Un Proceso de Big Data aplicado a datos de consumo cultural en Argentina*

Francisco Raúl Torres[1] and Agustina Buccella[1][0000−0002−8516−7453]

GIISCO Research Group - Departamento de Ingeniería de Sistemas
Facultad de Informática - Universidad Nacional del Comahue
Neuquen, Argentina
{francisco.torres,agustina.buccella}@fi.uncoma.edu.ar

Resumen El consumo cultural en la región nacional es tan diverso como la gente que lo consume. Esta actividad involucra interactuar con la radio, la música, la televisión, el teatro, o un recital, y esto luego impacta en la economía de la cultura, de la mano de las productoras que impulsan dichas actividades. Mediante el uso del Sistema de Información Cultural de la Argentina perteneciente al Ministerio de Cultura de la Nación, se pudieron conocer datos de una encuesta acerca de los consumos culturales de la población en diferentes áreas de la Argentina. En el presente trabajo abordamos un enfoque de proceso de Big Data completo, donde partimos de un conjunto de datos y lo procesamos de forma tal que podamos (1) analizar patrones sobre la edad y la cantidad de horas consumidas de radio, música y televisión y, (2) analizar segmentos de mercado relacionados al consumo de radio, música o televisión.

Keywords: Big Data · Consumo cultural · Another keyword.

1. Introducción

Big Data es un área de estudio que considera grandes volúmenes de datos, tanto estructurados como no estructurados, mecanismos de ingestión y procesamiento de los mismos, para centrarse luego en qué es lo que pueden hacer las organizaciones con ellos para aprovecharlos, especialmente en la toma de decisiones. Así, esta área abarca tareas de análisis, procesamiento y almacenamiento de grandes conjuntos de datos que se originan desde diferentes fuentes, con el objetivo de obtener conclusiones sobre esa información, ya sean patrones ocultos, correlaciones y otros tipos de perspectivas que pueden ser útiles, ayudando a las organizaciones a aprovechar sus datos y a utilizarlos para identificar nuevas oportunidades.

*Este trabajo es presentado en el marco de la Materia Electiva *Almacenamiento y Análisis para Big Data* perteneciente al 5to año de la Carrera Licenciatura en Sistemas de Información de la Facultad de Informática de la Universidad Nacional del Comahue. Cursado 2023 cuya docente es Agustina Buccella.

Existen en la actualidad muchas metodologías propuestas para llevar a cabo un proceso de desarrollo para Big Data [2,4]. En general, se converge en tres grandes actividades involucrando la limpieza y/o preprocesamiento, almacenamiento y la analítica de datos. Las dos primeras actividades pueden ser intercambiadas dependiendo del enfoque utilizado. Es decir, para realizar un desarrollo orientado al análisis de los datos, se debe primero seleccionar la/s fuentes de información que puedan resultar útiles, procesarlas para que tengan sentido, y luego almacenarlas en algún tipo de repositorio, en la forma de depósitos de datos [10] (enfoque ETL, extracción-transformación-carga) o lagos de datos [7] (enfoque ELT, extracción-carga-transformación). La característica principal de estos repositorios es que actúan como centros de información en donde se vuelcan, en algún formato específico, los datos de todas las fuentes que se deseen explotar [6,8,9]. Luego, la analítica de datos (data analytics) se dedica al proceso de crear información desde los datos fuente por medio de la contextualización, análisis y gobernanza de datos.

Se han presentado muchas aplicaciones de Big Data en la actualidad en dominios específicos, todas ellas realizando análisis interesantes que permiten explotar la información de manera de clasificarla, relacionarla y/o predecir nuevos comportamientos o sucesos. En particular, en el contexto de la cultura, los estudios sobre los datos de los consumos revelan segmentos de mercados por franjas etarias, indicando qué dispositivos son los mas utilizados para la escuchar o ver el producto, o los géneros musicales mas escuchados. También, los datos relacionados a la cultura son utilizados por la industria del turismo, donde su análisis revela qué patrones siguen los turistas, qué circuitos turísticos hay, y así poder tomar decisiones para potenciar esa oportunidad de negocio. [1,3].

Considerando este contexto, en el presente trabajo hemos aplicado un proceso de Big Data particular aplicado a fuentes de datos que contienen información de los consumos culturales de personas encuestadas, junto con datos de su edad, sexo y región. En particular se definieron dos objetivos específicos para el desarrollo del proceso a partir de las fuentes de datos seleccionadas: (1) analizar patrones sobre la edad y la cantidad de horas consumidas de radio, música y televisión y, (2) analizar segmentos de mercado relacionados al consumo de radio, música o televisión.

A continuación se describe como se organiza el contenido. En la sección siguiente describimos los trabajos relacionados en donde se aplican análisis sobre datos relacionados al dominio de cultura. En la Sección 3 describimos el proceso de Big Data que utilizamos a nivel general, para luego en la Sección 4 aplicarlo al contexto de las fuentes de datos seleccionadas y así lograr los objetivos propuestos. Finalmente se describen las conclusiones y trabajos futuros.

2. Trabajos Relacionados

Existen varios trabajos que aplican procesos de análisis de datos o de Big Data en el dominio de las actividades culturales. Por ejemplo, en el informe desarrollado por el Ministerio de Cultura [3], se detallan los resultados de una

encuesta, la cual constaba de 117 preguntas, en torno a las dimensiones: radio, música grabada y en vivo, diarios, libros, revistas, televisión, películas y series, cine, teatro, prácticas digitales, cultura comunitaria y videojuegos. Dentro de cada categoría se lograron distintas conclusiones, como por ejemplo, géneros musicales mas escuchados, los sitios, programas o aplicaciones que se usan para escuchar o descargar música, tipos de programas radiales mas escuchados, porcentaje de gente que fue al cine en el ultimo año, etc. Los datos de este informe muestran los contenidos mas consumidos y los que no lo son tanto, y puede ser usado para hacer análisis de mercado, ayudar al relevamiento del nivel de consumo de alguna industria, como por ejemplo el cine, y así ayudar a la toma de decisiones a la hora de tomar medidas para intentar mejorar algunos parámetros.

Por otro lado, en el articulo presentado por la Organización Mundial del Turismo [1], en el capitulo 2 se detallan 10 casos en los cuales se aplica un proceso de big data para hacer inteligencia sobre datos relacionados al consumo turístico de actividades culturales. Por ejemplo, en el caso 6 llamado *Administrando el turismo cultural a través de Big Data en Amsterdam, Holanda*, se utilizan diversos sistemas en conjunto con big data para procesar mucha información y brindarsela a los residentes y a los turistas. Por ejemplo, el sistema *Live Lines* es utilizado para saber, en tiempo real, el estado de las filas de los museos de la ciudad, advirtiendo a los visitantes que planeen sus visitas por fuera de los horarios mas concurridos. Por otra aparte, el monitoreo de los flujos de los turistas es relevante por los eventos culturales masivos. Por ejemplo, en el festival *Noord 24H* los datos del sistema global de comunicaciones móviles han demostrado que, en comparación con un fin de semana normal, durante el festival había entre 10.000 y 12.000 visitantes adicionales en la zona. Los resultados subrayaron la precisión de los datos GSM para este tipo de seguimiento, que genera datos útiles de marketing y gestión, ya que el sistema también puede medir la presencia de los visitantes en diferentes zonas de la ciudad. Esto es útil para planificar la organización de los eventos y mejorar la calidad de vida para los residentes y los turistas.

En este trabajo, a diferencia de los trabajos descriptos anteriormente, nos enfocamos en la aplicación de un proceso completo de Big Data sobre datos de consumo cultural en distintas regiones de la Argentina.

3. Proceso de Big Data

Como podemos ver en la Figura 1, el proceso de Big Data consta de una serie de fases que dividen las tareas a realizar. El enfoque elegido para la fase previa al análisis de los datos es el de ETL, el cual implica extraer los datos de la fuente, preprocesarlos para que tomen la forma en la que puedan ser consumidos, y posteriormente cargarlos en un repositorio para que sean accesibles por una aplicación. Creemos que este enfoque es el adecuado, dado que en este trabajo no implementamos aplicaciones que realizan varias llamadas hacia el repositorio donde están los datos, por lo que el costo de transformar los mismos es bajo.

Figura 1: Proceso de Big Data según un enfoque ETL

1. *Evaluación del Caso del Negocio*: se debe comenzar aprendiendo sobre el dominio del negocio, definiendo la motivación y las metas para realizar el análisis. También se deben analizar las tecnologías, costos y riesgos.
2. *Identificación y Recolección de Datos*: una vez realizada la evaluación del caso del negocio, se deben encontrar los conjuntos de datos apropiados para trabajar y las fuentes confiables de donde obtenerlos.
3. *Preparación de los datos*: una vez que se identifican las fuentes de datos, se deben extraer los datos y definir la forma y lugar donde almacenarlos. Esta fase involucra la construcción del conjunto de datos final, incluyendo la transformación, limpieza, normalización y filtrado. Todo previo a la carga en algún repositorio.
4. *Análisis de Datos*: dependiendo de la naturaleza del problema de Big Data, se lleva a cabo el análisis, el cual se puede clasificar como *confirmatorio* o *exploratorio* [4]. En el análisis confirmatorio, los datos se analizan para obtener respuestas definitivas a algunas preguntas específicas; mientras que en un análisis exploratorio, se examinan los datos para obtener información sobre por qué ocurrió un fenómeno.
5. *Modos de análisis*: luego de determinar el tipo de análisis, se debe elegir el modo de procesarlo, de acuerdo a la naturaleza del problema sobre el que estemos trabajando. Por ejemplo, podría ser un procesamiento por lotes, interactivo o en tiempo real.
6. *Visualización de los resultados*: las respuestas obtenidas suelen estar en una forma que no se puede presentar a los usuarios comunes, por lo que se requiere de representaciones gráficas para obtener valor o alguna conclusión del análisis.

En la sección siguiente describimos la aplicación de este proceso de Big Data a nuestro caso de estudio sobre datos de consumo cultural en Argentina.

4. Caso de Estudio

En la Figura 2 podemos observar gráficamente el proceso de Big Data aplicado aplicado a nuestro caso de estudio de fuentes de información que poseen datos relacionados con el consumo cultural en Argentina.

1) Evaluación del Caso del Negocio

Los objetivos definidos para la aplicación del proceso de Big Data para este dominio fueron: *(1) analizar patrones sobre la edad y la cantidad de horas consumidas de radio, música y televisión y, (2) analizar segmentos de mercado relacionados al consumo de radio, música o televisión.*

2) Identificación y Recolección de Datos

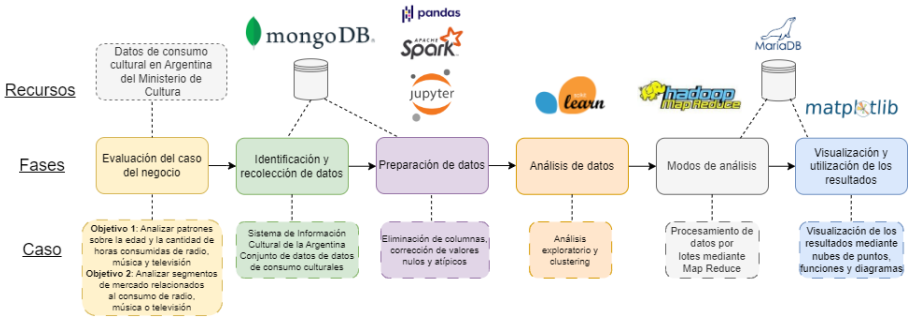


Figura 2: Actividades derivadas de las fases del proceso de Big Data, instanciado al dominio de consumos culturales en Argentina, junto con los recursos para la implementación de cada actividad

Con los objetivos planteados, recolectamos información de un sitio que provee datos públicos generados, guardados y publicados por organismos de gobierno de la República Argentina[1]. Este conjunto de datos registra los resultados de una encuesta que revela información de los consumos culturales de las personas en distintas regiones del país, en el año 2017.

Para esta fase la tecnología que utilizamos fue **MongoDB**[2], que es una base de datos categorizada como NoSQL, basada en documentos. Elegimos esta herramienta para aprovechar las ventajas que ofrecen este tipo de bases datos, que es la mayor capacidad de almacenamiento, necesaria en el contexto del presente trabajo, y su robustez ante fallos, necesario para mantener la integridad de los datos. Una vez que los datos fueron extraídos y transformados, se almacenaron en esta base de datos para que estén disponibles en las siguientes fases del proceso.

3) Preparación de los datos

Al observar y analizar las características del conjunto de datos, observamos que contenía 450 columnas, por lo que la primera decisión que se tomó fue definir cuales serian de utilidad para cumplir con los objetivos propuestos. De acuerdo a los mismos, las columnas resultantes son las que mostramos en la Tabla 1, donde por un lado tenemos datos personales del entrevistado, y por otro los datos de consumo cultural en relación a la *radio*, *música*, y *televisión*.

Luego, dentro de cada columna se realizo una limpieza y filtrado de datos, donde realizamos distintas tareas para mejorar la calidad de los mismos. Por ejemplo, analizamos la presencia de valores nulos, donde encontramos la ocurrencia solo 2 nulos en la columna *fecha*, por lo que procedimos a eliminarlas. Además, identificamos muchos nulos en las columnas donde se relevaba la cantidad de horas de consumo, debido a que si las personas contestaban 'NO' a si consumieron un determinado medio en el último año, este campo se define como

[1] `https://datos.gob.ar/dataset/cultura-encuesta-nacional-consumos-culturales/archivo/cultura_171494ce-e3cf-43fb-ad6e-f204bae1bb19`

[2] `https://www.mongodb.com/es`

Encuesta consumos culturales	
Atributo	**Descripción**
fecha	Fecha que se realizo la encuesta
región	Región donde se realizo la encuesta
sexo	Sexo del encuestado
edad	Edad del encuestado
p2	Escucho radio en el ultimo año
p3	Escucho radio por internet en el último año
p6horas	Horas por día que escucha radio
p10	Escucho musica en el último año
p12horas	Horas por día que escucha musica
p15	De diez canciones que escucha, cuantas son de artistas nacionales
p49	Miró televisión en el último año
p50horas	Horas por día que miró televisión

Cuadro 1: Características elegidas del conjunto de datos

nulo. Luego, agrupamos el contenido de cada columna por los datos que contenía, buscando inconsistencias, para su posterior corrección. Un ejemplo de todas las acciones que efectuamos para lograr lo mencionado anteriormente, es que encontramos que todos los datos que fueron cargados con acento, al mostrarlos por pantalla aparecían de la siguiente manera: 'VarÃ³n' en vez de 'Varón', por lo que se decidió renombrar todas las ocurrencias incorrectas por la misma cadena pero sin el acento.

Luego, también realizamos análisis para conocer como se distribuyen los datos, y poder observar gráficamente qué valores son los mas comunes dentro de un cierto rango, valores atípicos, etc. Por ejemplo, en la Figura 3 podemos ver la distribución de la cantidad de horas consumidas por las personas, donde observamos que la media de consumo de los 3 rubros se encuentra entre 2 y 3 horas diarias. Por otro lado, identificamos valores atípicos que se empiezan a ver después de las 7 horas de consumo, y para entender mejor su nivel de influencia sobre el conjunto de datos decidimos comparar su cantidad con respecto a la cantidad de muestras total del conjunto de datos. Luego de contar las filas, nos arrojo un resultado de 2802 muestras, y la cantidad de valores atípicos fue de 82, 165 y 113 para el consumo de radio, música y televisión respectivamente. Luego, como el numero de muestras de este tipo es pequeño en comparación al total, decidimos dejarlas para el posterior análisis.

Para la implementación de esta fase utilizamos **Pandas**[3] para aprovechar las estructuras de datos rápidas, flexibles y expresivas diseñadas para que el trabajo con datos estructurados sea más fácil e intuitivo; y **Apache Spark**[4] para el procesamiento distribuido. Todo lo trabajamos sobre el entorno **Jupyter**[5] que permitió desarrollar código en Python[6] de manera dinámica.

4) Análisis de Datos

[3] `https://pandas.pydata.org/docs/`
[4] `https://spark.apache.org/`
[5] `https://jupyter.org/`
[6] `https://www.python.org/`

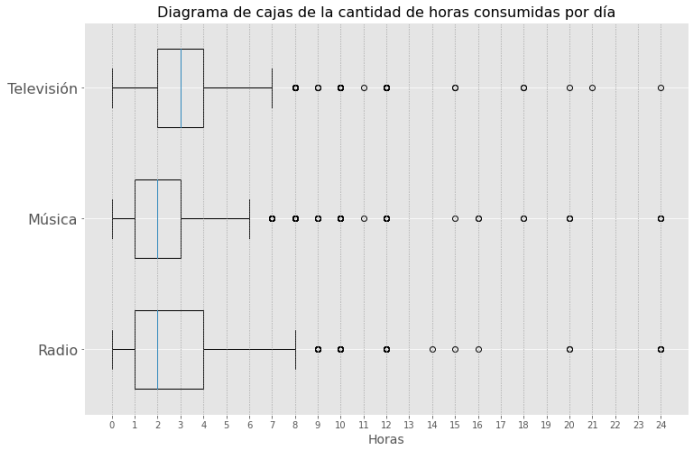


Figura 3: Distribución de los atributos horas consumidas de cada rubro

El análisis de los datos se realizó considerando los objetivos planteados en el primer paso del proceso.

Obj 1) Analizar patrones sobre la edad y la cantidad de horas consumidas de radio, música y televisión.

Para este objetivo se realizó un análisis exploratorio considerando investigar el conjunto de datos y resumir sus principales características.

Para comenzar este análisis, nos pareció interesante ver como se distribuían las preferencias de las personas en cuanto al consumo de los rubros seleccionados, lo que se puede visualizar en la Figura 4. Como podemos observar en el gráfico, el rubro musica y televisión es ampliamente consumido por la mayoría de las personas, mientras que el de radio esta un poco mas balanceado.

Por otro lado, dentro del grupo de personas que consumen cada rubro, analizamos la cantidad de horas de consumo (Figura 5). Este gráfico revela que la gran mayoría de personas mira televisión entre 3 y 4 horas, y para la musica entre 1 y 2 horas. En cambio la radio, que representa un conjunto mas reducido de personas, tiene un pico de consumo de entre 1 y 2 horas.

Siguiendo con el análisis sobre los datos, podemos observar en la Figura 6 que de las personas que escuchan radio, un 31,7 % lo hace mediante internet, por lo que vemos una transformación en la forma de consumo tradicional de este medio, que siempre fue mediante dispositivos electrónicos de sintonización de frecuencias AM o FM.

Como vimos en la Figura 4, el rubro radio tenía bastante menos consumo que la música o la televisión, por lo que decidimos analizar la edad de las personas para saber que relación tiene ese dato con este consumo. En la Figura 7 podemos observar las edades de las personas que no escuchan radio en la parte superior, y las que escuchan radio en la parte inferior. En el primer caso, vemos que las personas mas jóvenes son las que menos consumen este medio, y por otro lado, las que lo eligen tienden a ser personas de mayor edad. Si tomamos el *rango*

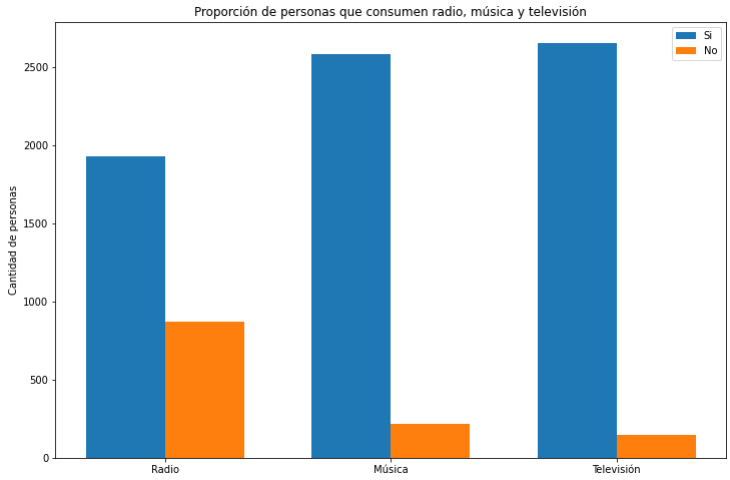


Figura 4: Distribución de las personas que consumen cada rubro

intercuartílico, la mitad de la evidencia muestra que tienen una edad de entre 25 y 60 años.

Para contrastar el anterior análisis con otro rubro, mostramos en la Figura 8 la misma distribución anterior pero con el rubro música. Podemos visualizar que hay dos distribuciones de personas bien marcados, por un lado los que no escuchan musica tienen entre 45 y 70 años, y los que si escuchan tienen entre 20 y 50 años. Por lo que hay evidencia a favor de que las personas mas jóvenes eligen consumir este rubro, por sobre la radio.

Obj 2) Analizar segmentos de mercado relacionados al consumo de radio, música o televisión.

Para analizar segmentos, decidimos usar una técnica de agrupamiento (clustering) que nos permitió identificar los diferentes grupos y sus características. Para aplicar una técnica de clustering a nuestro conjunto de datos, lo primero que tuvimos que hacer fue transformar los datos de tipo de caracteres a numérico, ya que la técnica eacepto solo ese formato y no acepta nulos. Luego quitamos algunas columnas cuyos datos no son de interés para este tipo de análisis, dichas columnas fueron: fecha, región, p3 y p15 (ver Tabla 1).

Luego iniciamos el análisis de clustering sobre los conjuntos propuestos, que son las personas que escuchan radio, música y miran televisión. Dicho análisis comenzó con la aplicación de una técnica llamada Principal Component Analisis (PCA) [5], la cual nos permitió reducir la dimensionalidad del conjunto de datos y así poder realizar gráficos en dos dimensiones para su posterior análisis. Una forma de medir que tan significativos son los componentes frente al conjunto de datos original es la varianza explicada por cada uno, la cual fue realizada en la Figura 9. La misma muestra que el primer componente explica el 40 % de la varianza y el segundo mas del 30 %, por lo que la varianza acumulada de los dos primeros supera el 70 %.

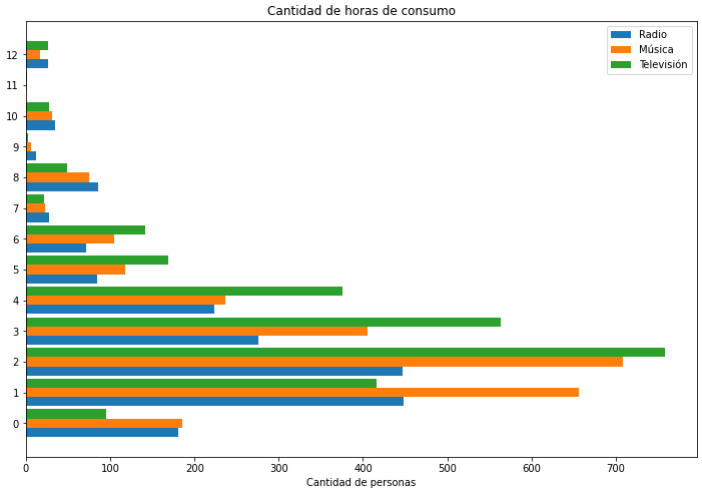


Figura 5: Cantidad de horas por día de consumo por rubro

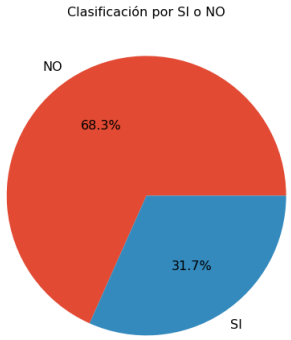


Figura 6: Cantidad de personas que escuchan radio por internet

Luego, definimos la cantidad de clusters óptimos para nuestro conjunto de datos. Para definir este valor denominado k, nos ayudamos con método del codo (elobw method), el cual nos permitió visualizar la cantidad de grupos adecuado para dividir los datos. Elegir el número de clusters es un paso fundamental para cualquier algoritmo no supervisado, ya que a priori no tenemos idea en cuántos clusters nos conviene dividir. Existen dos formas de obtener codos de los datos:

- *Usando Inercia:* Es la suma de los cuadrados de las distancias de las muestras al centro del cluster más cercano.
- *Usando Distorsión:* Es el promedio del cuadrado de las distancias desde el centro de los clusters. Usualmente se usa la métrica de la distancia euclidiana.

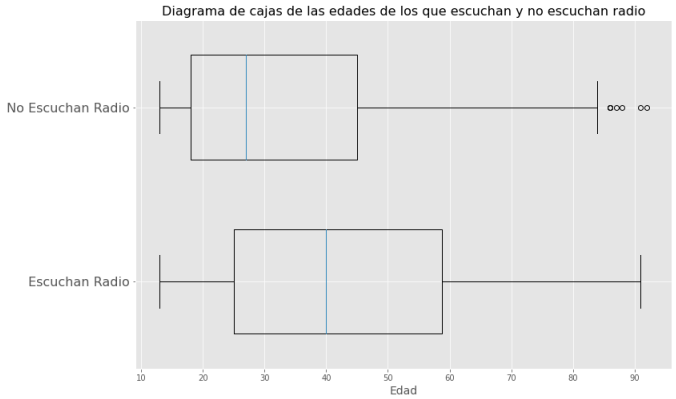


Figura 7: Distribución de las edades de las personas que consumen radio

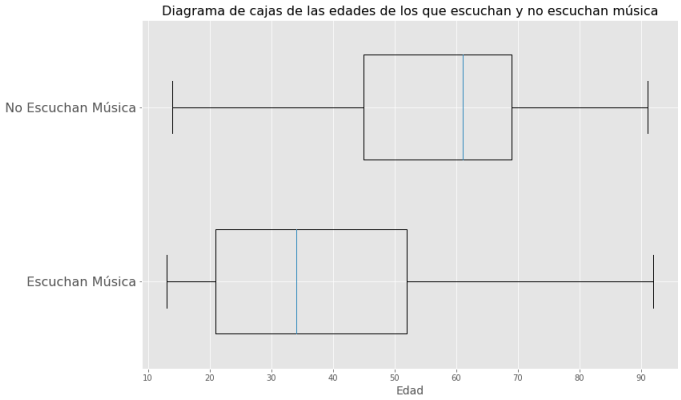


Figura 8: Distribución de las edades de las personas que consumen música

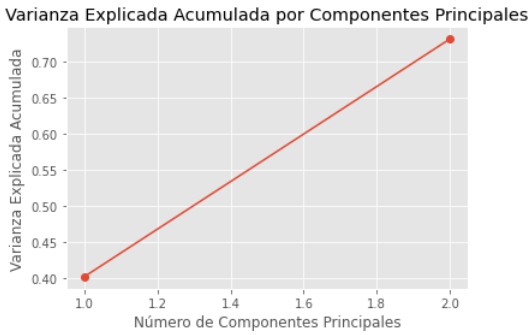
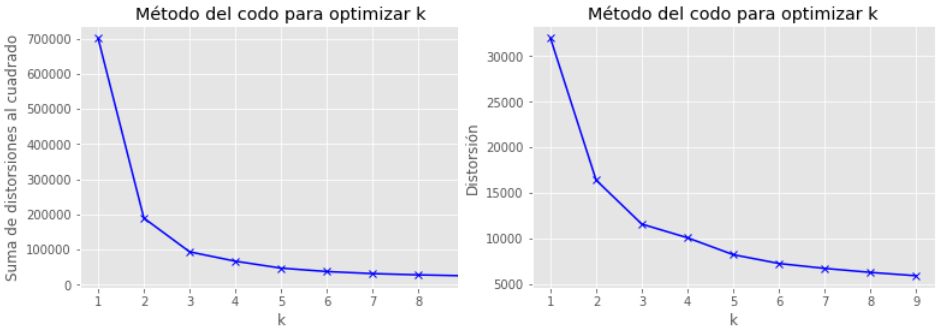


Figura 9: Varianza explicada por los componentes principales

En la Figura 10a podemos observar el resultado de método del codo utilizando inercia, y en la Figura 10b el resultado de aplicar distorsión. En ambos casos el numero óptimo de clusters es dos, por lo que el valor de k elegido fue de dos.



(a) Resultado inercia (b) Resultado distorsión

Figura 10: Método Elbow para conocer el valor óptimo de k utilizando inercia y distorsión

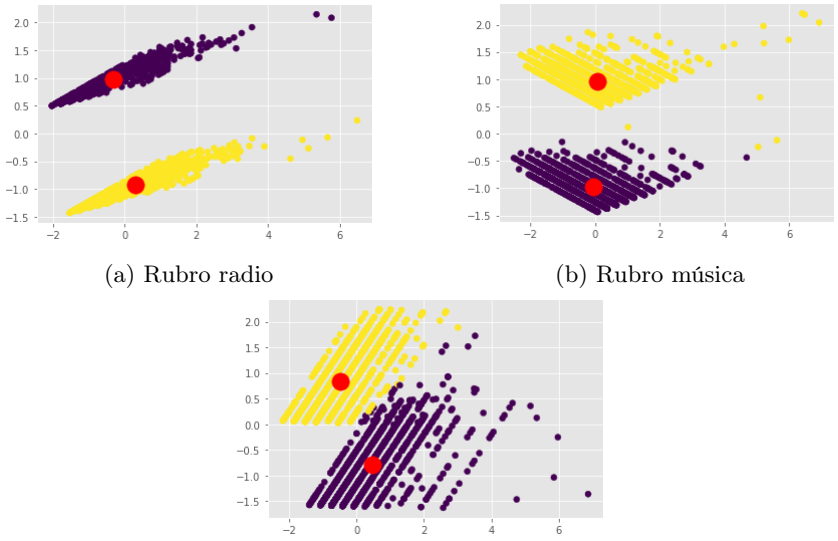
Luego, procedimos a realizar el análisis de clustering para el segmento de las personas que escuchan radio, utilizando el método *KMeans*. Este métdodo sirve para el agrupamiento o segmentación de n muestras, donde se define el valor de k (que ya obtuvimos previamente. Una muestra pertenece al cluster cuya media o "centroide" es más cercana. El algoritmo se detiene cuando cuando una nueva iteración no produce cambios. La complejidad del algoritmo es NP-Hard, por lo que se emplean diferentes heurísticas para resolverlo de manera más eficiente. El método estándar de *KMeans* comienza con un conjunto inicial de centroides ubicados utilizando partición aleatoria y se ejecutan dos acciones:

- *Paso de asignación:* Asigna cada muestra al cluster cuya media está más cercana.
- *Paso de actualización:* Calcula la nueva media del cluster con los nuevos valores asignados.

Con el valor de $k = 2$ ya definido, se entrenó el modelo con los datos, cuyo resultado se puede observar en la Figura 11. Allí vemos varios gráficos de dispersión bidimensional, donde los puntos representan los datos, los puntos rojos los centroides, y los colores violeta y azul representan dos clusters que concentran muestras con similitudes en la información que contienen. La Figura 11 muestra tres gráficos, uno por cada rubro analizado.

Considerando el rubro de radio, en la Figura Figura 11a podemos ver dos grupos bien separados. Analizando los puntos de cada grupo, las muestras pertenecientes al color violeta son personas de genero masculino, cuyo promedio de edad es de poco más de 42 años, y escuchan en promedio 3 horas continuas.

Por otro lado, las muestras amarillas son personas del genero femenino, cuyo promedio de edad es de 41 años y escuchan en promedio 2.8 horas continuas.



(a) Rubro radio

(b) Rubro música

(c) Rubro Televisión

Figura 11: Grupos creados luego de aplicar clustering para cada rubro

Luego, con respecto a las personas que escuchan música (Figura 11b), el color amarillo representa un grupo caracterizado por ser en su mayoría mujeres, con un promedio de edad aproximado de 38 años, y un consumo diario promedio de casi tres horas. El color violeta representa el otro grupo el cual se destaca por ser en su mayoría hombres, con un promedio de edad de 37 años y un consumo diario promedio de aproximadamente dos horas y media.

Finalmente, para completar el objetivo propuesto analizamos las personas que miran televisión, como vemos en la Figura 11c. El color amarillo representa un grupo caracterizado por contener en su mayoría mujeres, con un promedio de edad aproximado de 39 años, y un consumo diario promedio de poco más de tres horas. El color violeta representa el otro grupo el cual se destaca por ser en su mayoría hombres, con un promedio de edad de casi 40 años y un consumo diario promedio de aproximadamente dos horas y media.

Para el análisis de datos, utilizamos **scikit-learn**[7], que es una librería de machine learning, que brinda herramientas simples y eficientes para realizar análisis predictivo de datos.

5) Modo de Análisis

[7] `https://scikit-learn.org/stable/`

El modo de análisis elegido fue *por lotes*, el cual se produce fuera de linea y necesita que los datos estén cargados previamente en un repositorio.

6) Visualización y Utilización de los Datos.

Por último, los datos procesados se almacenaron en una base de datos **MariaDB Column Store**[8] para aprovechar los beneficios de rendimiento del almacenamiento en columnas, donde los datos se visualizarán con el uso de la librería **matplotlib**[9] que permitió la visualización de los datos de forma intuitiva y simple.

5. Conclusiones y Trabajo Futuro

En este trabajo hemos descripto las actividades realizadas sobre un proceso completo de Big Data aplicado a fuentes de información sobre datos de relevamiento de consumos culturales en Argentina. Nos centramos en los dos objetivos propuestos y en base a ellos hemos desarrollado cada fase, lo que nos permitió obtener resultados interesantes sobre el descubrimiento de patrones de consumo en relación a la edad y la cantidad de horas consumidas, y el análisis de segmentos de mercado sobre determinados subconjuntos. En relación a lo último, caracterizamos el perfil de las personas que consumían un determinado medio, con respecto a su edad, cantidad de horas de consumo y su género, obteniendo distribuciones de nubes de puntos con una dispersión variable, y fuertemente sesgada por el atributo *género* en los tres rubros propuestos.

Como trabajo futuro, usando las fuentes aplicadas aquí y posiblemente extendiendo a otras, se plantea un estudio en profundidad para la detección de nichos de consumo de determinados productos culturales.

Referencias

1. Big Data in Cultural Tourism – Building Sustainability and Enhancing Competitiveness. World Tourism Organization (UNWTO) (2021), https://www.e-unwto.org/doi/pdf/10.18111/9789284422937
2. Bahga, A., Madisetti, V.: Big data science analytics : a hands-on approach. Bahga, A. and Madisetti, V. (2016)
3. Catalano, F., Kunst, M., Mancinelli, E., Pérez, L., Laneri, P., Bonazzi, F., Castaño, A., Garido, A., Frutos, D., Grinberg, R., Unzain, D.Y., Zanabria, J.: Encuesta nacional de consumos culturales (2017), https://datos.gob.ar/dataset/cultura-encuesta-nacional-consumos-culturales/archivo/cultura_9a97dde5-3a33-4689-8333-24a2fa5b4a6e
4. Erl, T., Khattak, W., Buhler, P.: Big Data Fundamentals: Concepts, Drivers Techniques. Prentice Hall Press (2016)
5. Jolliffe, Ian T.; Cadima, J.: Principal component analysis: a review and recent developments (2016), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792409/

[8] https://mariadb.com/kb/en/columnstore-architectural-overview/

[9] https://matplotlib.org/

6. Luján-Mora, S., Vassiliadis, P., Trujillo, J.: Data mapping diagrams for data warehouse design with uml. In: Atzeni, P., Chu, W., Lu, H., Zhou, S., Ling, T.W. (eds.) Conceptual Modeling – ER 2004. pp. 191–204. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
7. Quix, C., Hai, R.: Data Lake, pp. 1–8. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-63962-8_7-1, https://doi.org/10.1007/978-3-319-63962-8_7-1
8. Simitsis, A., Skoutas, D., Castellanos, M.: Natural language reporting for etl processes. In: Proceedings of the ACM 11th International Workshop on Data Warehousing and OLAP. p. 65–72. DOLAP '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1458432.1458444, https://doi.org/10.1145/1458432.1458444
9. Trujillo, J., Luján-Mora, S.: A uml based approach for modeling etl processes in data warehouses. In: Song, I.Y., Liddle, S.W., Ling, T.W., Scheuermann, P. (eds.) Conceptual Modeling - ER 2003. pp. 307–320. Springer Berlin Heidelberg, Berlin, Heidelberg (2003)
10. Vaisman, A., Zimnyi, E.: Data Warehouse Systems: Design and Implementation. Springer Publishing Company, Incorporated, 1st edn. (2016)