluster correcto), asimismo se observa que para el cluster 1 hay una observación ligeramente por debajo de cero, lo que indica posiblemente que se trate de una observación no clasificada en el cluster correcto. El índice Dunn del modelo es 0.30.

Las Tablas 4-7 muestran la distribución de entidades por cluster.

Tabla 4. Entidades en cluster 1

Núm.	Entidad	Total de unidades económicas
1	Aguascalientes	26,392
2	CDMX	207,576
3	Guanajuato	103,243
4	Jalisco	153,664
5	Michoacán	81,321
6	Nuevo León	78,507
7	Puebla	106,448
8	Querétaro	37,940
9	San Luis Potosí	42,175
10	Yucatán	49,947

Tabla 5. Entidades en clusters 2

Núm.	Entidad	Total de unidades económicas
1	Baja California	62,162
2	Chihuahua	58,051
3	Coahuila	47,639
4	Sinaloa	51,438
5	Sonora	54,354
6	Tamaulipas	57,886

http://dx.doi.org/10.22201/fi.25940732e.2020.21n1.008

Tabla 6. Entidades en cluster 3

Núm.	Entidad	Total de unidades económicas
1	Baja California Sur	19,355
2	Campeche	19,353
3	Chiapas	70,100
4	Colima	17,009
5	Durango	26,638
6	Guerrero	58,286
7	Morelos	42,520
8	Nayarit	25,712
9	Oaxaca	72,806
10	Quintana Roo	32,378
11	Tabasco	33,502
12	Veracruz	122,477
13	Zacatecas	24,233

Tabla 7. Entidades en cluster 4

Núm.	Entidad	Total de unidades económicas
1	Hidalgo	48,486
2	Estado de México	222,712
3	Tlaxcala	27,293

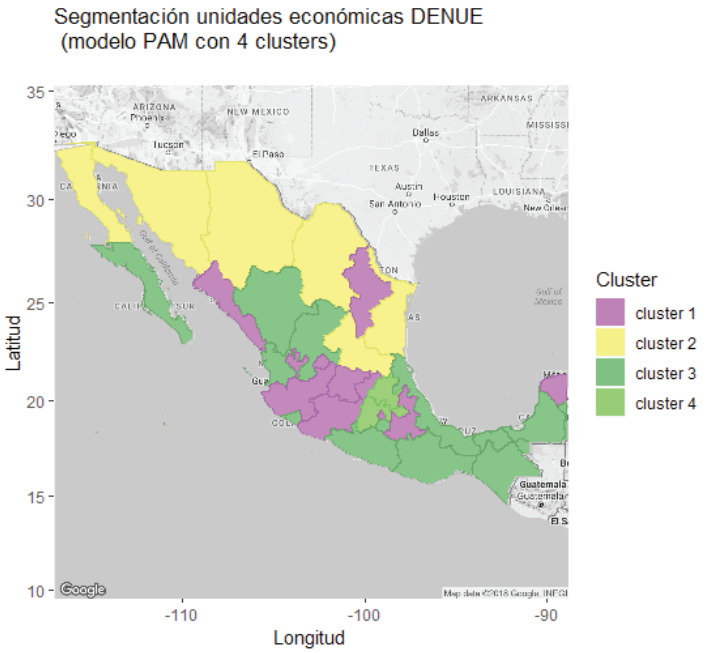

Figura 14. Distribución geográfica de modelo con 4 clusters

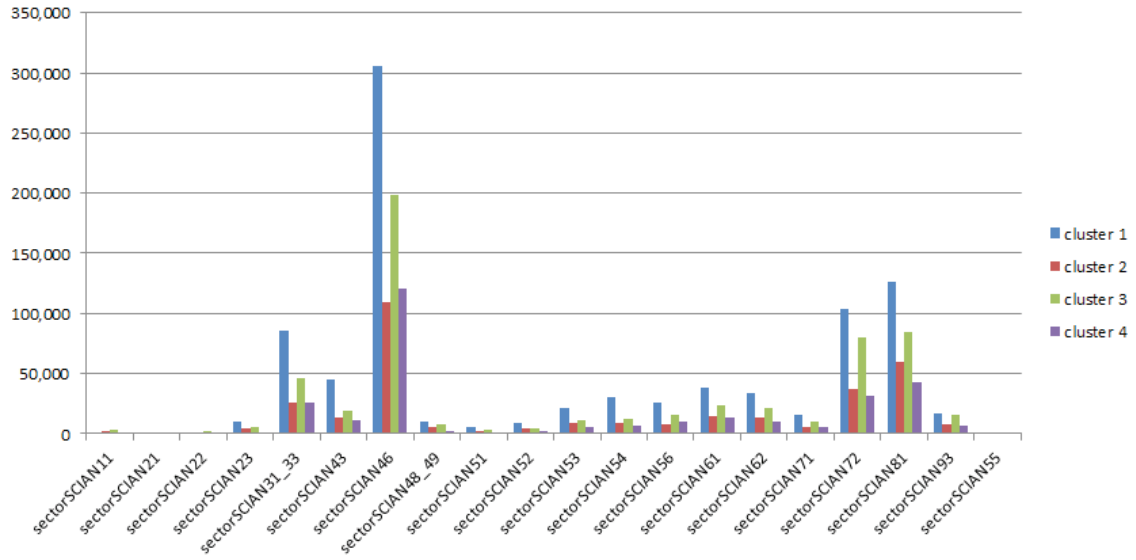

Figura 15. Comparativo de volumetría de unidades económicas DENUE por sector SCIAN para 7 clusters

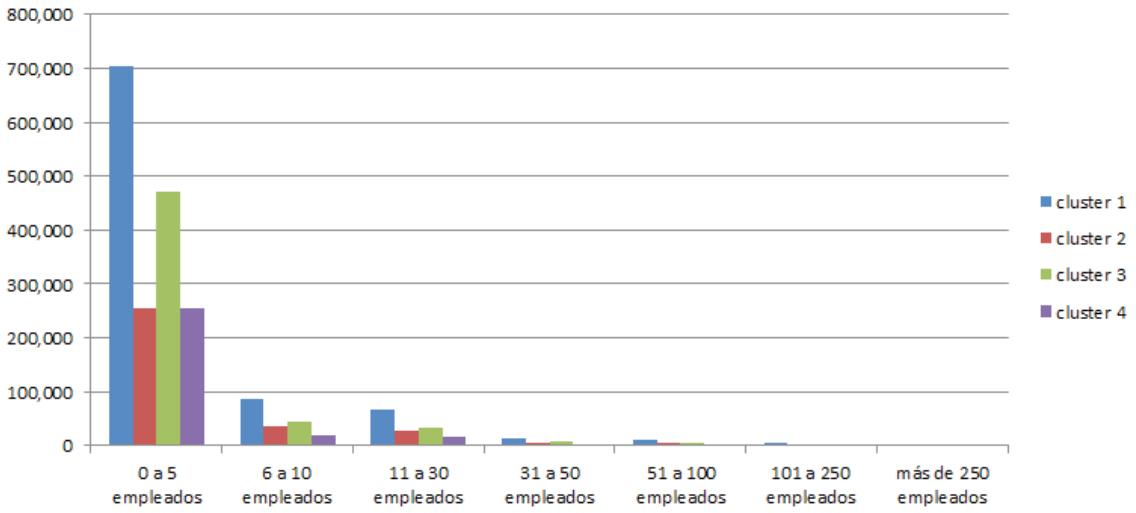

Figura 16. Comparativo de volumetría de unidades económicas DENUE por tamaño de empresa para 4 clusters

Implementación

Una forma de explotar los resultados del modelo es mediante su visualización en un mapa, la Figura 14 muestra la distribución geográfica de los clusters obtenidos.

La Figura 15 ejemplifica la comparación entre los clusters del volumen de unidades económicas por sector SCIAN, donde se observa que el sector económico 46 (Comercio al por menor) predomina en los 4 clusters.

La Figura 16 visualiza la comparación entre los cuatro clusters del volumen de unidades económicas por tamaño de empresa, donde se observa que el tamaño de empresa menor a 5 empleados predomina en todos los clusters.

Conclusiones

En el presente estudio se obtuvo un modelo de segmentación geográfica con la base pública del DENUE, aplicando la metodología CRISP-DM. El modelo agrupó las unidades económicas del país en 4 clusters donde se tiene que:

a) El cluster 1 agrupa, en general, entidades con alto desarrollo económico en el país (tal vez con excepción de Michoacán) ubicadas principalmente en el centro, occidente y sureste del país.

b) El cluster 2 apila entidades ubicadas en el norte del país que, en general, son de desarrollo económico medio y alto. Considerando el resultado de la validación del modelo (Figura 13) donde una observación del cluster 1 podría estar mal clasificada se supone que Nuevo León (clasificada en el cluster 1) bien podría pertenecer al cluster 2.

c) El cluster 3 concentra entidades que corresponden a un desarrollo económico medio y bajo, ubicadas en el sur y norte del país. Este es el cluster más numeroso (13 entidades) por lo que bien podría dividirse a su vez, en 2 sub-clusters.

d) El cluster 4 agrupa entidades del centro del país y es el cluster con menos entidades (3 en total, que son entidades de desarrollo económico bajo, excepto el Estado de México, que es una entidad de desarrollo económico alto).

e) El modelo agrupó las entidades conforme a tamaño de empresa y sector económico de sus unidades económicas, por lo que los resultados podrían considerarse inconsistentes en algunos casos, desde el punto de vista socioeconómico (por ejemplo, en el cluster 4 aparece Estado de México con Tlaxcala e Hidalgo). Para obtener una segmentación más precisa desde el enfoque socioeconómico sería necesario incluir otras variables en el modelo que no se encuentran actualmente en el DENUE, por ejemplo PIB estatal, índice de marginación estatal, etcétera.

Se proponen los siguientes pasos a fin de dar continuidad a este estudio:

- Obtener modelos específicos por sector económico (considerando en el modelo la rama, subrama y clase SCIAN).

- Aplicar alguna otra técnica en la etapa de modelado, por ejemplo: DBSCAN, para ver si las métricas de validación interna mejoran.

- Obtener el nivel de significancia de los clusters a fin de validar la calidad de los mismos con alguna métrica distinta a validación interna.

- Determinar cuáles son los sectores SCIAN con mejores perspectivas de crecimiento a mediano y largo plazo, lo que permitirá identificar los mejores clusters desde un enfoque de negocios.

- Aplicar alguna técnica (ya sea sobre todas o algunas de las variables descritas en el anexo I) a fin de determinar estadísticamente cuáles son las variables más significativas y utilizarlas en una segunda versión del modelo.

Agradecimientos

El autor agradece a Grupo Financiero Ve por Más S.A. de C.V. por su apoyo para realizar el presente artículo.

Anexo I

Lista de datos descargables desde sitio DENUE

1. ID asignado por el DENUE a la unidad económica
2. Nombre del establecimiento
3. Razón social
4. Código de actividad económica
5. Descripción de actividad económica
6. Tamaño del establecimiento
7. Tipo de vialidad donde se ubica la unidad económica

8. Nombre de vialidad donde se ubica la unidad económica
9. Tipo de vialidad 1 contigua a vialidad donde se ubica la unidad económica
10. Nombre de vialidad 1 contigua a vialidad donde se ubica la unidad económica
11. Tipo de vialidad 2 contigua a vialidad donde se ubica la unidad económica
12. Nombre de vialidad 2 contigua a vialidad donde se ubica la unidad económica
13. Tipo de vialidad 3 contigua a vialidad donde se ubica la unidad económica
14. Nombre de vialidad 3 contigua a vialidad donde se ubica la unidad económica
15. Número exterior donde se ubica la unidad económica
16. Letra exterior donde se ubica la unidad económica
17. Edificio donde se ubica la unidad económica
18. Número o letra de edificio
19. Número interior donde se ubica la unidad económica
20. Letra interior donde se ubica la unidad económica
21. Tipo de asentamiento donde se ubica la unidad económica
22. Nombre de asentamiento (colonia) donde se ubica la unidad económica
23. Tipo de centro comercial donde se ubica la unidad económica
24. Nombre de centro comercial donde se ubica la unidad económica
25. Número de local donde se ubica la unidad económica
26. Código postal donde se ubica la unidad económica
27. Clave de entidad donde se ubica unidad económica
28. Nombre de entidad donde se ubica unidad económica
29. Clave de municipio donde se ubica unidad económica
30. Nombre de municipio donde se ubica unidad económica
31. Clave de localidad donde se ubica unidad económica
32. Nombre de localidad donde se ubica unidad económica
33. AGEB (Área Geo Estadística Básica, extensión territorial correspondiente a la subdivisión de las áreas geo estadísticas municipales) donde se ubica la unidad económica
34. Manzana donde se ubica la unidad económica
35. Teléfono de contacto de la unidad económica
36. Correo electrónico de contacto de la unidad económica
37. Sitio Web de la unidad económica
38. Tipo de unidad económica (fijo, semifijo)
39. Coordenadas (latitud y longitud) donde se ubica la unidad económica
40. Fecha de alta en el DENUE de la unidad económica

Anexo II

Sectores económicos SCIAN

Código	Descripción
11	Agricultura, cría y explotación de animales, aprovechamiento forestal, pesca y caza
21	Minería
22	Generación, transmisión, distribución y comercialización de energía eléctrica, suministro de agua y de gas natural por ductos al consumidor final
23	Construcción
31	Industrias manufactureras
32	Industrias manufactureras
33	Industrias manufactureras
43	Comercio al por mayor
46	Comercio al por menor
48	Transportes, correos y almacenamiento
49	Transportes, correos y almacenamiento
51	Información en medios masivos
52	Servicios financieros y de seguros
53	Servicios inmobiliarios y de alquiler de bienes muebles e intangibles
54	Servicios profesionales, científicos y técnicos
55	Corporativos
56	Servicios de apoyo a los negocios y manejo de residuos, y servicios de remediación
61	Servicios educativos
62	Servicios de salud y de asistencia social
71	Servicios de esparcimiento culturales y deportivos, y otros servicios recreativos
72	Servicios de alojamiento temporal y de preparación de alimentos y bebidas
81	Otros servicios excepto actividades gubernamentales
93	Actividades legislativas, gubernamentales, de impartición de justicia y de organismos internacionales y extraterritoriales

Referencias

Alba, D. (2011). Detección de registros duplicados entre dos archivos digitales. CIMAT. Recuperado de https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/239/2/TE%20370.pdf

Amat, J. (Enero de 2016). Test estadísticos para variables cualitativas. Recuperado de https://rstudio-pubs-static.s3.amazonaws.com/220579_704e04beed0a499c8b4ac26dec006cd1.html

Amat, J. (2017). Clustering y heatmaps: Aprendizaje no supervisado. Recuperado de https://rpubs.com/Joaquin_AR/310338

Berzal, F. (2018). Clustering. Recuperado de https://elvex.ugr.es/decsai/intelligent/slides/dm/D3%20Clustering.pdf

DENUE-INEGI. (2018). Sitio oficial DENUE. Recuperado el 28 de noviembre de 2018 de DENUE http://www.beta.inegi.org.mx/app/mapa/denue/

Gallardo, J.A. (2018). Metodología para el desarrollo de proyectos en Minería de Datos CRIPS-DM. Recuperado de http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037

INEGI. (Septiembre de 2011). Directorio Estadístico Nacional de Unidades Económicas DENUE. Directorio Estadístico Nacional de Unidades Económicas DENUE. Recuperado de https://www.cepal.org/sites/default/files/events/files/mexico_denue.pdf

INEGI. (2012). Análisis de la demografía de los establecimientos. Recuperado de https://www.inegi.org.mx/investigacion/analisis/default.html#Metadatos

INEGI. (2017). DENUE Interactivo. Documento metodológico. Recuperado de http://www3.inegi.org.mx/contenidos/temas/economia/empresas/directorio/metodologias/DENUE_metodologia.pdf

Kassambara, A. (2018). Partitional Clustering in R: the Essentials. Data Novia. Recuperado de https://www.datanovia.com/en/lessons/clara-in-r-clustering-large-applications/

Marbán, M.S. (2018). A Data mining & knowledge discovery process model. Recuperado de http://cdn.intechopen.com/pdfs/5937/InTech-A_data_mining_amp_knowledge_discovery_process_model.pdf

Mathur, K. (2014). Comparative Study of K-Means and Hierarchical Clustering Techniques. Recuperado de https://www.researchgate.net/publication/293061584_Comparative_Study_of_K-Means_and_Hierarchical_Clustering_Techniques

Piatetsky, G. (Julio de 2002). What main methodology are you using for data mining? KD Nuggets. Recuperado de https://www.kdnuggets.com/polls/2002/methodology.htm

Piatetsky, G. (2004). Recuperado de https://www.kdnuggets.com/polls/2004/data_mining_methodology.htm

Piatetsky, G. (2007). Recuperado de https://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

R-CRAN. (2019). Tutorial to prepare train and test set using Data Preparation. Recuperado de https://cran.r-project.org/web/packages/dataPreparation/vignettes/train_test_prep.html

Secretaría de Comunicaciones y Transportes. (2016). Marco conceptual del Sistema de Clasificación Industrial de América del Norte. Recuperado de http://nats.sct.gob.mx/marco-conceptua-del-scian/

Wikipedia. (2018). Cross Industry Standrad Process for Data Mining. Recuperado de https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining