UNIVERSIDAD NACIONAL AGRARIA

LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"SEGMENTACIÓN DE LECTORES DIGITALES REGISTRADOS DE UN SITIO WEB INFORMATIVO CON EL ALGORITMO DE ANÁLISIS CLUSTER K-MEANS"

PRESENTADO POR

BRIAN ERICK CLEMENTE RIVERA

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO DE

INGENIERO ESTADÍSTICO INFORMÁTICO

Lima - Perú

2021

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

“SEGMENTACIÓN DE LECTORES DIGITALES REGISTRADOS DE UN SITIO WEB INFORMATIVO CON EL ALGORITMO DE ANÁLISIS CLUSTER K-MEANS”

PRESENTADO POR

BRIAN ERICK CLEMENTE RIVERA

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO DE INGENIERO ESTADÍSTICO INFORMÁTICO

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

…….	…….
Mg. Clodomiro Fernando Miranda Villagómez	Mg. Iván Dennys Soto Rodríguez
PRESIDENTE	ASESOR
…….	…….
Dr. Jorge Chue Gallardo	Mg. Sc. Ana Cecilia Várgas Paredes
MIEMBRO	MIEMBRO

Lima - Perú

2021

ÍNDICE GENERAL

I. INTRODUCCIÓN 1
1.1. Problemática 1
1.2. Objetivos 3
1.2.1. Objetivo General 3
1.2.2. Objetivos Específicos 3
II. MARCO TEÓRICO 4
2.1. Analítica Digital 4
2.2. Google Analytics 5
2.3. Google BigQuery 8
2.4. Análisis Clúster 10
2.5. K-means 12
III. MARCO METODOLÓGICO 14
IV. RESULTADOS Y DISCUSIÓN 20
V. CONCLUSIONES Y RECOMENDACIONES 27
5.1. Conclusiones 27
5.2. Recomendaciones 28
VI. REFERENCIAS BIBLIOGRÁFICAS 29

ÍNDICE DE TABLAS

Tabla 1. Las 10 herramientas de analítica web más populares 5

Tabla 2. Principales diferencias entre Google Analytics Standard y 360 8

Tabla 3. Homologación de campos de BigQuery con métricas en Google Analytics. 9

Tabla 4. Métodos de elección del número óptimo de agrupaciones para k-means 12

Tabla 5. Base de datos de navegación web obtenida con Google BigQuery 16

Tabla 6. Descripción de las variables del conjunto de datos de lectores digitales 17

Tabla 7. Base de datos final para aplicar algoritmo k-means. 18

Tabla 8. Análisis descriptivo de los segmentos de lectores obtenido por k-means 20

Tabla 9. Porcentaje de consumo de notas por secciones según segmentos de lectores 21

Tabla 10. Porcentaje de lectores digitales según características sociodemográficas 23

Tabla 11. Sesiones y porcentaje de lectores por segmento según el comportamiento web 23

Tabla 12. Perfil de lectores registrados según características de los segmentos encontrados. 25

Tabla 13. Porcentaje de lectores digitales que cuentan con suscripción al diario impreso 26

ÍNDICE DE FIGURAS

Figura 1. Embudo de participación de segmentos de lectores digitales. 2

Figura 2. Funcionamiento de Google Analytics. 6

Figura 3. Entorno de Google Analytics para la cuenta de demostración – Google Merchandise Store. 7

Figura 4. Entorno de Google BigQuery en GCP para la cuenta de demostración – Google Merchandise Store. 9

Figura 5. Distancias intra-clúster e inter-clúster finalizada la generación de grupos 10

Figura 6. Funcionamiento del algoritmo k-means cuando se seleccionan 2 grupos. 13

Figura 7. Proceso de integración de datos en RStudio para su posterior análisis. 15

Figura 8. Gráficos del índice D obtenido con la función NbClust. 19

RESUMEN

Debido a la gran diversidad de negocios de la empresa, un importante grupo de medios de comunicación, es que se genera mucha información, cada vez más específica y detallada; y es aquí en donde entra la gerencia de Inteligencia de Negocios, la cual centraliza todos estos datos provenientes de distintas fuentes y plataformas con la finalidad de analizarlos y brindar soporte a las diferentes áreas que requieran de algún análisis a detalle para sustentar una venta, una adquisición, el desarrollo de proyectos, etc. Una de estos requerimientos especializados tiene que ver con la integración de datos de distintos orígenes tanto digitales como los que se generan en los sitios web, las redes sociales; o las tradicionales como los ingresos que genera la publicidad para la compañía, base de datos de audiencia, entre otros, ya que permitirán abordar análisis más complejos para encontrar hallazgos más específicos, diferenciales y relevantes. La presente monografía aborda el desarrollo de una nueva metodología de segmentación de usuarios registrados en la página web, en la cual se ha planteado considerar el tipo de contenido que ellos visitan según la sección en la que están alojadas las notas y complementándolos con la información personal, sociodemográfica, de ubicación y otras disponibles en el negocio, todo esto apoyado en el análisis clúster, específicamente el algoritmo k-means. Para el preprocesamiento de datos, limpieza, construcción del conjunto de datos y ejecución de la metodología se utilizó el software R, que posee múltiples funciones que ayudaron con estas tareas. Estas seis agrupaciones encontradas permitirán ofrecer a los clientes un nuevo producto comercial, que además otorgará una ventaja para los clientes ya que podrán especificar la audiencia específica a la que quieren impactar mejorando significativamente los resultados que se obtendrían a diferencia del método tradicional de publicidad digital.

Palabras claves: analítica web, análisis clúster, k-means, segmentación, audiencia digital, google analytics

ABSTRACT

Due to the wide diversity of the company's business, an important group of media, a lot of information is generated, more specific and detailed each time; and this is where the Business Intelligence management comes in, which centralizes all these data from different sources and platforms in order to analyze them and provide support to the different areas that require a detailed analysis to support a sale, an acquisition, project development, etc. One of these specialized requirements has to do with the integration of data from different digital sources such as those generated on websites, social networks; or traditional ones such as advertising revenue generated for the company, audience database, among others, as they will allow to address more complex analysis to find more specific, differential and relevant findings. This monograph is about the development of a new methodology for segmentation of registered users on the website, in which it has been proposed to consider the type of content they visit according to the section in which the notes are hosted and complementing them with personal, sociodemographic, location and other information available in the business, all this supported by the cluster analysis, specifically the k-means algorithm. For data preprocessing, cleaning, construction of the data set and execution of the methodology, R software was used, which has multiple functions that helped with these tasks. These six clusterings found will allow offering clients a new sales product, which will also provide an advantage for consumers since they will be able to specify the specific audience they want to impact, significantly improving the results that would be obtained, unlike the traditional method of digital advertising.

Keywords: web analytics, clustering, k-means, segmentation, digital audience, google analytics

I. INTRODUCCIÓN

1.1. Problemática

A nivel digital, en la compañía se entregan mensualmente informes respecto al desempeño de los sitios web y de las redes sociales en función a ciertos indicadores digitales como el número de visitantes, el tiempo de permanencia promedio en la página o la cantidad de interacciones; que sirven de sustento para los ejecutivos de ventas al momento de ofrecer las plataformas digitales de la empresa a sus clientes; sin embargo, para otras áreas de negocio como prensa era importante conocer a los lectores digitales respecto a sus hábitos, es decir se necesitaba clasificarlos según ciertas especificaciones; además el poder ofrecer estos segmentos como una nueva propuesta comercial, permitiría una mejora en los resultados obtenidos en campañas publicitarias ya que estarían direccionadas a un público objetivo.

Inicialmente, la clasificación de los usuarios digitales empleada se basada en la metodología del News Consumer Insight de Google News Initiative, que utiliza el concepto de embudo de participación para segmentar en audiencias clave datos referentes a la frecuencia de visitas de cada uno al sitio web, que se obtienen desde Google Analytics, herramienta de analítica digital de Google, cuyo objetivo es maximizar el valor de estos a medida que los visitantes atraviesan un flujo de participación (Adams Harding & Gingras, 2018).

Los criterios de clasificación empleados para segmentar a la audiencia digital inicialmente fueron los siguientes:

- Lector casual: Usuarios identificados como nuevos según Google Analytics; es decir, quienes solo hayan visitado una vez el sitio durante cierto periodo.
- Lector habitual: Aquellos usuarios cuya frecuencia de visitas se encuentra entre 2 o 15 durante el periodo evaluado.
- Lector incondicional: Quienes han visitado el sitio web en más de 15 oportunidades en el periodo considerado.

Hay que tener presente que esta clasificación de lectores digitales se generaba mensualmente con los datos completos del mes, a diferencia del criterio de Google News Initiative que lo empleaba para los últimos 28 días de cada mes. La otra diferencia destacable era que ellos consideraban un cuarto nivel en su categorización, los llamados "suscriptores", aquellos usuarios que pagaban por acceder a contenido exclusivo (este es un plan a largo plazo que también se va a considerar). En la Figura 1 se visualiza el embudo de participación elaborado por Google.

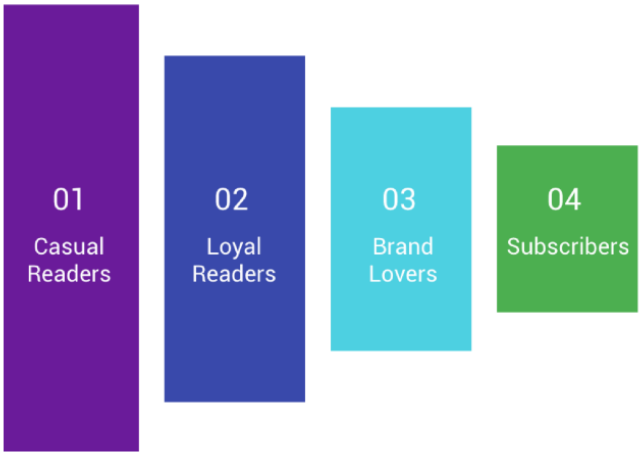

Figura 1. Embudo de participación de segmentos de lectores digitales.
Fuente: Google News Initiative – New Consumer Insights Playbook, 2018, pág. 6

Hasta cierto punto esta categorización era suficiente ya que todos los visitantes del sitio web eran anónimos, es decir no se podría hacer mayor rastreo respecto de ellos. Sin embargo, la necesidad de conocer a mayor detalle a estos usuarios digitales como conocer el tipo de dispositivos usan, desde donde se conectan, el tipo de contenido que leen, entre otros, llevó a implementar un sistema de registros a la página como siguiente paso.

En Ljubljana, Eslovenia se realizó un estudio de segmentación de audiencia digital basada en perfiles temáticos del contenido visitado, que señala que uno de los usos frecuentes de la categorización permite obtener información sobre los usuarios para respaldar acciones de marketing dirigido o personalizado y, en general, proporcionar recomendaciones en línea afines a sus intereses; sin embargo su experimento presentó problemas cuando los visitantes tenían intereses múltiples lo cual hacia más complicado la diferenciación de las agrupaciones, por lo

que proponen contar con datos adicionales como la demografía, la ubicación geográfica e incluso el área laboral (Kladnik, Stopar, Fortuna, & Mladenić, 2017).

Con el sistema de registros ya en marcha se hizo posible la integración de los datos digitales con otras fuentes de la empresa, como la de suscripciones al diario impreso, los concursos realizados anteriormente e incluso de proveedores externos. De esta manera se logró incrementar el nivel de detalle de los datos de casa usuario y se pudo generar una agrupación más prometedora.

Para el desarrollo e implementación del nuevo criterio de segmentación de lectores digitales en el sitio web informativo fue necesario recopilar e integrar los datos de todas las fuentes de información para con ellos construir agrupaciones de usuarios basadas en características a través del algoritmo k-means del análisis cluster, empleando las distintas librerías y funciones que contiene el software estadístico R para este fin.

1.2. Objetivos

1.2.1. Objetivo General

Segmentar a los lectores digitales registrados de un sitio web informativo en grupos diferenciales, teniendo en cuenta tanto los datos digitales como los no digitales, a través del análisis clúster empleando el algoritmo k-means.

1.2.2. Objetivos Específicos

- Obtener grupos diferenciados para los lectores digitales en función del tipo de contenido consumido e integrando datos de otras fuentes de la empresa como variables sociodemográficas o bases de terceros.
- Ofrecer grupos diferenciados como un nuevo producto comercial para clientes, el cual les otorgué una ventaja ya que permitirá especificar al grupo específico de usuarios al que quieren alcanzar.
- Identificar el tipo de contenido y la cantidad de notas que consumen en mayor proporción cada uno de estos segmentos encontrados.
- Conocer clientes que también cuenten con una suscripción al diario impreso y como se clasifican según este criterio de segmentación encontrado.

II. MARCO TEÓRICO

2.1. Analítica Digital

La analítica digital, específicamente la analítica web, fue definida por la Web Analytics Association en 2015 como la "medición, recopilación, análisis y presentación de informes de datos de Internet con el fin de comprender y optimizar el uso de un sitio web". La evolución de este proceso se produjo de forma simultánea con el desarrollo del marketing digital, y este a su vez, con el auge de la internet, la tecnología y la digitalización (Sponder & Khan, 2018).

Un sitio web informativo genera una cantidad infinita de datos diariamente que nos proporciona un detalle tanto sobre su audiencia como de su comportamiento dentro de él y que puede obtenerse mediante herramientas de analítica digital. Ellas están basadas en técnicas de seguimiento y algoritmos sofisticados que procesan y evalúan grandes volúmenes de datos capturados. Su uso brinda un soporte importante para un mejor reconocimiento de la actividad de los visitantes, permite identificar cuellos de botella y errores en el diseño de la interfaz, es capaz de medir el rendimiento del entorno y monitorear la disponibilidad del sitio web o recomendar contenido apropiado (Čegan & Filip, 2017).

Existen muchas herramientas diferentes de analítica web actualmente, tanto gratuitas como de pago, las cuales tienen como objetivo obtener datos cuantitativos y cualitativos como base para el proceso de toma de decisiones. Con estos datos se podrá averiguar cuántos visitantes llegan a un sitio web, cuánto tiempo pasan allí, qué porcentaje de ellos se suscriben y/o compran algo y mucha más información útil, lo que ayudará a ir perfeccionando un sitio (Skrba, 2020).

El portal First Side Guide ha listado el top 10 de herramientas de analítica web más populares basados en su uso en muchas empresas en todo el mundo para generar información valiosa sobre el rendimiento de su sitio web y su negocio, entre las que tenemos a Similarweb, Clicky, Woopra, Chartbeat y Google Analytics; está última destaca debido a que brinda información relevante y es principalmente una plataforma gratuita disponible para todos, sin embargo, también tiene una versión de pago con más funciones y características más avanzadas. La lista completa se puede ver en la Tabla 1.

Tabla 1. Las 10 herramientas de analítica web más populares

N°	Herramienta	Principal característica
1	Google Analytics	Convertir el conocimiento de los datos en acción
2	Similarweb	Mide el mundo digital
3	Clicky	Analiza el tráfico en tiempo real
4	Matomo	Análisis de sitios web con 100% de control
5	Finteza	Análisis y evaluación web avanzados
6	Woopra	Comprender el recorrido del cliente
7	Chartbeat	Medir el compromiso del cliente
8	Hotjar	Comprender el tráfico y los datos del sitio
9	Crazyegg	Análisis detallado de sitios web
10	Mixpanel	Análisis web y ciencia de datos

Fuente: (First Site Guide, 2021)

2.2. Google Analytics

Google Analytics es una de las principales herramientas de analítica digital hoy en día; es el servicio de análisis web gratuito de Google que permite analizar en profundidad los detalles de los visitantes de su sitio web como que hacen las personas cuando lo visitan, cuánto tiempo permanecen y qué páginas visitan en él (Thakur, 2017).

Para empezar a trabajar en Google Analytics se debe de crear una cuenta, esto generará una etiqueta o código de seguimiento (en JavaScript) que debe ser insertado en el sitio web, el cual le permitirá rastrear tanto a los visitantes de la página como cualquier acción que ellos realicen en ella. Cuando un usuario visita su sitio web, Google Analytics colocará una cookie en el navegador del usuario. Las cookies son pequeños archivos que contienen información sobre las actividades del usuario. Al usar estas cookies, Google Analytics sabrá cómo se comporta un usuario en su sitio web y luego recopila esta información para mostrarla a través de sus distintos informes (Akhtar, 2019). Este proceso lo podemos visualizar en la Figura 2.

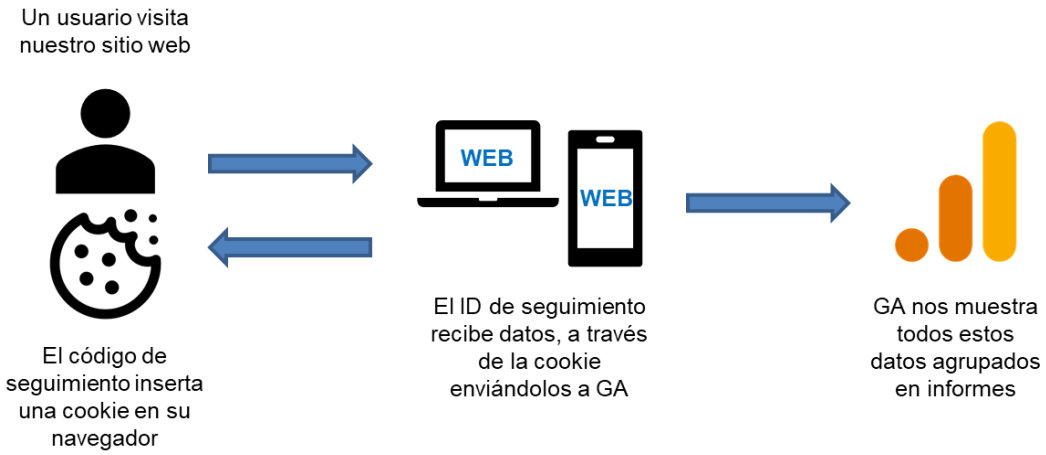

Figura 2. Funcionamiento de Google Analytics.

Fuente: Elaboración propia, 2020

Los informes de Google Analytics están compuestos de dimensiones y métricas. Las dimensiones son los atributos cualitativos de los datos, mientras que las métricas son los indicadores cuantitativos; ambos conceptos van de la mano.

Entre las principales dimensiones que genera Google Analytics tenemos:

- Tiempo: Año, mes, día, hora y fecha.
- Ubicación geográfica: País, región y ciudad
- Categoría de dispositivo: Desktop, mobile o Tablet
- Fuente/medio: El origen de la visita al sitio web y su categoría. Por ejemplo: Facebook, Google, o email.
- Atributos de página: Título, ruta de página (URL) y hostname, etc.

Entre las principales métricas (Marketing Analítico, 2019) que genera Google Analytics tenemos:

- **Usuario:** Un ID de usuario distinto identificada en la cookie de seguimiento (no una persona). Siempre que un usuario vuelva a acceder desde el mismo navegador y dispositivo y no elimine las cookies, se identificará como el mismo (usuario recurrente), caso contrario será considerado un nuevo usuario.
- **Sesión:** Es el período de navegación de un usuario. Esta comienza cuando el usuario accede por primera vez al sitio web y finaliza cuando ocurre una inactividad de 30

minutos, al llegar la medianoche ya que Google Analytics mide datos diariamente, y si la fuente de acceso a la página cambia.

- **Página Vista:** Es simplemente el acto de cargar una página. Si un usuario visita 3 páginas o vuelve a cargar la misma página 3 veces, eso cuenta como 3 páginas vistas, independientemente de las sesiones.
- **Duración media de la sesión:** Es el promedio de la duración total de todas las sesiones (en segundos) para el número de sesiones.

Podemos apreciar en la Figura 3 la plataforma de Google Analytics, en la que se nota que los usuarios son considerados como una métrica, es decir, no podríamos obtener mayor detalle sobre él, sin embargo, esto si es posible si cotamos con la versión de pago, Google Analytics 360, la cual posee integraciones con muchas otras herramientas de Google, entre ellas Google BigQuery.

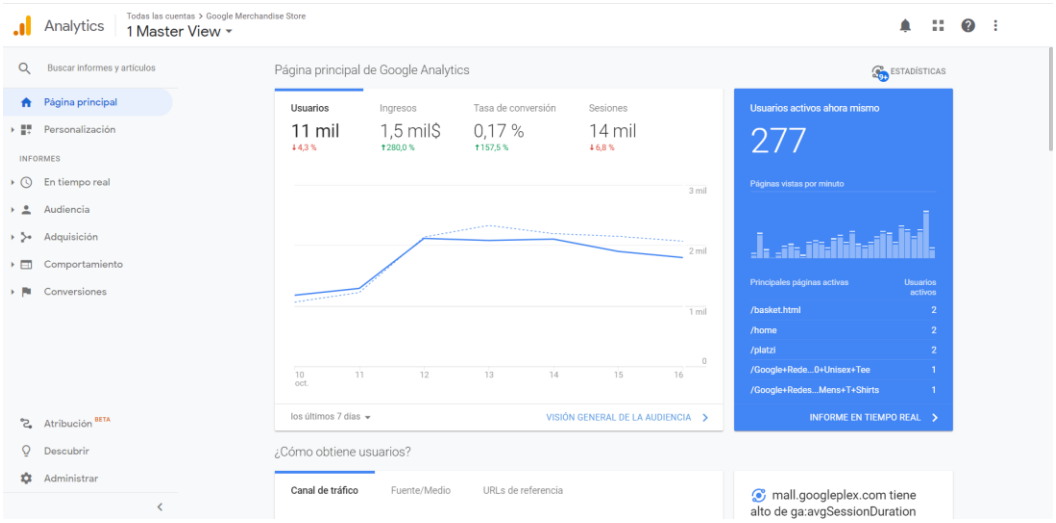

Figura 3. Entorno de Google Analytics para la cuenta de demostración – Google Merchandise Store.

Fuente: Elaboración propia, 2020

La principal diferencia entre la versión estándar de Google Analytics y la 360 se da en la precisión de la obtención de los datos: La gratuita aplica muestreo (sampling) en los informes por defecto a partir de 500,000 sesiones (por propiedad y período de análisis), sin opción de acceso a los datos sin procesar (raw data). Sin embargo, Analytics 360, eleva este umbral a los 100 millones y, también, dispone de los reportes sin muestreo (Unsampled Reports) para consultas puntuales y de conexión con BigQuery para trabajar con todos los datos a nivel de hit

(DBi Data Business Intelligence - Havas, 2019). En la Tabla 2 se resumen las principales diferencias entre ambas versiones:

Tabla 2. Principales diferencias entre Google Analytics Standard y 360

Característica	Google Analytics Standard	Google Analytics 360
Informes sin muestreo	No	Si
Acceso a datos sin procesar (raw data)	No	Si
Informes avanzados y segmentación	Si	Si
Integración con Google BigQuery	No	Si
Actualización de datos	No garantizada	4 horas garantizadas (según acuerdo)
Soporte	Centro de Ayuda y foros de la comunidad	Servicios a nivel de empresa y soporte
Opciones de pago	Gratuito	Facturación mensual

Fuente: (DBi Data Business Intelligence - Havas, 2019)

2.3. Google BigQuery

Google BigQuery es un data warehouse, un servicio que forma parte del Google Cloud Platform (GCP), es decir, un almacén de datos en la nube que permite el almacenamiento y la consulta rápida en lenguaje SQL estándar sobre grandes conjuntos de datos. BigQuery es rápido, devolviendo resultados en segundos incluso cuando se ejecutan consultas sobre datos a escala de terabytes. Lo que hace que esta velocidad sea uniforme y fácil de obtener es que no requiere de la creación (ni siquiera la especificación) de índices, es decir, cualquier campo puede ser consultado rápidamente, en comparación con MongoDB y los sistemas tradicionales de SQL como MySQL o Postgres, para los cuales las consultas son rápidas sólo en los campos con índices existentes (Lopez, Seaton, Ang, Tingley, & Chuang, 2017).

Se puede exportar datos de sesiones y de hits desde una cuenta de Google Analytics 360 a BigQuery sin problemas. Cuando un objeto de seguimiento envía datos a Google Analytics, se denomina enviar un hit, y cada hit debe tener un tipo. El fragmento de seguimiento JavaScript envía un hit del tipo page; otros tipos de hits son: screenview, event, transaction, item, social, exception y timing (Google Analytics Developers, 2019).

Entre las columnas del archivo de exportación de Google Analytics en BigQuery encontramos las métricas y dimensiones (Ayuda de Google Analytics, s.f.) de forma diferente, algunas homologaciones de estos campos se muestren en la Tabla 3.

Tabla 3. Homologación de campos de BigQuery con métricas en Google Analytics.

Nombre del campo en BigQuery	Tipo de dato	Descripción	Homólogo en Google Analytics
fullvisitorId	cadena	ID del usuario único	Usuario
visitId	entero	Identificador de esta sesión. Es la parte del valor que normalmente se almacena como la cookie _utmb. Solo es único para el usuario.	Sesión
totals.pageviews	entero	Número total de páginas vistas en la sesión.	Páginas Vistas

Fuente: Elaboración propia. 2020. Basado en el Esquema de BigQuery Export

Finalmente, la estructura de los datos que podemos obtener desde BigQuery es la que se muestra en la Figura 4. Ya con esta consideración de un usuario como código (fullvisitorId) será más sencillo para trabajarlo como una base de datos, donde cada fila representa un hit.

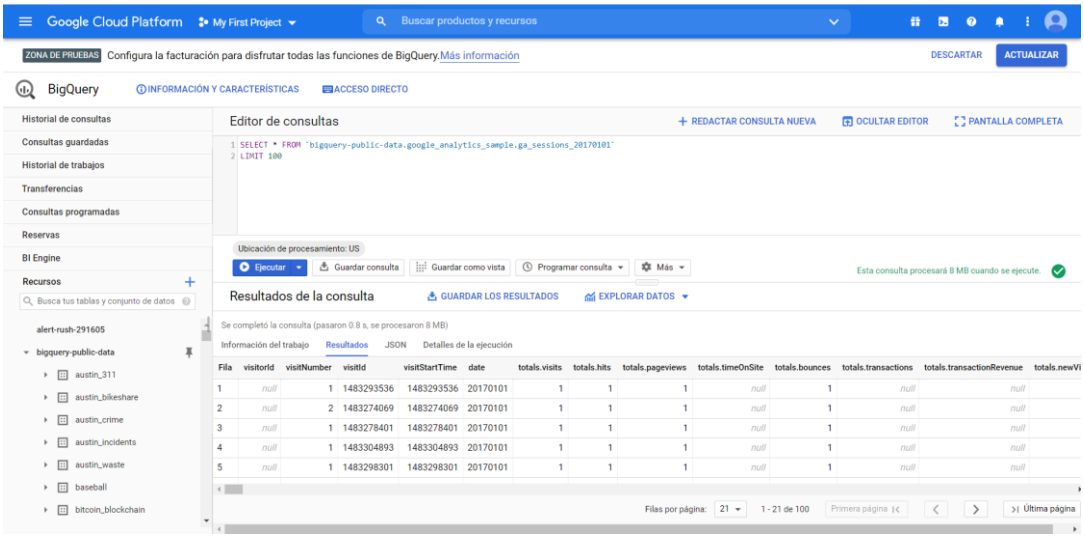

Figura 4. Entorno de Google BigQuery en GCP para la cuenta de demostración – Google Merchandise Store.

Fuente: Elaboración propia, 2020

2.4. Análisis Clúster

Las técnicas de análisis clúster se refieren a un conjunto muy amplio de algoritmos para encontrar subgrupos o grupos en un conjunto de datos. La idea de estos métodos es tratar de dividir los datos en grupos distintos para que las observaciones dentro de cada grupo sean bastante similares entre sí, mientras que las observaciones en diferentes grupos sean muy diferentes entre sí (James, Witten, Hastie, & Tibshirani, 2017).

La similitud entre las observaciones de un conjunto de datos se define utilizando algunas medidas de distancia entre ellas, incluidas medidas de distancia euclidianas y basadas en correlación, por lo que estas técnicas trabajan con datos de tipo cuantitativo.

Esta generación de grupos se basa en las distancias intra-clúster e inter-clúster:

- La distancia intra-clúster es la distancia entre los puntos de datos dentro del grupo. Si hay un fuerte efecto de agrupamiento, esta debería ser mínima (más homogéneo).
- La distancia inter-clúster es la distancia entre puntos de datos en diferentes grupos. Cuando exista un fuerte agrupamiento, estas deberían ser máximas (más heterogéneos).

En la Figura 5 se puede visualizar el efecto estas las distancias luego de la generación de los grupos.

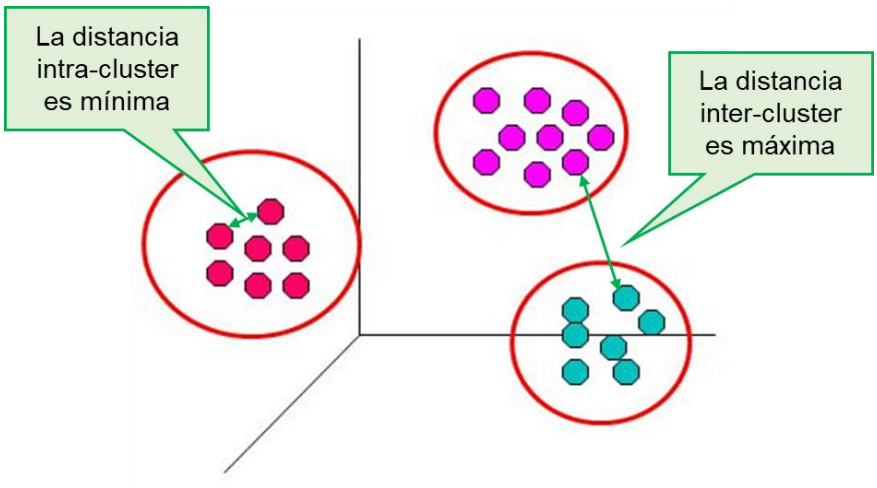

Figura 5. Distancias intra-clúster e inter-clúster finalizada la generación de grupos
Fuente: Elaboración propia, 2021. Adaptado de Medium.com

A diferencia de muchos otros métodos estadísticos, el análisis clúster se usa con frecuencia cuando no se hace una suposición sobre las posibles relaciones dentro de los datos; además, proporciona información sobre dónde existen asociaciones y patrones en los datos, pero no cuáles podrían ser estas o qué signifiquen.

El análisis clúster es utilizado en múltiples áreas:

- Salud: En la investigación del cáncer, para clasificar a los pacientes en subgrupos según su perfil de expresión génica.
- Marketing: Segmentación del mercado mediante la identificación de subgrupos de clientes con perfiles similares y que podrían ser receptivos a una forma particular de publicidad.
- Urbanismo: Identificar grupos de viviendas según su tipo, valor y ubicación.

Debido a que el análisis clúster es muy empleado en muchos campos, existe una gran cantidad de métodos de agrupación: los métodos jerárquicos y los no jerárquicos; la elección de alguno de ellos dependerá de los objetivos de estudio. Además, la limpieza de datos es un paso preparatorio esencial para un análisis de clúster exitoso; ya que esta técnica 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 (Qualtrics, 2020).

Los métodos no jerárquicos son usados para clasificar observaciones de un conjunto de datos en múltiples grupos basado en su similaridad, los cuales requieren que se defina a priori la cantidad de grupos a considerar. Se van reasignando las observaciones a los conglomerados de forma iterativa hasta que se satisfaga algún criterio de parada (por ejemplo, la suma de cuadrados de la varianza dentro de los grupos debe ser la más pequeña). En este grupo encontramos a los algoritmos K-means, PAM (Partitioning Around Medoids) y CLARA (Clustering Large Applications).

Para los métodos jerárquicos, no sabemos de antemano cuántos grupos queremos; de hecho, terminamos con una representación visual en forma de árbol de las observaciones, llamada dendrograma, que nos permite ver de una vez los agrupamientos obtenidos para cada número posible de agrupamientos, de 1 a n, siendo n la cantidad total de observaciones (James, Witten,

Hastie, & Tibshirani, 2017). En ese grupo encontramos a los algoritmos aglomerativos o AGNES (Agglomerative Nesting) y divisivos o DIANA (Divisive Analysis).

2.5. K-means

K-means es el método de agrupamiento más simple y común ya que tiene la capacidad de agrupar grandes cantidades de datos con un tiempo de cálculo relativamente rápido y eficiente (Syakur, Khotimah, Rochman, & Satoto, 2018).

Para realizar la agrupación por el método de k-means, primero se debe especificar el número deseado de agrupaciones (k) y con este valor determinado, el algoritmo asignará cada observación exactamente a uno de los k grupos (James, Witten, Hastie, & Tibshirani, 2017). Cada clúster se representa por el promedio de los datos que componen cada grupo; además, esta técnica es sensible a valores atípicos.

Elegir el número óptimo de agrupaciones es algo subjetivo ya que dependerá de muchos factores; sin embargo, existen dos tipos de métodos para la elección de esta cantidad.

Por un lado, los métodos directos, como codo y silueta, que consisten en optimizar la suma de cuadrados de la varianza dentro de cada grupo. Por otro lado, tenemos a los métodos de pruebas estadísticas, como las brechas, que compara la evidencia contra la hipótesis nula (Medium, 2019). El detalle de estos métodos se explica en Tabla 4:

Tabla 4. Métodos de elección del número óptimo de agrupaciones para k-means

Métodos directos		Método de prueba estadística
Método de codo	Método de silueta	Método de brecha
Consiste en identificar el número de grupos basándose en el supuesto de que el número óptimo de agrupaciones debe producir una pequeña inercia o una variación total dentro del clúster. Habrá una compensación entre la inercia y el número de grupos.	La puntuación de silueta mide qué tan bien está agrupada una observación y estima la distancia media entre agrupaciones. Trata de encontrar el número óptimo de clústeres que produzcan una subdivisión del conjunto de datos en bloques densos que estén bien separados entre sí.	Compara la variación total intra-clúster para diferentes valores de k con sus respectivos valores esperados bajo la distribución de referencia nula de los datos. Por lo tanto, la elección óptima de k es aquel valor que maximiza esta brecha.

Fuente: Elaboración propia, 2021. Basado en (Medium, 2019)

El funcionamiento detrás del algoritmo es bastante sencillo. Una vez elegida la cantidad de grupos a encontrar (k) se elige al azar un centroide inicial (coordenadas centrales) para cada grupo para luego aplicar un proceso en dos pasos:

- Paso de asignación: asigna cada observación a su centro más cercano.
- Paso de actualización: actualiza los centroides como el centro de su observación respectiva.

Estos dos pasos se repiten una y otra vez hasta que no haya más cambios en los grupos, es decir, estos sean lo más distintos posible. En este punto, el algoritmo ha convergido y podemos determinar los agrupamientos finales (Jeffares, 2019).

En la Figura 6 se muestra a manera gráfica el funcionamiento del algoritmo k-means, desde la designación de los centroides iniciales y como estos se van actualizando hasta llegar a la agrupación óptima, cuando se considera por defecto la generación de 2 agrupaciones.

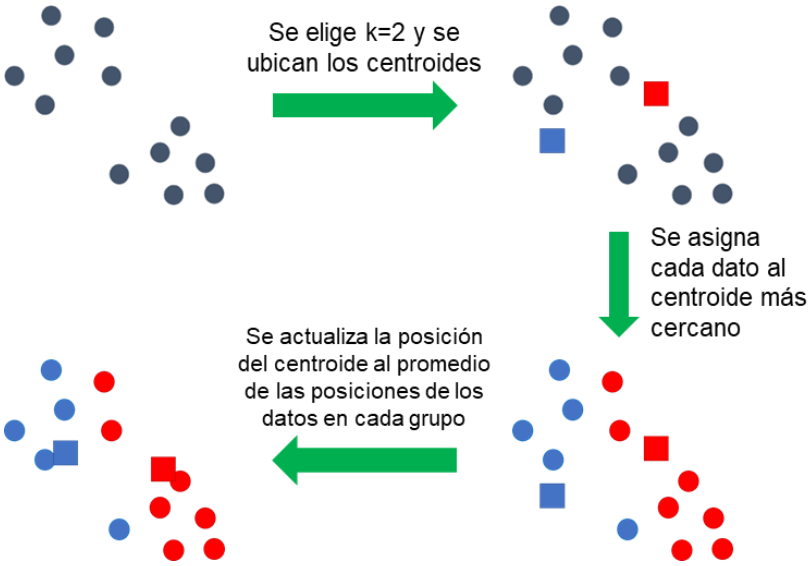

Figura 6. Funcionamiento del algoritmo k-means cuando se seleccionan 2 grupos.

Fuente: Elaboración propia, 2020. Adaptado de IArtificial.net

III. MARCO METODOLÓGICO

Como se mencionó anteriormente, este trabajo tuvo como objetivo generar un nuevo criterio de segmentación para los lectores digitales registrados en un sitio web informativo a través del algoritmo de análisis clúster k-means utilizando no solo los datos de navegación como la frecuencia de sesiones o el tipo de contenido sino también integrando otras fuentes con las que se cuenta en la empresa como por ejemplo la de suscripciones al diario impreso. Para esta aplicación se consideraron todos los datos de navegación web de tres meses del año (de julio a septiembre de 2018).

Por un lado, la información de navegación web del total de lectores fue capturada mediante el tag de seguimiento de Google Analytics insertado en el portal informativo; como la empresa cuenta con la versión de pago, Analytics 360, estos datos recolectados viajaban diariamente de forma automática al repositorio de BigQuery creado en Google Cloud Platform (GCP).

Sin embargo, este almacenamiento en GCP tenía un costo y debido a que se generaba una gran cantidad de datos mensualmente es que toda esta información fue copiada y guardada en una base de datos de navegación web dentro del data warehouse general de la compañía, el cual utilizaba el servicio de Amazon Web Services (AWS), aprovechando que se contaba una licencia de mayor capacidad.

Por otro lado, cuando un lector navegaba por el sitio web y llegaba a revisar más de 50 notas en un mes, le aparecía una ventana emergente (pop-up) con un mensaje indicándole que debía registrarse para continuar leyendo, con tres opciones de registro: usando su cuenta Gmail, su cuenta de Facebook o ingresando sus datos en un formulario (nombre, apellido, correo electrónico, contraseña, tipo de documento, estado civil, entre otros). Todos estos datos viajaban a la base de lectores registrados las cuales también se almacenaban en el data warehouse de la compañía en AWS. Hay que tener cuenta que cuando un usuario se registraba por primera vez solo le pide su correo y contraseña, luego debía completar los demás campos.

Una vez que un usuario completaba el registro, se le asignaba un código de registro, el cual capturaba Google Analytics y era asociado a su navegación siempre y cuando navegue por el sitio web habiendo iniciado sesión en su cuenta creada. Este código fue el nexo entre ambas fuentes de datos.

Con el fin de obtener mayor detalle de los lectores digitales registrados, se cruzaron estos datos con algunos no digitales con los que contaba la compañía, como los de las suscripciones al diario impreso o de concursos, y también se complementaron con algunos datos de terceros obteniendo un mejor detalle del usuario registrado. Todos estos datos se integraron con los digitales y así lograr un mayor porcentaje de completitud de todos los campos.

Un esquema a modo de resumen sobre las distintas fuentes de datos y el proceso de integración y desarrollo del análisis ejecutado en el software R se muestra en la Figura 7.

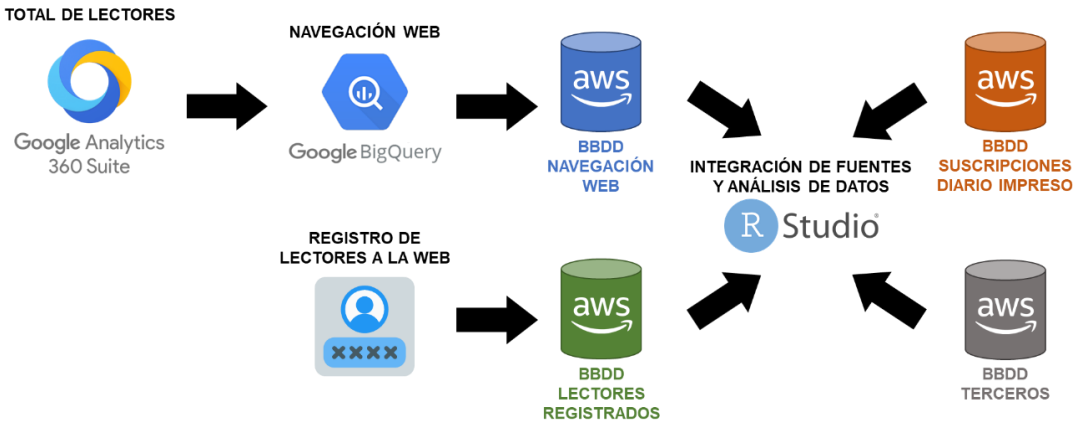

Figura 7. Proceso de integración de datos en RStudio para su posterior análisis.
Fuente: Elaboración propia, 2020.

El conjunto de datos de navegación web, obtenido desde Google BigQuery, empleado en el proyecto constaba de 20 millones de registros; donde cada fila hacía referencia a una página vista (un hit para cada acción de vista de página para un lector). Esta tabla contenía los datos del dispositivo de navegación, las secciones y url de notas leídas, los id de las visitas (con fecha y hora), la fuente de origen de la visita (por ejemplo, desde el buscador de Google o vía redes sociales), entre otros. Un breve esquema de la composición de esta base de datos se observa en la Tabla 5.

El trabajo de integración de todas estas fuentes de datos se ejecutó en el software libre R, específicamente en su interfaz de trabajo RStudio, el cual es una aplicación desarrollada para análisis estadístico que posee muchas librerías y funciones lo cual permitió realizar todo el procesamiento de la información recopilada.

Tabla 5. Base de datos de navegación web obtenida con Google BigQuery

user_id	fullvisitorId	visitId	pagepath	seccion	device	channel_ grouping
ID001	12352463	20200405132 743	/politica/congre so-envio-…	politica	tablet	Organic Search
ID002	23421459	20200405122 743	/	portada	desktop	Direct
ID003	22327531	20200405122 809	/futbol- peruano/uni…	deportes	mobile	Social

Fuente: Elaboración propia. 2020. Muestra de los datos exportados

El tratamiento y preparación de la base de datos final, donde se integraron todas las fuentes de datos, se realizó con las funciones join y group by de la librería dplyr. El campo que sirvió de llave entre la base de tráfico web y la del registro fue el user_id; y para conectar estos con la información no digital (suscripciones diario impreso y fuentes de terceros) se utilizaron los campos email e incluso el dni de los usuarios.

Antes de ejecutar la segmentación de lectores digitales con el algoritmo k-means, se filtraron a los usuarios que solo consumieron contenido, es decir, a aquellos que solo accedieron a notas y se excluyeron a los que accedieron solamente a la portada y a otras páginas que no tenían que ver contenido regular (por ejemplo, espacios patrocinados). Además, también se filtraron a aquellos lectores que visitaron menos de 5 notas en el periodo ya que representaba a aquellos visitantes casuales, es decir, a los que tenían nula interacción con el sitio.

De los 100,000 lectores registrados totales, se han considerado finalmente a 19,375 usuarios que en conjunto han consumido casi 4,000,000 de notas en los tres meses de estudio (un promedio mensual de 60 notas por lector).

Este conjunto de datos final, donde cada registro hace referencia a un lector único registrado e identificado que ha visitado el sitio web dentro de los tres meses de estudio, constó de 18 mil observaciones y de 25 variables, cuyo detalle se muestra en Tabla 6.

Tabla 6. Descripción de las variables del conjunto de datos de lectores digitales

Nombre	Tipo de variable	Descripción
user_id	cualitativa	Identificador del lector registrado (código generado)
nombre	cualitativa	Nombre(s) del lector registrado.
apellidos	cualitativa	Apellido(s) del lector registrado.
tipo_doc	cualitativa	Tipo de documento de identidad: DNI, CEX: Carnet de extranjería, PAS: pasaporte.
nro_doc	cualitativa	Número de documento de identidad (alfanumérico).
email	cualitativa	Email del lector registrado.
género	cualitativa	Género del lector registrado: M; F.
edad	cuantitativa	Edad del lector; calculado con la fecha de nacimiento.
estado_civil	cualitativa	Estado civil del lector registrado: S: soltero; C: casado; V: viudo; D: divorciado
sesiones	cuantitativa	Cantidad de sesiones (visitas) totales realizadas por el lector en el periodo seleccionado.
pag_vistas	cuantitativa	Cantidad de páginas vistas totales consumidas por el lector en el periodo seleccionado.
pais	cualitativa	País de conexión del lector registrado; obtenido mediante su IP o geolocalización al conectarse.
desktop	cuantitativa	Cantidad de páginas vistas consumidas por el lector desde una computadora u ordenador.
mobile	cuantitativa	Cantidad de páginas vistas consumidas por el lector desde un dispositivo móvil.
actualidad	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección actualidad.
deportes	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección deportes.
economia	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección economía.
espectaculos	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección espectáculos y moda.
mundo	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección mundo.
nacional	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección nacional.
politica	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección política.
redes-sociales	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección redes sociales.
tecnologia	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección tecnología.
viaje	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en la sección viaje y turismo.
otros	cuantitativa	Cantidad de páginas vistas (notas) consumidas por el lector en otras secciones más específicas.

Fuente: Elaboración propia, 2018

De todas las variables detalladas anteriormente, para este proyecto se emplearon solo aquellas referentes al consumo de páginas vistas en notas de contenido de las secciones de la web como actualidad, deportes, economía, espectáculos, mundo, nacional, política, redes sociales, tecnología, viajes y otros como variables predictoras (11) ya que el objetivo del mismo fue pronosticar grupos de lectores digitales utilizando el algoritmo k-means, que utiliza variables cuantitativas para determinar una respuesta a posteriori. En la Tabla 7 se muestra la estructura de la base de datos sobre la que se ejecutó el agrupamiento por k-means.

Tabla 7. Base de datos final para aplicar algoritmo k-means.

user_id	actualidad	deportes	economia	…	viajes	otros
ID001	5	10	0	…	7	12
ID002	7	20	6	…	0	5
ID003	28	13	1	….	5	30

Fuente: Elaboración propia. 2018. Muestra de los datos empleados

Para ejecutar este agrupamiento mediante el algoritmo k-means era necesario definir el número de grupos que se iban a considerar, para esto nos apoyamos, de aquí en adelante, en el software R, empleando la función NbClust, dentro del paquete del mismo nombre, ya que proporciona 30 índices para determinar el número óptimo de agrupaciones y propone al usuario el mejor esquema a partir de los diferentes resultados obtenidos al variar todas las combinaciones de número de clústeres, medidas de distancia y métodos de agrupamiento como el del codo, de silueta o de brecha (RDocumentation, s.f.). Esta función además brinda como resultado dos gráficos: el índice D que busca un codo significativo (el pico significativo en el gráfico de segundas diferencias del índice D) que corresponde a un aumento significativo del valor de la medida.

Luego de ejecutar la función NbClust para k-means se obtiene un gráfico donde podemos observar que para este caso se tiene que trabajar con 6 segmentos de lectores digitales, ya que es el número óptimo de grupos determinado luego de haber comparado internamente distintos métodos e índices de agrupamiento, 8 de los 30 índices validaron este resultado; además, para números mayores de clústeres se obtenían grupos con muy pocos individuos (hasta de un solo lector en algunos casos). El detalle del gráfico de los índices D entregado por la función se muestra en la Figura 8.

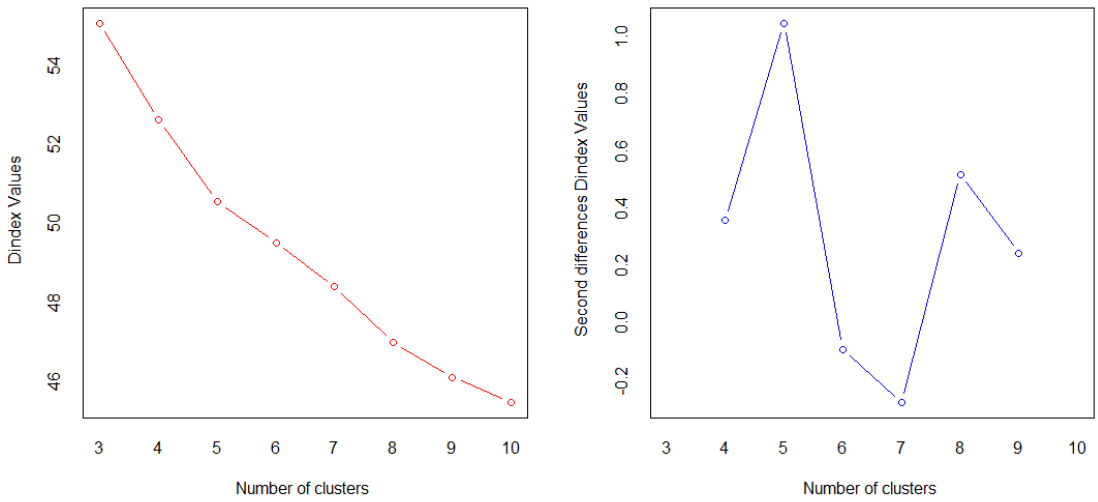

Figura 8. Gráficos del índice D obtenido con la función NbClust.
El gráfico de segundas diferencias del índice D determinó que 6 es el número óptimo de clústers para el proyecto, luego de comparar los resultados de 30 índices.
Fuente: Elaboración propia con el software R, 2018.

Ya determinado que se tuvo que trabajar con 6 grupos ejecutamos el agrupamiento con el método k-means en R, utilizando la función kmeans contenida en el paquete stats, que viene instalado por defecto en el software. Este comando requiere de ciertos parámetros como el conjunto de datos (x) y la cantidad de grupos a considerar (centers) (RDocumentation, s.f.).

Antes de su ejecución, se tuvo que estandarizar el conjunto de datos, ya que k-means trabaja con distancias. La estandarización es un paso central del preprocesamiento para que los datos sean limpios, libres de ruido y consistentes. La normalización de datos estandariza los datos sin procesar convirtiéndolos en un rango específico mediante una transformación lineal que puede generar agrupaciones de buena calidad y mejorar la precisión de los algoritmos de agrupación (Mohamad & Usman, 2013).

Para este caso se estandarizó el conjunto de datos (de 11 variables referentes al consumo de notas por secciones de la web) con la función scale, dentro del paquete base en R, que nos ayudó con esta tarea. Una vez estandarizados los datos se procedió a la ejecución de la segmentación de los lectores digitales en 6 grupos según el contenido leído utilizando k-means.

IV. RESULTADOS Y DISCUSIÓN

Luego de haber ejecutado el agrupamiento de los lectores digitales en 6 grupos utilizando el algoritmo k-means sobre el conjunto de datos del consumo del tipo de contenido, se obtuvieron segmentos con distinta cantidad de usuarios, unos muy grandes y otros muy pequeños. A nivel descriptivo, podemos observar algunas diferencias tanto sobre la composición de las agrupaciones como del promedio de notas que revisan al mes en Tabla 8.

Tabla 8. Análisis descriptivo de los segmentos de lectores obtenido por k-means

Segmentos	Lectores	% Lectores	Promedio de notas mensuales	% Consumo
1	397	2.1%	7.6	4.8%
2	221	1.1%	13.8	8.8%
3	45	0.2%	20.6	13.1%
4	16,821	86.8%	78.0	49.5%
5	420	2.2%	15.4	9.8%
6	1,471	7.6%	22.2	14.1%

Fuente: Elaboración propia. 2018.

Si bien, se tuvo un segmento con muchos lectores (16,821) y otro con muy pocos (45), el también analizar el promedio de notas mensuales por grupo, indistintamente del tipo de contenido consumido, permitió obtener algunos hallazgos:

- El grupo 4 es el que está compuesto por la mayor cantidad de lectores (16,821, que representa el 86.8% del total) y también el que consume en promedio la mayor cantidad de notas mensuales (78.0).
- El grupo 3, a diferencia del 4, es aquel compuesto por la menor cantidad de lectores registrados (solo 45); sin embargo, consumen en promedio 20.6 notas mensuales, consumo similar al del grupo 6, que es el segundo más grande (1,471 lectores que consumen en promedio 22.2 notas al mes).

- Si bien el grupo 1 y el grupo 5 están compuestos por cantidades similares de lectores (397 y 420), existe una diferencia respecto al consumo de notas, ya que en el primero se visitan en promedio la mitad de notas que en el quinto al mes (7.6 y 15.4).
- A su vez, el grupo 5 posee un consumo promedio mensual de notas similar al grupo 2 (15.4 y 13.8); sin embargo, el quinto está compuesto por el doble de la cantidad de lectores que el segundo (420 y 221 respectivamente).

Desde esta primera vista ya podemos notar algunas diferencias entre estas agrupaciones, pero no son del todo clasificables entre sí, por esta razón se analizó en segundo lugar el porcentaje de consumo de notas según las secciones analizadas, esto con la finalidad de obtener un mayor detalle sobre los hábitos de los lectores dentro de cada uno de los seis segmentos encontrados. Este detalle se muestra en Tabla 9.

Tabla 9. Porcentaje de consumo de notas por secciones según segmentos de lectores

Sección	Segmentos					
	1	**2**	**3**	**4**	**5**	**6**
Actualidad	9.8%	4.1%	2.1%	8.0%	2.9%	7.7%
Deportes	2.7%	4.3%	3.7%	**14.9%**	**70.7%**	8.9%
Economía	6.1%	1.4%	0.8%	2.6%	1.5%	2.8%
Espectáculos	**21.8%**	**64.5%**	**14.8%**	**19.3%**	7.8%	8.9%
Mundo	**17.8%**	7.8%	5.8%	**15.4%**	4.9%	**10.3%**
Nacional	9.4%	3.2%	2.1%	5.7%	2.5%	6.4%
Política	1.4%	2.6%	1.7%	**18.9%**	2.9%	**44.0%**
Redes Sociales	0.2%	1.1%	**62.6%**	2.4%	0.6%	0.8%
Tecnología	6.9%	3.3%	2.4%	4.8%	2.4%	3.5%
Viaje	**10.5%**	2.8%	1.3%	2.7%	1.4%	1.6%
Otros	**13.5%**	5.0%	2.6%	5.3%	2.2%	5.0%

Fuente: Elaboración propia. 2018.

Algunos de los hallazgos encontrados luego de analizar el porcentaje de consumo de notas por secciones para cada uno de los seis segmentos de lectores encontrados fueron:

- El segmento 1 presenta una participación similar para la mayoría de secciones, destacando las notas de espectáculos y moda, mundo (noticias internacionales), viajes y otros temas (que representan el 63.5% del total de consumo del grupo). Estos lectores son afines a contenido útil, aquel que no pierde valor en el tiempo.
- Para el caso de los segmentos 2, 3, 5 y 6 se observa un mayor porcentaje de consumo de notas en las secciones espectáculos, redes sociales, deportes y política (con 64.5%, 62.6%, 70.7% y 44.0% del total de notas respectivamente), lo que ya nos da una idea del tipo de contenido más afín a estos tres grupos.
- En el segmento 4 podemos observar una mayor participación en las notas de las secciones de espectáculos, política, mundo y deportes (un 68.5% del consumo total del grupo). Ellos muestras mayor interés en temas informativos y coyunturales.

De manera adicional, y como apoyo en la etapa de la definición y caracterización de los segmentos se emplearon las otras variables que se dejaron de lado para el cálculo el algoritmo k-means, como las sociodemográficas (género, edad o estado civil) y las de comportamiento en la web (sesiones, país o dispositivo de conexión).

Respecto a los datos sociodemográficos, que podemos visualizar en la Tabla 10, hay que tener en cuenta que estos no estuvieron disponibles para el total de lectores, ya que no todos suelen brindar este tipo de información considerada como sensible; por lo que los porcentajes de usuarios según estos criterios están basados en el porcentaje de completitud de cada uno de las variables (Entre 70% y 75% para género y edad; y de menos del 40% para el estado civil). Además, la edad se trabajo con rangos, los cuales son los trabajados por muchas de las plataformas digitales existentes.

Para la información de comportamiento web, a diferencia de lo sociodemográfico, si se cuenta con la información del total de lectores, ya que estos datos fueron obtenidos de la misma forma que las secciones, mediante Google Analytics y Google BigQuery. En se observan el promedio de sesiones mensuales que realizan los usuarios en cada segmento, y el porcentaje de notas que fueron consumidas desde un ordenador o laptop (desktop) y un dispositivo móvil (mobile); o si se conectaron desde Perú o de algún lugar en el extranjero.

Tabla 10. Porcentaje de lectores digitales según características sociodemográficas

Características sociodemográficas		Segmentos					
		1	**2**	**3**	**4**	**5**	**6**
Género	Femenino	30.8%	73.3%	49.3%	36.6%	20.0%	38.9%
	Masculino	69.2%	26.7%	50.7%	63.4%	80.0%	61.1%
Edad	0 - 17	1.9%	0.0%	0.0%	0.5%	0.5%	0.2%
	18 - 24	7.5%	4.0%	21.6%	2.5%	2.3%	4.9%
	25 - 34	11.3%	14.4%	33.3%	19.6%	26.1%	10.5%
	35 - 44	29.6%	30.8%	29.2%	25.4%	27.5%	15.1%
	45 - 54	39.9%	42.0%	12.1%	35.5%	27.1%	47.2%
	55 - 64	7.0%	6.4%	3.8%	11.3%	13.7%	14.9%
	65 a más	2.8%	2.4%	0.0%	5.2%	2.8%	7.2%
Estado Civil	Soltero	54.6%	58.4%	70.6%	54.2%	58.7%	50.0%
	Casado	38.7%	39.0%	29.4%	40.7%	36.5%	42.6%
	Viudo	0.8%	0.0%	0.0%	0.6%	1.9%	1.2%
	Divorciado	5.9%	2.6%	0.0%	4.4%	2.9%	6.2%

Fuente: Elaboración propia. 2018

Tabla 11. Sesiones y porcentaje de lectores por segmento según el comportamiento web

Comportamiento Web		Segmentos					
		1	**2**	**3**	**4**	**5**	**6**
Sesiones mensuales prom.		2.2	3.6	5.0	18.8	5.1	7.1
Dispositivo de conexión	desktop	70.0%	62.8%	85.4%	62.0%	60.8%	61.7%
	mobile	30.0%	37.2%	14.6%	38.0%	39.2%	38.3%
País de conexión	Perú	81.1%	80.1%	73.3%	85.8%	67.1%	80.7%
	Extranjero	18.9%	19.9%	26.7%	14.2%	32.9%	19.3%

Fuente: Elaboración propia. 2018

Luego de analizar estas otras variables como complemento, se determinaron más detalles que nos permiten diferenciar aún más cada uno de los seis segmentos de lectores encontrados, los cuales fueron:

- Con respecto al género, se puede observar que, a pesar de contar en la base con una mayor proporción de lectores masculinos identificados, hay 2 grupos totalmente diferentes: el segundo segmento es en su mayoría femenino (73.3%) y el quinto está compuesto por un 80% de varones.
- A nivel de los rangos etarios, hay que destacar a dos segmentos distintos al resto: el tercero está compuesto por gente joven en su mayoría, de entre 18 y 44 años (84.1% en total del grupo) y el sexto grupo compuesto por más adultos, de entre 35 y 65 años (un 77.2% en conjunto).
- Por parte del estado civil, la base de datos está compuesta en su mayoría por solteros, sin embargo, para el tercer segmento esta diferencia es mayor, ya que consta de un 70.6% de lectores con esta condición.
- Para las sesiones mensuales promedio se observa un comportamiento similar a la del consumo de notas, en términos de proporción. Los lectores que más visitan el sitio web son los que conforman el segmento 4, con un promedio de 18.8 sesiones mensuales (podría decirse que realizan visitas interdiarias en el mes).
- La mayor parte de los lectores registrados se conecta desde un dispositivo de escritorio, como una computadora o una laptop. Esta característica es transversal a todos los segmentos.
- Finalmente, respecto al país desde donde se conectan los lectores, obtenidos mediante el IP de los dispositivos, se observa que estos son principalmente nacionales (más del 80% accede desde Perú); sin embargo, para el quinto segmento esta diferencia es menor, ya que contiene a un 32.9% de lectores que se conectan desde el extranjero.

Ya teniendo todo este detalle, resultó mucho más sencillo caracterizar estos seis segmentos, basándonos no solo en el comportamiento web y el tipo de consumo que realizan los lectores, sino también apoyados en las variables sociodemográficas para perfilar con más detalle estas agrupaciones.

Definimos a nuestros seis segmentos bajo los nombres de utilitarios, faranduleros, modernos, informados, futboleros y politiqueros respectivamente; ya que describen en una sola palabra las características de las agrupaciones. Este detalle se explica en Tabla 12.

Tabla 12. Perfil de lectores registrados según características de los segmentos encontrados

Segmento	Nombre	Definición / Características
1	Utilitarios	• Son lectores de entre 25 y 54 años, solteros y casados. • Realizan 2 visitas y consumen 8 notas al mes. • Conectados principalmente desde Perú. • Interesados en temas utilitarios, aquellos que no pierden valor en el tiempo, de las secciones de espectáculos, mundo, turismo y las otras secciones de menor tráfico.
2	Faranduleros	• En su mayoría lectoras mujeres de entre 25 y 54 años, solteras y casadas. • Conectados principalmente desde Perú. • Realizan 4 visitas y consumen 14 notas al mes. • Interesados en notas de espectáculos, farándula local, moda y cine mayormente.
3	Modernos	• Son lectores jóvenes, de entre 18 y 44 años, solteros. • Realizan 5 visitas y consumen 21 notas al mes. • Conectados también desde el extranjero (26.7%) • Interesados en temas de tendencia por su consumo de notas en las secciones de redes sociales y espectáculos.
4	Informados	• Lectores de entre 25 y 54 años, solteros y casados. • Realizan 19 visitas y consumen 78 notas mensuales, es decir, realizan visitas interdiarias en el mes. • Conectados principalmente desde Perú • Interesados en temas de coyuntura nacional e internacional porque consumen notas de política, deportes, espectáculos, mundo y actualidad.
5	Futboleros	• En su mayoría lectores varones, de entre 25 y 54 años, solteros y casados. • Realizan 5 visitas y consumen 15 notas mensuales, es decir, en cada sesión consumen 3 noticias al mes. • Conectados también desde el extranjero (32.9%) • Interesados principalmente en notas de deportes, relacionadas al futbol tanto local como internacional.
6	Politiqueros	• Lectores adultos mayores, de entre 35 y 64 años, solteros y casados. • Realizan 7 visitas y consumen 22 notas mensuales. • Conectados principalmente desde Perú • Interesados en notas de política, mundo y opinión.

Fuente: Elaboración propia. 2018

Luego de obtenidos los seis segmentos de los lectores digitales registrados que visitan el sitio web informativo, se decidió conocer cuántos de estos usuarios tienen una suscripción al diario impreso, para tomar algunas acciones futuros sobre estos. En Tabla 13 se muestra estos porcentajes respecto a cada una de las agrupaciones.

Tabla 13. Porcentaje de lectores digitales que cuentan con suscripción al diario impreso

Segmentos	Nombre	Lectores	Lectores con suscripción print	% Lectores con suscripción
1	Utilitarios	397	47	11.8%
2	Faranduleros	221	20	9.0%
3	Modernos	45	7	15.6%
4	Informados	16,821	786	4.7%
5	Futboleros	420	31	7.4%
6	Politiqueros	1,471	99	6.7%

Fuente: Elaboración propia. 2018.

Dentro de los principales hallazgos de este análisis se encontró:

- Los suscriptores al diario impreso son alrededor de 85,000; sin embargo, de estos solo 990 cuentan también con una cuenta de acceso para acceder al contenido del sitio web, es decir, solo el 1.2% del total.

V. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

Como parte de la ejecución del siguiente proyecto se puede concluir que, para los lectores digitales registrados al sitio web informativo, se encontraron seis segmentos de usuarios a través de la técnicas de análisis clúster llamada k-means, los cuales se basan en variables tanto digitales como sociodemográficas, como el tipo de contenido al que acceden según las secciones, las visitas mensuales, el país y el dispositivo desde el que se conectan, así como el género, rango de edad y estado civil de los mismos. Este procedimiento y toda la etapa de preprocesamiento de datos se ejecutó utilizando el software R mediante la función kmeans del paquete stats. Las agrupaciones fueron nombradas como utilitarios, faranduleros, modernos, informados, futboleros y politiqueros.

Estos segmentos diferenciados permitirán a la compañía ofrecerlos como un nuevo producto digital comercial para los clientes en términos de publicidad, los cuales ayudarían a incrementar los ingresos que se perciben aprovechando un mercado no explorado, el de la publicidad segmentada respaldado en el análisis de los datos del sitio web obtenidos con Google Analytics y Google BigQuery, además de los datos de tercero para complementar los perfiles. Esto también será una ventaja para los consumidores ya que podrán especificar el grupo de usuarios al que quieren alcanzar, mejorando significativamente los resultados que se obtendrían a diferencia del método tradicional de publicidad digital.

Para el caso del segmento de los informados, se encuentra que hay un gran porcentaje de lectores que no cuentan con una suscripción al diario impreso, por lo que se aplicarían acciones de mailing (envío de correos electrónicos) con los beneficios que podrían obtener de contar con ambos productos, como vales de consumo y descuento en restaurantes, tiendas por departamento y demás. Del mismo modo para aquellos suscriptores al diario impreso que no cuentan con una cuenta en el sitio web, se les enviaría junto con sus entregas de periódicos, afiches publicitarios que indiquen los beneficios de contar con dicha suscripción o algún descuento para convencer a los lectores.

5.2. Recomendaciones

Hay que tener en cuenta como se clasificarían los lectores no registrados según los segmentos encontrados, ya que a diferencia de los 19,375 registrados, estos son más de 2.5 millones de usuarios. Estos hallazgos permitirían establecer diferencias en la conducta de consumo de contenido de ambos tipos de usuarios y ayudaría a complementar este análisis, con el fin de impactar a posibles personas que navegan de forma anónima y que se registren en la web para que no tengas límites al visitar las notas del sitio web.

También se podrían considerar para una segunda etapa del proyecto analizar otras variables digitales como, por ejemplo, la fuente de tráfico desde donde se originan las visitas. Como se sabe, cuando una persona accede a una nota del sitio web lo puede hacer a través de distintos canales o fuentes como la búsqueda orgánica (en buscadores como Google), las redes sociales (desde Facebook o Instagram), las notas referidas (que provienen de otras páginas) o si acceden de forma directa (digitando la URL). También se podrían considerar hábitos de consumo digital como las horas en las que se conectan, los días de la semana de mayor actividad, entre otras.

Finalmente, hay que recordar que antes de la ejecución de la segmentación de lectores registrados, se dejó de lado a los usuarios casuales, aquellos que visitaron en una sola oportunidad el sitio y consumieron menos de 5 notas en promedio al mes, es decir, aquellos que solo ingresaron de forma esporádica al sitio web informativo ya que tenían nula interacción en él. Este grupo podría tratarse como un séptimo clúster sobre el cual se podría analizar su conducta y nos permitiría complementar los hallazgos de este proyecto.

VI. REFERENCIAS BIBLIOGRÁFICAS

Adams Harding, A., & Gingras, R. (2018). *Google News Initiative.* Obtenido de News Consumer Insights Playbook: https://newsinitiative.withgoogle.com/training/states/consumer_insights/pdfs/gni-new-consumer-insights-playbook.pdf

Akhtar, A. (Setiembre de 2019). *MonsterInsights*. Obtenido de How Does Google Analytics Work? (Complete Beginner's Guide): https://www.monsterinsights.com/how-does-google-analytics-work-beginners-guide/

Ayuda de Google Analytics. (s.f.). Obtenido de Esquema de BigQuery Export: https://support.google.com/analytics/answer/3437719?hl=es

Čegan, L., & Filip, P. (2017). Webalyt: Open Web Analytics Platform . *27th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 1.

DBi Data Business Intelligence - Havas. (2019). Obtenido de Google Analytics: ¿Y tú qué necesitas? ¿la versión gratuita o 360?: https://dbibyhavas.io/es/blog/google-analytics-y-tu-que-necesitas-la-version-gratuita-o-360/

First Site Guide. (2021). Obtenido de The Best Web Analytics Tools 2021: https://firstsiteguide.com/best-website-analytics-tools/

Google Analytics Developers. (2019). Obtenido de Enviar datos a Google Analytics: https://developers.google.com/analytics/devguides/collection/analyticsjs/sending-hits?hl=es-419

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R.* California: Springer.

Jeffares, A. (Noviembre de 2019). *Towards Data Science*. Obtenido de K-means: A Complete Introduction: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c

Kladnik, M., Stopar, L., Fortuna, B., & Mladenić, D. (2017). Audience Segmentation Based on Topic Profiles. *Jožef Stefan Institute and Jožef Stefan International Postgraduate School*, 1.

Lopez, G., Seaton, D. T., Ang, A., Tingley, D., & Chuang, I. (2017). Google BigQuery for Education: Framework for Parsing and Analyzing edX MOOC Data. *L@S '17: Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*.

Marketing Analítico. (2019). Obtenido de Usuario, Sesión y Visitas a Páginas en Google Analytics: https://www.marketing-analitico.com/analitica-web/usuario-sesion-visitas-a-paginas-en-google-analytics

Medium. (2019). Obtenido de Get the Optimal K in K-Means Clustering: https://medium.com/towards-artificial-intelligence/get-the-optimal-k-in-k-means-clustering-d45b5b8a4315

Mohamad, I. B., & Usman, D. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology 6*, 3299-3300.

Qualtrics. (2020). Obtenido de What is cluster analysis? When should you use it for your survey results?: https://www.qualtrics.com/experience-management/research/cluster-analysis/

RDocumentation. (s.f.). Obtenido de kmeans: K-Means Clustering: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans

RDocumentation. (s.f.). Obtenido de NbClust Package For Determining The Best Number Of Clusters: https://www.rdocumentation.org/packages/NbClust/versions/3.0/topics/NbClust

Skrba, A. (Agosto de 2020). *First Site Guide*. Obtenido de The Best Website Analytics Tools 2020: https://firstsiteguide.com/tools/analytics/

Sponder, M., & Khan, G. F. (2018). *Digital Analytics for Marketing.* New York: Routledge.

Syakur, M. A., Khotimah, B. K., Rochman, E. M., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, 1.

Thakur, D. (Julio de 2017). *Medium*. Obtenido de 10 Good Reasons Why You Should Use Google Analytics: https://medium.com/@dineshsem/10-good-reasons-why-you-should-use-google-analytics-699f10194834