Ciencia Latina
Revista Multidisciplinar

DOI: https://doi.org/10.37811/cl_rcm.v7i1.5255

Uso de datamining y analisis de clúster para mejorar la productividad de Mipymes de la Ciudad de Pilar

Alberto Luis Ríos Vargas
riosvar2001@gmail.com
https://orcid.org/0000-0002-7064-5870
Fac. de Ciencias Aplicadas.
Universidad Nacional de Pilar.
Pilar - Paraguay

Brian Marin Rios Nicoli
brios95@protonmail.com
https://orcid.org/0000-0001-7191-828X
Fac. de Ciencias y Tecnología.
Universidad Católica Campus Itapúa.
Encarnación - Paraguay

RESUMEN

El objetivo de este trabajo es proporcionar a los usuarios empresariales conocimientos de las características operativas y de rendimiento en los aspectos de producción del negocio. El mismo presenta un enfoque metodológico para desarrollar una métrica práctica de recopilación de datos para la productividad basada en factores de influencia establecidos en los emprendimientos de carácter industrial.

Este estudio se realizará utilizando una técnica de **Datamining** denominada "**Clúster Analysis**" o análisis de clústeres, que permite identificar dentro de un conjunto de datos, un determinado grupo de usuarios según características comunes.

Palabras claves: *datamining; análisis de cluster; producción; industria.*

Correspondencia: riosvar2001@gmail.com
Artículo recibido 26 enero 2023 Aceptado para publicación: 26 febrero 2023
Conflictos de Interés: Ninguna que declarar

Cómo citar: Ríos Vargas, A. L., & Rios Nicoli, B. M. (2023). Uso de datamining y analisis de clúster para mejorar la productividad de Mipymes de la Ciudad de Pilar. Ciencia Latina Revista Científica Multidisciplinar, 7(1), 10793-10804. https://doi.org/10.37811/cl_rcm.v7i1.5255

Use of datamining and cluster analysis to improve the productivity of MSMEs in the city of Pilar

ABSTRACT

The goal of this paper is to provide business users with insights into operational and performance characteristics in production aspects of business. It presents a methodological approach to develop a practical data collection metric for productivity based on influence factors established in industrial companies.

This study will be carried out using a Datamining technique called "Cluster Analysis", which allows a certain group of users to be identified within a data set according to common characteristics.

Palabras claves: *datamining; cluster analysis; production; industry.*

1. INTRODUCCIÓN

Al considerar una empresa, el análisis a través de datamining se refiere a comprobar y explorar grandes cantidades de conjuntos de datos para extraer y verificar los patrones, conexiones y tendencias, y también para tener una mejor idea de la cadena de suministro. Eso ayuda a mantener la fabricación y productividad en el camino correcto y también para averiguar las tendencias de los empleados y relaciones con los clientes.

El presente estudio se llevará a cabo en tres fases: en la primera, se llevará a cabo la recopilación de datos a través de encuestas a trabajadores industriales, la segunda, en la que este conjunto de datos se agrupará en distintos clústeres, y la tercera, en donde se analizará el volumen de los distintos clústeres para determinar los factores que puedan contribuir a la mejora de la producción y, posteriormente, la productividad de la manufacturera.

Datamining

El desarrollo de la tecnología de la información ha generado una gran cantidad de bases de datos y enormes datos en varias áreas. Un ejemplo claro es el área de marketing digital, donde se utilizar la minería de datos para diseñar campañas de marketing personalizadas para cada tipo de cliente, a partir de una gran cantidad de datos, como información de facturación, correo electrónico, mensajes de texto, transmisiones de datos web y servicio al cliente.

La investigación en bases de datos y tecnologías de la información ha dado lugar a un enfoque para almacenar y manipular estos datos valiosos para tomar decisiones adicionales. La minería de datos es un proceso de extracción de información útil y patrones de datos enormes. También se denomina proceso de descubrimiento de conocimientos, minería de conocimientos a partir de datos, extracción de conocimientos o análisis de datos / patrones.

La minería de datos es un proceso lógico que se utiliza para buscar en una gran cantidad de datos con el fin de encontrar datos útiles.

El objetivo de esta técnica es encontrar patrones que antes se desconocían. Una vez estos se encuentran se pueden utilizar además para tomar ciertas decisiones para el desarrollo de sus negocios u otras áreas.

Análisis de Clústeres

El análisis de clústeres es un método estadístico para procesar datos. Funciona organizando elementos en grupos, o agrupaciones, sobre la base de cuán estrechamente asociados están.

El análisis de clústeres, al igual que el análisis de espacio reducido (análisis factorial), se ocupa de matrices de datos en las que las variables no se han dividido de antemano en subconjuntos de criterio versus predictores. El objetivo del análisis de clústeres es encontrar grupos similares de sujetos, donde la "similitud" entre cada par de sujetos significa alguna medida global sobre todo el conjunto de características.

El análisis de clústeres es un algoritmo de aprendizaje no supervisado, lo que significa que no sabe cuántos clústeres existen en los datos antes de ejecutar el modelo. A diferencia de muchos otros métodos estadísticos, el análisis de clústeres se utiliza normalmente cuando no se hace una suposición sobre las posibles relaciones entre los datos. Proporciona información sobre dónde existen asociaciones y patrones en los datos, pero no cuáles podrían ser o qué significan.

¿Cómo es utilizado?

El uso más común del análisis de clústeres es la clasificación. Los sujetos se separan en grupos para que cada sujeto sea más similar a otros sujetos de su grupo que a sujetos fuera del grupo.

En un contexto de investigación de mercado, esto podría usarse para identificar categorías como grupos de edad, tramos de ingresos, ubicación urbana, rural o suburbana.

En marketing, el análisis de clústeres se puede utilizar para la segmentación de la audiencia, de modo que los diferentes grupos de clientes puedan dirigirse con los mensajes más relevantes.

Los investigadores de la salud pueden usar el análisis de conglomerados para averiguar si diferentes áreas geográficas están vinculadas con niveles altos o bajos de ciertas enfermedades, de modo que puedan investigar los posibles factores locales que contribuyen a los problemas de salud.

Cualquiera que sea la aplicación, la limpieza de datos es un paso preparatorio esencial para un análisis de clúster exitoso. La agrupación funciona a nivel de conjunto de datos donde cada punto se evalúa en relación con los demás, por lo que los datos deben ser lo

más completos posible. La agrupación se mide mediante la distancia entre grupos y entre grupos.

En el trabajo "E-Learning Challenges Faced by Academics in Higher Education" de la Universidad de Sheffield Hallam, se dice lo siguiente: Al revisar la literatura sobre e-learning, hay varias críticas a la calidad de los sistemas de e-learning que se están utilizando actualmente. Se han planteado problemas que incluyen: problemas de usabilidad, mal desempeño, instituciones que no pueden personalizar de acuerdo con sus requisitos y, a veces, críticas por tener un sistema centrado en el maestro en lugar de estar centrado en el alumno (Chua y Dyson, 2004).

Justificación en términos de Necesidades y Pertinencia.

Las empresas requieren una revisión analítica en profundidad de los datos de producción para comprender mejor su entorno empresarial y su capacidad de competitividad. En la era de la información, las empresas deben ver los datos recopilados como una ventaja competitiva. La minería de datos y el análisis de clústeres son técnicas prometedoras para aprovechar el valor potencial de los datos que se encuentran en las organizaciones.

La aplicación de minería de datos y otras herramientas de análisis de datos produce información útil o funciones relacionales que ayudan a los gerentes de manufactureras a tomar decisiones positivas para la empresa.

2. METODOLOGÍA

Un proceso de revisión debe guiarse por preguntas de revisión, con el fin de organizar y definir el conocimiento existente sobre el tema elegido.

En la bibliografía consultada se define este tipo de revisión como una síntesis de la investigación realizada de manera sistemática, transparente y reproducible, con el objetivo de mejorar la base de conocimiento existente e informar sobre la practicidad y las prácticas existentes en la educación superior.

La metodología consta de las siguientes etapas:

En la primera etapa, se determinarán las principales razones para buscar investigar sobre este tema, que se expresará en forma de preguntas de revisión.

Luego, se desarrollará un protocolo de revisión para utilizarlo para extraer información relevante al tema.

En la segunda etapa, se extraerán los artículos de los que se extraerá la información relevante. El protocolo de revisión se centra en la identificación y selección de artículos

de acuerdo con su relevancia para el tema estudiado, así como la eliminación de fuentes no relacionadas con el tema. Una vez hecho esto, el proceso de síntesis de la información relevante continuará.

En la tercera etapa, se realizará el estudio de campo. Utilizando una herramienta de encuestas, se hará un sondeo en las distintas empresas, a modo de recopilar el mayor volumen de datos posible, para poder realizar el procesamiento de los mismos.

En la etapa final, los resultados de la revisión se informarán mediante un informe escrito con la ayuda de tablas explicativas, así como las conclusiones y recomendaciones pertinentes para académicos y profesionales interesados en este tema.

Para el análisis de los datos se utilizará el lenguaje y entorno de desarrollo R. R proporciona una amplia variedad de técnicas estadísticas (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupamiento, …) y técnicas gráficas, y es altamente extensible. El lenguaje S suele ser el vehículo elegido para la investigación en metodología estadística, y R proporciona una ruta de código abierto para participar en esa actividad.

El entorno R.

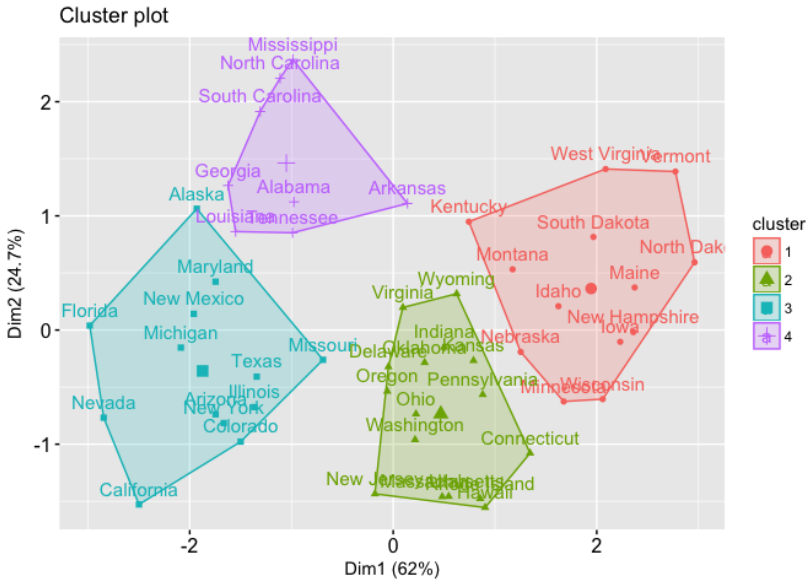


Figura 1: Ejemplo de salida en R.

R es un conjunto integrado de instalaciones de software para la manipulación de datos, el cálculo y la visualización gráfica. Incluye:

- Una instalación efectiva de manejo y almacenamiento de datos,
- Un conjunto de operadores para cálculos en arreglos, en particular matrices,

- Una colección grande, coherente e integrada de herramientas intermedias para el análisis de datos,
- Facilidades gráficas para el análisis y visualización de datos, ya sea en pantalla o en papel, y
- Un lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario y facilidades de entrada y salida.

Preparación de Datos.

Tabla 1: *Ejemplo de conjuntos de datos (Producción industrial de la ciudad de Pilar)*

Departamento, distrito, sector y rama de actividad económica	Unidades económicas	Personal ocupado			Total de Remuneraciones	Gastos por compras de bienes y servicios	Ingresos por suministro de bienes y servicios
		Total	Hombre	Mujer	(En miles de Gs)	(En miles de Gs)	(En miles de Gs)
Industria	**43**	**1,335**	**1,112**	**223**	**33,358,542**	**149,955,349**	**196,026,977**
Elaboración de productos lácteos	3	10	7	3	55,406	208,153	306,277
Elaboración de productos de molinería, almidones y productos derivados del almidón	3	3	2	1	-	65,902	87,840
Elaboración de otros productos alimenticios n.c.p.	34	196	124	72	1,506,510	5,870,135	8,644,907
Hilandería, tejeduría y acabado de productos textiles	3	1,126	979	147	31,796,626	143,811,159	186,987,954

Fuente: INE

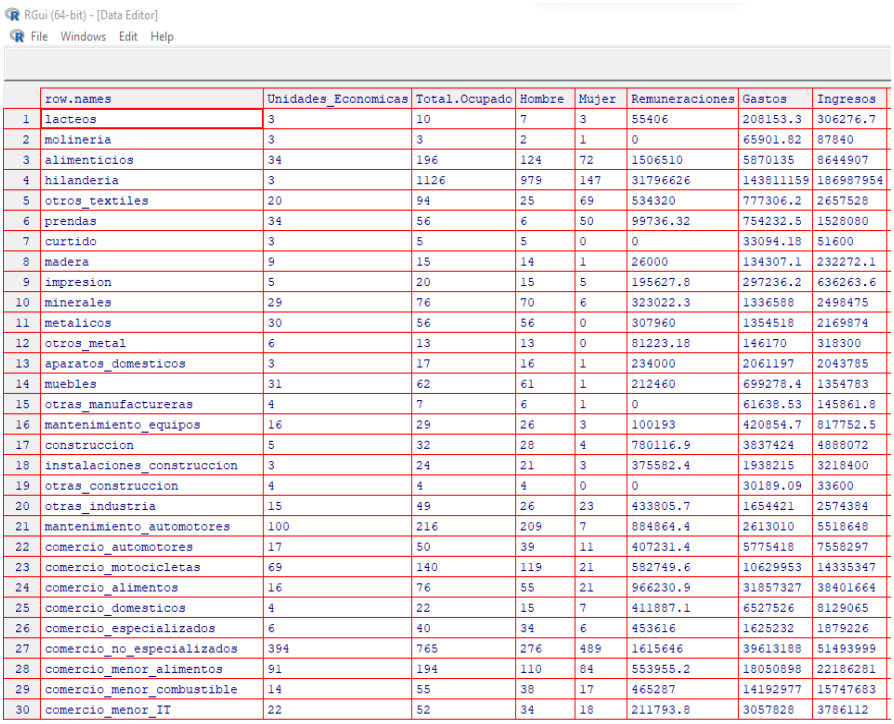
Para realizar un análisis de clústeres en R, generalmente, los datos deben prepararse de la siguiente manera:

- Las filas son observaciones (individuos) y las columnas son variables
- Cualquier valor faltante en los datos debe ser eliminado o estimado.
- Los datos deben estar estandarizados (es decir, escalados) para que las variables sean comparables. La estandarización consiste en transformar las variables de manera que tengan media cero y desviación estándar uno.

Los datos industriales de a ser analizados serán recopilados por los alumnos del 1er curso de la carrera de Ingeniería Industrial, y serán ordenados teniendo en cuenta la siguiente tabla:

Nombre de la Industria (Rama de actividad económica)	Personal Ocupado	Gastos por compras de bienes y servicios	Ingresos por suministro de bienes y servicios

Figura 2: Datos *de la industria pilarense extraídos del Instituto Nacional de Estadística.*



RGui (64-bit) - [Data Editor]
File Windows Edit Help

	row.names	Unidades_Economicas	Total.Ocupado	Hombre	Mujer	Remuneraciones	Gastos	Ingresos
1	lacteos	3	10	7	3	55406	208153.3	306276.7
2	molineria	3	3	2	1	0	65901.82	87840
3	alimenticios	34	196	124	72	1506510	5870135	8644907
4	hilanderia	3	1126	979	147	31796626	143811159	186987954
5	otros_textiles	20	94	25	69	534320	777306.2	2657528
6	prendas	34	56	6	50	99736.32	754232.5	1528080
7	curtido	3	5	5	0	0	33094.18	51600
8	madera	9	15	14	1	26000	134307.1	232272.1
9	impresion	5	20	15	5	195627.8	297236.2	636263.6
10	minerales	29	76	70	6	323022.3	1336588	2498475
11	metalicos	30	56	56	0	307960	1354518	2169874
12	otros_metal	6	13	13	0	81223.18	146170	318300
13	aparatos_domesticos	3	17	16	1	234000	2061197	2043785
14	muebles	31	62	61	1	212460	699278.4	1354783
15	otras_manufactureras	4	7	6	1	0	61638.53	145861.8
16	mantenimiento_equipos	16	29	26	3	100193	420854.7	817752.5
17	construccion	5	32	28	4	780116.9	3837424	4888072
18	instalaciones_construccion	3	24	21	3	375582.4	1938215	3218400
19	otras_construccion	4	4	4	0	0	30189.09	33600
20	otras_industria	15	49	26	23	433805.7	1654421	2574384
21	mantenimiento_automotores	100	216	209	7	884864.4	2613010	5518648
22	comercio_automotores	17	50	39	11	407231.4	5775418	7558297
23	comercio_motocicletas	69	140	119	21	582749.6	10629953	14335347
24	comercio_alimentos	16	76	55	21	966230.9	31857327	38401664
25	comercio_domesticos	4	22	15	7	411887.1	6527526	8129065
26	comercio_especializados	6	40	34	6	453616	1625232	1879226
27	comercio_no_especializados	394	765	276	489	1615646	39613188	51493999
28	comercio_menor_alimentos	91	194	110	84	553955.2	18050898	22186281
29	comercio_menor_combustible	14	55	38	17	465287	14192977	15747683
30	comercio_menor_IT	22	52	34	18	211793.8	3057828	3786112

Como no queremos que el algoritmo de agrupamiento dependa de una unidad variable arbitraria, comenzamos por escalar/estandarizar los datos usando la función *scale* de R. La función scale () en lenguaje R es una función genérica que centra y escala las columnas de una matriz numérica. El parámetro central toma un vector numérico similar o un valor lógico. Si se proporciona el vector numérico, entonces a cada columna de la matriz se le resta el valor correspondiente desde el centro. Si el valor lógico es VERDADERO, las medias de las columnas de la matriz se restan de sus columnas correspondientes. La escala toma un vector numérico similar o un valor lógico.

Figura 3: *Datos escalados listos para procesamiento.*

	row.names	Unidades_Economicas	Total.Ocupado	Hombre	Mujer	Remuneraciones	Gastos	Ingresos	v
1	lacteos	-0.465811	-0.4426895	-0.3641067	-0.4248149	-0.2116465	-0.3072084	-0.3176162	
2	molineria	-0.465811	-0.4839738	-0.4036185	-0.4541278	-0.2257269	-0.3144	-0.3261611	
3	alimenticios	0.1097019	0.654294	0.5604681	0.5864799	0.157126	-0.02096585	0.008578569	
4	hilanderia	-0.465811	6.139211	7.316977	1.685713	7.854823	6.95267	6.985093	
5	otros_textiles	-0.1502072	0.05272241	-0.2218644	0.5425106	-0.08993897	-0.2784348	-0.2256388	
6	prendas	0.1097019	-0.1713925	-0.3720091	0.2640381	-0.2003807	-0.2796013	-0.2698211	
7	curtido	-0.465811	-0.4721783	-0.3799114	-0.4687842	-0.2257269	-0.3160586	-0.3275788	
8	madera	-0.3544214	-0.4132007	-0.3087903	-0.4541278	-0.2191195	-0.3109418	-0.3205112	
9	impresion	-0.4286812	-0.3837119	-0.3008879	-0.395502	-0.1760116	-0.3027048	-0.3047076	
10	minerales	0.01687721	-0.05343729	0.1337413	-0.3808455	-0.1436365	-0.2501602	-0.2318607	
11	metalicos	0.03544214	-0.1713925	0.02310838	-0.4687842	-0.1474644	-0.2492538	-0.244715	
12	otros_metal	-0.4101162	-0.4249962	-0.3166926	-0.4687842	-0.2050855	-0.310342	-0.3171459	
13	aparatos_domesticos	-0.465811	-0.4014052	-0.2929856	-0.4541278	-0.16626	-0.2135274	-0.2496475	
14	muebles	0.05400708	-0.1360059	0.06262013	-0.4541278	-0.171734	-0.2823795	-0.2766002	
15	otras_manufactureras	-0.4472461	-0.4603828	-0.3720091	-0.4541278	-0.2257269	-0.3146155	-0.3238914	
16	mantenimiento_equipos	-0.2244669	-0.330632	-0.2139621	-0.4248149	-0.2002647	-0.2964553	-0.297608	
17	construccion	-0.4286812	-0.3129388	-0.1981574	-0.4101584	-0.02747404	-0.1237299	-0.1383832	
18	instalaciones_construccion	-0.465811	-0.3601208	-0.2534738	-0.4248149	-0.1302793	-0.2197448	-0.2036983	
19	otras_construccion	-0.4472461	-0.4780761	-0.3878138	-0.4687842	-0.2257269	-0.3162055	-0.3282829	
20	otras_industria	-0.2430318	-0.2126768	-0.2139621	-0.131686	-0.1154829	-0.2340921	-0.2288912	
21	mantenimiento_automotores	1.334987	0.7722492	1.232168	-0.3661891	-0.0008543133	-0.1856304	-0.113716	
22	comercio_automotores	-0.205902	-0.2067791	-0.1112316	-0.3075633	-0.1222363	-0.02575427	-0.03392799	
23	comercio_motocicletas	0.7594745	0.3240194	0.5209564	-0.1609988	-0.07763144	0.2196677	0.2311801	
24	comercio_alimentos	-0.2244669	-0.05343729	0.01520604	-0.1609988	0.01982355	1.292822	1.172619	
25	comercio_domesticos	-0.4472461	-0.3719164	-0.3008879	-0.3661891	-0.1210531	0.01226872	-0.01160038	
26	comercio_especializados	-0.4101162	-0.2657567	-0.1507433	-0.3808455	-0.1104485	-0.2355677	-0.2560848	
27	comercio_no_especializados	6.793078	4.01012	1.761625	6.698218	0.1848609	1.684921	1.684771	
28	comercio_menor_alimentos	1.167903	0.6424985	0.4498352	0.7623573	-0.08494904	0.5948351	0.5382969	
29	comercio_menor_combustible	-0.2615968	-0.1772903	-0.1191339	-0.2196246	-0.1074825	0.3997971	0.2864286	
30	comercio_menor_IT	-0.1130773	-0.1949835	-0.1507433	-0.2049682	-0.1719033	-0.1631426	-0.1814902	

K-means.

La agrupación en clústeres de K-medias es el algoritmo de aprendizaje automático no supervisado más utilizado para dividir un conjunto de datos dado en un conjunto de k grupos (es decir, k clústeres), donde k representa la cantidad de grupos especificados previamente por el analista. Clasifica objetos en múltiples grupos (es decir, clústeres), de modo que los objetos dentro del mismo conglomerado son lo más similares posible (es decir, alta similitud intraclase), mientras que los objetos de diferentes conglomerados son lo más diferentes posible (es decir, baja inter-clase). semejanza de clase). En el agrupamiento de k-medias, cada grupo está representado por su centro (es decir, centroide) que corresponde a la media de los puntos asignados al grupo.

Cálculo de k-means en *R*

La función k-means.

Los datos proporcionados por 'x' se agrupan mediante el método k-means, que tiene como objetivo dividir los puntos en grupos de modo que se minimice la suma de los cuadrados de los puntos a los centros de conglomerados asignados.

Como mínimo, todos los centros de conglomerados están en la media de sus conjuntos de Voronoi (el conjunto de puntos de datos que están más cerca del centro de conglomerados).

Podemos calcular k-means en R con la función kmeans. Aquí se agrupará los datos en dos grupos (centers = 2). La función kmeans también tiene una opción nstart que intenta

múltiples configuraciones iniciales e informa sobre la mejor. Por ejemplo, agregar nstart = 25 generará 25 configuraciones iniciales

En la siguiente figura se utilizaron 3 centroides (k) y arrojo como resultado 3 clústeres de dimensión 1, 62 y 3. Abajo se denotan los vectores y a que grupo pertenecen.

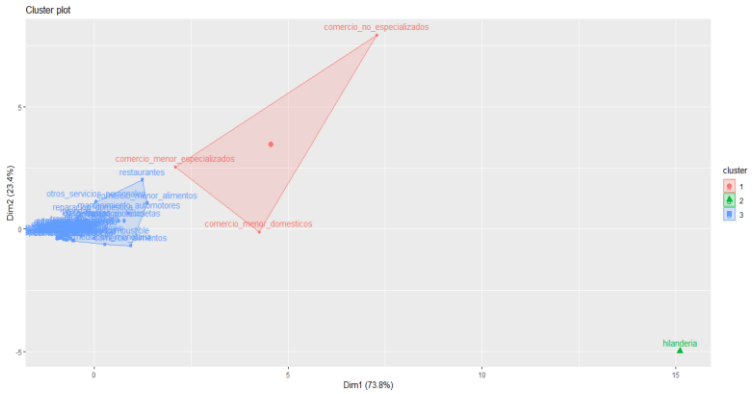
Figura 4: K-means clustering en *R*

```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
> k3 <- kmeans(ind, centers = 3, nstart = 25)
> print(k3)
K-means clustering with 3 clusters of sizes 1, 62, 3

Cluster means:
  Unidades_Economicas Total.Ocupado     Hombre      Mujer Remuneraciones
1          -0.4658110     6.1392114  7.3169767  1.6857134      7.8548228
2          -0.1496083    -0.2032594 -0.1707541 -0.1884206     -0.1459845
3           3.2471754     2.1542911  1.0899255  3.3321209      0.3987388
      Gastos   Ingresos
1  6.9526703  6.9850926
2 -0.1909758 -0.1914986
3  1.6292775  1.6292725

Clustering vector:
                      lacteos                     molineria
                            2                             2
                  alimenticios                   hilanderia
                            2                             1
                otros_textiles                      prendas
                            2                             2
                       curtido                       madera
                            2                             2
                     impresion                     minerales
                            2                             2
                     metalicos                   otros_metal
                            2                             2
           aparatos_domesticos                      muebles
                            2                             2
         otras_manufactureras         mantenimiento_equipos
                            2                             2
                  construccion instalaciones construccion
```

También podemos ver nuestros resultados usando fviz_cluster. Esto proporciona una buena ilustración de los grupos. Si hay más de dos dimensiones (variables), fviz_cluster realizará un análisis de componentes principales (PCA) y trazará los puntos de datos de acuerdo con los dos primeros componentes principales que explican la mayor parte de la varianza.

3. RESULTADOS Y DISCUSIÓN

Figura 5: *Resultado del análisis de clústeres.*



Se estos resultados, y atendiendo a nuestro dataset podemos extraer, por ejemplo, como la industria de Hilandería mueve muchísimo más capital que el resto, y que las ramas de actividad económica *Comercio al por menor en comercios no especializados, Comercio al*

por menor de otros artículos y equipos de uso doméstico en comercios especializados y *Comercio al por menor de otros artículos en comercios especializados* poseen una mayor cantidad de unidades económicas, pero no tanto movimiento de capital como la industria de hilandería.

4. CONCLUSIONES

En pocas palabras, los clústeres industriales son aglomeraciones regionales de industrias relacionadas. Los clústeres están formados por empresas, proveedores y prestadores de servicios, así como instituciones estatales y otras instituciones que brindan educación, información, investigación y apoyo técnico a la economía regional. Se puede decir entonces que un clúster es una red de relaciones económicas que crea una ventaja competitiva para las empresas relacionadas en una región determinada. Crear nuevos clústeres económicos se convierte en un incentivo para que industrias similares y sus proveedores se desarrollen o se trasladen a la región.

Una forma de crear clústeres completamente nuevos en la región es desarrollar estrategias que mejoren el entorno comercial general, mejoren las habilidades, el acceso a financiamiento e infraestructura, simplifiquen las regulaciones gubernamentales, respalden las necesidades locales y abran la inversión y la competencia extranjeras.

5. LISTA DE REFERENCIAS

Nurul Islam, Martin Beer, Frances Slack. 2015. E-Learning Challenges Faced by Academics in Higher Education: A Literature Review. Sheffield Hallam University, UK. Publicado por Redfame Publishing.

Bharati M. Ramageri. DATA MINING TECHNIQUES AND APPLICATIONS. Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305. 2010

Documentación NCSS. Clustering in NCSS. 2020https://www.ncss.com/software/ncss/clustering-in-ncss/

Sunil Kumar. 5 Common Problems Faced By Students In eLearning And How To Overcome Them. 2015https://elearningindustry.com/5-common-problems-faced-by-students-in-elearning-overcome

Explorium Data Science Team. Clustering — When You Should Use it and Avoid It. 2020https://www.explorium.ai/blog/clustering-when-you-should-use-it-and-avoid-it/

Tabla 1. Instituto Nacional de Estadística (INE)

Figura 1. scikit-learn.org

Fabián Pedregosa; Gael Varoquaux; Alejandro Gramfort; Vicente Michel; Bertrand Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vicente Dubourg; Jake Vanderplas; Alejandro Passos; David Cournapeau; Matthieu Perrot; Édouard Duchesnay (2011). "Scikit-learn: aprendizaje automático en Python" . Revista de investigación de aprendizaje automático.

Tabla 1. Instituto Nacional de Estadística (INE)

Figura 1. scikit-learn.org

Fabián Pedregosa; Gael Varoquaux; Alejandro Gramfort; Vicente Michel; Bertrand Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vicente Dubourg; Jake Vanderplas; Alejandro Passos; David Cournapeau; Matthieu Perrot; Édouard Duchesnay (2011). "Scikit-learn: aprendizaje automático en Python" . Revista de investigación de aprendizaje automático.

Manjarrés Betancourt, Juan Carlos. 8 algoritmos de agrupación en clústeres en el aprendizaje automático que todos los científicos de datos deben conocer.https://www.freecodecamp.org/espanol/news/8-algoritmos-de-agrupacion-en-clusteres-en-el-aprendizaje-automatico-que-todos-los-cientificos-de-datos-deben-conocer/

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). Sthda.

George Seif (2018). *The 5 Clustering Algorithms Data Scientists Need to Know.* (Towards Data Science)

RDocumentation, Nick Carchedi, https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans