Natura@economía 6(1): 60-77 (2021)

Natura@economía

ISSN 2226-9479 (Versión electrónica)
Website: http://revistas.lamolina.edu.pe/index.php/neu

ARTÍCULO ORIGINAL

Caracterización de los lectores registrados de una web informativa con el algoritmo K-Means para incrementar las ventas

Characterization of the registered readers of an informative website with the K-Means algorithm to increase sales

Iván Soto Rodríguez[1*]; Brian Erick Clemente Rivera[2]

[1]Universidad Nacional Agraria La Molina, Lima, Perú. ID de ORCID: 0000-0002-4928-8362, ivans@lamolina.edu.pe

[2]Universidad Nacional Agraria La Molina, Lima, Perú. ID de ORCID: 0000-0002-8491-8657, brianerickcr@gmail.com

Recepción: 11/5/2021; Aceptación: 16/6/2021

Resumen

El presente artículo abordó el desarrollo de un nuevo método de caracterización para los lectores registrados de un sitio web de noticias basado en su comportamiento de uso y temasde interés que leen con frecuencia para crear productos adicionales y encontrar oportunidades comerciales en una compañía peruana de medios de comunicación. Se utilizaron variablesdel entorno digital como el tipo de notas a la que acceden según la sección en la que estasse encuentran alojadas, complementándose con datos no digitales como la información sociodemográfica (edad, género), de ubicación (país), entre otras que disponibles en la empresa.Se han considerado la información de 19,375 lectores correspondientes a un periodo de tres meses de análisis, y fueron obtenidos mediante las herramientas de Google, Analytics 360 y BigQuery. Para la generación de las agrupaciones se empleó el algoritmo k-means, un método de análisis clúster no jerárquico y una técnica de aprendizaje no supervisado; además, todos los cálculos computacionales, así como la construcción del conjunto de datos final se efectuaron con el software R, que contiene múltiples funciones que facilitaron estas actividades. Como conclusión se obtuvieron y definieron seis segmentos, los cuales permitieron ofrecer una ventajacompetitiva a los clientes en términos de publicidad ya que podrán seleccionar la audiencia

Forma de referenciar el artículo: Rodríguez, I. y Clemente, B. (2021). Caracterización de los lectores registrados de una web informativa con el algoritmo K-Means para incrementar las ventas, Natura@economía, 6(1), 60-77. http://dx.doi.org/10.21704/ne.v6i1.1731

DOI: http://dx.doi.org/10.21704/ne.v6i1.1731
* Autor de correspondencia: Soto, I. Email: ivans@lamolina.edu.pe

específica a la que quieran impactar, mejorando en gran medida los resultados que se obtendrían comparado con las estrategias digitales convencionales; además, se realizarán acciones para incrementar el número de suscriptores con el fin de repotenciar estos grupos e incrementar los ingresos obtenidos anualmente.
Palabras claves: caracterización, audiencia digital, k-means, analítica web, google analytics

Abstract

This article was concerned with the development of a new characterization method for registered users in a news and information website based on their usage behaviour and topics of interest they usually read in order to create additional products and find business oportunities for an important peruvian media company. Some digital features were used such as kind of content readers access depending on the name of the section which the note was hosted, complementing with non-digital data like sociodemographic (gender ora ge), location (country), among others that are available in the enterprise. Data from 19,375 registered users were considered which corresponding to the web activity of three months, and were obtained through Google digital tools, Analytics 360 and Google BigQuery. The k-means algorithm, a non-hierarchical cluster analysis method and an unsupervised learning technique were used to generate the clusters; in addition, all computational calculations, as well as the construction of the final data set, were performed with R software, which contains multiple functions that facilitated these activities. In conclusion, six groups were found and defined which allow to offer a competitive advantage to clients in terms of advertising because they could select the target audience they want to reach, improving greatly the obtained results compared to traditional digital strategies; in addition to that, some actions will be taken to increase the number of subscribers in order to strengthen these groups and increase annual revenues.
Keywords: clustering, digital audience, k-means, web analytics, google analytics

Introducción

La creación de sitios web como una alternativa de consumo de contenido para los lectores y como estrategia de captar un nuevo grupode audiencia fue un paso importante en el proceso de digitalización de la compañía, y luego se fueron integrando las redes sociales para apoyar el alcance y difusión de las noticias publicadas. Estas nuevas plataformas generan mucha información como la cantidad de visitantes que poseen, el tiempo que estos permanecen leyendo notas o el número de interacciones que las publicaciones reciben, etc.

Para analizar el desempeño de estos entornos digitales se elaboran y entregan informes mensuales sobre sus tendencias, los cuales sirven, por ejemplo, como sustento al momento de ofrecerlos a los clientes a nivel de ventas. Sin embargo, esta información no era suficiente para otras áreas en la compañía, por ejemplo, para la jefatura de prensa era importante conocer aún con más detalle asu audiencia respecto a su comportamientoy hábitos para poder clasificarlos, lo que permitirá no solo conocer mejor a la comunidad sino también podría ser aprovechada por la gerencia comercial para convertirla en una nueva propuesta comercial especializada para sus clientes publicitarios.

En un inicio, se emplearon los criterios del News Consumer Insight (NCI) de Google News Initiative, el cual analizaba los datos de la frecuencia de visitas de los usuariosdel sitio web a través de Google Analytics, herramienta de analítica digital con la que cuenta la compañía, y los clasificaba bajo

el concepto del embudo de participación de forma mensual (últimos 28 días del mes). Este criterio segmenta a los lectores como casuales, habituales e incondicionales (Adams Harding & Gingras, 2018) y se definen así:

- Lector casual: Aquellos que solo visitaron el sitio web en una oportunidad.
- Lector habitual: Los que visitaron el sitio web entre 2 y 15 veces.
- Lector incondicional: Quienes han visitado más de 15 veces el sitio web.

Esta clasificación es suficiente si se consideran a todos los lectores digitales como anónimos, ya que no podremos obtener mayor información que no sea del tráfico web como desde que tipo de dispositivo se conectan, que clase de notas leen, cuanto tiempo permanecen conectados, entre otros. Por lo que, un siguiente paso para conocer esta información a nivel persona, fue la implementación de los registros en el sitio, lo que permitirá integrar a los datos digitales con los no digitales, por ejemplo, la información sociodemográfica.

Ya implementado el plan de registros en el sitio web fue posible la integración de estos datos, para el cual los usuarios requieren de un correo electrónico de acceso. Esta variable fue el enlace entre el entorno digital con la información de otras fuentes de datos disponibles en la compañía como la de suscripciones a los newsletters, los concursos y premiaciones y también las adquiridasa través de proveedores externos. Así, se incrementa el nivel de detalle para cada lector registrado para generar agrupaciones más potentes.

Por lo tanto, el objetivo del estudio es desarrollar un nuevo criterio que permita caracterizar a lectores digitales del sitio web informativo en grupos diferenciados, apoyado en la recolección, integración y análisis de todas las fuentes de datos para la construcción de agrupaciones basadas en características, utilizando el algoritmo k-means de análisis clúster con el software estadístico R.

Analítica web

La Web Analytics Association define a la analítica web como el proceso de recopilación, medición, análisis y presentación de informes de datos de Internet para comprender y optimizar la usabilidad de una página web. Hay que tener en cuenta tanto los datos cuantitativos como los cualitativos al analizar el sitio web, ya que esto permite mejorar de forma continua la experiencia de usuario lo cual es un aspecto clave ya que conlleva a un eficiente cumplimiento de objetivos para una compañía (Bekavac & Garbin Praničević, 2015).

Una manera de poder disponibilizar todo ese gran volumen de información generado por los sitios web, que dan indicios sobre quienes componen su audiencia o como estos interactúan en ella, es utilizando herramientas de analítica web. Estas plataformas han sido desarrolladas teniendo como bases técnicas de seguimiento y algoritmos sofisticados para poder procesar las grandes cantidades dedatos que almacenan. El contar con alguna deellas permiten reconocer mejor que hacen los visitantes, identificar acciones que generen problemas en la usabilidad (cuellos de botella) o errores en el diseño y, además, puede medir el desempeño del sitio y supervisar la disponibilidad del sitio web e incluso podría recomendar contenido de afinidad para los usuarios (Čegan & Filip, 2017).

En la actualidad existen muchas herramientas de analítica web que ayudan con este proceso de recopilación y análisis de datos digitales, desde gratuitas hastade pago. Hotjar, empresa privada que crea soluciones de investigación y optimización para empresas web, en su informe “State of web analytics” presentó aquellas plataformas en las que más de 2,000 profesionales en el mundo confían para hacerle seguimiento asus sitios web, entre las que destacan Google Analytics, Adobe Analytics, Mixpanel, Matomo, HubSpot, entre otros.

Google Analytics

Producto de Google Marketing Platform desarrollado como solución de análitica web. Es la herramienta más popular y líder del mercado digital actualmente, ya que es utilizada en al menos 30 millones de sitios web en el mundo según BuiltWith. Además, más del 75% de profesionales la usa debido a que cuenta con una versión gratuita y es intuitiva de manejar.

Para comenzar a hacerle seguimiento al sitio web a través de Google Analytics (en adelante GA) se debe crear y configurar una cuenta, y ya con esto se puede generar un código de seguimiento o etiqueta en JavaScript, que debe ser insertado en el código fuente (HTML). Con esta integración correcta se podrá rastrear a los usuarios que lleguen, así como toda su actividad realizada dentro del sitio sin problemas. Cuando un visitante ingresaa la página web, GA insertará una cookie en su navegador y mediante él podrá recolectar toda la información de su actividad para luego visualizarla a través de distintos informes: en tiempo real, de audiencia, de adquisición, de comportamiento y conversiones; los cualesse emplean según los distintos objetivos que como negocio se planteen (Akhtar, 2019).

Estos informes de GA están elaborados en base a dimensiones y métricas, la cuales hacen referencia a las características cualitativas y cuantitativas de los datos respectivamente.
Dentro de las dimensiones que recopila GA se encuentran:

- Tiempo: Fecha de conexión, se puede separar en año, mes, día e incluso hora.
- Ubicación geográfica: Mediante IP o GPS, registra país, región o ciudad de conexión.
- Categoría de dispositivo: Desktop, mobile o tablet
- Fuente/medio: Desde donde se origina la visita al sitio web. Por ejemplo: directo, Facebook o Google.
- Atributos de página: Título de la nota, ruta de página (URL), etc.

Para entender mejor a las principales métricas que reporta GA nos apoyamos en las definiciones que el portal Marketing Analítico ha realizado sobre ellas:

- **Usuario:** Un código de usuario distinto detallado en la cookie de seguimiento (no una persona, ya que cada dispositivo y cada buscador generan distintos códigos). Cada vez que un usuario acceda nuevamente al sitio web utilizando el mismo navegador y dispositivo y no elimine sus cookies, GA reconoce que se trata del mismo (usuario recurrente); de no identificarlo, lo clasifica como nuevo.
- **Sesión:** Periodo de navegación en el sitio de un usuario. Empieza cuando sucede el primer contacto con la web y finaliza cuando ocurre alguna de estas tres situaciones: una inactividad de 30 minutos, si cambia la fuente de tráfico y si llega la medianoche porque GA mide datos de forma diaria.
- **Página Vista:** Simplemente la acción de carga de una página. Por ejemplo, si un usuario revisa 4 páginas distintas o visita la misma página 4 veces, GA considera que este usuario generó 4 páginas vistas, independiente de sus sesiones.
- **Duración media de la sesión:** Es el tiempo promedio de todas las visitas (en segundos) para la cantidad total de sesiones generadas por todos los usuarios.

Todo lo expuesto anteriormente está disponible tanto en la versión gratuita de GA como en Analytics 360 (la versión de pago). Una gran diferencia entre ambas versionesse observa al momento de la precisiónal exportar los datos; la versión estándar (gratuita) aplica muestreo por defecto en los informes más detallados, esto con el fin de generar ahorro en costo computacional; por el contrario, con Analytics 360 se dispone de los reportes sin muestreo basados en el 100% de la información y además, posee una conexión integrada con BigQuery, solución de Google Cloud, para trabajar con los datos a nivelde hit (transacciones) (DBi Data Business Intelligence - Havas, 2019).

Google BigQuery

Es un servicio de data warehouse de Google Cloud Platform (GCP); es decir, se trata de un almacén de datos en la nube que permite almacenar grandes volúmenes de datos y consultarlos de forma rápida mediante lenguaje SQL estándar. Debido a esto, BigQuery entrega resultados rápidamente aun tratándose de consultas sobre datos a nivel de terabytes ya que no requiere de la creación (ni de la especificación) de índices, por lo que cualquier campo puede ser consultadoen cuestión de segundos, comparado con otros sistemas como MongoDB e incluso los tradicionales como MySQL o Postgres, que requieren de campos con índices existentes para una ejecución rápida (Lopez, Seaton, Ang, Tingley, & Chuang, 2017).

Es posible obtener los datos a nivel de hit de una cuenta de Analytics 360 en Bigquery fácilmente, debido a que esta integración ya viene incluida en el servicio premium. Un hit hace referencia al envío de los datos de las acciones que se recopilan desde el fragmento de código de seguimiento a GA. Entre los principales tipos de hits encontramos al de tipo page (la carga de página en el sitio web), screenview, transaction, item, social, timing y evento (para medir acciones como clicks o descargas) (Google Analytics Developers, 2019).

Lo que finalmente se obtiene en BigQuery es una tabla, en formato base de datos, con filas y columnas que hacen referencias a los hits, métricas y dimensiones que se encuentran en GA, pero de forma distinta. Teniendo como base el esquema de BigQuery Export podemos homologar algunos de los principales campos a utilizar entre ambas plataformas, las cuales son:

- **fullvisitorId:** Un campo de tipo cadena, que muestra el código que GA genera para cada visitante, es decir, un ID de usuario único. La métrica homóloga en GA es usuarios.
- **visitId:** Campo de tipo numérico, que devuelve el identificador de cada visita. Este código está basado en la fecha yhora en la que inicia la sesión en el sitio web, que normalmente se guarda como la cookie _utmb. Solo es único a nivelde usuario. Para generar un ID de visita completamente único se deben combinar las variables fullVisitorId y visitId. La métrica homóloga en GA es sesiones.
- **totals.pageviews:** Expresa el número total de páginas vistas generadas por los usuarios. La métrica homóloga en GA es páginas vistas.

Como se ha expuesto, BigQuery a diferencia de GA, considera a los usuarios como una dimensión y no como una métrica. Esto será importante porque es posible trabajar sobre los visitantes y explorar los atributos que lo caracterizan como las sesiones y páginas que visitan cada uno independientemente.

Análisis Clúster

Son un conjunto de técnicas de aprendizaje no supervisado utilizadas para clasificar a un conjunto de individuos en grupos o subgrupos ya que permite encontrar asociaciones que no son evidentes en principio pero que pueden ser de gran utilidad una vez detectadas. El objetivo de estos métodos es tratar de dividir los individuos de un conjunto de datos en agrupaciones distintas para que estos sean muy similares entre ellos dentro de cada grupo, y a su vez, estas sean diferentes alas que pertenecen a otros clústeres (James, Witten, Hastie, & Tibshirani, 2017).

Esta similitud entre los individuos se basa en el cálculo de medidas de distancia entre ellos, por ejemplo, la euclideana o algunas basadas en correlaciones, razón por la cual es necesario trabajar con datos cuantitativos para aplicar el análisis clúster.

Al momento de generar las agrupaciones a las cuales pertenecerán los individuos de un conjunto de datos se debe tener en cuenta dos principios:

- La distancia entre los individuos que componen una agrupación (intra-clúster) debe ser mínima para que exista una

fuerte asociación entre ellos, es decir, sea homogénea.
- La distancia entre los individuos que forman parte de diferentes agrupaciones (inter-clúster) debe ser máxima para quienes conforman cada grupo distinto, es decir, sean lo más heterogéneos posibles.

El análisis clúster es empleado en múltiples campos laborales como en salud, marketing e incluso urbanismo. Por ejemplo, en el campo de la investigación sobre enfermedades como el cáncer esta técnica facilitaría la clasificación de los pacientes en segmentos según su perfil de expresión génica; también permitiría segmentar mercados mediante el reconocimiento de pequeños grupos de clientes potenciales afines a recibir algún tipo de publicidad; y en el rubro de la urbanización se podrían identificar distintos sectores de viviendas según su distribución, tipo, precio e incluso su ubicación.

Debido a los múltiples campos en los que esta metodología se emplea es que existen un gran número de técnicas de agrupación clasificadas como métodos jerárquicos y no jerárquicos; cuya elección dependerá también de los objetivos del trabajo a realizar.

En el caso de los métodos no jerárquicos, estos se emplean cuando se clasifican individuos de un conjunto de datos en cierto número de agrupaciones basados en su similaridad, pero requieren con anterioridad de la definición de la cantidad de grupos a utilizar. Las observaciones se van reasignando a los clústeres iterativamente hasta cumplir con algún criterio de parada como por ejemplo contar con una mínima suma de cuadrados de la varianza entre grupos. Entre los algoritmos más populares encontramos el K-means, PAM (Partitioning Around Medoids) y CLARA (Clustering Large Applications).

Para los métodos jerárquicos no se conoce de antemano cuántas agrupaciones se deben generar; sino esto se decide a través de un gráfico llamado dendograma, dondese puede identificar los grupos generados para cada número de clúster, y van desde1 a n, considerando a n como la cantidad total de individuos (James, Witten, Hastie, & Tibshirani, 2017). Aquí encontramos a los algoritmos aglomerativos o AGNES (Agglomerative Nesting) y a los divisivos o DIANA (Divisive Analysis).

K-means

Fue propuesto por MacQueen en 1967 y es considerado el método de segmentación más popular en el campo del análisis clúster debido a que permite generar agrupaciones de grandes cantidades de datos de forma rápida y eficiente (Syakur, Khotimah, Rochman, & Satoto, 2018).

Al tratarse de un método de agrupación no jerárquico es necesario indicar la cantidad deseada de agrupaciones (k) antes de iniciar para que el algoritmo le asigne a cada individuo uno de los k grupos a generar (James, Witten, Hastie, & Tibshirani, 2017). Cada clúster es representado por la media de los datos que pertenecen a dicha agrupación; sin embargo, no hay que dejar pasar por alto que esta técnica es susceptible a valores atípicos u outliers.

La elección del número óptimo de clústeres a considerar es algo subjetivo, sobretodo en el ámbito laboral, pues depende de los procedimientos utilizados para medir las similitudes y los parámetros utilizados durante la agrupación. Entre los métodos más empleados encontramos los directos y los de pruebas estadísticas (Medium, 2019):

Los métodos directos optimizan la suma de cuadrados de la varianza intra-clúster. Entre los más empleados encontramos:
- Método del codo: Identifica el número óptimo de agrupaciones teniendo como base la producción de una pequeña variación total o inercia dentro del clúster para generar un equilibrio entre la inercia y la cantidad de grupos.
- Método de silueta: Mide el grado de agrupación de una observación y estima la distancia promedio entre grupos mediante puntuaciones. Su objetivo es definir la cantidad adecuada de clústeres que generen bloques del conjunto de datos que estén correctamente separados entre sí.

Los métodos de pruebas estadísticas comparan una evidencia científica contra la hipótesis nula. Dentro de estos tenemos:

- Método de brecha: Fue desarrollado por Tibshirani, Walther y Hastie en 2001 y consiste en comparar la varianza total intra-clúster para distintos valores de k con sus correspondientes valores esperados bajo la distribución de la hipótesis nula de los datos. Por esta razón, la elección adecuada del valor de k será el que maximice esta brecha, es decir, que la composición del agrupamiento está lejos de una distribución uniforme aleatoria de puntos.

Si bien estas no son las únicas técnicas existentes para encontrar la cantidad adecuada de agrupaciones, podemos encontrar algunas adicionales como el índice de Calinski-Harabasz o la inestabilidad de clúster, y cada una tiene sus ventajas y desventajas.

Una vez seleccionada la cantidad óptima k de grupos, ejecutar el algoritmo k-means es sencillo. Se elige aleatoriamente un centroide inicial (coordenadas centrales) en cada una de las k agrupaciones y luego se aplican los siguientes pasos:

- Paso de asignación: asigna a cada individuo el centroide más cercano.
- Paso de actualización: actualiza nuevamente los centroides en base al centro de su observación respectiva.

Este proceso se repite hasta que los grupos sean lo más diferentes posibles, es decir, hasta que ya no se produzcan más cambios de individuos en las agrupaciones. Aquí, el algoritmo ha concurrido y se puede determinar finalmente los clústeres (Jeffares, 2019).

Materiales y métodos

Un estudio anterior de caracterización de la audiencia digital basada en la temática de los contenidos visitados en un sitio web realizado en Ljubljana, Eslovenia, obtuvo información relevante sobre los usuarios que respaldaron acciones como campañas de marketing dirigido e incluso personalizado y también bridaron un sistema de recomendaciones basado en las preferencias de los visitantes; sin embargo esta investigación también presentó complicaciones cuando estos lectores tenían múltiples intereses, lo que complicaba el proceso de diferenciación de las agrupaciones, por lo que parte de sus propuestas fueron incluir información adicional como los datos demográficos, la ubicación (UBIGEO) e incluso variables como sus ingresos o su profesión como una medida de solución (Kladnik, Stopar, Fortuna, & Mladenić, 2017).

Después de analizar los antecedentes en estudios similares, la presente investigación tuvo como objetivo generar criterios de caracterización de los lectores registrados de un sitio web informativo con el algoritmo k-means, cuya principal ventaja es ser un algoritmo con un procesamiento sencillo y rápido pero que depende de la cantidad de agrupaciones a elegir y de los centroides iniciales, pero no solo utilizando los datos digitales como la cantidad de visitas generados o el contenido que leen sino integrando también datos de otros orígenes como los generados por las suscripciones a los newsletters, concursos y actividades, e incluso datos externos. De esta manera se podrá identificar perfiles de usuarios, razón por la que se consideró un estudio descriptivo, tanto cuantitativo como cualitativo.

En este proyecto se consideraron los datos de navegación en el sitio web de los lectores registrados durante tres meses de 2018 (entre julio y septiembre). Esta información de actividad digital de los usuarios fue obtenida a través del tag de seguimiento de Google Analytics insertado dentro del mismo. Debido a que la empresa contaba con Analytics 360, la versión premium, todo este detalle se almacena diariamente a nivel de hit en el repositorio de BigQuery en Google Cloud Platform (GCP).

En un inicio, como era necesario fomentar que los usuarios se registrarán en la página web para poder obtener información cualitativa sobre ellos se estableció un límite de notas

visitadas al mes, una vez superada esta cantidad se mostraba una ventana emergente (pop-up) que indicaba que debían registrarse si querían continuar leyendo contenido. Para esto contaban con tres opciones: usando una cuenta Gmail, de Facebook o completando sus datos directamente en un formulario. Entre las variables que se solicitaban encontramos: nombres y apellidos, correo electrónico, contraseña, género, edad, entre otros. Todo este detalle era almacenado en el repositorio de la compañía alojado en Amazon Web Services (AWS).

Cuando el lector completaba el procedimiento de suscripción, se generaba un código de registro, que también se incluía en Google Analytics para asociarlo a su navegación cada que acceda a la página web luego de haber iniciado sesión en su cuenta creada. Este código será el enlace entre la información de GA y la de registros.

Para lograr un mayor porcentaje de nivel de detalle de todos los campos en el conjunto de datos se integraron estas fuentes con otras que posee la empresa y algunos de proveedores terceros. Esta tarea se ejecutó con el software libre R utilizando la interfaz RStudio que posee múltiples librerías y funciones de análisis estadístico lo que permitió realizar todo este procesamiento sin problemas.

El preprocesamiento y preparación de la base de datos final a utilizar en el estudio se realizaron empleando las funciones join y group by de la librería dplyr en R. La variable código de registro (user_id) funcionó como enlace entre la base de navegación web y la del sistema de registros; y para relacionar estos con los datos no digitales se utilizaron tanto los campos dni como el email ingresado al momento de la inscripción.

Como la caracterización de lectores digitales se realizará sobre el contenido que visitan en la página web se excluyeron a quienes hayan ingresado solo a la portada y a los espacios patrocinados, que normalmente son exclusivos para clientes publicitarios. Adicional, se filtraron también a los visitantes casuales, es decir, a los que visitaron muy poco el sitio (menos de 5 notas al mes).

La base de datos final empleada para la ejecución del clustering k-means consideróla información de un total de 19,375 lectores que consumieron cerca de 4 millones de notas durante los tres meses analizados (60 notas por usuario aproximadamente) teniendo 25 variables separadas en tres grupos, para facilitar la tarea de segmentación.

El primer grupo de 9 variables hace referencia a la información personal de los lectores registrados, obtenido del grupo de varias fuentes de datos:

- user_id: Código generado al momento del registro de un usuario (almacenado también en GA).
- nombre: Nombre(s) ingresado(s) por el lector al registrarse.
- apellidos: Apellido(s) ingresado(s) por el lector al registrarse.
- tipo_doc: Tipo de documento de identidad (DNI, Carnet de extranjería o pasaporte) ingresado por el lector al registrarse.
- nro_doc: Número de documento de identidad (alfanumérico) ingresado por el lector al registrarse.
- email: Correo electrónico de acceso del lector registrado.
- género: Género (masculino o femenino) ingresado por el lector al registrarse.
- edad: Edad calculada del lector registrado. Determinada por la fecha de nacimiento ingresada al momento del registro.
- estado_civil: Estado civil (soltero, casado, viudo o divorciado) ingresado por el lector al registrarse.

El segundo grupo de variables, compuesto por 5 características que se refieren a los datos de navegación web de los lectores registrados:

- sesiones: Número de sesiones totales (visitas) realizadas por el lector registrado durante el periodo analizado.
- pag_vistas: Número de páginas vistas totales (notas) cargadas por el lector registrado durante el periodo analizado.
- pais: País de conexión registrado al momento de la navegación del lector registrado; obtenido por su dirección IP o geolocalización (GPS).
- desktop: Número de páginas vistas

cargadas por el lector registrado a través de un laptop u ordenador de escritorio.

- mobile: Número de páginas vistas cargadas por el lector registrado a través de un dispositivo móvil como un smartphone o una tablet.

El tercer grupo de variables compuesto de 11 características que hacen referencia al número de notas web consumidas por los lectores registrados en las 10 principales secciones de contenido dentro del sitio web informativo (la última hace referencia en conjunto a las menos visitadas), y que fueron consideradas para la ejecución de la caracterización con el algoritmo k-means debido a su naturaleza (datos cuantitativos):

- actualidad: Número de notas consumidas de la sección "Actualidad" durante la navegación de un lector registrado.
- deportes: Número de notas consumidasde la sección "Deportes" durante la navegación de un lector registrado.
- economía: Número de notas consumidas de la sección "Economía" durante la navegación de un lector registrado.
- espectaculos: Número de notas consumidas de la sección "Espectáculos y moda" durante la navegación de un lector registrado.
- mundo: Número de notas consumidas de la sección "Mundo" (coyuntura internacional) durante la navegación de un lector registrado.
- nacional: Número de notas consumidasde la sección "Nacional" (coyuntura local) durante la navegación de un lector registrado.
- politica: Número de notas consumidas de la sección "Política" durante la navegación de un lector registrado.
- redes-sociales: Número de notas consumidas de la sección "Redes sociales" (tendencias) durante la navegación de un lector registrado.
- tecnologia: Número de notas consumidas de la sección "Tecnología" durante la navegación de un lector registrado.
- viaje: Número de notas consumidas dela sección "Viaje y Turismo" durante la navegación de un lector registrado.
- otros: Número de notas consumidas en otras secciones de menor tráfico durante la navegación de un lector registrado.

Un paso importante, previo a la ejecucióndel agrupamiento k-means de los lectores digitales considerando el grupo de 11 variables de contenido, fue definir el número de grupos a considerar. Para esto se empleó la función NbClust, perteneciente al paquete del mismo nombre instalado en el software R, debido a que proporciona el mejor escenario basándose en los resultados obtenidos al combinar la cantidad de clústeres con las medidas de distancia y 30 índices de agrupamiento.

Luego de ejecutar dicha función para el caso del algoritmo k-means, se determinó que el número adecuado de agrupaciones a considerar debe ser de 6, ya que 8 de los 30 índices de agrupamiento señalaron a esta cantidad como la óptima; además, si se consideraban más de 7 u 8 clústeres se obtenían grupos compuestos de un solo lector en algunos casos.

Otro paso necesario para esta ejecuciónfue la estandarización de la base de datosporque este método está basado en el cálculo de distancias. La estandarización es crucial dentro del preprocesamiento de los datos para que estos queden limpios y consistentes para analizar. Con la normalización de los datos se busca estandarizar aquellos valores sin procesar para convertirlos a través de una transformación lineal en rangos específicos y de esta manera se puedan generar buenas agrupaciones y se mejore la precisión obtenida al aplicar algoritmos de agrupación, como el k-means (Mohamad & Usman, 2013).

En R se estandarizaron los datos conla función scale, dentro del paquete base para luego proceder con la ejecución de la caracterización de los lectores digitales en 6 grupos según el tipo de contenido consumido empleando la función k-means, disponible también en el mismo software.

Resultados y discusión

Como resultado de la caracterización de los lectores registrados sobre el conjunto de datos del consumo de notas según el tipo de contenido (11 variables de secciones del sitio web) empleando el algoritmo k-means para generar 6 grupos según lo determinado en la etapa anterior como el valor adecuado de agrupaciones, se encontraron tanto segmentos con muchos lectores (16,821) como otros con muy pequeños (45). En Tabla 1 podemos apreciar el detalle del número de usuarios asignado a cada uno de los clústeres luego de la ejecución del clustering.

Tabla 1. Composición de lectores digitales según clúster obtenido por k-means.

Clúster	Número de lectores	Porcentaje de lectores por segmento
1	397	2.1%
2	221	1.1%
3	45	0.2%
4	16,821	86.8%
5	420	2.2%
6	1,471	7.6%

Una forma de visualizar a los clústeres es utilizandoelmétodomáspopulardereducciónde dimensionalidad, los componentes principales (CP). Esta técnica funciona mediante el uso de transformaciones ortogonales convirtiendo variables correlacionadas en un conjunto de componentes no correlacionados linealmente. Lo que queda son las características que contienen la mayor variación posible. En la Figura 1 podemos ver la distribución de las observaciones en cada uno de los seis clústeres, considerando un gráfico de 3 dimensiones para mejor interpretación. Este fue desarrollado en el software R con las funciones prcomp del paquete base y plot3d de rgl.

A nivel del número de lectores registrados, se pudo apreciar que el que los clústeres 3 y 4 son los más extremos y diferentes (representan el 0.2% y el 86.8% de los usuarios totales. Por el contrario, los grupos 1, 2 y 5 están compuestos por una cantidad de entre 1% y 2% de visitantes. Es también cierto que, a nivel de tamaño de agrupaciones no se podrían obtener mayores conclusiones, por lo que se analizó en segundo lugar el porcentaje de consumo de notas en cada una de las 11 variables de secciones consideradas en el análisis. En la Tabla 2 se muestra el detalle sobre los hábitos de los visitantes en cada uno de las agrupaciones calculadas.

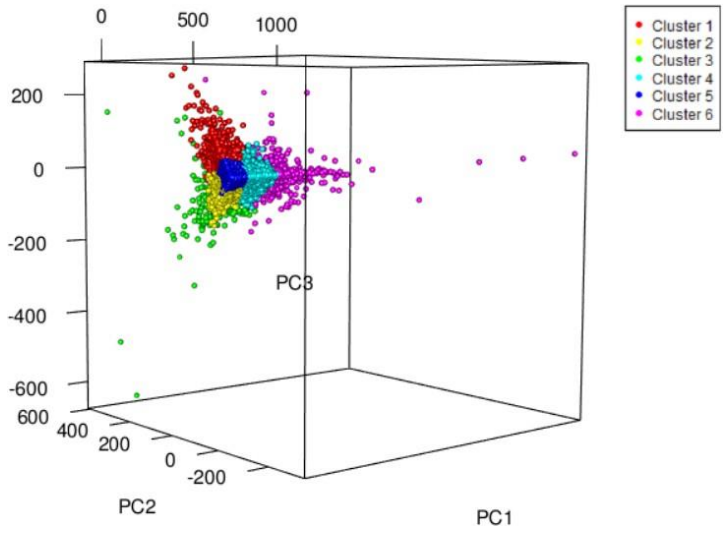

Figura 1. Visualización de los 6 segmentos generados en 3 dimensiones.
Fuente: Elaboración propia, con el software R

Tabla 2. Porcentaje de consumo de notas por sección según clúster.

Sección	Clúster					
	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

Entre los más importantes descubrimientos encontrados después de analizar esta información se observan:

- Los lectores del segmento 1 presentaron afinidad en información variada como noticias de espectáculos y moda, de coyuntura internacional (mundo), tipsde viajes y otros temas menores. Este mix de contenido utilitario (atemporal)representa más del 60% del consumo para esta agrupación.
- En el caso de los clústeres 2, 3, 5 y 6 se observó mayor afinidad de los lectoresen temas de las secciones espectáculos y moda, redes sociales, deportes y política respectivamente. Solamente una sección representa más del 60% del consumo en estos casos.
- Finalmente, el grupo 4, que contiene a más de 86% del total de lectores, mostró tener mayor interés en temas informativos y coyunturales, independiente del género (desde espectáculos a política).

A este nivel ya se tenía más detalle sobreel perfil de los lectores de los seis clústeres encontrados; sin embargo, como apoyo para la caracterización se consideraron las otras variables de los grupos de información personal como el género o edad, y de navegación web como sesiones, páginas vistas, o el país de conexión.

Como observación, los datos sociodemográficos no estuvieron disponibles para el total de lectores, ya que el porcentaje de completitud en estos casos se encontraba entre el 70% y 80% para el género y la edad, y era del 38% para el estado civil. Entre las diferencias más destacadas respecto al perfil demográfico de los segmentos se encontró que:

- Respecto al género del lector registrado, dos segmentos resultaron muy distintos: el segundo, compuesto en su mayoría por mujeres (75% de total femenino) yel quinto en gran proporción por varones (80% usuarios masculinos).
- En el caso de la edad, casi todos los clústeres estuvieron compuestos por usuarios de entre 25 y 54 años; sin embargo, hubo dos agrupaciones un tanto distintas. La tercera estuvo compuesta por usuarios un poco más jóvenes (desde los 18 años) y la sexta por visitantes de hasta 64 años.
- Para el estado civil de los lectores, solo se encontró un segmento diferente al resto, el tercero, compuesto mayoritariamente por usuarios solteros. Por lo tanto, se obtiene que esta agrupación es la conformada por visitantes jóvenes y solteros.

Finalmente, respecto a las variables de navegación web generadas por los visitantes registrados se encontró:

- En las sesiones mensuales promedio se obtuvieron valores similares, de manera proporcional, a los del consumo de notas de contenido. Sin embargo, los lectores del segmento 4 son aquellos que visitan la página web con mayor frecuencia, ya que realizan un promedio mensual de 18.8 visitas, es decir, de forma interdiaria.
- El mayor porcentaje de nuestros lectores visitaron el sitio web iniciando sesión con sus credenciales a través de una computadora, ya que desde un dispositivo móvil tiene distintas formas de acceder al portal sin requerir de su cuenta. Este comportamiento se repite en todas las agrupaciones.
- Más del 80% de nuestros lectores accedieron al sitio web desde dispositivos en Perú, se conoce esto porque a nivel digital se conoce desde donde proviene una visita rastreando su ubicación a través de la IP de navegación y/o el GPS. Sin embargo, el tercer y quinto segmento contienen a más de 27% de visitantes extranjeros.

Tabla 3. Caracterización de los seis clústeres de lectores registrados encontrados.

Segmento	Nombre	Descripción	Número de Lectores
1	Utilitarios	• Tienen entre 25 y 54 años y son solteros y casados principalmente. • Leen en promedio 8 notas al mes. • Gustan de leer contenido de utilidad, aquel que no pierden valor en el tiempo, como los de las secciones de mundo, viajes, espectáculos y otros.	397
2	Chismosos	• Principalmente mujeres solteras y casadas de entre 25 y 54 años. • Leen 14 notas en promedio al mes. • Consumen en gran cantidad notas de farándula local e internacional y cine.	221
3	Modernos	• Jóvenes solteros de entre 18 y 44 años. • Leen 21 notas en promedio al mes. • Se conectan también del extranjero. • Buscan contenido de tendencia dentro de la sección de novedades de redes sociales y algo de espectáculos.	45
4	Informados	• Entre 25 y 54 años, solteros y casados. • En promedio visitan 19 veces el sitio y leen 78 notas al mes. • Consultan notas de coyuntura nacional e internacional puesto que leen las secciones deportes, mundo, política, espectáculos y actualidad.	16,821
5	Futboleros	• Mayormente varones solteros y casados de entre 25 y 54 años. • Visitan el sitio 5 veces al mes y leen aproximadamente 15 notas. • Revisan principalmente las notas de la relacionadas al futbol local e internacional en la sección deportes.	420
6	Políticos	• Adultos un poco mayores solteros y casados de entre 35 y 64 años. • Visitan el sitio 7 veces al mes y leen aproximadamente 22 notas. • Interesados en lo que ocurre con la política, notas internacionales y gusta de las columnas de opinión.	1,471

Fuente: Elaboración propia

Una vez analizadas todas las características expuestas de los lectores digitales registrados en cada uno de los 6 clústeres determinados, se etiquetaron a estas agrupaciones bajo los nombres de utilitarios, chismosos,modernos, informados, futboleros y políticos respectivamente; ya que bajo estas denominaciones se pueden comprender con una sola palabra las características de estas agrupaciones. El detalle completo de estas denominaciones y su explicación se muestran en Tabla 3.

Estas 6 agrupaciones de lectores registrados ya etiquetados nos permiten determinar que existen ciertas características que comparten un grupo de visitantes, las cuales pueden ser utilizadas tanto para mejorar su experiencia e interacción con la página web, así como también en cuestión de publicidad personalizada, lo cual es una ventaja para el negocio.

Una vez puesto en marcha el proyecto, se tiene pronosticado alcanzar un total de más de 100,000 lectores registrados a la web y un incremento trimestral de los ingresos por publicidad digital considerando el nuevo producto comercial de publicidad segmentada por clústeres. Por trimestre se registra en promedio un ingreso de S/ 150,000.00, el cual se estima llegué a S/ 240,000.00 al término del tercer año. Este tendrá el doble del costo de la publicidad tradicional ya que permitiráa los clientes elegir el target óptimo para sus campañas, lo que se traduce en mejores resultados para ellos. A medida que la cantidad de los usuarios vayan aumentando, estos grupos serán más relevantes para los compradores, lo que se traduce en mayor interés de su parte, dejando de lado la compra de publicidad tradicional. El detalle del proyectado trimestral de incremento de ingresos percibidos se muestra en Tabla 4.

Tabla 4. Proyectado de ingresos digitales percibidos luego de la implementación del producto digital “Publicidad Segmentada”.

			Ingresos Digitales		
Año	**Trimestre**	**Lectores Registrados**	**Publicidad Tradicional**	**Publicidad Segmentada**	**Total**
1	Q1	19,375	S/150,000.00	S/0.00	S/150,000.00
	Q2	20,375	S/150,000.00	S/0.00	S/150,000.00
	Q3	21,375	S/140,000.00	S/20,000.00	S/160,000.00
	Q4	22,375	S/140,000.00	S/20,000.00	S/160,000.00
2	Q1	24,875	S/130,000.00	S/40,000.00	S/170,000.00
	Q2	27,375	S/120,000.00	S/60,000.00	S/180,000.00
	Q3	32,375	S/110,000.00	S/80,000.00	S/190,000.00
	Q4	37,375	S/100,000.00	S/100,000.00	S/200,000.00
3	Q1	47,375	S/90,000.00	S/120,000.00	S/210,000.00
	Q2	62,375	S/80,000.00	S/140,000.00	S/220,000.00
	Q3	82,375	S/70,000.00	S/160,000.00	S/230,000.00
	Q4	**107,375**	**S/60,000.00**	**S/180,000.00**	**S/240,000.00**

Fuente: Elaboración propia

Dentro de este análisis hay que considerar también los gastos que conlleva la puesta en marcha de este proyecto, como las acciones para captar a nuevos lectores, el costo de las plataformas donde se alojan los datos y el costo de personal tanto de la fuerza de ventas como del equipo encargado del mantenimiento del modelo de segmentación de lectores. Se estima que a partir del segundo año se contrate a una persona más en el equipo de ventas con el fin de abarcar la mayor cantidad de clientes posibles dispuestas a adquirir este nuevo producto. Esta información se muestra en Tabla 5.

Tabla 5. Proyectado de gastos realizados para la puesta en marcha del proyecto de "Publicidad Digital Segmentada".

Año	Trimestre	Costo Plataformas	Acciones Captación	Mantenimiento Modelo	Fuerza de Ventas	Total de Egresos
1	Q1	S/75,000.00	S/5,000.00	S/12,000.00	S/18,000.00	S/110,000.00
	Q2	S/75,000.00	S/5,000.00	S/12,000.00	S/18,000.00	S/110,000.00
	Q3	S/75,000.00	S/3,000.00	S/12,000.00	S/18,000.00	S/108,000.00
	Q4	S/75,000.00	S/3,000.00	S/12,000.00	S/18,000.00	S/108,000.00
2	Q1	S/75,000.00	S/2,000.00	S/12,000.00	S/24,000.00	S/113,000.00
	Q2	S/75,000.00	S/2,000.00	S/12,000.00	S/24,000.00	S/113,000.00
	Q3	S/75,000.00	S/1,000.00	S/12,000.00	S/24,000.00	S/112,000.00
	Q4	S/75,000.00	S/1,000.00	S/12,000.00	S/24,000.00	S/112,000.00
3	Q1	S/75,000.00	S/500.00	S/12,000.00	S/24,000.00	S/111,500.00
	Q2	S/75,000.00	S/500.00	S/12,000.00	S/24,000.00	S/111,500.00
	Q3	S/75,000.00	S/500.00	S/12,000.00	S/24,000.00	S/111,500.00
	Q4	**S/75,000.00**	**S/500.00**	**S/12,000.00**	**S/24,000.00**	**S/111,500.00**

Fuente: Elaboración propia

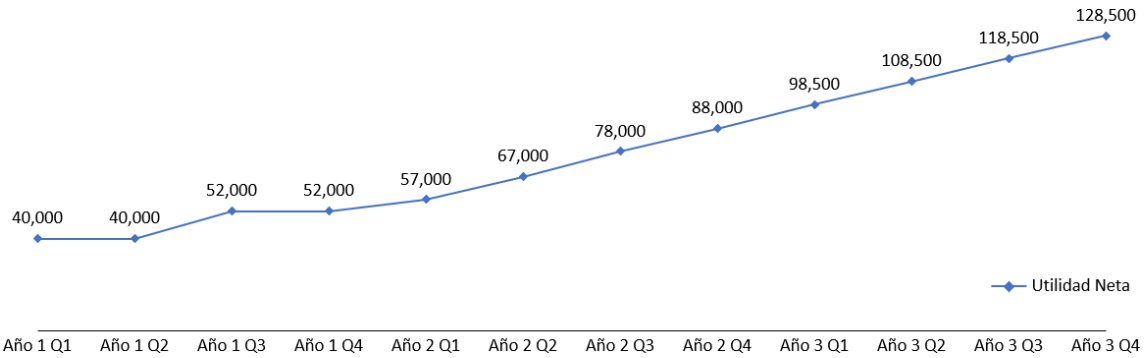

Figura 2. Evolutivo trimestral de proyección de utilidad neta percibida luego de la implementación del proyecto de "Publicidad digital segmentada".
Fuente: Elaboración propia

Finalmente, la utilidad neta percibida desde las plataformas y acciones digitales en la compañía durante el primer trimestre de implementado este proyecto es de S/ 40,000.00. Se tiene pronosticado que, a fines del tercer año, esta utilidad neta percibida se incremente al triple, alcanzando más de S/ 120,000.00, lo cual hace que esta implementación sea rentable para la empresa con el pasar del tiempo. El detalle de los ingresos, egresos y utilidad neta percibida se muestra en Tabla 6. En Figura 2 se observa el evolutivo de las utilidades.

Tabla 6. Utilidad neta trimestral percibida luego de la implementación del nuevo producto "Publicidad Digital Segmentada".

Año	Trimestre	Ingresos	Egresos / Gastos	Utilidad Neta
1	Q1	S/150,000.00	S/110,000.00	S/40,000.00
	Q2	S/150,000.00	S/110,000.00	S/40,000.00
	Q3	S/160,000.00	S/108,000.00	S/52,000.00
	Q4	S/160,000.00	S/108,000.00	S/52,000.00
2	Q1	S/170,000.00	S/113,000.00	S/57,000.00
	Q2	S/180,000.00	S/113,000.00	S/67,000.00
	Q3	S/190,000.00	S/112,000.00	S/78,000.00
	Q4	S/200,000.00	S/112,000.00	S/88,000.00
3	Q1	S/210,000.00	S/111,500.00	S/98,500.00
	Q2	S/220,000.00	S/111,500.00	S/108,500.00
	Q3	S/230,000.00	S/111,500.00	S/118,500.00
	Q4	**S/240,000.00**	**S/111,500.00**	**S/128,500.00**

Fuente: Elaboración propia

Conclusiones

Si bien se desarrolló esta caracterización basada en el tipo de contenido que los lectores registrados consumían según la cantidadde notas que visitaban en cada sección dela web, obtenido desde Google Analyticsy BigQuery, debido a la naturaleza de las mismas variables (datos numéricos), otras variables no consideradas, de tipo cualitativo, como el rango de edad, género e incluso el estado civil (información demográfica) ayudaron a determinar las etiquetas para cada uno de las agrupaciones generadas bajo los nombres de utilitarios, chismosos, modernos, informados, futboleros y políticos. Estos clústeres diferenciados facultarán a la compañía de medios de comunicación para ofrecer a sus clientes publicidad digital segmentada por target (público objetivo) la cual estará respaldada en todo este desarrollo analítico del algoritmo k-means.

Esto será de provecho tanto para los clientes como para la empresa, ya que serán capaces de especificar con mayor precisión a qué tipo de usuarios quieren impactar con sus campañas publicitarias, lo cual mejoraría en gran medida a los resultados que se obtendrían si se empleara el método tradicional de publicidad digital, y por consiguiente la inversión que realicen será mayor al tratarse de un producto especializado. Además, luego de analizar los ingresos y egresos de poner en marcha el lanzamiento de este proyecto, se observa que al término del segundo año ya

se va duplicando la utilidad neta percibida, y para el tercer año ya se obtiene una utilidad neta trimestral de casi S/ 130,000.00; lo cual ya hace sustentable dicho proyecto.

Recomendaciones

En el caso particular del grupo de los informados, donde encontramos a casi el 87% de los lectores registrados y que representan casi el 50% del consumo de contenido, se podrían realizar sobre ellos algunas acciones de captación o fidelización, basado en promociones y beneficios que la empresa pueda ofrecer. Y hacer efectivas estas comunicaciones a través de canales como el envío de correos electrónicos (mailing) e incluso los SMS. También en base a este perfil se podrían explorar algunos productos adicionales que se manejen en la compañía, como la suscripción a un newsletter e incluso mejorar su experiencia de usuario a través de publicidad que sea afín lo que les interesa.

Por último, para la caracterización de estos perfiles se excluyeron a los lectores no registrados, por lo que se podría enriquecer este análisis considerando como se comportan estos usuarios anónimos respecto de las agrupaciones halladas. Este descubrimiento permitiría comprar la conducta de ambos tipos de usuarios basados en la cantidady tipo de contenido que consumen. Otro complemento para este análisis podría ser el considerar incluir variables digitales adicionales como el día y hora de conexión, la fuente de tráfico e incluso la duración dela sesión. Con todas estas consideracionesse podría elevar el proyecto a una segunda etapa que permita mejorar el grado de detalle descriptivo de caracterización en cada una de las agrupaciones ya establecidas.

Referencias

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 GoogleAnalyticsWork? (Complete Beginner's Guide): https://www.monsterinsights.com/how-does-google-analytics-work-beginners-guide/

Bekavac, I., & Garbin Praničević, D. (2015). Web analytics tools and web metrics tools: An overview and comparative analysis. *Croatian Operational Research Review*, 373-386.

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

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/

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*.

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.

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