ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELOS ESTADÍSTICOS PARA LA DETECCIÓN DE PATRONES EN MEDIO AMBIENTE Y FINANZAS

SEGMENTACIÓN DE LOS ANUNCIANTES EN UNA REVISTA UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

MARIO ALEJANDRO ALDEÁN JIMÉNEZ
mario.aldean@epn.edu.ec

Director: PH. D. MIGUEL ALFONSO FLORES SÁNCHEZ
miguel.flores@epn.edu.ec

QUITO D.M. , FEBRERO 2022

CERTIFICACIONES

Yo MARIO ALEJANDRO ALDEÁN JIMÉNEZ, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

Mario Alejandro Aldeán Jiménez

Certifico que el presente trabajo de integración curricular fue desarrollado por MARIO ALEJANDRO ALDEÁN JIMÉNEZ, bajo mi supervisión.

Ph. D. Miguel Alfonso Flores Sánchez
Director del Proyecto

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

MARIO ALEJANDRO ALDEÁN JIMÉNEZ

PH. D. MIGUEL ALFONSO FLORES SÁNCHEZ

AGRADECIMIENTOS

A mis padres, los cuales hicieron posible la construcción de este trabajo y a mi director, el cual dió guía y claridad al estudio.

Índice general

Resumen

En el presente trabajo se estudia la construcción de grupos de empresas similares a través de la publicidad que estos publican. Segmentando una base de datos de anuncios en clases, permitiéndonos caracterizar estos grupos. Para esto se empieza por estudiar teóricamente el algoritmo necesario para separar estos anuncios en grupos, posteriormente se realiza un ejemplo en R que incluye en sus conclusiones una interpretación basada en minería de datos y una caracterización de cada grupo.

Abstract

In the present work the construction of groups of similar companies is studied through the ads they publish, segmenting a database of ads into classes, which allow us to understand these ad groups. To do this, a theoretical study of the necessary algorithm is conducted, and example is coded into R and conclusions are given based on data mining and group characterization.

Capítulo 1

Introducción

1.1. Objetivo General

Desarrollar e implementar una segmentación de anuncios en un medio de comunicación escrito, tomando en cuenta sus productos o servicios ofertados y las características de las empresas que los anuncios representan; a través del algoritmo de agrupamiento U-k-medias.

1.2. Objetivos Específicos

- Construir una base de datos de anuncios que contenga información tanto del anuncio como de la empresa que paga por publicarlo.
- Estudiar e implementar el método de U-k-medias en R y segmentar la base de datos en grupos de anuncios similares.

1.3. Alcance

Elaborar una segmentación de anunciantes a partir de una previa construcción de una base de datos que incluya empresas anunciantes en un medio escrito, esta base de datos describirá de cada empresa las líneas de negocio en las que esta participa, además de su cantidad de empleados y sus ingresos. Esta segmentación será evaluada mediante distintos criterios de validez como el índice de Dunn DNo, el índices Davies-Bouldin DB, y el valor de silueta.

1.4. Marco Teórico

Empresas como revistas, periódicos y otros medios de comunicación tienen como modelo de negocios el difundir información al público (Iankova, 2019). Este tipo de compañías poseen principalmente dos formas de ingresos, el dinero obtenido por las suscripciones de los lectores y los ingresos por anuncios publicitarios (Ellman, 2009).

Los anuncios que estas empresas de comunicación venden pueden ser entendidos como productos. Se propone entonces desarrollar un método que permita identificar posibles ediciones que publicará un medio de comunicación, siendo estos pertenecientes a una misma categoría; por ejemplo, en una edición de cosméticos se pueden ofertar anuncios de maquillaje y artículos para la piel, en una edición de electrónicos se pueden ofertar anuncios de computadoras, periféricos, entre otros.

Las distintas compañías que anuncian sus productos o servicios en los medios son considerados como clientes B2B del inglés Business to Business, esto es, son una compañía que es cliente de otra. Este tipo de clientela es la que se analizará en este proyecto. Estos anunciantes tienen en su mayoría diversas líneas de negocios (Medina, 2005), ofreciendo diversos servicios tanto a consumidores como a otras empresas.

Ahora, puesto que el medio de comunicación deberá elegir qué temática será la base de su siguiente edición es necesario conocer que tema será atractivo para los anunciantes (Mehta, 2000). Se propone encontrar grupos de empresas con similares líneas de negocios o productos ofertados, las cuales podrían comprar anuncios publicitarios en una misma edición de la revista con una temática determinada debido a la influencia que esta genera entre sus lectores (Hamouda, 2018).

1.4.1. Planificación de Anuncios y Agrupación de Anunciantes

Los métodos para la elección del tema principal del medio son puramente empíricos, es decir, basados en las experiencias anteriores de los publicistas, gerentes comerciales y creativos del medio (Abrahamson, 2015). Es por tanto que generar grupos de empresas similares es importante para el publicista, pues nos permite saber que empresas podrían aparecer en una determinada edición mediante el pago de anuncios.

Por esta razón se propone un método de *clustering* que permita identificar secto-

res de empresas similares, que puedan anunciarse en el medio. Los métodos *clustering* tratan de agrupar elementos en clústeres o grupos (Xu, 2008), teniendo como objetivo que estos grupos de empresas similares sean tomados en cuenta al diseñar la edición. Por ejemplo, si es un grupo de empresas de venta de aparatos electrónicos, la edición tendrá como temática central las utilidades de estos aparatos.

Este tipo de métodos son conocidos como aprendizaje no supervisado puesto no se conoce el número de grupos de empresas que habrán, sin embargo, es posible que el algoritmo determine cuántos de estos grupos son necesarios mediante el método de clustering no supervisado de U-k-medias (Sinaga, 2020) usando medidas como la disimilaridad en los grupos. Esto nos libera de la necesidad de encontrar el número óptimo de anuncios existentes en la base de datos (Ienco, 2012)

Este tipo de métodos encontrará que empresas tienen líneas de negocios similares (Sinaga, 2020) de una lista de líneas de negocios generada para cada empresa. El método de K-medias ha sido estudiado con varias extensiones en la literatura y aplicado en distintas áreas (Alhawarat, 2018). Sin embargo este tipo de algoritmos son usualmente afectados por inicializaciones y la necesidad de recibir un número inicial de colecciones. De no conocerse estas colecciones índices de validez pueden ser usados para encontrar un número de clústeres adecuado (Sinaga, 2020). Varios índices de validez han sido propuestos como el criterio de información bayesiano (BIC) (Kass, 1995), el criterio de información de Akaike (Bozdogan, 1987), entre otros.

Al desconocer cuantos conjuntos de empresas existen en realidad es necesario usar un método que no nos restrinja al encontrar el número real de grupos, por lo que se considera el algoritmo de U-k-medias. Este procedimiento de aprendizaje puede encontrar automáticamente el número de clústeres sin ningún tipo de inicialización o de selección de parámetros (Sinaga, 2020).

Para el desarrollo de un producto, en este caso una edición, es necesario conocer a sus patrocinadores, por tanto la investigación de los anunciantes es un paso importante en la planeación de una publicación (Abrahamson, 2015). Se tiene entonces que identificar los posibles clientes del ejemplar que se pretende lanzar al público, sin embargo, todo medio periodístico está en contacto con cientos de empresas, algunas prestando servicios similares o iguales.

La identificación o caracterización de los grupos formados por el algoritmo vendrá entonces de los atributos más frecuentes de los elementos del conjunto (Alhawarat, 2018). Ahora, al estar nuestro algoritmo enfocado en encontrar cuan parecidos son los elementos de un conjunto (Xu, 2008), el método clustering agrupará las em-

presas con líneas de negocios similares.

Al agrupar estas empresas se tienen grupos homogéneos que pueden ser identificados como un conjunto de clientes similares (Ienco, 2012), para los cuales estará enfocada una determinada publicación; teniendo así una forma de definir las posibles temáticas con las que se podría construir una edición del medio. Es importante notar que así las temáticas presentes en la revista se pueden acoplar a los distintos canales o servicios que ofrecen las compañías anunciantes (Iankova, 2019).

La planificación de los anuncios se puede entender como una planificación de productos, puesto que son elementos que la empresa vende a terceros, sujetos a una demanda que depende del tamaño del público (Kahn, 2000). Es por tanto que se debe entender este proyecto como una planificación de la producción, proveyendo una programación de la publicidad a ofertar, lo cual nos permitirá ajustar planes operativos anuales, presupuestos entre otros tipos de organización (Kahn, 2020).

Este tipo de planificación ha venido tomando fuerza desde los años setenta y son parte del proceso organizacional de diversas empresas a lo largo y ancho del mundo (Feldman, 1984) usando métodos de agrupación y clustering para indicar productos que optimizarán ingresos y aumentarán la satisfacción de los compradores (Natchwey, 2009).

La publicidad es un tema de debates en el área de psicología y relaciones interpersonales, puesto que influyen de gran manera en los contenidos del medio (Burgoon, 1982). Por tanto la selección de la temática mensual de una revista o publicación depende del público al que esta está dirigida (Buckinx, 2005). Con esto en mente se tiene como objetivo caracterizar a estos clientes, de manera que la temática seleccionada esté relacionada con los anunciantes.

Hoy en día muchas plataformas dependen de los anuncios para sobrevivir, desde compañías de internet hasta medios impresos (Ellman, 2009), siendo estos últimos los beneficiarios de ésta investigación al poder controlar la temática de las publicaciones y quien anuncia en cada una de ellas. Por otro lado el valor de un anuncio depende de cuan interesados estén sus lectores en el tema y de cómo el lector perciba la utilidad de aquello que se anuncia para sí mismo (Ducoffe, 1996), es por tanto que al elegir una temática interesante, que tenga una propuesta de valor atractiva se puede crear una publicación llamativa para los anunciantes y así lograr aumentar los dividendos percibidos por el publicador.

1.4.2. El algoritmo U-k-medias

Para buscar un tipo no supervisado de algoritmo necesitamos solucionar el problema de la inicialización, puesto que este afecta al algoritmo y requiere un número *a priori*. Para construir un algoritmo de agrupación libre de inicializaciones y que encuentre automáticamente el número de clústeres usamos el concepto de la entropía. Primero consideraremos proporciones α_k en las cuales el término α_k es visto como la probabilidad de que uno de los puntos pertenezca a la k-ésima clase. Por tanto, usaremos $-\ln \alpha_k$ como la información en la ocurrencia de que un punto de los datos pertenezca a la k-ésima clase, de tal manera $-\sum_{k=1}^{c} \alpha_k \ln \alpha_k$ se convierte en el *promedio de la información* al ser la entropía sobre las proporciones α_k.

Cuando $\alpha_k = 1/c, \forall k = 1, 2, \ldots, c$ diremos que no hay información acerca de α_k. En este punto la entropía habrá alcanzado su máximo valor, por tanto añadiremos el valor a la función objetiva $J(z, A)$ como un castigo. Entonces construimos un método para estimar α_k minimizando la entropía de manera que podamos obtener la mayor información para α_k.

El minimizar $-\sum_{k=1}^{c} \alpha_k \ln \alpha_k$ es equivalente a maximizar $\sum_{k=1}^{c} \alpha_k \ln \alpha_k$, por esta razón usaremos $\sum_{k=1}^{c} \alpha_k \ln \alpha_k$ como un término de penalización para la función objetivo propuesta de la siguiente manera

$$J_{UKM_1}(\mathrm{z}, A, \alpha) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \left\| x_i - a_k \right\|^2 - \beta n \sum_{k=1}^{c} \alpha_k \ln \alpha_k \tag{1}$$

Además para determinar el número de clústeres consideremos otro término de la entropía. Combinando las variables de membresía z_{ik} y la proporción α_k y usando la base de la teoría de la entropía sugerimos un nuevo término en la forma de $z_{ik} \ln \alpha_k$ proponiendo entonces la siguiente función objetivo:

$$\begin{aligned} J_{U-k-means}(\mathrm{z}, A, \alpha) = & \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \left\| x_i - a_k \right\|^2 - \beta n \sum_{k=1}^{c} \alpha_k \ln \alpha_k \\ & - \gamma \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \ln \alpha_k \end{aligned} \tag{2}$$

Sabemos también que cuando β y γ en (2) son cero este se vuelve el método k-medias

original. El lagrangiano de (2) es

$$\begin{aligned}\tilde{J}(\mathrm{z}, A, \alpha, \lambda) = &\sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik} \left\|x_i - a_k\right\|^2 - \beta n \sum_{k=1}^{c} \alpha_k \ln \alpha_k \\ &- \gamma \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik} \ln \alpha_k - \lambda \left(\sum_{k=1}^{c} \alpha_k - 1\right)\end{aligned} \tag{3}$$

Que derivando parcialmente con respecto a z_{ik} e igualando a 0 podemos obtener que la ecuación de actualización para z_{ik} es:

$$z_{ik} = \begin{cases} 1 & \text{if } \left\|x_i - a_k\right\|^2 - \gamma \ln \alpha_k = \min\limits_{1 \le k \le c} \left\|x_i - a_k\right\|^2 \\ & \qquad\qquad - \gamma \ln \alpha_k \\ 0, & \text{de otra manera.} \end{cases} \tag{4}$$

De manera que la ecuación de la actualización para el centro del clúster α_k es:

$$a_k = \textstyle\sum_{i=1}^{n} z_{ik} x_{ij} \Big/ \sum_{i=1}^{n} z_{ik} \tag{5}$$

Luego se procede a derivar parcialmente el lagrangiano con respecto a α_k, obteniendo

$$\frac{\partial \tilde{J}}{\partial \alpha_k} = -\beta n \left(\ln \alpha_k + 1\right) - \gamma \sum_{i=1}^{n} \frac{z_{ik}}{\alpha_k} - \lambda = 0 \text{ y } -\beta n \alpha_k \left(\ln \alpha_k + 1\right) - \gamma \sum_{i=1}^{n} z_{ik} - \lambda \alpha_k = 0$$

Por lo que tenemos que

$$-\sum_{k=1}^{c} n\beta\alpha_k \ln \alpha_k - \sum_{k=1}^{c} n\beta\alpha_k - \gamma \sum_{k=1}^{c}\sum_{i=1}^{n} z_{ik} - \sum_{k=1}^{c} \lambda\alpha_k = 0$$

con

$$\lambda = -n\beta \sum_{k=1}^{c} \alpha_k \ln \alpha_k - n\beta - n\gamma$$

Obteniendo

$$-\beta n \alpha_k \left(\ln \alpha_k + 1\right) - \gamma \sum_{i=1}^{n} z_{ik} - \left(-n\beta \sum_{k=1}^{c} \alpha_k \ln \alpha_k - n\beta - n\gamma\right) \alpha_k = 0$$

Por lo que la ecuación de actualización para α_k será:

$$\alpha_k^{(t+1)} = \sum_{i=1}^{n} z_{ik}/n + (\beta/\gamma)\alpha_k^{(t)} \left(\ln \alpha_k^{(t)} - \sum_{s=1}^{c} \alpha_s^{(t)} \ln \alpha_s^{(t)} \right) \tag{6}$$

con t el número de la iteración en el algoritmo.

Se debe de mencionar además que (6) es de gran importancia para el algoritmo, al ser $\sum_{k=1}^{c} \alpha_k \ln \alpha_k$ la media ponderada del $\ln \alpha_k$ con pesos $\alpha_1, \ldots, \alpha_c$. Para la k-ésima proporción $\alpha_k^{(t)}$, si $\ln \alpha_k^{(t)}$ es menor al promedio ponderado, entonces la nueva proporción de mezcla $\alpha_k^{(t+1)}$ será más pequeña el antiguo $\alpha_k^{(t)}$. Eso es, la proporción más pequeña disminuirá y la más grande se incrementará en la próxima iteración y entonces la competencia ocurrirá. Si $\alpha_k \leq 0$ o $\alpha_k \leq 1/n$ para algún $1 \leq c^{(t)}$, se las considera proporciones ilegítimas. Ante esta situación descartamos esos clúster y luego actualizamos el número de clúster, esto es:

$$c^{(t+1)} = c^{(t)} - \left| \left\{ \alpha_k^{(t+1)} \middle| \alpha_k^{(t+1)} < 1/n, k = 1, \ldots, c^{(t)} \right\} \right| \tag{7}$$

donde $|\,|$ denota la cardinalidad del conjunto . Después de actualizar el número de clúster c, las proporciones de mezcla α_k y sus correspondientes z_{ik} deberán ser normalizadas por:

$$\alpha_k^* = \alpha_k^* \Big/ \sum_{s=1}^{c^{(t+1)}} \alpha_s^* \tag{8}$$

$$z_{ik}^* = z_{ik}^* \Big/ \sum_{s=1}^{c^{(t+1)}} z_{is}^* \tag{9}$$

Es importante notar que los parámetros de aprendizaje γ y β incluidos en la función objetivo. Basados en algunas tasas de crecimiento de aprendizaje de número de clúster con $e^{-c^{(t)}/100}$, $e^{-c^{(t)}/250}$, $e^{-c^{(t)}/500}$, $e^{-c^{(t)}/750}$ y $e^{-c^{(t)}/1000}$ podemos ver que $e^{-c^{(t)}/100}$ disminuye a mayor velocidad, pero que $e^{-c^{(t)}/500}$, $e^{-c^{(t)}/750}$, $e^{-c^{(t)}/1000}$ disminuyen más lento. Para evitar que sea muy rápido o muy lento fijamos el parámetro γ como

$$\gamma^{(t)} = e^{-c^{(t)}/250} \tag{10}$$

Por otra parte, el parámetro β nos puede ayudar a controlar la evolución del algoritmo. Empezamos por aplicar la noción de que si $0 < \alpha_k \leq 1 \forall k$ entonces se tiene que

$e^{-1} \leq \alpha_k \ln \alpha_k < 0$, poniendo $E = \sum_{k=1}^{c} \alpha_k \ln \alpha_k < 0$, tenemos que

$$\alpha_k E = \alpha_k \sum_{k=1}^{c} \alpha_k \ln \alpha_k < 0$$

Con lo que obtenemos que

$$-e^{-1}\beta < \beta\alpha_k \left(\ln \alpha_k - \sum_{s=1}^{c} \alpha_s \ln \alpha_s \right) < \beta(-\alpha_k E) \tag{11}$$

Bajo la restricción de que $\sum_{k=1}^{c} \alpha_k = 1$, si y sólo si $\alpha_k < 1/2$, se tiene que $(\ln \alpha_k - \sum_{s=1}^{c} \alpha_s \ln \alpha_s) < 0$. Ahora, para evitar la situación donde todos los $\alpha_k \leq 0$, la parte izquierda de (14) debe ser mayor a $-\max\{\alpha_k \mid \alpha_k < 1/2, k = 1, 2, \cdots, c\}$. Por lo que se tiene la siguiente condición de β: $-e^{-1}\beta > -\max\{\alpha_k \mid \alpha_k < 1/2, k = 1, 2, \cdots, c\}$. Por lo que $\beta < \max\{\alpha_k e \mid \alpha_k < 1/2, k = 1, 2, \cdots, c\} < e/2$. Motivo por el cual usaremos $\beta \in [0, 1]$. Por otro parte, si la diferencia entre $\alpha_k^{(t+1)}$ y $\alpha_k^{(t)}$ es pequeña β debe agrandarse para que el algoritmo avance. Si la diferencia entre $\alpha_k^{(t+1)}$ and $\alpha_k^{(t)}$ es grande, entonces β deberá disminuir para mantener estabilidad en el algoritmo. Por lo que definimos β como

$$\beta = \textstyle\sum_{k=1}^{c} \exp\{-\eta n |\alpha_k^{(t+1)} - \alpha_k^{(t)}|\}/c \tag{12}$$

donde $\eta = min(1, 1/t^{\lfloor d/2-1 \rfloor})$. Por otra parte considerando que

$$\begin{aligned} \max_{1\leq k\leq c} \alpha_k^{(t+1)} &\leq \max_{1\leq k\leq c} \left(\tfrac{1}{n} \textstyle\sum_{i=1}^{n} z_{ik} \right) \\ &+ \tfrac{\beta}{\gamma} \max_{1\leq k\leq c} \alpha_k^{(t)} \left(\ln \max_{1\leq k\leq c} \alpha_k^{(t)} - \textstyle\sum_{s=1}^{c} \alpha_s^{(t)} \ln \alpha_s^{(t)} \right) \end{aligned}$$

y que

$$\begin{aligned} &\max_{1\leq k\leq c} \left(\tfrac{1}{n} \textstyle\sum_{i=1}^{n} z_{ik} \right) + \tfrac{\beta}{\gamma} \max_{1\leq k\leq c} \alpha_k^{(t)} \\ &\times \left(\ln \max_{1\leq k\leq c} \alpha_k^{(t)} - \textstyle\sum_{s=1}^{c} \alpha_s^{(t)} \ln \alpha_s^{(t)} \right) < \max_{1\leq k\leq c} \left(\tfrac{1}{n} \textstyle\sum_{i=1}^{n} z_{ik} \right) \\ &+ \beta \left(- \left(\max_{1\leq k\leq c} \alpha_k^{(t)} \textstyle\sum_{s=1}^{c} \alpha_s^{(t)} \ln \alpha_s^{(t)} \right) \right). \end{aligned}$$

Si se tiene que

$$\max_{1\leq k\leq c} \left(\tfrac{1}{n} \textstyle\sum_{i=1}^{n} z_{ik} \right) - \beta \max_{1\leq k\leq c} \alpha_k^{(t)} \textstyle\sum_{s=1}^{c} \alpha_s^{(t)} \ln \alpha_s^{(t)} \leq 1,$$

Entonces la restricción de que $\max\limits_{1\leq k\leq c} \alpha_k^{(t+1)} \leq 1$ se cumple, por lo cual tenemos que

$$\beta \leq \left(1 - \max_{1\leq k\leq c} \left(\frac{1}{n}\sum_{i=1}^{n} z_{ik}\right)\right) \Bigg/ \left(-\max_{1\leq k\leq c} \alpha_k^{(t)} \sum_{s=1}^{c} \alpha_s^{(t)} \ln \alpha_s^{(t)}\right) \tag{13}$$

Por lo que de las ecuaciones (12) y (13) podemos obtener que β en la iteración $t+1$ será

$$\beta^{(t+1)} = \text{mín}\left(\frac{\sum_{k=1}^{c} \exp(-\eta n \left|\alpha_k^{(t+1)} - \alpha_k^{(t)}\right|)}{c}, \frac{1 - \max\limits_{1\leq k\leq c} \left(\frac{1}{n}\sum_{i=1}^{n} z_{ik}\right)}{\left(-\max\limits_{1\leq k\leq c} \alpha_k^{(t)} \sum_{k'=1}^{c} \ln \alpha_{k'}^{(t)}\right)}\right) \tag{14}$$

Ahora, puesto que β puede cambiar en cualquier momento, pondremos $\beta = 0$ cuando el número de clústeres c es estable, esto es, cuando c ya no disminuye.

En lo que se refiere a inicialización usaremos todos los puntos iniciales como medias iniciales, es decir que $a_k = x_k$, por lo que $c^{(0)} = n$ y usando $\alpha_k = 1/c^{(0)}$ como proporciones de mezcla iniciales.
En cuanto a complejidad computacional, el algoritmo U-k-medias puede ser dividido en tres partes:

1. Computar la membresía z_{ik} con complejidad $\mathcal{O}(ncd)$
2. Calcular la proporción de mezcla α_k con $\mathcal{O}(nc)$
3. Actualizar los centros de los clúster, con complejidad $\mathcal{O}(n)$.

Por tanto el algoritmo tendrá una complejidad total de $\mathcal{O}(ncd)$. Donde n es el número de puntos, c el número de clúster y d la dimensión de los puntos.

Capítulo 2

Metodología

Este trabajo se caracteriza por ser de investigación *aplicada*, esto es, que está caracterizada por que busca la aplicación o utilización de los conocimientos adquiridos. Es también de carácter documental pues obtiene los anuncios de ediciones existentes de la revista, por lo que se apoya en fuentes de carácter documental, además de las líneas de negocio ofrecidas por el CITEC y disponibles en su página web.

Esta investigación también se puede considerar como descriptiva, pues utiliza un método de análisis (clustering), e intenta caracterizar un objeto de estudio, en este caso los anuncios de las ediciones de la revista.

Para realizar esta segmentación es necesario tener un conjunto de datos que comprenda información suficiente acerca de los anuncios con los que se trabajará, detallando cada anuncio como un punto. Este grupo de puntos conformará una base de datos que será analizada mediante el algoritmo U-k-medias, el cual será implementado en R. Posteriormente se estudiarán los resultados provenientes de diversos índices de validez para determinar la verosimilitud de las conclusiones y se darán recomendaciones para estas últimas.

2.1. La Base de Datos

El primero de los objetivos comprende la creación de una base de datos de anuncios, con el objetivo encontrar clústeres de anuncios que permitan encontrar anunciantes similares, al ser cada punto o objeto de la base de datos un anuncio, es importante conocer a que clase de producto o servicio corresponde puesto que diferentes sectores empresariales tienen diferentes necesidades al anunciar (Bagwell, 2007).

Otro aspecto fundamental al analizar una compañía es su tamaño, existen varios problemas entorno a la definición del tamaño de una compañía, el amplio espectro de definiciones existentes no solo reflejan estos problemas de definición sino también el amplio rango de disciplinas y negocios cubiertas, además de los diferentes tipos de investigación (Brooksbank, 1991). En general el tipo de enfoque que se le da al tamaño puede ser categorizado como cuantitativo o cualitativo , siendo el primero el más popular para motivos de investigación.

La ventaja de un enfoque cuantitativo es que provee una base teórica en la que basarse al ser números, criterios cuantitativos efectivos a la hora de categorizar empresas son el número de trabajadores y el volumen total de ventas (Brooksbank, 1991) pero una mezcla de ambos enfoques es necesario para una definición exitosa, un aspecto cualitativo que se encontró importante al clasificar empresas es la cantidad de productos o servicios diferentes que estos poseen (Brooksbank, 1991), atributo que también se incluirá en la base de datos.

Se construyó una base de datos de los anuncios de una revista ecuatoriana, que comprende un total de 254 anuncios, en donde cada entrada incluye las siguientes características del anuncio de acuerdo a lo antes mencionado:

1. ID o número de anuncio:
   Carácter numérico que permite identificar al anuncio.

2. Nombre de la empresa anunciante:
   Razón social de la entidad que publica el anuncio, lo que permite identificar al cliente.

3. Patrimonio de la empresa:
   Comprende al valor monetario de todos los bienes que posee la empresa

4. Líneas de negocio incluidas en el anuncio:
   Incluye en una o más columnas todos los servicios o productos ofrecidos en el anuncio, proveyendo información de que servicios son los más populares para anunciar, estas líneas de negocios son basadas en aquellas propuestas por el CITEC en su directorio de socios, así como otras de otros sectores de producción.

5. Número de empleados en la empresa:
   Basado en datos del 2020 publicados por la Superintendencia De Compañías del Ecuador.

6. Ingreso por venta:
   Cantidad de ingresos reportada por el ejercicio económico de la empresa.

7. Utilidad, utilidad Neta, utilidad antes de impuestos:
   Comprende la suma de los ingresos menos los egresos, incluyendo el pago de nóminas, impuestos, entre otros.

8. IR Causado:
   Valor de impuesto a la renta causado por el ejercicio económico.

9. Ingreso Total:
   Cantidad monetaria total que comprende todo ingreso, desde aquel motivado por el ejercicio económico hasta aquellos réditos externos como venta de bienes.

10. Costo del anuncio:
    Compensación económica recibida por el medio por la publicación del anuncio.

2.2. Pseudocódigo e Implementación

En el anterior capítulo se justificó teóricamente la validez del algoritmo, el cual procede de acuerdo al siguiente algoritmo, se incluye bajo cada paso el código en R que se implementa para ello

1. Fijar la tolerancia $\epsilon > 0$. Inicializar $c^{(0)} = n$, $\alpha_k^{(0)} = 1/n$, $a_k^{(0)} = x_i$ y de parámetros iniciales a $\gamma^{(0)} = \beta^{(0)} = 1$. Poner $t = 1$

```
eps=.0001
n=nrow(X)
d=ncol(X)
c0=c=n
At1=At=X
alphat=alphat1=rep(1/(n),c)
Z=matrix(data = NA, nrow = n,ncol = c)
gamma0=beta0=1
t=1
```

2. Calcular la membresía de los grupos $z_{ik}^{(t+1)}$ mediante (4); esto nos permite saber a que grupo pertenecerá cada elemento en la iteración.

```
Z=matrix(data = 0, nrow = n,ncol = c[t])
 for (i in 1:n) {
 rx<-matrix(rep(as.numeric(X[i,]),c[t]),ncol=d,byrow=T)
 aux2=apply(rx-At, 1, norm,type="2")^2-
  gamma0*log(alphat,exp(1))
  if (t==1){
  aux2[i]=NA
  }
  k0=which(aux2 == min(aux2, na.rm = TRUE))
  Z[i,k0]=1
 }
```

3. Calcular $\gamma^{(t+1)}$ usando (10), este parámetro nos permite controlar la velocidad con la que los grupos disminuyen, valores mayores a 250 disminuirán el número de grupos más lentamente en cada iteración. De manera similar valores menores a 250 disminuirán más rápido el número de grupos.

```
gamma0=exp(-c[t]/250)
```

4. Actualizar las proporciones de mezcla $\alpha_k^{(t+1)}$ usando (6), esto nos proporciona una idea de cuan probable es que un elemento pertenezca a un determinado grupo.

```
for (k in 1:c[t]) {
aux6= sum(alphat*log(alphat,exp(1)))
alphat1[k]=sum(Z[,k])/n + beta0/gamma0 * alphat[k] *
(log(alphat[k], exp(1)) -aux6)
}
```

5. Calcular $\beta^{(t+1)}$ a partir de (14), este parámetro ayuda al algoritmo a determinar las proporciones de mezcla, si éstas proporciones no cambian mucho, beta aumentará para descartar más grupos y que no se estanque el algoritmo.

```
eta=min(1,1/t^floor(ncol(X)/2 -1))
beta0=min( sum(exp(-eta*n*abs(alphat1-alphat)))/c[t],
          (1-max(colSums(Z)/n))/
            -max(alphat*sum(log(alphat,base = exp(1)))))
```

6. Actualizar el número de clúster $c^{(t+1)}$ descartando aquellos cuya proporción de mezcla cumpla $\alpha_k < 1/n$ y ajustar los asociados α_k y z_{ik} de acuerdo a (8) y (9), asignando a los elementos en grupos descartados a los grupos que aún permanezcan.

```
ct1=c[t] - length(alphat1[alphat1<1/n])
c=c(c,ct1)
Z=Z[,alphat1>=1/n]
auxc=which(alphat1>=1/n)
alphat1=alphat1[alphat1>=1/n]
alphat1=alphat1/sum(alphat1)
alphat1
if(c[t+1]==1){
  break
}
auxZ=rowSums(Z)
for (i in 1:nrow(Z)) {
  for (k in 1:ncol(Z)) {
    if(auxZ[i]!=0){
      Z[i,k]=Z[i,k]/auxZ[i]
    }
    else {
      rx <- matrix(rep(as.numeric(X[i,]),c[t+1]),
      ncol = d,byrow = T)
      aux2=apply(rx-At[auxc,], 1, norm,type="2")^2-
        gamma0*log(alphat[auxc],exp(1))
      if (t==1){
        aux2[i]=NA
      }
      k0=which(aux2 == min(aux2, na.rm = TRUE))
      Z[i,k0]=1
    }
  }
}
```

7. Si se cumple que $t \geq 60$ y $c^{(t-60)} - c^{(t)} = 0$ Hacer $\beta^{(t+1)} = 0$, esto nos permite que cuando existan 60 iteraciones sin un cambio en el número de grupos el algoritmo finalice.

```
if(t>60 && (c[t-60]-c[t])==0){
    beta0=0
}
```

8. Actualizar los centros de los grupos $a_k^{(t+1)}$ por (5), sumando los elementos en el grupo y dividiendo tal cantidad por el número de elementos.

```
for (k in 1:c[t+1]) {
      if (sum(Z[,k]==1)>1) {
       At1[k,]=colSums(X[which(Z[,k]==1),])/sum(Z[,k]==1)
}
else {
  if(sum(Z[,k]==1)==1){
    At1[k,]=X[which(Z[,k]==1),]
  }
}
}
At1=At1[1:c[t],]
```

9. Comparar $a_k^{(t+1)}$ con $a_k^{(t)}$, de manera que si $\max_{1\leq k\leq c^{(t)}} ||a_k^{(t+1)} - a_k^{(t)}|| < \epsilon$ el algoritmo para, de otra manera regresar al paso 2. Esto nos dice que si los centros de los grupos no cambian, el algoritmo debe acabar. Este paso se forma utilizando un laso *while* de la siguiente manera:

```
while(c[t]>1 &&
max(apply(At1-At[auxc,], 1, norm, type="2"))>eps)
```

El algoritmo completo y comentado, incluyendo el tratamiento de datos y los índices calculados se puede encontrar en los anexos.

2.3. Validez de los Clúster e Índices Internos

El objetivo de esta investigación es encontrar grupos de anuncios similares, pues estos pueden ser presentados en una misma edición; la cual incluye material periodístico o informativo estrechamente relacionado a aquellos productos o servicios incluidos en estos anuncios.

Por tanto es necesario investigar a estos grupos, para esto usaremos índices *internos*, esto es, índices usados para evaluar la bondad de la estructura del clúster usando solamente la información intrínseca de los datos originales (Deborah, 2010). En este trabajo se han elegido algunos de los índices internos más comúnmente usados, tales como para evaluar el número de grupos o clúster c, el índice original de Dunn DNo (Dunn, 1973), índice Davies-Bouldin DB (Davies, 1979) además del valor de silueta o *Silhouette value* (Rousseeuw, 1987).

2.3.1. Índice de Dunn

El índice DNo busca comparar el tamaño de los clúster con la distancia existente entre ellos y es computado como la razón entre la mínima distancia entre dos clúster y el tamaño del clúster más grande (Dunn, 1973) de la siguiente manera:

$$D(\mathcal{C}) = \frac{\text{mín}_{C_k, C_l \in \mathcal{C}, C_k \neq C_l} \left(\text{mín}_{i \in C_k, j \in C_l} \text{dist}(i,j) \right)}{\text{máx}_{C_m \in \mathcal{C}} \text{diam}\left(C_m\right)}$$

dónde $diam(C_m)$ es la distancia máxima entre observaciones en el clúster C_m. El índice Dunn tiene un valor entre cero e infinito y debe ser maximizado para mejorar la calidad de los grupos. Sin embargo, esta formulación tiene un problema peculiar; en el caso de que uno de los grupos sea problemático, y aunque los demás estén muy bien formados, puesto que el denominador contienen un término de máximo, el índice Dunn para ese conjunto de grupos será bajo.

Para su implementación en R se utilizará el paquete *clValid* (Brock, 2008), el cual recibe de entrada un vector que incluye la pertenencia de los individuos a cada grupo, además de una matriz de distancias que incluye cada una de las distancias entre individuos.

2.3.2. Índice Davies-Bouldin

El índice DB Davies-Bouldin mide la similitud promedio entre cada clúster y su clúster más parecido. Este último índice de validez intenta maximizar las distancias entre clústeres mientras minimiza la distancia entre el centroide de cada clúster con los otros puntos (Davies, 1979), este se calcula mediante

$$DB = \frac{1}{c} \sum_{r=1}^{c} \underset{\substack{s \\ r \neq s}}{\text{máx}} \left(\frac{S_r + S_s}{d_{rs}} \right)$$

donde:

- c es el número de clúster,
- $r, s \in \{1, \ldots, c\}$,
- n es el número de individuos o sujetos,
- $i, k \in \{1, \ldots, n\}$,
- $d_{rs}, = \sqrt[p]{\sum_{j=1}^{m} \left| z_{rj} - z_{sj} \right|^p}$ representa la distancia entre centroides de los clúster P_r y P_s,
- $z_r = (z_{r1}, \ldots, z_{rm})$ es el centroide del clúster P_r,
- m el número de columnas de los individuos ,
- $j \in 1, \ldots, m$,
- $S_r = \sqrt[q]{\frac{1}{n_r} \sum_{i \in P_r} \sum_{j=1}^{m} \left| x_{ij}^r - z_{rj} \right|^q}$ es la medida de dispersión del clúster P_r. Para $q = 1$ esto es la distancia media de los objetos del clúster P_r al centroide del clúster P_r, para $q = 2$ esto es la desviación estándar de los objetos del clúster P_r,
- n_r es el número de objetos en el clúster P_r

Al estar este índice definido como la razón entre la dispersión dentro de los clúster y la dispersión entre clústeres, indica que un valor menor significará que la agrupación es mejor. Se usan los valores de $q = 1$ y $p = 2$ en el método de las *U-k-medias* (Sinaga, 2020). Datos bidimensionales aleatorios producen valores de este índice cercanos al 0.6, valor que se considera satisfactorio (Davies, 1979).

Es apropiado comunicar también que un conjunto de datos debe ser particionado en al menos dos grupos con centros distintos para que este índice tenga sentido. Esto puesto que debido a la que la medida de distancia en el denominador debe ser distinta de cero para que esta esté definida. El alcance de éste índice también está limitado por la necesidad de que los clústeres tengan más de un elemento, puesto que tales clústeres tendrían una dispersión intra-clúster de 0.

Para la implementación del índice en R se utiliza la librería *clusterSim* que contiene diversos índices externos e internos (Walesiak, 2020), se usa en especial la función *index.DB* con parámetros: *centrotypes = 'centroids'* para usar los centroides en vez de los medoides y además $p = 2$, $q = 1$ como se mencionó previamente.

2.3.3. Valor de Silueta

El valor de silueta *SW* es una medida de cuan similar es un punto a su propio grupo comparado a otros grupos (Rousseeuw, 1987). Este valor varía desde -1 hasta 1, donde un valor alto indica que el objeto se ajusta bien al grupo en el que esta incluido y además que no se ajusta bien a los otros grupos. De esta manera valores de silueta pueden indicar si un objeto está o no bien agrupado.

A los puntos con un valor de silueta SW cercanos al cero se los considera como objetos que podrían estar en varios clúster. Supongamos que los datos están divididos en C clústeres, para cada punto $i \in C_I$ (punto i en el clúster C_I):

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

es la distancia media entre el punto i y todos los demás puntos en el mismo grupos, con $|C_I|$ la cardinalidad del clúster correspondiente. La distancia a utilizar será la euclidiana. Es posible interpretar a $a(i)$ como una medida de que tan bien está la asignación de i a su grupo, por lo que a menores valores habrá una mejor asignación.

Definamos además la disimilaridad media del punto i a un grupo C_J como el promedio de la distancia de i hacia los demás puntos en C_J, llamaremos al mínimo de estos $b(i)$

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

Es entonces posible definir al valor de silueta como

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ si } |C_I| > 1$$

y

$$s(i) = 0, \text{ si } |C_I| = 1$$

El cual también puede ser expresado como:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{si } a(i) < b(i) \\ 0, & \text{si } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{si } a(i) > b(i) \end{cases}$$

De donde es posible ver que

$$-1 \leq s(i) \leq 1$$

Para que $s(i)$ este cerca de 1, es necesario que $a(i) \ll b(i)$. Puesto que $a(i)$ es una medida de la disimilitud de i con su propio grupo, un valor pequeño significará una buena asignación. De manera similar, un valor grande de $b(i)$ implica que i estaría mal emparejado con los demás miembros de los otros grupos. Un valor de $s(i)$ cercano a cero indica que el individuo está en el borde de dos grupos naturales.

La media de los valores $s(i)$ en un grupo es una medida de que tan estrechamente agrupados están todos los puntos en tal clúster. De manera similar la media de todos los valores $s(i)$ existentes en la base o el conjunto de datos es una medida de que tan bien están agrupados los elementos.

La implementación del valor de silueta se hará usando el paquete *cluster* que contiene la función *silhouette* que computa la información de la silueta dado una agrupación en k clústeres (Maechler, 2021) la cual posteriormente será graficada haciendo uso del paquete *factoExtra* que permite mostrar los valores de silueta de manera intuitiva además de mostrar las secciones que presentan problemas en los grupos (Kassambara, 2020)

2.4. Limpieza y Adecuación de los Datos

Se utiliza el paquete *readxl* para la entrada de datos, se cambian los nombres de las columnas a nombres más apropiados, carentes de espacios y caracteres especiales. Se transforma además a las líneas de negocios en variables binarias, es decir, creando una columna para cada línea de negocios. Se adjuntan y renombran estas nuevas columnas en ves de los identificadores anteriores. Además se retiran las variables de identificación del anuncio y de la empresa a la que pertenece.

El algoritmo a utilizar U-k-medias incluye la distancia euclidiana para calcular las distancias de cada elemento al centroide de su clúster, es por tanto que si una de las características del individuo es más grande (al menos en términos de valor absoluto) que las demás, esa variable pasará a ser predominante en el algoritmo (Zheng, 2018). En nuestro caso, las variables tienen rangos muy distintos como se puede notar en el siguiente gráfico:

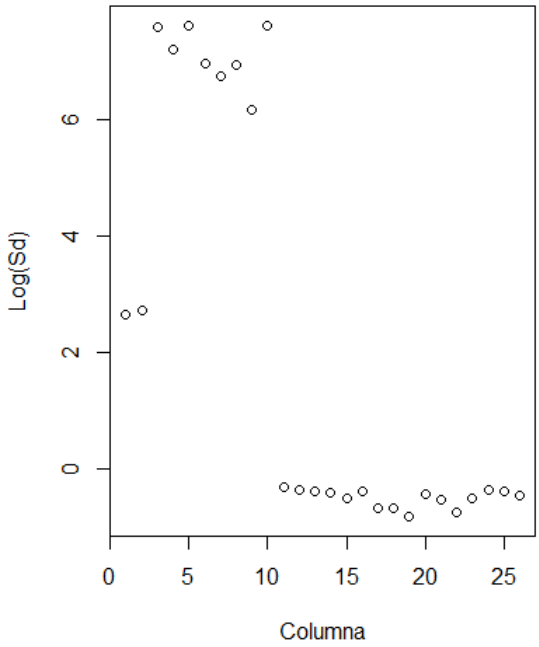

Figura 2.1: Logaritmo base 10 de la desviación estándar de las columnas

Con el objetivo de darle a todas las variables un peso similar se estandariza la variable usando el método la normalización en donde a cada columna x se la transforma en x^* mediante

$$x^* = \frac{x - \bar{x}}{\sigma}$$

donde $\bar{x}$ es la media de la columna y $sigma$ es la desviación estándar. Se eliminan filas con datos faltantes y se asegura la naturaleza numérica de los datos.

Capítulo 3

Resultados, Conclusiones y Recomendaciones

3.1. Resultados

Se obtuvieron seis grupos de anuncios en trece iteraciones del algoritmo, a continuación una gráfica del número de grupos evolucionando con cada iteración hasta estabilizarse en seis.

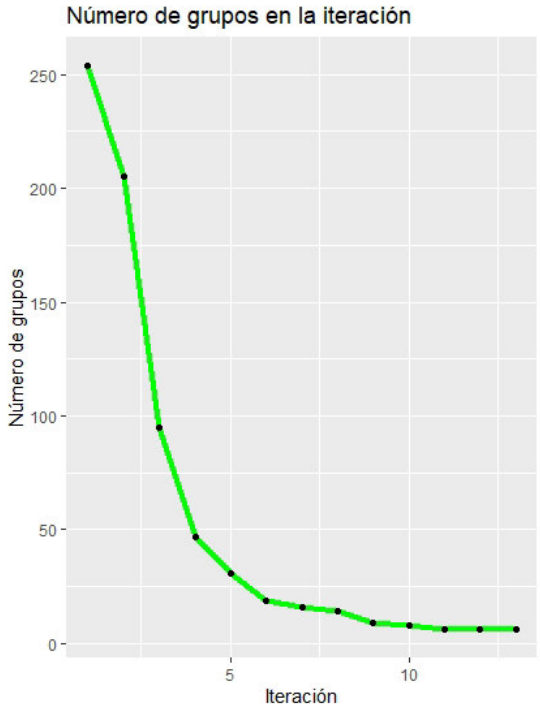

Figura 3.1: Número de grupos en cada iteración

Se puede notar que el algoritmo empieza con el mismo número de grupos que de individuos, la caída más rápida en una sola iteración se presenta cuando $t = 3$, reduciendo de 205 grupos a 95. A partir de la décimo primera iteración el algoritmo se encarga de poner todos los individuos en su grupo más cercano.

El error, previamente definido como el máximo de las normas de la diferencia de los centroides en la t-ésima y t+1-ésima iteración cambia de acuerdo al siguiente gráfico

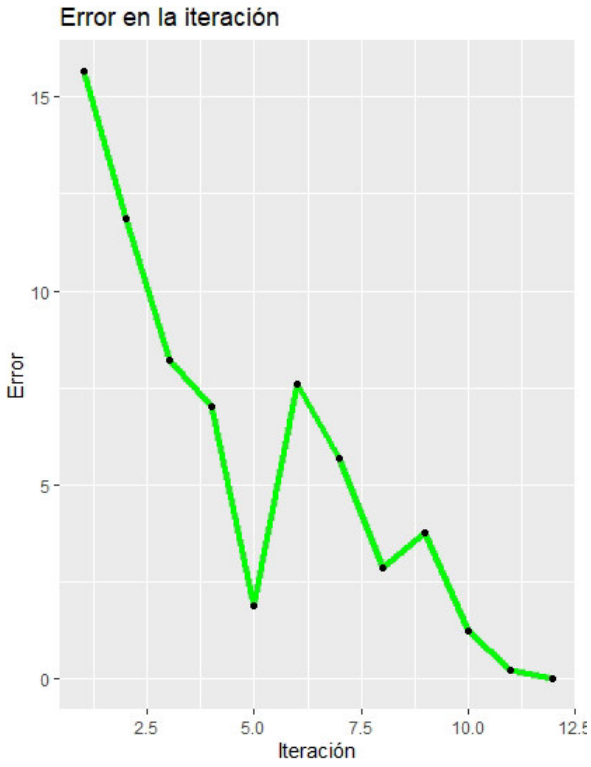

Figura 3.2: Error en cada iteración

Se puede notar como una consecuencia del gráfico anterior como el error baja a partir de la décimo primera iteración, puesto que una vez está fijo el número de grupos que el algoritmo usará, este se encarga de asignar correctamente a los individuos. Los grupos quedan entonces distribuidos de la siguiente manera:

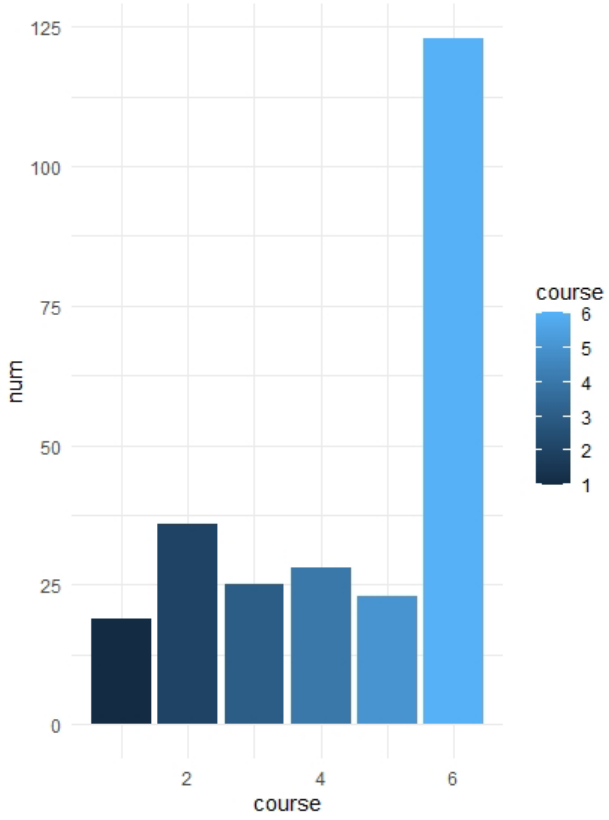

Figura 3.3: Número de individuos por grupo

Es importante notar que los primeros grupos tienen números semejantes de individuos, pero casi la mitad de toda la base de datos se aglutina en el último grupo, explicaremos esto posteriormente describiendo las características de cada grupo, sugiriendo razones del por qué de esta agrupación.

En cuanto a los índices de Dunn y Davies-Bouldin estos toman valores de

Dunn	Davies-Bouldin
0.151	2.105

Donde se puede ver que los grupos no están completamente definidos, esto es de esperarse debido a que muchas empresas actúan en lineas de negocios compartidas con otras empresas, por lo cual no pueden estar completamente en un solo grupo.

Analicemos entonces el valor de silueta, recordando que este está entre -1 y 1, este valor está dado por el siguiente gráfico

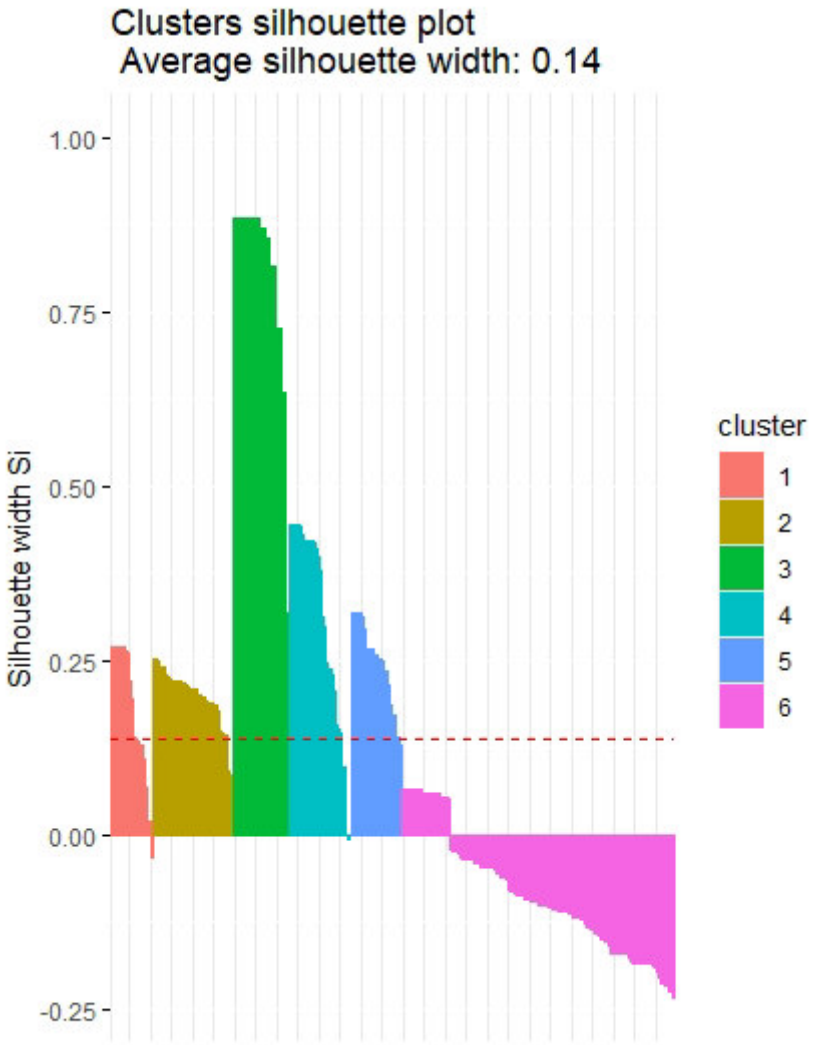

Figura 3.4: Valor de silueta con seis grupos

En este gráfico se pudo apreciar que los primeros cinco grupos se encuentran bien asignados, en dónde el tercero es el mejor definido y solo dos valores son negativos, para dar un mayor contexto a estos grupos los analizaremos uno a uno.

- Grupo 1: El grupo más pequeño con 19 individuos, está caracterizado por incluir a los anuncios del área legal con 12 ocurrencias y del área comercial con 8 ocurrencias. Este presenta un elemento con un valor de silueta negativo, al analizar este anuncio se pudo evidenciar que pertenece al área comercial, pero también al área de datos y electrónica, áreas que no están bien representadas en el grupo, a continuación los promedios del grupo, entre los que se puede destacar que las variables actividad, patrimonio e ingreso por ventas son las más bajas entre todos los grupos. Este es posiblemente el grupo menos redi-

tuable entre todos.

Valor	Empleados	Actividad	Patrimonio	IngresoVentas
813,6842	53,57895	2010940	586259,2	4107522
UtilidadSI	**Utilidad**	**Uneta**	**IR**	**IngresosTotales**
33311,46	106742,7	-8498,76	43282,61	4151867

- Grupo 2: Este grupo cuenta con 36 individuos, en los que predomina el área de tecnologías de la información o TI, todos sus individuos pertenecen a este sector. Otras áreas importantes para el grupo son el área de recopilación de datos, el área de integración y el área de electrónica; áreas profundamente relacionadas con el sector TI, a continuación un resumen de sus otras características, entre las cuales cabe destacar que la utilidad neta y la utilidad antes de impuestos son las más altas de todas, empresas en este grupo son las que más ganancias producen.

Valor	Empleados	Actividad	Patrimonio	IngresoVentas
1064,722	68,58333	8122072	2765028	21381195
UtilidadSI	**Utilidad**	**Uneta**	**IR**	**IngresosTotales**
389399,4	721880,7	263611	126565,5	21764961

- Grupo 3: Este grupo está fuertemente caracterizado por el área que comprende todas las líneas de negocios que no caben en otras áreas, empresas de agroindustria, ecoplaneamiento entre otras se encuentran agrupadas aquí. Éstas empresas destacan por ser aquellas que más gravan impuestos, estos anuncios son los menor valuados.

Valor	Empleados	Actividad	Patrimonio	IngresoVentas
610	26,08	6253445	1143587	4608183
UtilidadSI	**Utilidad**	**Uneta**	**IR**	**IngresosTotales**
92364,51	88753,08	72210,07	36718,37	4700362

- Grupo 4: Este grupo está conformado por tres pilares, el área de analítica de datos con 26 participantes, el área de recopilación de datos con 25 y el área de transformación digital o digitalización con 18, estás tres áreas están intrínsecamente relacionadas y generan los anuncios más valiosos. Está conformado en total por 28 individuos. A continuación sus características.

Valor	Empleados	Actividad	Patrimonio	IngresoVentas
1093,03571	71,21429	5179590	2921815	4590167
UtilidadSI	**Utilidad**	**Uneta**	**IR**	**IngresosTotales**
-222466	139302,7	-269474	47008,3	4700512

- Grupo 5: Este grupo está conformado completamente por empresas del área de la humanística, esto incluye entidades de capacitación, entrenamiento y otras actividades relacionadas. Este grupo presenta el menor número de empleados de todos. Sus otras características se presentan a continuación.

Valor	Empleados	Actividad	Patrimonio	IngresoVentas
1093,03571	71,21429	5179590	2921815	4590167
UtilidadSI	**Utilidad**	**Uneta**	**IR**	**IngresosTotales**
-222466	139302,7	-269474	47008,3	4700512

- Grupo 6: Este grupo presenta 123 empresas de diferentes áreas y conforma el grupo más diverso de todos al incluir a aquellas empresas que no están especializadas en una sola línea de negocios. Este clúster comprende a las empresas más grandes y con más líneas de negocios, esto se puede evidenciar en la gráfica a continuación, donde es posible evidenciar que tiene el mayor número de empleados, la mayor actividad comercial, el mayor patrimonio así como también el mayor ingreso por ventas y el mayor ingreso total. Debido a la pandemia estas han sido las empresas que más han sufrido, teniendo la utilidad neta más negativa de entre todos los grupos. Este es el grupo que más inconvenientes presenta y es debido a la gran diversidad de empresas grandes que contiene, de las cuales la mayoría son del área de comunicación, seguridad y soluciones empresariales. A continuación sus características:

Valor	Empleados	Actividad	Patrimonio	IngresoVentas
1013,38122	359,1301	26616837	10435603	28924812
UtilidadSI	**Utilidad**	**Uneta**	**IR**	**IngresosTotales**
-2179551	3780982	-3101402	918495,9	29090220

3.2. Conclusiones

- Se generaron seis grupos de anuncios, de los cuales cinco presentan candidatos perfectos para ser anunciantes en una edición de la revista. La temática principal de la edición estará dadas por la línea de negocios más recurrente en el grupo.

- Los tamaños de los grupos fueron similares para los cinco primeros grupos, siendo respectivamente 19, 36, 26, 28 y 23, mientras que para el último fue de 123.

- El sexto grupo incluyó a las empresas más grandes y diversas, esto se evidencia no solo en los atributos económicos del grupo como los ingresos totales o el número de empleados sino también en la gran cantidad de líneas de negocio que estas empresas ofrecen.

- El método de las U-k-medias presenta una alternativa viable para la agrupación de anuncios, creando grupos de empresas similares como se pedía, sin embargo empresas muy grandes presentan complicaciones al poder estar en más de un grupo, al no estar bien definidas las empresas grandes eventualmente forman un solo grupo heterogéneo.

- El índice de Dunn DN_0 presenta complicaciones justo como se mencionó, al estar un grupo compuesto de elementos no tan similares el índice baja significativamente. De manera similar el índice Davies-Bouldin manifiesta aquello mencionado anteriormente, el último grupo es demasiado amplio y tergiversa los resultados.

- El valor de silueta es quizá el más útil de los índices, indicando una buena agrupación de los cinco primeros grupos y mostrando que gran parte del sexto grupo podría estar en otros grupos o conformar uno nuevo. Este tiene un valor promedio de 0.14.

3.3. Recomendaciones

- El algoritmo U-k-medias usa como medida de distancia a la distancia euclidiana, por lo cual solo admite variables numéricas y es sensible al tamaño de los valores en cada columna y puede ser necesario reescalar las variables. Es

por tanto que es recomendable modificarlo para que use la distancia de Gower en vez de la euclidiana. Esto volviera posible incluir columnas categóricas, aumentando el poder de agrupamiento el algoritmo y eliminando la necesidad de reescalar.

- Es también posible utilizar métodos supervisados, dando un intervalo en el cual el número de grupos debe de estar incluido, entonces, se pueden estudiar los grupos mediante índices internos, eligiendo al número de grupos a través de sus índices.

- Al conformar la base de datos se sugiere usar métodos de *web scraping* para obtener las líneas de negocio de la empresa, puesto que la investigación de cada anuncio requiere de información recuperada de la página web de la empresa o de alguna página web que actúe como repositorio o directorio de empresas.

- El uso de matrices *sparse* es recomendable para las líneas de negocio puesto que son columnas binarias donde la mayor parte son ceros. Este tipo de matrices es codificado de manera que solo se guardan las ubicaciones de los elementos cuyo valor sea distinto al 0 o al vacío, ahorrando una importante cantidad de memoria.

- Se recomienda además dividir el último grupo en varias ediciones, al ser el más diverso se pueden aprovechar todas sus líneas de negocio si más de una edición se trata de los sectores más populares de las empresas incluidas en éste grupo.

- Escoger los índices internos correctos es de vital importancia, índices que presenten grandes sesgos como el índice de Dunn cuando uno de los grupos es diverso dan conclusiones que no son objetivas en casos concretos.

Anexos

Código en R

```
library(readxl)
library(clValid)
library(clusterSim)
library(factoextra)
#Paso 1: inicializacion
#Fijar la tolerancia
eps=.0001
#Entrada de datos
X <- read_excel("D:/Documentos/Libros/UIC/AnunciosF.xlsx")
names(X) <- c("ID", "Empresa", "Valor", "A o", "Linea1",
"Linea2", "Linea3", "Linea4", "Linea5","Empleados","Actividad",
"Patrimonio","IngresoVentas","UtilidadSI","Utilidad","Uneta",
"IR","IngresosTotales")
n=nrow(X)
d=ncol(X)
auxX=matrix(0,nrow = n,ncol = 16)
#Transformar las columnas de lineas de negocio en binarias
for (i in 1:n) {
  lineas=X[i,5:9]
  nlinea=5-sum(is.na(lineas))
  for (k in 1:nlinea) {
    auxX[i,lineas[[k]]] <- 1
  }
}
X <- cbind(X[,- c(1,2,4:9)],auxX)
names(X) <- c(names(X)[1:10],"RecopilacionDatos","Integracion",
"SeguridadyRiesgos","TI","Electronica",
"SolucionesEmpresariales","Legal","Finanzas","SolucionesOficina",
"Industrial","Humanistica","Comercial","Otros","Comunicacion",
"Digitalizacion","AnaliticaDatos")

n=nrow(X)
d=ncol(X)
minimos <- NULL
```

```
maximos <- NULL
medias <- NULL
desv <- NULL
#Reescalamiento de X
for (i in 1:d) {
  minimos <- c(minimos,min(X[,i]))
  maximos <- c(maximos,max(X[,i]))
  medias <- c(medias,mean(X[,i]))
  desv <- c(desv,sd(X[,i]))
  X[,i] <- (X[,i]-mean(X[,i])) /(sd(X[,i]))
}


n=nrow(X)
d=ncol(X)
#Eliminar nombres y variables de identificacion
#A_t es centro de los clusteres, hay c clusteres
#A_t se inicializa como X puesto que a_k=x_k
#Incializacion
c0=c=n
At1=At=X
alphat=alphat1=rep(1/(n),c)
#X es data frame de anuncios
#Z indica si X_i pertenece al k-esimo cluster, k=1,...c
Z=matrix(data = NA, nrow = n,ncol = c)
#Paso 1
gamma0=beta0=1
#No se inicializa con 0 puesto que c[0] no existe
t=1
err=30
clust <- NULL
er <- NULL

#Condiciones del paso 8,
#Se eligen de A_t aquellos centros de los grupos que no
#fueron desechados
```

```
while(c[t]>1 && err>eps){
  #Actualiza elementos de iteraci n
  alphat=alphat1
  At=At1
  #Paso 2
  Z=matrix(data = 0, nrow = n,ncol = c[t])
  for (i in 1:n) {
    rx <- matrix(rep(as.numeric(X[i,]),c[t]),ncol = d,byrow = T)
    aux2=apply(rx-At, 1, norm,type="2")^2-
      gamma0*log(alphat,exp(1))
    if (t==1){
      aux2[i]=NA
    }
    k0=which(aux2 == min(aux2, na.rm = TRUE))
    Z[i,k0]=1
  }
  #Paso 3
  gamma0=exp(-c[t]/250)
  #Paso 4
  for (k in 1:c[t]) {
    aux6= sum(alphat*log(alphat,exp(1)))
    alphat1[k]=sum(Z[,k])/n + beta0/gamma0 * alphat[k] *
      (log(alphat[k], exp(1)) -aux6)
  }
  alphat1
  #Paso 5
  eta=min(1,1/t^floor(ncol(X)/2 -1))
  beta0=min( sum(exp(-eta*n*abs(alphat1-alphat)))/c[t],
             (1-max(colSums(Z)/n))/
               -max(alphat*sum(log(alphat,base = exp(1)))))
  #Paso 6
  #Actualizar el numero de clusteres c
  ct1=c[t] - length(alphat1[alphat1<1/n])
  c=c(c,ct1)
  c
  Zt=Z
```

```
Z=Z[,alphat1>=1/n]
Z
auxc=which(alphat1>=1/n)
alphat1=alphat1[alphat1>=1/n]
#Actualizar proporciones de mezcla alfa*
alphat1=alphat1/sum(alphat1)
alphat1
#Actualizar Z
if(c[t+1]==1){
  break
}
##Cuenta el numero de elementos por grupo
auxZ=rowSums(Z)
for (i in 1:nrow(Z)) {
  for (k in 1:ncol(Z)) {
    if(auxZ[i]!=0){
      Z[i,k]=Z[i,k]/auxZ[i]
    }
    else {
      rx <- matrix(rep(as.numeric(X[i,]),c[t+1]),ncol = d,
      byrow = T)
      aux2=apply(rx-At[auxc,], 1, norm,type="2")^2-
        gamma0*log(alphat[auxc],exp(1))
      if (t==1){
        aux2[i]=NA
      }
      #k0=which.min(aux2)
      k0=which(aux2 == min(aux2, na.rm = TRUE))
      Z[i,k0]=1
    }
  }
}
if(t>60 && (c[t-60]-c[t])==0){
  beta0=0
}
#Paso 7 actualizar $At_1$
```

```
  for (k in 1:c[t+1]) {
    if (sum(Z[,k]==1)>1) {
      At1[k,]=colSums(X[which(Z[,k]==1),])/sum(Z[,k]==1)
    }
    else {
      if(sum(Z[,k]==1)==1){
        At1[k,]=X[which(Z[,k]==1),]
      }
    }
  }
  t=t+1
  At1=At1[1:c[t],]
  err=max(apply(At1-At[auxc,], 1, norm, type="2"))
  clust <- c(clust,c[t])
  er <- c(er,err)
}
c
colSums(Zt)

for (i in 1:d) {
  X[,i] <- (X[,i])*(desv[i])+medias[i]
}

grupos <- 1:n
for (i in 1:c[t]) {
  grupos[which(Zt[,i]==1)]=i
}

#Encuentra los elementos del primer grupo
g1 <- X[grupos==1,]
colSums(g1[11:26])
#Calculo indices Dunn, Davies-Bouldin, valor de silueta
dn0 <- dunn(clusters = grupos, Data = X,)
DB <- index.DB(X,grupos,centrotypes = "centroids",p=2,q=1)
DB
sil <- silhouette(grupos,dist =dist(X))
```

```
fviz_silhouette(sil)
```

Tabla de iteraciones, número de grupos y error

t	c	Error
1	254	100
2	205	15.66
3	95	11.87
4	47	8.21
5	31	7.03
6	19	1.88
7	16	7.59
8	14	5.69
9	9	2.86
10	8	3.75
11	6	1.25
12	6	0.24
13	6	0

ID/Sector	Nombre		Tags
1	Sector de Recopilación y Acceso a Información.	Sector enfocado en captar información, guardarla y asegurar su disponibilidad para su futuro uso.	Servicios en la Nube
			Estudios / informaciòn de Mercados
			Agencias de Marketing
			Data Center
2	Sector de Integración	Sector enfocado en la conectividad de una empresa, comprende las redes físicas dentro de una empresa así como el diseño necesario para que todos los departamentos	Redes (Físicas)
			Diseño de Bases de Datos
			IoT / Domótica
3	Sector de Seguridad y Riesgos	accedan a la información que necesiten. Provee protección de factores externos de manera física y virtual, reduciendo riesgos y desastres futuros.	Cyber Seguridad y Riesgos
			Rastreo Satelital / Tracking
			Biometría y Seguridad y Riesgos Física
			Calidad de software QA
4	Sector de Tecnologías de la información	Soluciona problemas de las tecnologías de la información proveyendo además software desarrollado por terceros .	Licenciamiento de Software
			IT / Asesoría
5	Sector de Objetos Electrónicos	Instrumentos físicos compuestos puramente de electrónicos, provee con computadoras y teléfonos a las diversas empresas.	Venta de Computadoras
			Manufactura instrumentos con electrònicos
			Venta Electrónicos
6	Sector de Soluciones Empresariales	Relacionados con el manejo empresarial en sus diversos niveles, incluye CRM, BI, ERP, gestión de proveedores y la consultoría acerca de procesos empresariales	CRM Manejo Relacion Cliente / Atencion al cliente
			Soluciones de negocios / BI / sistemas de manejo contable
			Soluciones administrativas / ERP / BPM / ECM
			Ingeniería y consultoria de negocios / Impulsadora e Incubadora
			SRM Sistema Gestion de proveedores / Supply Chain Management
7	Sector de Soluciones Legales	Facilita los procesos requeridos por la ley, tanto mediante la informática y el software como mediante la representación legal.	Planificacion y consultoria tributaria / arancelaria
			LegalTech
			Asesoría legal
			Certificados electrónicos / firma electrónica / Facturación Electrónica
8	Sector de Soluciones Financieras	Provee facilidades en el manejo del dinero, el crédito y demás. Contiene soluciones como las pasarelas de pagos, así como entidades financieras como bancos y entidades de recuperación de cartera.	Entidad Financiera / Fondo de Inversiòn
			Pasarela de pagos
			Red de pagos / Billetera electrónica / Punto de venta / Pagos
			Fondo de Inversiòn / Administrador de Inversiones
			Software Financiero - FINTECH
			Blockchain
			Gestion de cobranza y recuperaciòn de cartera
			Buro de credito
9	Sector de Soluciones de Oficina	Contiene los materiales de oficina y entidades que proveen un lugar donde tener estas.	Muebles y materiales de oficina
			Immueble y construcción
10	Sector de Soluciones Industriales	Provee maquinaria, sistemas y herramientas necesarias para una industrialización de primer nivel, contiene soluciones eléctricas, maquinaria, herramientas, movilidad y logística para la industria.	Sistemas y Soluciones Elèctricas
			RPA / Automatización Industrial
			Maquinaria, Industria, Herramientas y Laboratorio
			Tecnologia Industrial, Seguridad y Riesgos Industrial
			Soluciones en Movilidad
			Transporte y envíos / Logistica
			Energía y generadores eléctricos
11	Sector de Talento Humano	Dirigido a los empleados de la empresa, el sector de talento humano provee capacitaciones y coaching al personal.	Capacitación Software
			Liderazgo / Coaching / Relaciones Interpersonales
			Capacitación Proyectos / negocios
			Capacitación Idiomas, contabilidad y otros
12	Sector Comercial	Sector enfocado en los clientes, como la información llega a ellos y como estos se sienten tanto con el producto como por la empresa, incluye la experiencia de usuario y los Contact Center	Service Desk / Contact Center/ Help desk
			Tiendas Online E-Store
			Streaming / Televisión / Audiovisuales / Medio Digital
			CX / Customer Experiencie / UX
13	Otros	No tecnológicos, incluyen sectores como la agricultura entre varios otros.	Manejo de activos digitales
			Agricultura
			Ganadería
			Impulsadora/ Aceleradora / Incubadora
			Medio Digital / Revista
			Anuncios
			Otros
14	Comunicación	Facilita la comunicación entre los empleados de una empresa, tanto entre ellos como con el exterior.	Proveedores de Internet
			Conectividad (no física)
			Telefonia
15	Sector de Digitalización	Transforma los procesos físicos en digitales, construye la presencia en la web, facilitando al usuario del producto un contacto mediante internet con la empresa, también incluyen las tecnologías específicas para determinadas áreas como EdTech, HealthTech y E-gobern.	Diseño y Desarrollo de Software
			Diseño de Páginas Web / apps web
			Seguros / INSURTECH
			SaaS / Paas / IaaS
			EdTech
			Agrotech
			HealthTech / MedTech/ Medicina
			Plataforma / Software de RRHH
			E-gobern
16	Sector de aprovechamiento de datos	Da acceso a nueva información de una empresa basada en diversos datos de ésta	Analítica de Acceso a la información / datos
			AI

Cuadro 3.1: ID de los diferentes servicios ofrecidos

Bibliografía

[1] Abrahamson, D. . P.-M. M. R. *The Routledge Handbook of Magazine Research: The Future of the Magazine Form* . Routledge, US, 2015.

[2] Ackerman, M. Characterization of Linkage-based Clustering. . *Journal of Machine Learning Research*, 17:1–17, 2016.

[3] Alhawarat, M. Revisiting K-Means and topic modeling a comparison study to cluster arabic documents . *IEEE*, 6, 2018.

[4] Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370, 1987.

[5] Brock, G., P. V.-D. S. y Datta, S. clValid: An R Package for Cluster Validation . *Journal of Statistical Software*, 25, 2008.

[6] Ducoffe, R. Advertising value and advertising on the web . *Journal of Advertising Research,*, 36, 1996.

[7] Ellman, M. What do the papers sell? A model of advertising and media bias. *The Economic Journal*, 119:680–704, 2009.

[8] Feldman, L. P. Principles vs. practice in new product planning. *Journal of Product Innovation Management*, 1:43–55, 1984.

[9] Hamouda, M. Understanding social media advertising effect on consumers' responses: An empirical investigation of tourism advertising on Facebook . . *Journal of Enterprise Information Management*, 2018.

[10] Iankova, S. . A comparison of social media marketing between B2B, B2C and mixed business models . *Industrial Marketing Management*, 81:169–179, 2019.

[11] Ienco, D. From context to distance: Learning dissimilarity for categorical data clustering. . *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6:1–25, 2012.

[12] Kahn, K. B., . M. M. *Innovation and New Product Planning* . Routledge, 2020.

[13] Kahn, K. B. *Product planning essentials*. Sage Publications, 2000.

[14] Kass, R. Bayes factors. *J. Amer. Stat. Assoc*, 90:773–795, 1995.

[15] Kassambara, A. y Mundt, F. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. .

[16] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., y Hornik, K. *cluster: Cluster Analysis Basics and Extensions*, 2021. R package version 2.1.2.

[17] Media, C. Characteristics of innovative companies: A case study of companies in different sectors . *Creativity and Innovation Management*, 14:272–287, 2005.

[18] Mehta, A. Advertising attitudes and advertising effectiveness . *Journal of advertising research*, 40:67–72, 2000.

[19] Natchwwey, A. Cluster analysis as a method for the planning of production systems. *IEEE*, págs. 725–728, 2009.

[20] Sinaga, K. P. Unsupervised K-means clustering algorithm. *IEEE*, 8, 2020.

[21] Walesiak, M. y Dudek, A. *The Choice of Variable Normalization Method in Cluster Analysis*. International Business Information Management Association (IBIMA), 2020.

[22] Xu, R. *Clustering* . John Wiley Sons, 2008.

[23] Zheng, A., . C. A. *Feature engineering for machine learning: principles and techniques for data scientists.* . O'Reilly Media, Inc., 2018.