Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

GUSTOS DIVERSOS, ESTRATEGIAS ÚNICAS

MEMORIA PROYECTO I

Grado en Ciencia de Datos

Autores: Imad Rifai, Mihai Cristian Mihalache, Nouh Khouyi Etbar, Raúl Ruiz Sánchez y Rubén Tormo Piles

Tutor: José Miguel Carot Sierra

1º A1

Tabla de contenidos

1. Introducción

En un entorno actual donde todo está altamente digitalizado y saturado de datos, las empresas y marcas deben afrontar el desafío de captar y llamar la atención de sus clientes potenciales, para convertirlos en clientes leales que lleguen a consumir su producto. La publicidad desempeña un rol esencial en este proceso, sirviendo como puente entre las marcas y su audiencia objetivo. No obstante, en un mundo donde la atención es un recurso escaso y valioso, las estrategias publicitarias tradicionales habitualmente no son suficientes para generar el impacto deseado en el consumidor.

Como respuesta a esta necesidad de maximizar la efectividad de las campañas publicitarias, ha surgido un enfoque innovador y altamente prometedor: la publicidad personalizada basada en los datos recopilados de cada cliente. Esta estrategia implica el uso de estos, que abarcan desde información demográfica y comportamiento en línea, hasta historial de compras y preferencias personales. Para así, crear mensajes publicitarios relevantes y personalizados según las preferencias y gustos de cada segmento de la población global. Todo esto para atraer clientes potenciales y maximizar el retorno de inversión.

Volver a la tabla de contenidos

2. Motivación

La idea final llegaría mediante una conversación espontánea en clase con nuestro tutor José Miguel Carot. De una forma u otra, el diálogo desembocaría en un tema que, además de ser sumamente interesante, fue trascendente en la historia de los últimos años: “El Escándalo *Facebook – Cambridge Analítica”.*

Destacado por el papel crucial de Cambridge Analytica en la victoria de Donald Trump. La empresa británica utilizó un test de personalidad desarrollado por Aleksandr Kogan para obtener perfiles de 50 millones de usuarios de Facebook sin su consentimiento explícito, influyendo en las elecciones estadounidenses.

Christopher Wylie, ex empleado de Cambridge Analytica, reveló cómo usaron estos datos para crear perfiles psicológicos detallados y personalizar

mensajes que influyeron en las decisiones de los votantes de manera individualizada. Además de la publicidad personalizada, Cambridge Analytica también difundió noticias falsas a través de diversas plataformas, distorsionando la percepción pública y afectando las decisiones electorales.

Entendido ya todo el contexto de la noticia, ver el poder de la influencia y del conocimiento de la población mediante la recopilación de datos, nos motivó a llevar a cabo nuestro proyecto, para hacer lo mismo, pero de una manera ética.

Poder apreciar el control que tienen los datos de las personas sobre ellas mismas, nos motiva de una manera positiva a conseguir crear publicidad personalizada, pero **sin sobrepasar los límites éticos y morales**, en contraposición con *Cambridge Analitica.*

Asimismo, ayudar mediante la ciencia de datos directamente a las empresas y a las personas: las empresas pudiendo hacer unas campañas publicitarias efectivas, y la gente sintiéndose identificada con lo anunciado; llegando a agradar de esta manera a las 2 partes: vendedor y cliente.

Finalmente, poder comprobar la capacidad de los datos para influir sobre la humanidad, sus comportamientos y la toma de decisiones nos ha motivado a querer saber más del tema y desarrollar nuestro proyecto.

Es por ello por lo que hemos elegido realizar este trabajo.

Volver a la tabla de contenidos

3. Estado del arte

En los últimos años, la publicidad personalizada ha ido evolucionando significativamente hasta alcanzar el estado actual siendo impulsada, en gran medida, por los avances en tecnología, la recolección de datos y su posterior análisis. Aquí hay algunos aspectos clave del estado actual:

- ***Big Data y Análisis Avanzado:*** Las empresas recopilan infinidad de datos sobre gustos, preferencias, situación personal, localización, etc. para comprender mejor a los consumidores y ofrecer anuncios más relevantes para ellos.

  Una empresa que emplea estos grandes volúmenes de datos sería *Amazon* ya que utiliza Big Data para analizar el comportamiento de compra

de sus usuarios. Ofreciendo, así, recomendaciones personalizadas basadas en compras anteriores y en las búsquedas recientes.

- ***Inteligencia Artificial y Aprendizaje Automático:*** Estos tipos de tecnología desempeñan un papel crucial en la publicidad personalizada. Los algoritmos de aprendizaje automático pueden analizar datos para identificar patrones y predecir el comportamiento del usuario.

Una empresa que utiliza bastante la IA es *Spotify*, ya que lo hace con el fin de crear playlists personalizadas y recomendaciones de música basadas en los hábitos de escucha de sus usuarios.

- ***Segmentación Avanzada:*** La publicidad personalizada se basa en una segmentación precisa de la población. Esto permite a los anunciantes crear mensajes que pueden llegar a influir mejor en ciertos clientes específicos. Un ejemplo de esto sería *Facebook*, que utiliza datos demográficos y de comportamiento para segmentar su audiencia y ofrecer anuncios altamente personalizados a sus usuarios.

- ***Conexión Marca – Cliente:*** La publicidad personalizada puede ayudar a construir relaciones más fuertes entre la marca y el consumidor. Cuando una marca muestra que comprende las necesidades y deseos de un usuario, el usuario se siente más conectado con la marca y es más probable que regrese para futuras compras.

- ***Privacidad y Ética:*** Con el aumento de la preocupación por la privacidad de los datos, las empresas están adoptando enfoques más transparentes y éticos para la publicidad personalizada. Esto incluye la obtención del consentimiento del usuario para recopilar y utilizar sus datos, así como el cumplimiento de regulaciones como el Reglamento General de Protección de Datos (GDPR) en la Unión Europea y leyes de privacidad similares en otras jurisdicciones.

Microsoft, por ejemplo, implementa prácticas de privacidad rigurosas y cumple con el GDPR, proporcionando a los usuarios control sobre sus datos.

Volver a la tabla de contenidos

4. Objetivos

Nuestro objetivo principal es demostrar y explicar que mediante la publicidad personalizada basada en los datos recopilados de distintos segmentos de la población se puede llegar a captar la atención de un mayor número de clientes de forma más eficiente y efectiva. Proporcionando una atención prácticamente individualizada para cada segmento de la población.

Para esto, buscamos identificar patrones que nos permitan dividir la población analizada de forma precisa con el fin de encontrar los gustos y preferencias que unen o separan a cada colectivo. Además, es necesario encontrar correlaciones significativas que nos puedan ayudar a relacionar estos gustos y preferencias con cada grupo y, así, poder llegar a unas conclusiones claras.

En cuanto al alcance, este puede ser inconmensurable. La publicidad está prácticamente en todas partes y nos acompaña en nuestro día a día. Por ejemplo, cuando vemos la televisión o utilizamos el teléfono móvil siempre nos aparecen anuncios de ciertas marcas. Muchos de estos anuncios, como pueden ser los de YouTube, son personalizados en base a las preferencias o búsquedas en Google de cada usuario.

Por ende, esta estrategia publicitaria puede ser mucho más potente que otras, ya que el consumidor se sienta más identificado con el producto. Aumentando así sus ganas o necesidad de adquirirlo.

Finalmente, tenemos los limitantes que nos han ido surgiendo. Uno de estos es que no tenemos conocimientos tan avanzados sobre el tema como los que podría tener un científico de datos experto.

Un factor limitante fue el tiempo disponible, dado que solo contábamos con un cuatrimestre para completar el proyecto. Esta restricción nos obligó a reducir el alcance original y nos impidió expandirlo tanto como hubiéramos deseado. Nuestro objetivo final se centró en demostrar cómo el análisis de datos puede proporcionar ventaja competitiva, en lugar de abordar un caso específico de una empresa en particular pata demostrarlo.

Volver a la tabla de contenidos

5. Impacto esperado

Este proyecto sirve para entender los beneficios de la publicidad personalizada basada en datos recopilados. Esta puede traer numerosas ventajas a las empresas ya que podrán anunciar sus productos atrayendo de forma más eficaz a posibles compradores.

Además, también es beneficioso para los clientes ya que, la mayoría de las veces, visualizaran publicidad relacionada con sus gustos y preferencias.

El hecho de ofrecer una propuesta personalizada, enfocada directamente en las preferencias y necesidades de los usuarios afecta positivamente a la eficacia y eficiencia de las campañas publicitarias. De esta forma, todo el mundo sale ganando ya que las empresas podrán alcanzar a mucha más gente con su publicidad. Asimismo, el resto de las personas podrán sentirse mejor identificadas con las marcas que consumen.

Por ello, con un análisis completo de los datos, su respectiva interpretación y conclusiones, se buscaría definir diferentes tipos de comunidades, mejorar la experiencia publicitaria de la población y maximizar la rentabilidad de las empresas.

Volver a la tabla de contenidos

6. Metodología

a) Obtención de la Base de Datos

Dados nuestros objetivos y nuestra motivación, hemos tenido que buscar alguna base de datos que satisfaga nuestras necesidades para poder llevar a cabo un buen análisis. Es decir, necesitábamos una base de datos que nos diera la información suficiente sobre la gente y que nos permitiera llegar a conclusiones sobre los gustos y preferencias de las personas.

Terminamos optando por una base de datos de Kaggle que consta de más de 1000 individuos y 150 variables que reúnen preferencias musicales, de películas, hobbies, fobias, hábitos saludables, hábitos de gastos, y características demográficas. Al contener tantas variables, nos resultó una

buena opción como base de datos, ya que recopila características que pueden resultar útiles al crear perfiles o modelos de usuarios.

b) Tratamiento de la base de datos

Nuestra base de datos, como era de esperar, no estaba preparada para ser analizada al contener datos alfanuméricos, valores faltantes y datos anómalos. Teníamos que arreglar una serie de inconvenientes para poder analizar y trabajar con ella.

b.1) Localización de valores atípicos y tratamiento

En un primer lugar, tratamos de encontrar aquellos valores (outliers) que se desvían considerablemente del resto. Un ejemplo de esto lo encontramos en la variable 'Height' donde podemos observar un individuo con 62 cm de altura. Para estos datos asumimos error de escritura y lo sustituimos por el valor esperado basándonos en variables asociadas como 'Weight' y 'Gender', de donde se extrae que el individuo es una mujer de 55kg de peso. Por tanto, reemplazamos 62cm con 162cm.

Siguiendo con la búsqueda de valores atípicos, buscamos datos no coherentes comparando valores entre varias variables estrictamente relacionadas. El ejemplo más claro lo encontramos al comparar las variables 'Only Child' y 'Number of Siblings'. Observamos que hay hijos únicos con hermanos, para estos casos utilizamos una de las dos variables para poner el valor esperado en la otra (si es hijo único, numero de hermanos = 0).

b.2) Procesamiento de datos ausentes

En segundo lugar, nos ocupamos de los datos ausentes, la base de datos cuenta con pocos ausentes, pero al estar presentes en distintos individuos, si tratamos de eliminarlos, perderíamos mucha información valiosa. Por ello, empleamos las siguientes técnicas:

- Para variables como *'Smoking', 'Alcohol', 'Lying'*, etc imputamos la media.

- En cambio, para variables como el género, podemos basarnos en los valores de peso, edad y altura de estos individuos, de tal manera que dependiendo del peso y la altura del individuo se le imputa un género u otro.

- Para el nivel de educación, imputamos valores basándonos en la edad del individuo y, de manera similar, imputamos la edad basándonos en el nivel de educación.

Para ciudad o pueblo, vemos que aquel que vive en la ciudad tiene muchas más probabilidades de vivir en un bloque de pisos que en una casa, por lo tanto, dependiendo de si vive en un bloque de pisos o no, imputamos.

b.3) Recodificado de algunas variables

Finalmente, con tal de hacer más intuitivo el trabajo de la base de datos, modificamos el nombre de sus columnas, esta decisión se tomó debido a la difícil interpretación de los nombres de las variables, por ejemplo, cambiamos el nombre de la variable ‘Shoping Centre’ a ‘I enjoy going to large shoping centres’ que facilita su comprensión.

Para facilitar el análisis de la base de datos, recodificamos algunas variables categóricas cambiando sus valores a valores numéricos. Por ejemplo, recodificamos categorías como 'Fumo siempre', 'Fumo a veces', 'No fumo nunca', etc., a valores numéricos (0, 1, 2, 3, ...).

c) Descripción de la base de datos.

Como ya mencionamos, nuestra base de datos cuenta con 150 variables. Para facilitar el análisis, las agrupamos en conjuntos temáticos. Al observar la proporción de cada conjunto temático en la base de datos, obtuvimos lo siguiente:

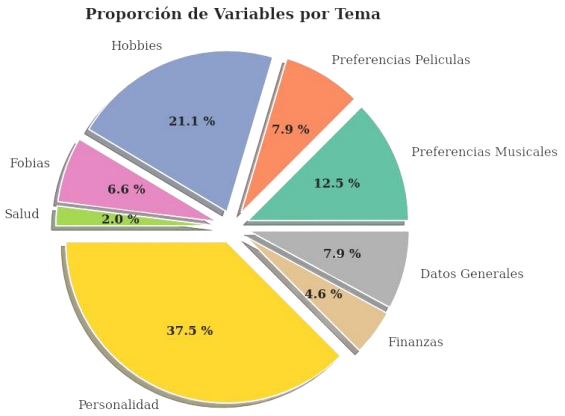

Ilustración 1 - Gráfico de Tarta sobre la proporción de cada tipo de variable

La mayoría de las variables en nuestra base de datos son cuantitativas, con valores que van del 1 al 5 (formato de encuesta con escala Likert). También

incluimos variables cuantitativas como edad, altura y otras características demográficas.

No hemos considerado relevante realizar una descripción paramétrica detallada de las variables en nuestra base de datos. Esto se debe a que, en procesos más complejos, estos parámetros son utilizados directamente, haciendo innecesaria su descripción inicial.

d) Análisis Exploratorio de Datos (AED) (Python)

Dividimos el Análisis Exploratorio de Datos (AED) en 3 partes:

d.1) Visualización segmentada:

Una buena manera de visualizar la distribución de nuestros individuos frente a cada una de nuestras variables es segmentar nuestra base de datos por género y edad. Esto nos permite ver, de un vistazo, si hay diferencias significativas entre estos grupos.

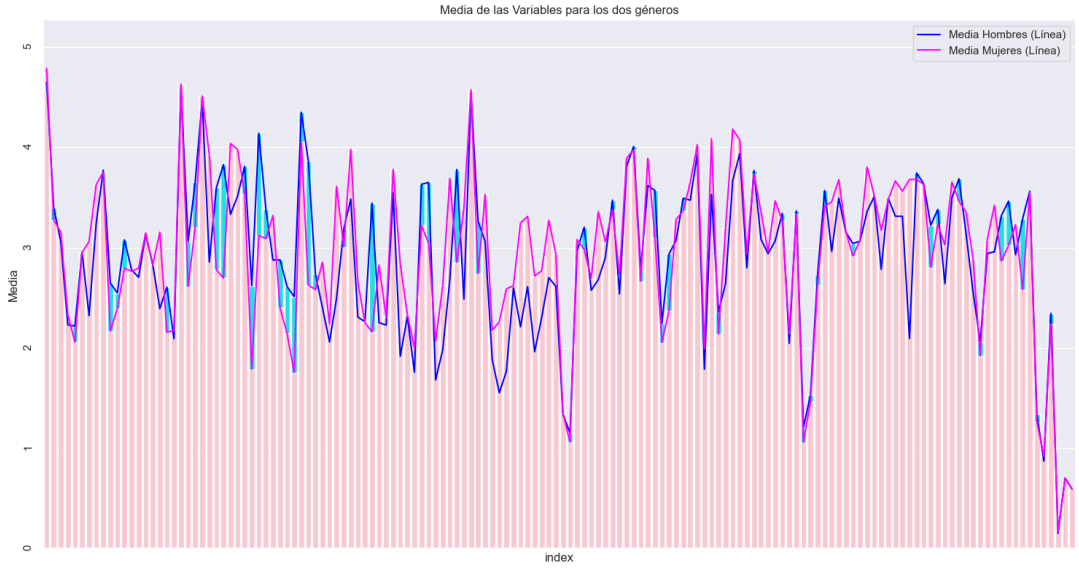

Ilustración 2 – Comparación de la media de las variables según el género

A primera vista, ya podemos observar diferencias en la distribución de nuestros individuos en base a género. Esto nos impulsará a estudiar estadísticamente qué tan grande es esta diferencia, lo que podría proporcionar un mayor conocimiento si una empresa desea dirigirse a un género específico.

En cambio, para los dos segmentos de edad, vemos que se distribuyen de manera similar (podría deberse a que los individuos tienen entre 15 a 30 años):

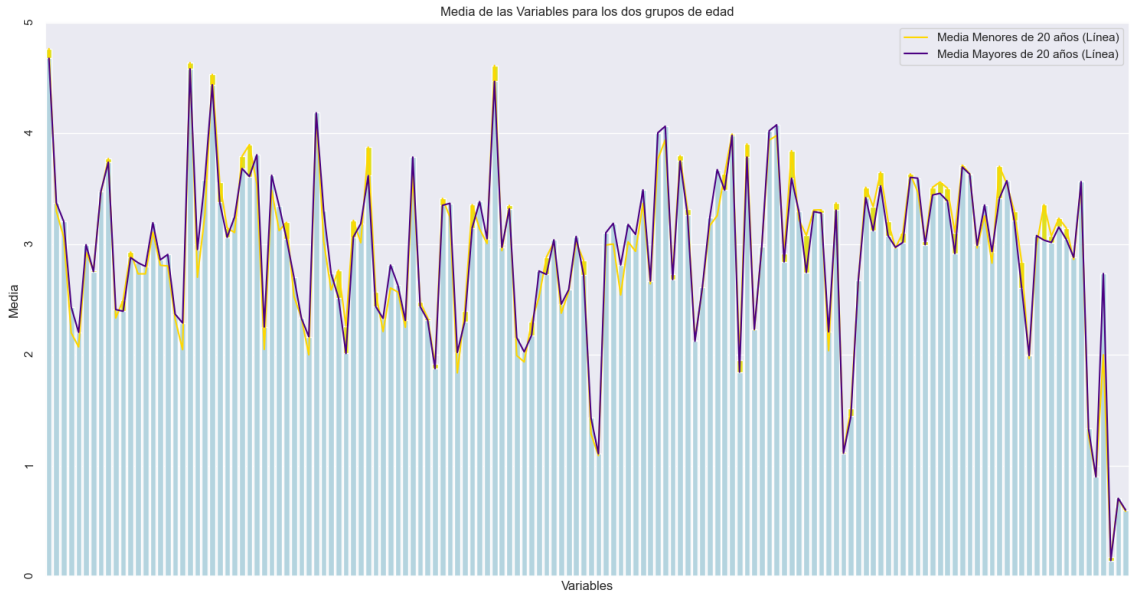

Ilustración 3 Comparación de la media de las variables según la edad

Esto podría indicar a las empresas que enfocar sus recursos en estrategias diferenciadas por edad puede no ser tan relevante, permitiéndoles ahorrar recursos y enfocarse en otros factores más significativos.

d.2) Correlaciones entre variables:

Estudiar la correlación entre las variables de nuestra base de datos es fundamental porque nos permite identificar relaciones estadísticas significativas y patrones ocultos entre diferentes aspectos de los encuestados.

Esto es interesante porque, al entender cómo se interrelacionan variables como preferencias musicales, hábitos de consumo, y características demográficas, podemos desarrollar estrategias de segmentación (***clustering***, realizado más adelante mediante grafos), dirigidas específicamente a subgrupos similares, aumentando el conocimiento del público objetivo de las campañas publicitarias.

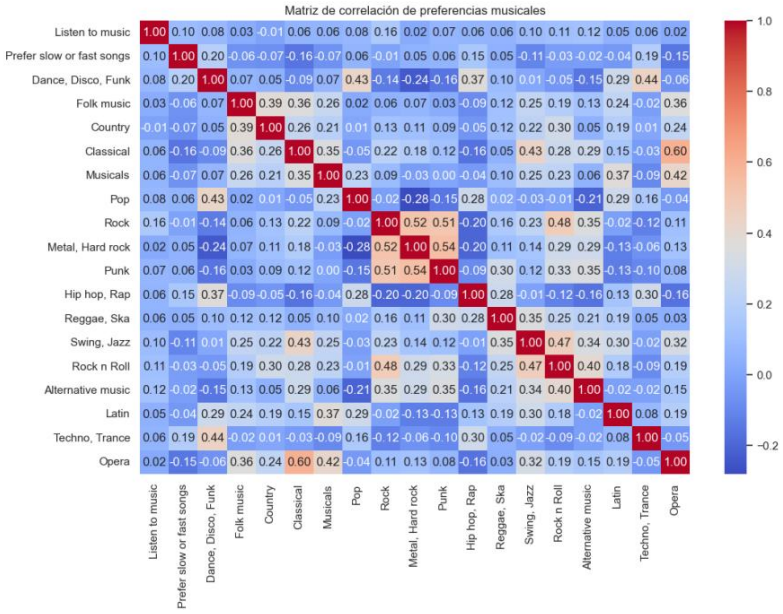

Ilustración 4 – Matriz de Correlaciones de gustos musicales

Para llevar a cabo el estudio de las correlaciones entre nuestras variables, utilizamos las librerías de *matplotlib* y *seaborn* de Python para imprimir matrices de correlación de Pearson para cada conjunto temático (agrupación de variables). Esto nos permitió visualizar claramente las relaciones entre variables dentro de cada grupo temático. Por ejemplo, obtuvimos matrices que mostraban correlaciones significativas y fáciles de interpretar, como:

También encontramos matrices que no mostraron las correlaciones fácilmente debido a la gran cantidad de variables en algunos conjuntos temáticos. Estos conjuntos temáticos, al tener muchas variables, producían matrices densas y difíciles de interpretar, como:

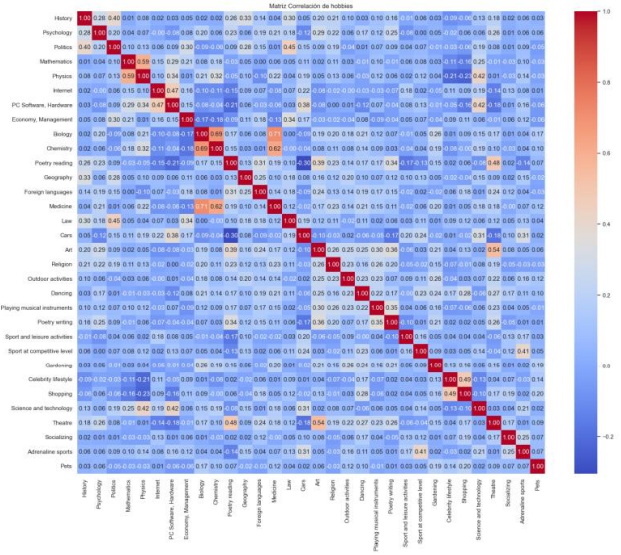

Ilustración 5 – Matriz de correlación de hobbies

d.3) Análisis Gráfico de Correlaciones de Variables (Visualización de Datos)

Para mejorar la visualización y análisis de las correlaciones entre variables en nuestra base de datos, adoptamos un enfoque basado en grafos utilizando la biblioteca NetworkX en Python, y definiendo una función que nos devuelva el grafo, para así poder iterar y mostrar el grafo para cada conjunto temático.

Visualización mediante Grafos:

Agrupamos las variables en conjuntos temáticos y creamos grafos para cada grupo:

- **Nodos:** Representan variables individuales.
- **Aristas:** Reflejan la correlación entre variables, calculada mediante la matriz de correlaciones de Pearson. (Solo tenemos en cuenta una arista si su valor es mayor que 0.25 (correlación mínima)).

Para mejorar la interpretación visual, establecimos que el grosor de las aristas entre dos nodos reflejara la magnitud de la correlación entre esas variables. Además, cuando un conjunto temático contenía más de 15 variables, optamos por numerar los nodos y proporcionar una leyenda que vinculara cada número con la variable correspondiente.

Este enfoque nos permite una representación más clara y legible de las relaciones entre las variables en nuestra base de datos, facilitando así el análisis de las correlaciones y sus implicaciones.

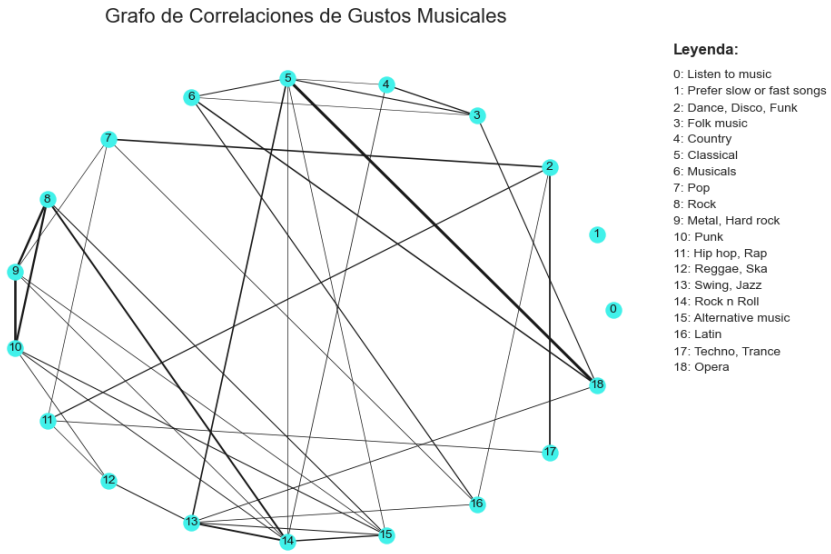

Ilustración 6 – Grafo de Correlación de Gustos Musicales

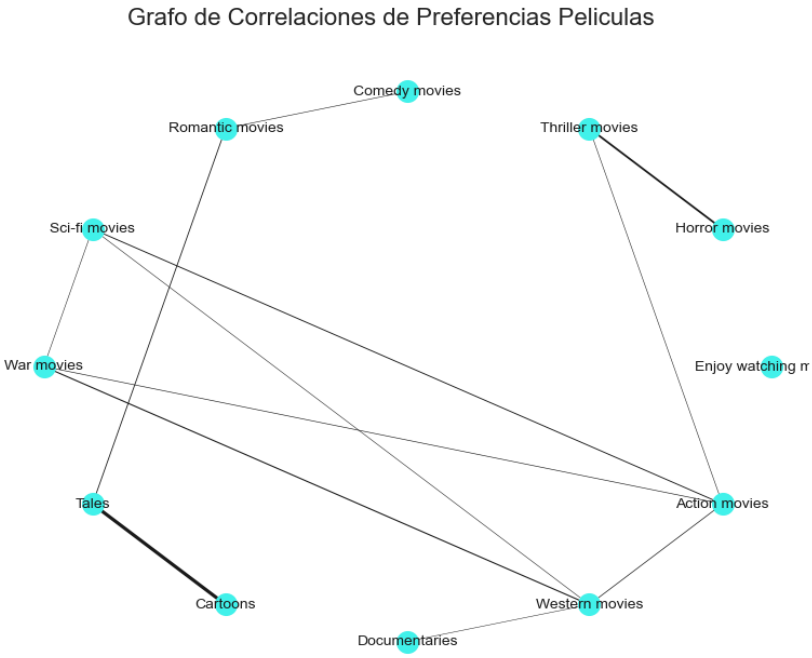

Ilustración 7 – Grafo de Correlación de Gustos Peloculas

e) Detección de Comunidades Mediante el Algoritmo Girvan-Newman (Grafos)

Para identificar comunidades de variables, utilizamos el algoritmo de Girvan Newman. Este algoritmo elimina iterativamente los enlaces con mayor

centralidad de intermediación (betweenness centrality), fragmentando la red en comunidades más pequeñas y muy cohesionadas entre sí.

Pasos del algoritmo Girvan-Newman:

• Calcular la centralidad de intermediación para todos los enlaces.

• Eliminar el enlace con mayor centralidad.

• Recalcular y repetir hasta que la red se divida en comunidades.

Para ello definimos una función en Python, que devuelve dos elementos:

- Un diccionario con claves las comunidades a las que referencia, y como valor para cada clave, una lista en la que podemos observar las variables que describen dicha comunidad.
- Imprime por pantalla un grafo, que muestra visualmente lo descrito anteriormente:

Ejemplo para Gustos Musicales:

{'Comunidad 3': ['Dance, Disco, Funk', 'Pop', 'Hip hop, Rap', 'Techno, Trance'], 'Comunidad 4': ['Folk music', 'Country', 'Classical', 'Musicals', 'Latin', 'Opera'], 'Comunidad 5': ['Rock', 'Metal, Hard rock', 'Punk', 'Reggae, Ska', 'Swing, Jazz', 'Rock n Roll', 'Alternative music']}

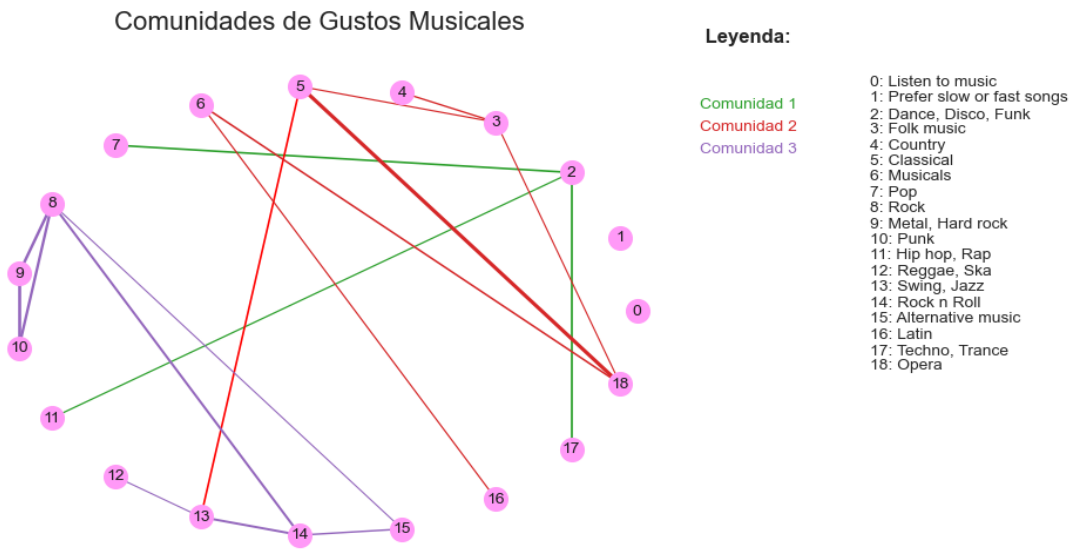

Ilustración 8 – Grafo con Comunidades para Gustos Musicales

f) PCA

Se ha empleado la técnica de PCA para simplificar la complejidad de los datos, conservando la mayor cantidad de información posible. Esto se logra proyectando los datos en un espacio de menor dimensión, donde las nuevas variables (componentes principales) son combinaciones lineales de las variables originales. Además, estas están ordenadas de manera que las primeras componentes capturan la mayor parte de la varianza de los datos. Así, hemos conseguido, para cada grupo de variables, poder visualizar grupos de individuos.

Para entender el procedimiento aplicado para cada grupo de variables, veremos el ejemplo del grupo de variables hobbies que consta de 32 variables que van del 1 al 5, al aplicar esta técnica, comenzamos por estudiar el porcentaje de varianza que explican sus componentes, en este caso:

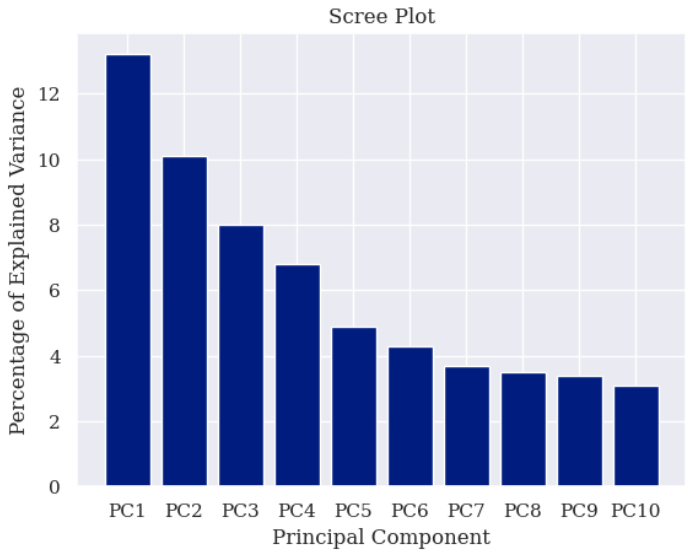

Ilustración 9 – Histograma sobre las Componentes Principales

La gráfica muestra que los primeros componentes principales (PC1 a PC4) explican una mayor proporción de la varianza en los datos, lo que sugiere que se puede reducir la dimensionalidad del conjunto de datos sin perder mucha información. Los componentes adicionales contribuyen cada vez menos a la explicación de la varianza total. Sabiendo esto, procedemos a observar los pesos de cada variable en las dos primeras componentes mediante un histograma para cada componente:

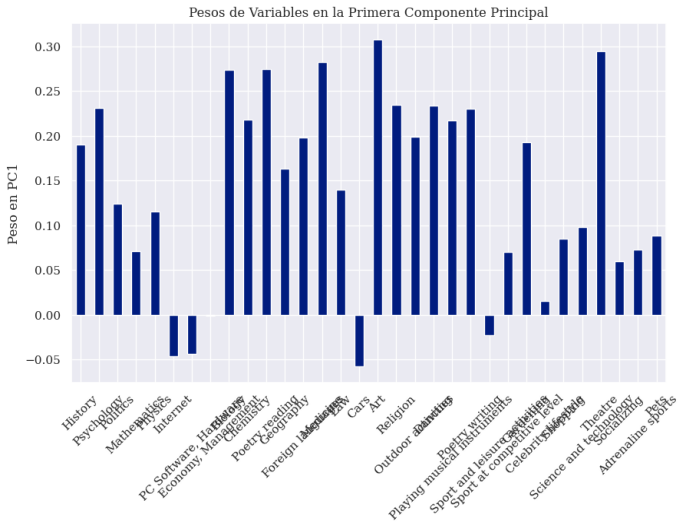

Ilustración 10 – Peso de las variables para la Primera componente principal (Hobbies)

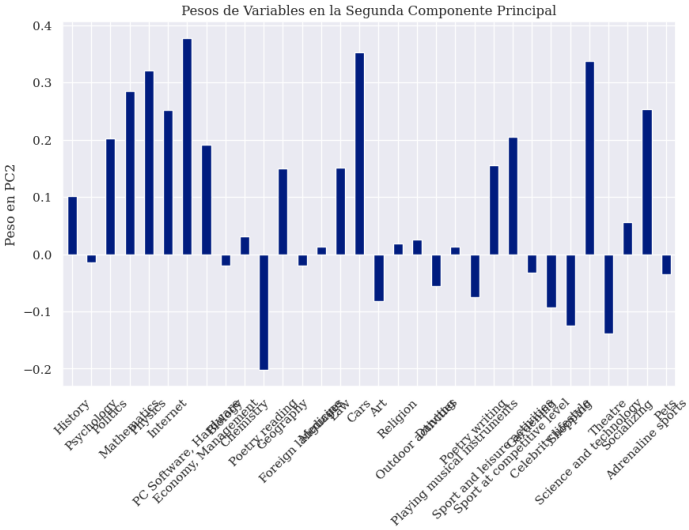

Ilustración 11- Peso de las variables para la Segunda componente principal (Hobbies)

Como se puede ver la primera componente principal (PC1) parece estar influenciada principalmente por intereses académicos y artísticos, incluyendo Matemáticas, Historia, Política, Psicología, Arte y Religión. En contraste, la segunda componente principal (PC2) está influenciada por intereses técnicos y actividades de alto riesgo, como Física, Economía, Software de PC, Deportes de adrenalina e Instrumentos musicales.

El siguiente paso es ver como se distribuyen los individuos entre estas componentes, para ello se emplea el uso del gráfico de dispersión, podemos ver:

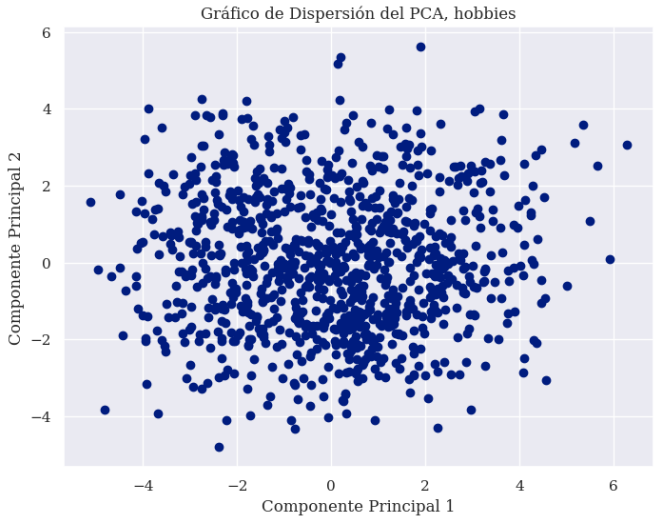

Ilustración 12 - Gráfico de Dispersión del PCA para hobbies

Y ahora finalmente filtrar ese gráfico con otras variables para ver cómo se comportan en esas componentes, vemos ejemplos filtrando por '*Gender*':

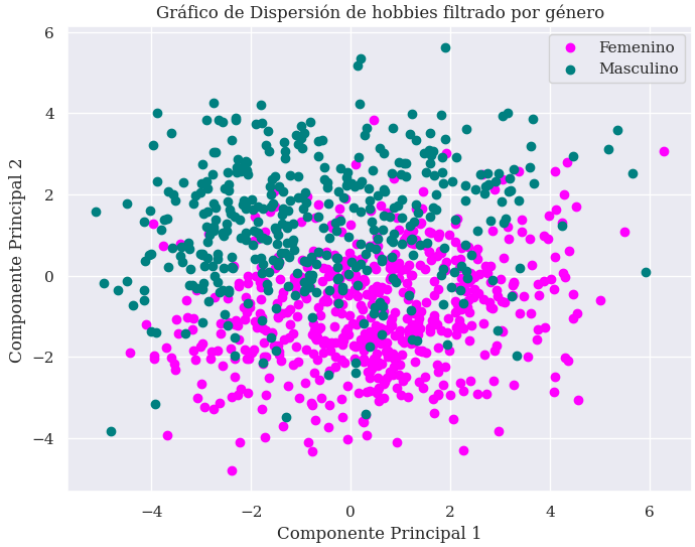

Ilustración 13 - Gráfico de Dispersión del PCA para hobbies filtrado por género

Observamos como claramente el género femenino ha obtenido menos puntación en la segunda componente, retornando a lo mencionado anteriormente, la segunda componente principal (PC2) está influenciada por intereses técnicos y actividades de alto riesgo, como Física, Economía, Software de PC, Deportes de adrenalina e Instrumentos musicales.

Por otro lado, probamos también con otras variables como la edad, pero no se observan comportamientos inusuales:

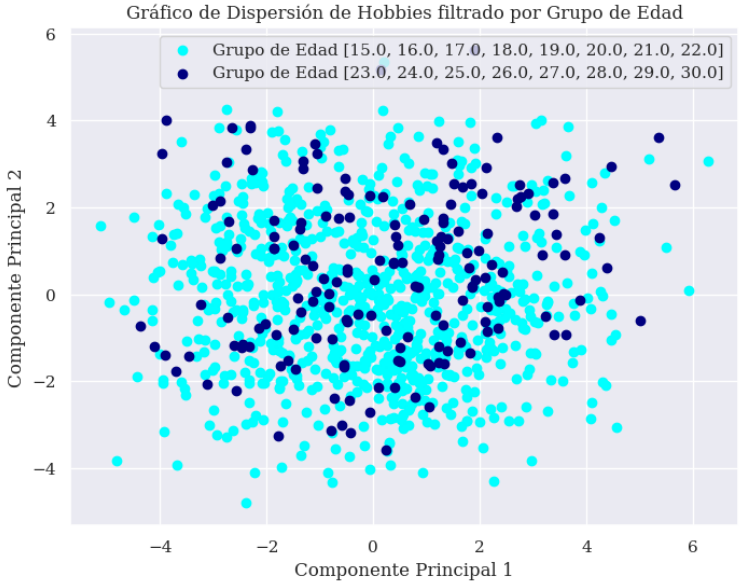

Ilustración 14 - Gráfico de Dispersión del PCA para hobbies filtrado por Edad

A su vez, hemos filtrado los gráficos de dispersión de los diversos análisis de PCA que hemos hecho para cada variable por edad y por género, para poder ver así subgrupos de individuos, y a su vez, cómo se comporta.

g) Contraste de Hipótesis:

Realizar un contraste de hipótesis en estadística es fundamental para obtener evidencia sobre las diferencias entre grupos en un estudio. Este proceso comienza con la formulación de una hipótesis nula (H0), que establece que no hay efecto o diferencia entre los grupos, y una hipótesis alternativa (H1), que sugiere que sí existe alguna diferencia significativa.

Es importante realizar este tipo de análisis porque nos permite tomar decisiones basadas en evidencia estadística y no depender solo de observaciones o intuiciones. Para llevar a cabo este estudio, hemos desarrollado un algoritmo basado en tres funciones esenciales.

1) La primera función, 'media_dicotomica', recibe una base de datos y una variable dicotómica como parámetros. Esta función devuelve una tupla con dos diccionarios: uno para cada valor único de la variable dicotómica. Cada diccionario contiene la media y la varianza de cada variable en la base de datos, así como el tamaño de la muestra correspondiente a ese valor único.

2) La segunda función, 'test_contraste', toma como parámetros dos diccionarios y opcionalmente un nivel de significancia (alpha, predeterminado en 0.05). En esta función, se realiza un contraste de hipótesis sobre las varianzas para evaluar la homogeneidad de varianza entre los grupos. Si se cumple la homocedasticidad, se aplica un test t para comparar las medias entre grupos; de lo contrario, se utiliza el test t de Welch. La función devuelve dos listas: una con las proporciones de variables en las que ambos grupos muestran comportamientos similares y otra con las variables donde muestran diferencias significativas.

3) La tercera función, 'grafico_tarta', recibe la lista de proporciones y la variable dicotómica como parámetros. Esta función genera un gráfico de sectores que visualiza la proporción de variables en la base de datos donde los individuos de la variable dicotómica se comportan de manera similar y el porcentaje de variables donde se observan diferencias significativas.

Comenzamos identificando las variables con dos poblaciones distintas, lo que nos permitió seleccionar automáticamente las variables dicotómicas e iterar sobre ellas. Utilizamos la función 'media_dicotomicas' para calcular las medias, varianzas y el tamaño muestra de estas variables divididas por cada población. Luego, aplicamos la función 'test_contraste' para contrastar estadísticamente las diferencias entre las dos poblaciones en términos de medias, considerando también la homogeneidad de varianzas. Finalmente, visualizamos los resultados utilizando 'grafico_tarta', que nos proporcionó una representación gráfica clara de las proporciones de variables donde las poblaciones mostraban comportamientos similares y diferentes.

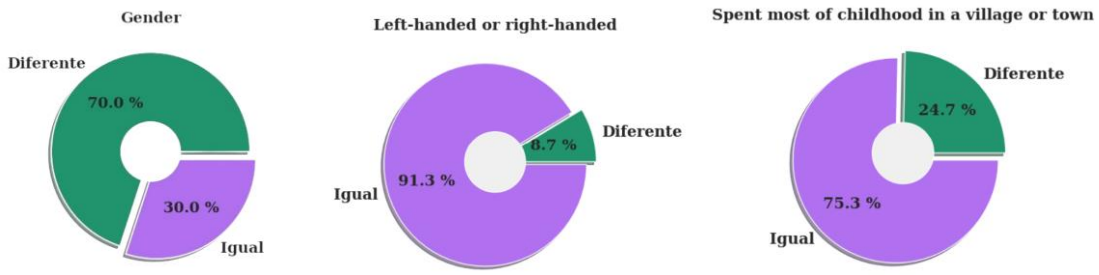

Ilustración 15 – Ejemplos de gráficas devueltas por el algoritmo

Volver a la tabla de contenidos

7. Resultados y Conclusiones

En conclusión, el proyecto "Gustos Diversos, Estrategias Únicas" demostró el potencial de la publicidad personalizada basada en el análisis de datos. El objetivo principal era demostrar que, mediante el análisis de datos de distintos segmentos de la población, se puede captar la atención de un mayor número de clientes de forma más efectiva. A lo largo del proyecto, implementamos varias metodologías y técnicas para tratar y analizar la base de datos.

- **Metodología y Tratamiento de Datos**

Obtención de la Base de Datos: Seleccionamos una base de datos de Kaggle que contenía una amplia variedad de variables.

Tratamiento de la Base de Datos: Identificamos y corregimos valores atípicos y datos ausentes.

Análisis Exploratorio de Datos (AED): Realizamos un análisis visual segmentado por género y edad, lo que nos permitió identificar diferencias significativas a primera vista. Por ejemplo, observamos que las preferencias de las variables analizadas variaban entre los géneros, pero no tanto entre los distintos grupos de edad.

Resultados del Análisis:

Mediante el análisis observamos varios patrones y correlaciones significativas que ayudaron a entender mejor los gustos y preferencias de los diferentes segmentos de la población:

- **Segmentación por Género:** La visualización de la distribución de variables según el género mostró diferencias notables, esto nos llevó a analizar mediante contrastes de hipótesis esta diferencia. Por ejemplo, los hombres y mujeres tenían preferencias distintas en cuanto a hábitos de compra y consumo de medios. Esto se evidenció en los análisis estadísticos realizados, donde se utilizó la función test_contraste para validar estas diferencias. Aplicamos pruebas de hipótesis, como el test t de Welch y el test t de Student, para determinar si las diferencias observadas eran estadísticamente significativas. Los resultados

mostraron que, en la mayoría de los casos (70%), las diferencias eran significativas con un nivel de confianza del 95%.

- **Correlaciones entre Variables**: El estudio de las correlaciones entre variables nos permitió identificar patrones ocultos. Utilizando matrices de correlación y grafos, se encontraron relaciones significativas entre preferencias musicales, hábitos de consumo y características demográficas. Por ejemplo, se observó una fuerte correlación entre los gustos musicales y las preferencias de películas, lo cual puede ser utilizado para crear campañas publicitarias cruzadas entre estas dos áreas.

- Adoptamos un **enfoque basado en grafos para mejorar la visualización y análisis de las correlaciones**. Los grafos ayudaron a identificar comunidades dentro de los datos que compartían características similares (mediante el algoritmo de Girvan-Newman), facilitando una segmentación más precisa de la población. El análisis de componentes principales (PCA) y los grafos de redes evidenciaron agrupaciones significativas, validando así la segmentación realizada.

Ejemplo de Resultados:

Un ejemplo específico de los resultados obtenidos es el análisis de las preferencias musicales y su relación con otras variables. La matriz de correlación de gustos musicales mostró que aquellos individuos que disfrutaban de la música clásica también tenían una mayor propensión a disfrutar de actividades culturales como visitar museos y leer libros. Esta información es valiosa para empresas que desean orientar sus campañas publicitarias hacia un público que aprecia actividades culturales, permitiendo así una personalización más efectiva.

En resumen, el proyecto "Gustos Diversos, Estrategias Únicas" demuestra cómo el análisis de datos puede revolucionar la publicidad personalizada, ofreciendo una herramienta poderosa para captar la atención del consumidor.

8. Glosario de términos

- Algoritmo: conjunto de instrucciones o reglas bien definidas y ordenadas que se utilizan para realizar una tarea específica o resolver un problema.

- Big Data: conjuntos de datos extremadamente grandes y complejos que no pueden ser gestionados, procesados ni analizados con las herramientas de procesamiento de datos tradicionales. Se caracterizan por su gran volumen, variedad y velocidad.
- Correlaciones: relación estadística entre dos variables. Cuando dos variables están correlacionadas, significa que existe una tendencia consistente en cómo cambian juntas.

- Inteligencia Artificial (IA): es la capacidad de una máquina para realizar tareas que normalmente requieren inteligencia humana, como aprender, razonar, resolver problemas y tomar decisiones.

- Patrones: tendencias, estructuras o regularidades que pueden observarse en conjuntos de datos. Estos patrones pueden ser identificados mediante técnicas estadísticas, algoritmos de aprendizaje automático o métodos de visualización de datos.

- Publicidad personalizada: estrategia publicitaria basada en adaptar los mensajes publicitarios y marketing para ajustarse a las necesidades, intereses y preferencias individuales de los consumidores.
- Segmento de la población: grupo homogéneo de personas que comparten características similares o tienen necesidades, comportamientos o características comunes en relación con un producto, servicio o mercado específico.

Volver a la tabla de contenidos

9. Bibliografía

- https://www.appinio.com/es/blog/investigacion-de-mercados/margen-error-tamano-muestra (teoría para el tamaño de la muestra y su cálculo)
- https://tools4success.es/tag/big-5/#:~:text=En%20el%20estudio%20de%20la,(dimensiones%20de%20la%20personalidad).
- https://www.celag.org/cambridge-analytica-el-big-data-y-su-influencia-en-las-elecciones/
- https://www.bbc.com/mundo/noticias-43472797
- https://www.bbc.com/mundo/noticias-internacional-43500891
- https://www.eldiario.es/internacional/theguardian/documento-cambridge-analytica-estrategia-trump_1_2207870.html
- https://www.puromarketing.com/25/18744/publicidad-personalizada-mayor-impacto-mayor-conversion

Volver a la tabla de contenidos