Aplicación de un modelo de clusterización para clasificar las audiencias que asisten a conciertos y espectáculos en Colombia.

Lizette Ximena Gil González & Jorge Aurelio Herrera Cuartas

Facultad de Ciencias Naturales e Ingeniería

Universidad Jorge Tadeo Lozano. Bogotá, Colombia

Abstract: Este artículo representa la aplicación de un modelo de clusterización para clasificar los asistentes a conciertos, recitales y/o presentaciones musicales con base en la encuesta cultural realizada por el DANE en todas las cabeceras urbanas de Colombia en el año 2019. Como resultado de esta investigación, se encontró que los clústeres 0 y 1 tienen características similares mientras que el *clúster* 2 tiene diferencias muy particulares. Se muestra la concentración de las respuestas principalmente en los clústeres 0 y 1, especialmente en el clúster 1, haciendo este su mayor aparición en los grupos de audiencias clasificados tanto por ingresos agrupados por edades como por el gasto en boletos para conciertos y presentaciones musicales. Por su parte el clúster 2 se encontró compuesto por una pequeña audiencia en su mayoría mujeres que asisten, pero gastan en menor proporción en tiques para conciertos en las edades entre 15 y 50 años. Se concluye que el objetivo de llenar un vacío en el conocimiento de investigaciones relacionadas con el entretenimiento en el área de *Machine Learning* se ha cumplido a cabalidad, sentando esta, la base para futuros trabajos de exploración en estos ámbitos.

Keywords

Aprendizaje de máquina, clústeres, análisis de datos, aprendizaje no supervisado, Crisp-DM.

1. INTRODUCCIÓN

Colombia se caracteriza por ser un país que ofrece muchas actividades culturales producidas a nivel Nacional y traídas del exterior, especialmente para el público de sus ciudades capitales. Sin embargo, la asistencia a los mismos está determinada por el tipo de actividades que se ofrece a los asistentes y sus condiciones financieras que limita a la audiencia pues es importante tener en cuenta que este es un país en desarrollo donde los ingresos de los colombianos son distintos a los de otros países más desarrollados y donde se priorizan para otras necesidades vitales más "importantes" que para la asistencia a los diferentes eventos ofrecidos diariamente en el país.

Por otro lado, la pandemia de COVID golpeó considerablemente la industria, afectando miles de empleos directos e indirectos y especialmente a los diferentes artistas locales que viven de sus expresiones artísticas. Además, el estallido social y la situación sociopolítica no favoreció la asistencia a los eventos locales. Con el fin de esclarecer el comportamiento de los colombianos en el año 2019-2020, período de estos acontecimientos, se pretendió realizar un análisis predictivo expuesto en este artículo, tomando los datos de la encuesta de Consumo Cultural realizada anualmente por el DANE (Departamento Administrativo Nacional de Estadística) y aplicando un modelo no supervisado de aprendizaje de máquina de clusterización.

Lo anterior con el fin de agrupar en audiencias los diferentes asistentes a eventos y espectáculos musicales con base a sus datos demográficos, ingresos, gastos en conciertos y preferencias basadas en los resultados de la encuesta. El presente trabajo de investigación pretende aportar a los involucrados en la industria cultural un análisis que les dé a conocer qué tipo de audiencias en Colombia asisten a eventos culturales y así mismo, darles un aporte para que tomen las mejores decisiones con el fin de que ofrezcan a cada audiencia el evento más adecuado conforme a sus preferencias y características demográficas.

A modo de revisión de algunos trabajos anteriores relacionados con la presente investigación, en [1] analizaron la experiencia de varios jugadores de videojuegos a partir de medidas psicofisiológicas y entrenaron dichos modelos utilizando métodos de aprendizaje automático. Utilizaron medidas como la frecuencia cardíaca, la actividad electro dérmica y la actividad respiratoria, en combinación con los autoinformes (encuestas, calificaciones, grupos focales, etc.) para preparar conjuntos de entrenamiento para algoritmos de aprendizaje automático.

Los datos de entrenamiento se recopilaron de 31 participantes durante sesiones experimentales de una hora de duración, en las que jugaron varios videojuegos. Luego, entrenaron y compararon los resultados de cuatro modelos diferentes de aprendizaje automático, como red neuronal profunda de retroalimentación y un bosque de decisión aleatoria, de los cuales el mejor produjo aproximadamente un 96% de precisión. Los resultados sugieren que, de hecho, se pueden utilizar medidas psicofisiológicas para evaluar el disfrute de los consumidores de entretenimiento digital.

En [2] Describen en su trabajo como llevaron a cabo varios modelos de clasificación de videos extrayendo sus subtítulos. En esta investigación utilizaron modelos como regresión logística, k-NN, Árbol de decisión, clasificador de cresta, GBDT, *Random Forest*, Bernoulli-NB, entre otros, teniendo el mejor rendimiento el de *Random forest.* En este trabajo, se demostró cómo se pueden clasificar los vídeos extrayendo los subtítulos y se demostró que el clasificador de bosque aleatorio proporciona un rendimiento de clasificación bueno y estable.

En el artículo promoción y posicionamiento de los artistas del entretenimiento utilizando enfoques de clasificación y segmentación [3], los autores buscan desarrollar un enfoque para evaluar el potencial de los artistas del entretenimiento y construir mapas profesionales individualizados. Se utilizaron técnicas de minería de datos para correlacionar las noticias de entretenimiento en Internet con el grado de exposición y éxito de los artistas. Utilizaron segmentación de artistas del entretenimiento con el método k-*means* y construyeron modelos de predicción con ANN para comparar los resultados del enfoque k-*means* para pronosticar las categorías de artistas potenciales.

En su trabajo, [4] apunta a ayudar a los músicos independientes a planificar sus conciertos desde las perspectivas de precio y selección de la locación mediante técnicas de clasificación como *Random Forest* el cual obtuvo el mayor rendimiento mejorando el resultado de clasificación un 316%.

[5] en su proyecto de grado hizo uso de técnicas de aprendizaje automático no supervisado y supervisado para identificar patrones en los datos de pacientes que sufren trastornos cognitivos leves, a quienes se les hizo un tratamiento con juegos desarrollados para la estimulación de la atención. Utilizó la metodología CRISP-DM para poder encontrar patrones, tendencias y características en los datos de los juegos y jugadores. De los 7 algoritmos de clasificación que utilizó, encontró que el mejor fue el clasificador SVM radial (support vector Machine), el cual tuvo una exactitud del 73%.

En [6] analizaron los ratings de programas y número de episodios mediante varios modelos de predicción como regresión lineal, regresión logística, LASSO, bosques aleatorios, aumento de gradiente y máquina de vectores de soporte. Los resultados del análisis muestran que el rating promedio de los programas antes de la primera emisión se ve afectado por la compañía emisora, el rating promedio de la temporada anterior y el año de inicio.

También encontraron que el rating promedio de un programa después de la primera emisión está influenciado por el rating de la primera emisión, la empresa emisora y el tipo de programa. Adicionalmente, descubrieron que las calificaciones promedio previstas, el año de inicio, el tipo y la compañía de transmisión son variables importantes para predecir el número de episodios.

En [7] analizaron las canciones más populares de 52 países en Spotify en sus características como bailabilidad, positividad e intensidad. Para ello, propusieron un modelo de series de tiempo multivariado para predecir el tipo de música preferido en esos países en función de sus patrones previos de escucha musical y los factores contextuales. Los resultados muestran algunos cambios de comportamiento relevantes en estos patrones por el efecto de la pandemia. Además, el modelo de predicción resultante permite pronosticar el tipo de música que se escuchará en tres grupos diferentes de países en los próximos 4 meses con un error de alrededor del 1%. Estos resultados pueden ayudar a comprender mejor el consumo de música en *streaming* en empresas relacionadas con la industria de la música y el *marketing*.

Los autores del artículo Papel de diferentes factores en la predicción del éxito de una película [8], proponen un modelo de aprendizaje automático de regresión lineal multivariada para estimar el éxito

de una película considerando características como la clasificación de *IMDb*, el recuento de visitas de *Youtube* y el número de salas en las que se estrena la película. Actores, financistas, directores, etc., pueden hacer uso de estas predicciones para tomar decisiones más informadas.

Como su título lo indica, los autores del artículo Predicción del éxito de taquilla de las películas con redes neuronales [9], exponen como hicieron uso de este algoritmo de *machine learning* para determinar el desempeño financiero de una película antes de su estreno en cines. Los resultados muestran que las redes neuronales empleadas en este estudio pueden predecir la categoría de éxito de una película antes de su estreno en cines con una precisión milimétrica de 36,9 % y dentro de una categoría con un 75,2 % de precisión.

En comparación con los otros tipos de modelos (es decir, regresión logística, análisis discriminante y árboles de clasificación y regresión), utilizando exactamente las mismas condiciones experimentales, las redes neuronales funcionaron significativamente mejor.

En [10] implementaron dos enfoques de aprendizaje automático como redes neuronales (NN) y regresión de vectores de soporte (SVR) para investigar la previsión energética basada en sensores en el contexto de los lugares de organización de eventos. El enfoque se aplicó a un gran lugar de entretenimiento ubicado en Ontario, Canadá. Con datos diarios, el modelo NN logró mayor precisión que el SVR; sin embargo, con datos cada hora y cada 15 minutos, no hubo un dominio definitivo de un enfoque sobre otro. La precisión de la predicción de la demanda máxima diaria fue significativamente mayor que la precisión de la predicción del consumo.

En [11] los autores presentaron un enfoque de predicción temprana multimodal para modelar la participación de los visitantes en las exhibiciones interactivas de los museos de ciencias. Utilizaron datos de sensores multimodales que incluyen mirada, expresión facial, postura y datos de registro de interacción capturados durante las interacciones de los visitantes con una exhibición interactiva del museo para la educación en ciencias ambientales, para inducir modelos predictivos del tiempo de permanencia de los visitantes.

Investigaron técnicas de aprendizaje automático como bosque aleatorio, máquina de vectores de soporte, regresión *LASSO*, árboles de aumento de gradiente y perceptrón multicapa para inducir modelos predictivos multimodales de participación de los visitantes con datos de 85 asistentes del museo. Los modelos muestran un rendimiento predictivo mejorado a lo largo del tiempo, lo que demuestra que se pueden lograr predicciones cada vez más precisas del tiempo de permanencia de los visitantes a medida que se dispone de más evidencia de las interacciones de los visitantes con las exhibiciones interactivas de los museos de ciencias.

En [12] los autores realizaron encuestas a los individuos del área metropolitana de Chicago para analizar las actividades diarias de los encuestados por hora al día en la ciudad. Se recogieron datos detallados sobre las actividades por hora del día de más de 30.000 individuos (y 10.552 hogares). Luego determinaron que la población se pudo dividir en 8 y 7 grupos en función de sus actividades en días laborales y fines de semana. Los conglomerados resultantes combinados con información demográfica social aportan gran valor para la planificación urbana y del transporte, para la respuesta a emergencias y la dinámica de propagación, al abordar cuándo, dónde y cómo interactúan los individuos con los lugares en las áreas metropolitanas.

En [13] se analizan y discuten los clusters resultantes de tomar las reseñas en línea de las atracciones turísticas al agrupar los destinos turísticos de Florida para determinar finalmente que en cada mercado de origen surgieron los mismos tres grupos de atracciones (ocio, patrimonio y naturaleza). En [14] se analizan los diferentes comportamientos de gamers en computadores con hallazgos útiles para los desarrolladores de videojuegos. Se analizó el conjunto de datos de dos grandes títulos de juegos comerciales a los que se le aplicó k-means y Simplex Volume Maximization, combinados con consideraciones sobre el diseño de los juegos, lo que dio lugar a perfiles de comportamiento procesables.

En [15] analizan los diferentes perfiles de usuarios en las redes sociales. Los resultados mediante una segmentación latente es que los usuarios se clasifican en: introvertidos, pasivos, el versátil y el experto comunicador. Este enfoque podrá ser muy útil para las empresas que rastrean la relación entre los usuarios y sus marcas. En [16] se empleó un análisis de conglomerados para clasificar a los viajeros en función de su uso de redes sociales y otras características demográficas ya que se ha determinado que estas son cruciales para planificación de los viajes y que sus decisiones de viaje están altamente determinadas por estas redes.

En [17] se utilizó la metodología supervisada de conglomerados al utilizar los hashtags como indicadores de temas en twitter para clasificar los diferentes comentarios o también conocidos "tweets" en diferentes categorías con el fin de hacer un seguimiento de todos los mensajes publicados en microblogs y publicaciones de los amigos o contactos en la red social. En esta misma línea, en [18] realizan un estudio donde se analizan durante un año tres plataformas donde la gente se informa continuamente sobre tendencias actuales como Twitter, Wikipedia y Google. Se basaron en una técnica de pronóstico del vecino más cercano donde los autores suponen que los temas semánticamente similares tienen un comportamiento similar. Demostraron que, en un conjunto de datos de estadísticas de visitas a Wikipedia, los pronósticos realizados con el enfoque propuestos son entre 9 y 48k visitas comparando con otros enfoques de referencia y se logró un error porcentual promedio del 45 al 19% para períodos de tiempo de hasta 14 días.

Se encontró un estudio muy relacionado con el presente trabajo de investigación, donde en [19] se exploraron los diferentes perfiles y variables de calidad del entretenimiento para eventos en vivo como lo son calidad de la actuación, personal e instalaciones para eventos de 13.428 encuestados para determinar su satisfacción, intenciones y recomendaciones para asistir a un evento en cada uno de los lugares donde se llevan a cabo. Se utilizó la técnica de conglomerados k-means para determinar que resultaron cuatro segmentos de clientes los cuales diferían significativamente y lo cual puede ayudar a los administradores de los lugares donde se llevan a cabo estos espectáculos a comprender mejor cómo se diferencian sus clientes para el marketing estratégico futuro.

Finalmente, y a modo de revisión teórica, se consultó un artículo [20] en el cual se explican los pasos para realizar el análisis de conglomerados los cuales son: selección de variables, gestión de datos, selección de métodos de agrupamiento, obtención de soluciones de conglomerados, validación de resultados e interpretación de los resultados. El artículo ofrece al lector un ejemplo de cómo se procesó cada uno de los pasos en la identificación y descripción de tres segmentos de mercado basados en las preferencias de entretenimiento. El propósito del artículo es fomentar un mayor uso de este valioso medio para revelar nichos de mercado por parte de los investigadores en el campo de la hospitalidad.

2. PLANTEAMIENTO DEL PROBLEMA

Actualmente la oferta cultural que se distribuye por todo el país no tiene un análisis de audiencias estructurado. Se ofertan eventos supremamente costosos donde cada día aumenta el valor de los boletos. Se desconocen las características demográficas de los asistentes a los diferentes tipos de conciertos musicales y eventos culturales que se ofertan por todo el país.

Por otro lado, la pandemia de COVID-19 y su consecuente crisis económica, afectó en gran medida la industria, al cancelar la oferta y asistencia a estos conciertos. Miles de empleos directos e indirectos se vieron fuertemente afectados incluyendo el de los artistas. Por ello, la presente investigación se pretende centrar en el comportamiento de los asistentes a eventos previo a la aparición del virus, antes de la "nueva normalidad".

La idea al hacer este trabajo es esclarecer las diferentes características poblacionales de los diversos grupos de colombianos según sus preferencias y asistencia a eventos en todas las cabeceras urbanas del país. Esto para futuras bases investigativas en el sector cultural del país aportará para que

instituciones públicas y privadas de la industria, tomen decisiones sobre como ofertar de la manera más acertada eventos para toda la clasificación de audiencias sugeridas.

Otro objetivo al realizar el presente análisis es abordar desde la disciplina del análisis de datos el hallazgo de nichos de mercado que beneficien a diferentes industrias en este caso la del entretenimiento mediante técnicas de *machine learning* que ayuden a los diferentes Gerentes en las organizaciones y CEOs a tomar las mejores decisiones donde los equipos de data los ayuden a dejar de lado labores operativas y que se puedan enfocar en ofrecer verdadero valor a los diferentes consumidores de sus productos y servicios.

Por otra parte, se pretende llenar un vacío en el conocimiento sobre las técnicas de clusterización de aprendizaje de máquina para aquellos que hasta ahora se están sumergiendo en el mundo del análisis de datos y la Inteligencia Artificial. Se quiere también aportar al movimiento de artistas a que sus expresiones culturales sean vistas por el sector de consumidores más interesados en sus ofertas y a que mejoren su calidad de vida en el país a pesar de las problemáticas sociales, políticas y económicas que encara el país. También se pretende establecer una base investigativa en el país que aporte para futuras investigaciones en el sector del entretenimiento.

3. METODOLOGÍA

La presente investigación se trabajó bajo el modelo CRISP-DM siguiendo los diferentes pasos que propone esta metodología para lo que corresponde al análisis de datos.

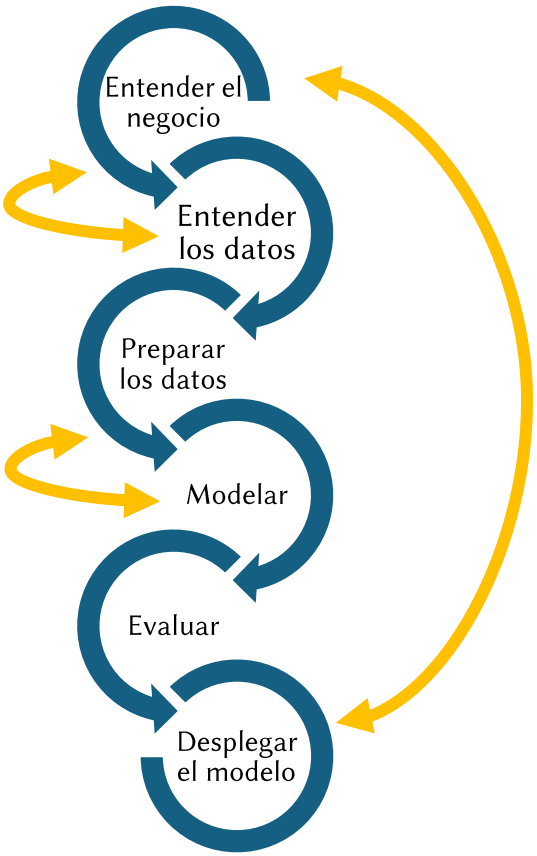


Figura 1. Diagrama de la metodología CRISP-DM para proyectos de análisis de datos

3.1 Comprensión del negocio

Durante el 2023 y lo que va de este año, el sector del entretenimiento en vivo ha crecido en gran medida rompiendo récords de los años anteriores hasta convertirse en el tercer segmento que más

impulsa la economía en el país. Esto debido a la recuperación del sector después de su declive proveniente de la pandemia y el interés de los asistentes por retomar sus actividades recreativas.

Según datos de Tuboleta, la empresa líder en Colombia vendedora de *tickets* para eventos como espectáculos musicales y deportes, el 27% de los compradores destinan su dinero a la categoría de conciertos. Por su parte, los géneros más consumidos en cuanto a espectáculos se tratan son: pop (24.2%), reggaetón (19.1%) y Rock (18.4%).

Por su parte, también analizaron las edades que más asisten a este tipo de conciertos encontrando que el 63% de compradores de entradas a este tipo de eventos están entre los 18 y 34 años. Los artistas internacionales cada vez apetecen más venir al país a presentarse lo que ha hecho que el 2023 haya habido un aumento del 11% en el recaudo de la industria en su categoría de conciertos, representado en $900,992 millones de pesos en este aspecto para Tuboleta. [21]

3.2 Comprensión de los datos

Para esta etapa de la metodología, se seleccionaron las bases de datos a trabajar. En primera instancia, se tomaron las bases relacionadas con la información demográfica de los encuestados y se seleccionaron las columnas con las características que se querían tomar en cuenta. El primer archivo que se importó fue el de la base *Características generales.csv* que se muestra en la Tabla 1.

	DIRECTORIO	PERSONA_NUMERO	P6020	P5785	P5465	P5501	P6070	P6170	P260	P261
0	3086229	1	2	71	6	1	6	2	3	5
1	3086230	1	1	34	6	1	3	2	5	2
2	3086230	2	2	41	6	2	3	2	5	2
3	3086230	3	1	2						
4	3086231	1	1	36	6	1	4	2	5	2

Tabla 1. Datos sobre las características generales de la población encuestada

Cada código o número acompañado con la letra P corresponde a las siguientes preguntas de la encuesta:

P6020 Sexo 1. Hombre 2. Mujer

P5785 ¿Cuántos años cumplidos tiene?

P5465 De acuerdo con su cultura, pueblo o rasgos físicos usted se reconoce como: 1. Indígena 2. Gitano 3. Raizal del archipiélago de San Andrés, Providencia y Santa Catalina 4. Palenquero de San Basilio o descendiente 5. Negro, mulato, afrocolombiano o Afrodescendiente 6. Ninguna de las anteriores

P5501 ¿Cuál es el parentesco de...con el(la) jefe(a) del hogar? 1. Jefe del hogar 2. Pareja, esposo, cónyuge, compañero 3. Hijo o hijastro 4. Nieto 5.Otro pariente 6. Empleado del servicio doméstico y sus parientes 7.Pensionista, compañero del pensionista 8.Trabajador 9.Otro no pariente

P6070 Actualmente: 1. No está casado y vive en pareja hace menos de dos años 2. No está casado y vive en pareja hace dos años o más 3. Está casado 4.Está separado o divorciado 5.Está viudo 6.Está soltero

P6170 Actualmente asiste al preescolar, escuela, colegio o universidad? 1.Sí 2. No

P260 Cuál es el nivel educativo más alto alcanzado por 1. Ninguno 2. Preescolar 3.Básica primaria(1-5) 4.Básica secundaria (6-9) 5.Media(10-13) 6.Superior(Técnica, Tecnológica, Universitaria-Pregrado) 7.Posgrado(especialización, maestría, doctorado) 99. No sable / No informa

P261 ¿Cuál es el último año o grado aprobado en este nivel?

La segunda base que se tomó fue la de *hogares.csv* con las siguientes columnas:

	DIRECTORIO	P6008	P259
0	3086229	1	1
1	3086230	3	2
2	3086231	1	1
3	3086238	3	2
4	3086239	5	4

Tabla 2. Columnas seleccionadas con la información de los hogares de los encuestados

P6008 Total de personas en el hogar

P259 Total de personas de 12 años y más en el hogar

En tercer lugar, se seleccionó la base *viviendas.csv* con las siguientes columnas:

	DIRECTORIO	REGION	P4000	P4031S1A1
0	3086229	4	2	1
1	3086230	4	2	1
2	3086231	4	2	1
3	3086238	4	2	1
4	3086239	4	2	1

Tabla 3. Columnas seleccionadas con las características de las viviendas de los encuestados

REGION 1. Bogotá 2.Caribe 3.Oriental 4.Central 5.Pacífica 6.Amazonía/Orinoquía

P4000 Tipo de vivienda 1. Casa 2.Apartamento 3.Cuartos 4.Vivienda indigena 5.Otra vivienda(carpa, vagón, embarcación, cueva, refugio natural, etc.)

P4031S1A1 Estrato para la tarifa 0. conexión ilegal-pirata 1. Estrato1 2. Estarto2 3. Estrato3 4. Estrato4 5. Estrato5 6. Estrato6 9. No se puede establecer estrato.

Para realizar un análisis descriptivo de los datos fue necesario unir los anteriores dataframes con la base que contenía las preguntas de la encuesta relacionada con la asistencia a eventos y espectáculos culturales para lo cual se utilizó la función merge en Python tomando como columna de referencia la de DIRECTORIO.

```
#Unión de Dataframes demográficos con espectaculos

caracteristicas_hogares=pd.merge(caracteristicas_generales_filtradas,hogares_filtradas, on='DIRECTORIO',how='outer')
plus_viviendas=pd.merge(caracteristicas_hogares,viviendas_filtradas, on='DIRECTORIO',how='outer')
demograficos_espectaculos=pd.merge(plus_viviendas,espectaculos1, on='DIRECTORIO',how='outer')
```

Algoritmo 1. Unión de las tablas hogares, viviendas y tabla espectáculos

Los datos quedaron con duplicados por lo que fue necesario realizar el procedimiento de eliminación de datos duplicados y de conversión de las respuestas numéricas a variables categóricas y booleanas en el caso de las respuestas de verdadero y falso para las preguntas relacionadas con la asistencia a los eventos.

```
#Eliminación de duplicados

index_def5=demograficos_espectaculos[demograficos_espectaculos['P5785']<12].index
Espectaculos_mayoresde12=demograficos_espectaculos.drop(index_def5)

index_def6=Espectaculos_mayoresde12[Espectaculos_mayoresde12['PERSONA_NUMERO']!= Espectaculos_mayoresde12['PERSONA_NUMERO_ESP']].index
Espectaculos_sinduplicados=Espectaculos_mayoresde12.drop(index_def6)

Espectaculos_final1=Espectaculos_sinduplicados.drop_duplicates(subset=['DIRECTORIO','PERSONA_NUMERO','P5785','PERSONA_NUMERO_ESP'],keep='first')
Espectaculos_final = Espectaculos_final1.replace(r'^\s*$', np.nan, regex=True)

pd.set_option('display.max_columns', None)
#demograficos_espectaculos
Espectaculos_final
```

Algoritmo 2. Código para la eliminación de filas duplicadas en el *Dataframe* consolidado

Se procedió a realizar diferentes gráficos que mostraran un análisis descriptivo de los datos de las bases unidas en un solo DataFrame. El primer gráfico que se proyectó fue una nube de palabras para determinar en qué actividad ocupa la mayor parte del tiempo los encuestados. Este fue el resultado:

La mayor parte del tiempo los encuestados permanecían realizando oficios del hogar o trabajando.

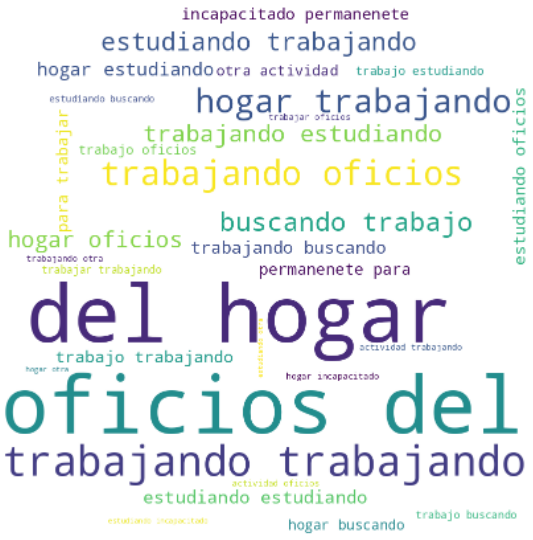
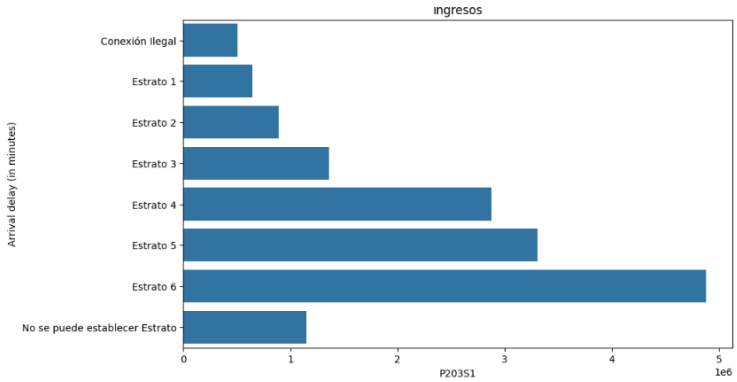


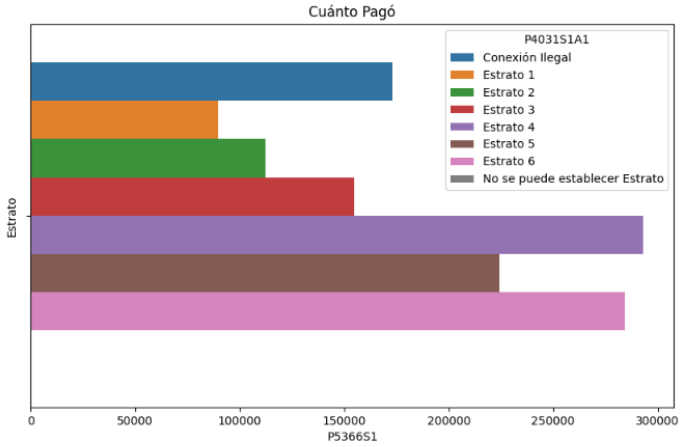
Figura 2. Nube de palabras con la actividad que los encuestados realizaron la mayor parte de su tiempo.

Por otro lado, se analizaron los ingresos basados en el estrato de los encuestados. Se pudo evidenciar que los estratos 4, 5 y 6 son los que más ingresos reciben según la gráfica resultante:



Gráfica 1. Ingresos de los encuestados distribuidos por estratos

A la pregunta de cuánto pagó por asistir a conciertos se desglosó por estratos dando como resultado que los estratos 4 y 6 son los grupos que más invierten en conciertos y espectáculos según la gráfica:



Gráfica 2. Pago de tiques para conciertos distribuido por estratos

3.3 Preparación de los datos

En esta fase fue necesario trabajar con los datos vacíos y tomar decisiones sobre cómo gestionarlos para poder prepararlos para el siguiente modelo de *clustering* que se implementó.

Lo primero que se hizo fue crear una copia con la unión de los DataFrames de los datos demográficos y las respuestas a la encuesta con sus respuestas en formato numérico. Para los espacios vacíos en las respuestas de las encuestas se reemplazaron por el número 100 y en el caso de las columnas donde las respuestas eran datos numéricos (monetarios), se complementaron los espacios en blanco con el promedio de los datos de cada columna, como se muestra en el siguiente ejemplo, se replicó en las demás columnas con valores de dinero:

```
Copia_Espectaculos = Espectaculos_final1.copy()

Copia_Espectaculos = Copia_Espectaculos.replace(r'^\s*$',int(100), regex=True)

Copia_Espectaculos.replace(np.nan, 100, inplace=True)



promedio1=sum(Copia_Espectaculos['P203S1'])/len(Copia_Espectaculos['P203S1'])
promedio=round(promedio1,0)
Copia_Espectaculos['P203S1'].replace(100, promedio, inplace=True)
Copia_Espectaculos['P203S1'].replace(np.nan, promedio, inplace=True)
Copia_Espectaculos['P203S1'] = Copia_Espectaculos['P203S1'].replace(r'^\s*$', promedio, regex=True)
Copia_Espectaculos['P203S1']=pd.to_numeric(Copia_Espectaculos['P203S1'],errors='coerce')
```

Algoritmo 3. Código de realización de la copia de los datos y su preparación para el modelo.

A continuación, se convirtieron todos los datos en formato numérico para evitar errores en el momento de aplicar el modelo de *clustering* mediante el siguiente código:

```
Copia_Espectaculos['DIRECTORIO']=pd.to_numeric(Copia_Espectaculos['DIRECTORIO'],errors='coerce')
Copia_Espectaculos['PERSONA_NUMERO']=pd.to_numeric(Copia_Espectaculos['PERSONA_NUMERO'],errors='coerce')
Copia_Espectaculos['P6020']=pd.to_numeric(Copia_Espectaculos['P6020'],errors='coerce')
Copia_Espectaculos['P5785']=pd.to_numeric(Copia_Espectaculos['P5785'],errors='coerce')
Copia_Espectaculos['P5465']=pd.to_numeric(Copia_Espectaculos['P5465'],errors='coerce')
Copia_Espectaculos['P5501']=pd.to_numeric(Copia_Espectaculos['P5501'],errors='coerce')
Copia_Espectaculos['P6070']=pd.to_numeric(Copia_Espectaculos['P6070'],errors='coerce')
Copia_Espectaculos['P6170']=pd.to_numeric(Copia_Espectaculos['P6170'],errors='coerce')
Copia_Espectaculos['P260']=pd.to_numeric(Copia_Espectaculos['P260'],errors='coerce')
Copia_Espectaculos['P261']=pd.to_numeric(Copia_Espectaculos['P261'],errors='coerce')
Copia_Espectaculos['P6008']=pd.to_numeric(Copia_Espectaculos['P6008'],errors='coerce')
Copia_Espectaculos['P259']=pd.to_numeric(Copia_Espectaculos['P259'],errors='coerce')
Copia_Espectaculos['REGION']=pd.to_numeric(Copia_Espectaculos['REGION'],errors='coerce')
Copia_Espectaculos['P4000']=pd.to_numeric(Copia_Espectaculos['P4000'],errors='coerce')
Copia_Espectaculos['P4031S1A1']=pd.to_numeric(Copia_Espectaculos['P4031S1A1'],errors='coerce')
Copia_Espectaculos['PERSONA_NUMERO_ESP']=pd.to_numeric(Copia_Espectaculos['PERSONA_NUMERO_ESP'],errors='coerce')
Copia_Espectaculos['P6240']=pd.to_numeric(Copia_Espectaculos['P6240'],errors='coerce')
Copia_Espectaculos['P203']=pd.to_numeric(Copia_Espectaculos['P203'],errors='coerce')
Copia_Espectaculos['P203S1']=pd.to_numeric(Copia_Espectaculos['P203S1'],errors='coerce')
Copia_Espectaculos['P5355']=pd.to_numeric(Copia_Espectaculos['P5355'],errors='coerce')
Copia_Espectaculos['P5355S1']=pd.to_numeric(Copia_Espectaculos['P5355S1'],errors='coerce')
Copia_Espectaculos['P5355S2']=pd.to_numeric(Copia_Espectaculos['P5355S2'],errors='coerce')
Copia_Espectaculos['P204S1']=pd.to_numeric(Copia_Espectaculos['P204S1'],errors='coerce')
Copia_Espectaculos['P204S2']=pd.to_numeric(Copia_Espectaculos['P204S2'],errors='coerce')
Copia_Espectaculos['P204S3']=pd.to_numeric(Copia_Espectaculos['P204S3'],errors='coerce')
Copia_Espectaculos['P204S4']=pd.to_numeric(Copia_Espectaculos['P204S4'],errors='coerce')
Copia_Espectaculos['P204S5']=pd.to_numeric(Copia_Espectaculos['P204S5'],errors='coerce')
Copia_Espectaculos['P204S6']=pd.to_numeric(Copia_Espectaculos['P204S6'],errors='coerce')
Copia_Espectaculos['P204S11']=pd.to_numeric(Copia_Espectaculos['P204S11'],errors='coerce')
Copia_Espectaculos['P204S7']=pd.to_numeric(Copia_Espectaculos['P204S7'],errors='coerce')
Copia_Espectaculos['P204S9']=pd.to_numeric(Copia_Espectaculos['P204S9'],errors='coerce')
Copia_Espectaculos['P204S10']=pd.to_numeric(Copia_Espectaculos['P204S10'],errors='coerce')
Copia_Espectaculos['P204S8']=pd.to_numeric(Copia_Espectaculos['P204S8'],errors='coerce')
Copia_Espectaculos['P5360']=pd.to_numeric(Copia_Espectaculos['P5360'],errors='coerce')
Copia_Espectaculos['P5360S1']=pd.to_numeric(Copia_Espectaculos['P5360S1'],errors='coerce')
```

Algoritmo 4. Conversión de los datos a formato numérico para aplicarlos al modelo

Paso seguido, se procedió a eliminar las columnas donde sus celdas con valores iguales a 100 es decir nulos, fuera mayor a 10 espacios con este valor mediante el siguiente código:

```
count_100 = Copia_Espectaculos.eq(100).sum()

# Obtener nombres de columnas que tienen más de 10 repeticiones de 100
columnas_a_eliminar = count_100[count_100 > 10].index


# Eliminar las columnas del DataFrame original
Copia_Espectaculos = Copia_Espectaculos.drop(columns=columnas_a_eliminar)
```

Algoritmo 5. Eliminación de espacios mayores a 10 con valores de 100 en sus celdas que correspondían a valores nulos.

El número de columnas se redujo a 44, resultando las de mayor valor para aplicarlas al modelo.

3.4 Modelo

El modelo que se implementó para la clasificación de los diferentes encuestados en audiencias fue el de clusterización. El primer paso fue obtener la lista de variables categóricas del Dataframe resultante del preprocesamiento de los datos con las siguientes líneas de código:

```
#MODELO MACHINE LEARNING CLUSTERING

#Obtener lista de variables categóricas
s = (Copia_Espectaculos.dtypes == 'object')
object_cols = list(s[s].index)

print("Variables categóricas en el conjunto de datos:", object_cols)
```

```
Variables categóricas en el conjunto de datos: []
```

Algoritmo 6. Obtención de variables categóricas en el modelo

En segunda instancia se realizaron las siguientes líneas para determinar que todos los datos fueran numéricos:

```
#Label Encoding los tipos de objeto.
LE=LabelEncoder()
for i in object_cols:
    Copia_Espectaculos[i]=Copia_Espectaculos[[i]].apply(LE.fit_transform)

print("Todas las funciones ahora son numéricas")
```

```
Todas las funciones ahora son numéricas
```

Algoritmo 7. Validación de datos numéricos

En tercer lugar, se realizó el algoritmo para escalar las variables a partir de una copia de los datos:

Todas las características ahora están escaladas

	DIRECTORIO	PERSONA_NUMERO	P6020	P5785	P5465	P5501	P6070	P6170	P260	P6008	P259	REGION	P4000	PERSONA_NUMERO_ESP	P6240	P203	P203S1	P5355	P5360	P5360S1
0	-1.698475	-0.845505	0.931005	1.699947	0.350487	-0.886774	1.136311	0.510302	-0.578446	-1.492120	-1.480579	0.36135	1.062232	-0.845505	1.121196	-0.374039	-0.324689	0.368842	-0.068127	-0.085534
1	-1.698154	-0.845505	-1.074108	-0.287584	0.350487	-0.886774	-0.598558	0.510302	0.105270	-0.425909	-0.801900	0.36135	1.062232	-0.845505	-0.948711	-0.374039	0.082269	0.368842	-0.068127	-0.085534
2	-1.698154	-0.108638	0.931005	0.088435	0.350487	-0.193883	-0.598558	0.510302	0.105270	-0.425909	-0.801900	0.36135	1.062232	-0.108638	1.121196	-0.341859	-0.254318	0.368842	-0.068127	-0.085534
3	-1.697834	-0.845505	-1.074108	-0.180150	0.350487	-0.886774	-0.020268	0.510302	0.105270	-1.492120	-1.480579	0.36135	1.062232	-0.845505	-0.948711	-0.374039	-0.467649	0.368842	-0.068127	-0.085534
4	-1.695590	-0.845505	-1.074108	-0.395018	0.350487	-0.886774	-1.176847	0.510302	0.447128	-0.425909	-0.801900	0.36135	1.062232	-0.845505	-0.948711	-0.374039	0.420416	0.368842	-0.068127	-0.085534
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
22792	2.020866	-0.108638	0.931005	0.410737	0.350487	0.499008	1.136311	0.510302	0.105270	-0.959014	-0.801900	0.36135	1.062232	-0.108638	1.121196	-0.374039	-0.658262	0.368842	-0.068127	-0.085534
22793	2.021186	-0.845505	-1.074108	0.357020	0.350487	-0.886774	-0.020268	0.510302	-0.578446	-1.492120	-1.480579	0.36135	-0.790821	-0.845505	-0.948711	2.779576	-0.254318	0.368842	-0.068127	-0.085534
22794	2.021507	-0.845505	0.931005	-0.341301	0.350487	-0.886774	1.136311	0.510302	0.105270	-0.425909	-0.801900	0.36135	1.062232	-0.845505	1.121196	-0.374039	-0.505771	0.368842	-0.068127	-0.085534
22795	2.021507	-0.108638	-1.074108	-1.093340	0.350487	0.499008	1.136311	0.510302	0.447128	-0.425909	-0.801900	0.36135	1.062232	-0.108638	1.121196	-0.341859	-0.254318	0.368842	-0.068127	-0.085534
22796	2.021828	-0.845505	-1.074108	1.162776	0.350487	-0.886774	-0.020268	0.510302	0.788986	-1.492120	-1.480579	0.36135	-0.790821	-0.845505	-0.948711	-0.374039	2.105624	-2.711190	-0.068127	-0.085534

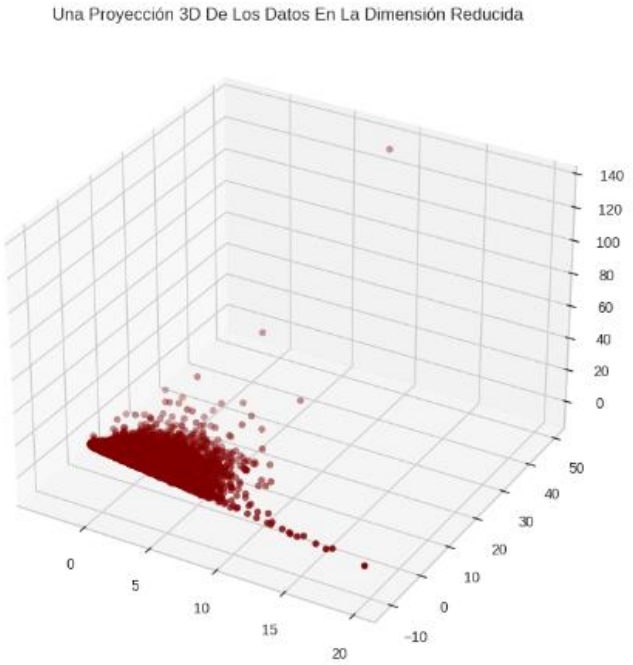
22797 rows × 42 columns

Tabla 4. Variables escaladas

A continuación, se realizó un PCA para reducir la dimensionalidad de los datos escalados esto con el fin de poderlos graficar en un esquema 3D como se mostrará a continuación:

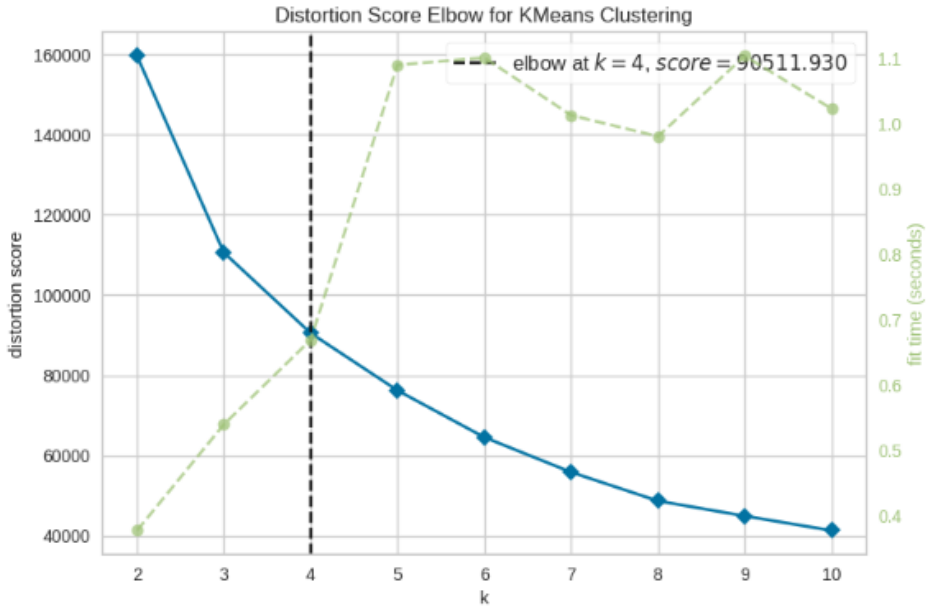
	count	mean	std	min	25%	50%	75%	max
col1	22797.0	-1.221796e-16	2.036307	-3.478646	-1.599409	-0.522231	1.388482	19.837334
col2	22797.0	4.986921e-17	1.964994	-11.211102	-1.108899	-0.381503	0.664351	49.422854
col3	22797.0	4.986921e-17	1.384245	-7.092347	-0.411462	0.030385	0.373842	133.998289

Tabla 5. Reducción de la dimensionalidad de los datos escalados a tres columnas



Gráfica 3. Proyección gráfica de la dimensionalidad reducida en 3D

Se aplicó el método del codo en Python para determinar el número de clústeres necesarios para el presente análisis de clasificación, resultando 4 grupos:

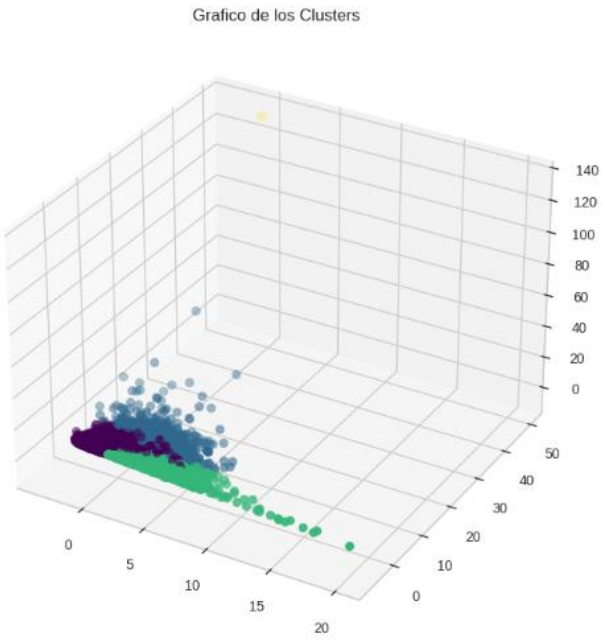


Gráfica 4. Número de grupos para el análisis de clasificación

Después, se incia el modelo de Clustering Aglomerativo para determinar los conglomerados y para agregar la función de los clústeres al marco de datos original y para luego graficarlos

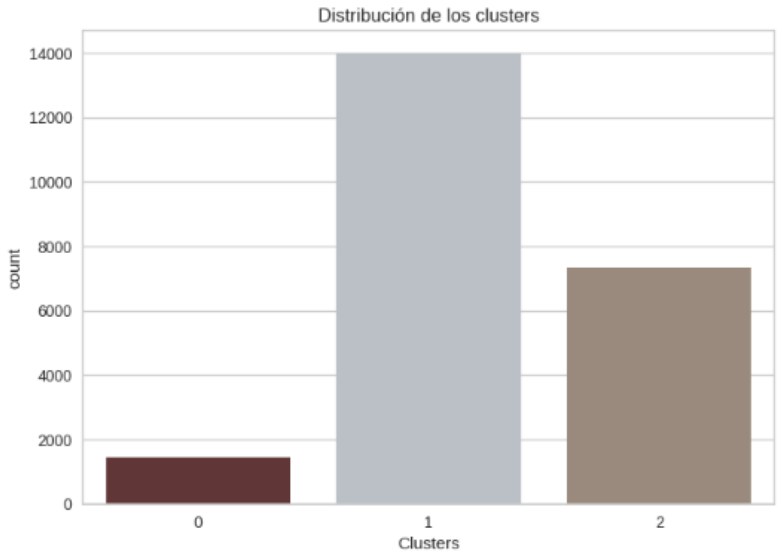
```
[69] #Iniciando el modelo de Clustering Aglomerativo
     AC = AgglomerativeClustering(n_clusters=3)
     # ajuste el modelo y estimación de los conglomerados
     yhat_AC = AC.fit_predict(PCA_ds)
     PCA_ds["Clusters"] = yhat_AC
     #Agregando la función de clústeres al marco de datos original.
     Copia_Espectaculos["Clusters"]= yhat_AC
```

Algoritmo 8. Determinación de los conglomerados para graficarlos



Gráfica 5. Proyección en 3D de los Conglomerados determinados

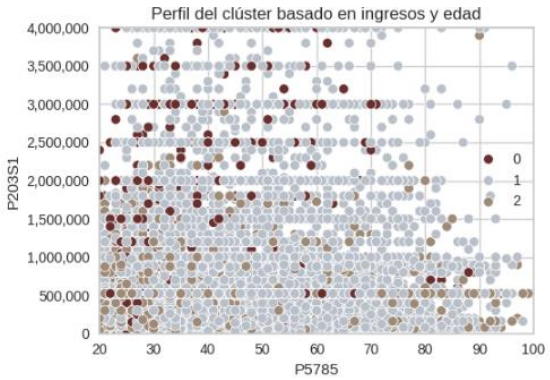
Luego, se hace la distribución de los clústeres en los datos originales



Gráfica 6. Distribución de los conglomerados en los datos de la encuesta

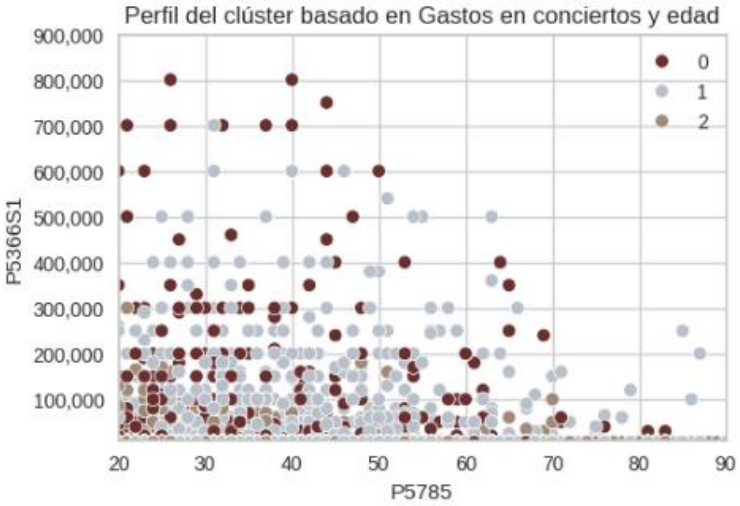
Para saber cómo se distribuyen los clústeres en los datos originales se realizaron diferentes gráficas de dispersión para saber a qué grupo corresponde cada rango de los datos:

Por ejemplo, en la siguiente gráfica se representa el ingreso distribuido por edades donde se puede ver que el clúster predominante es el 1 el cual cubre casi todas las edades y va desde los cero ingresos hasta los 4 millones acotando los datos ya que es el ingreso promedio que reciben los colombianos.



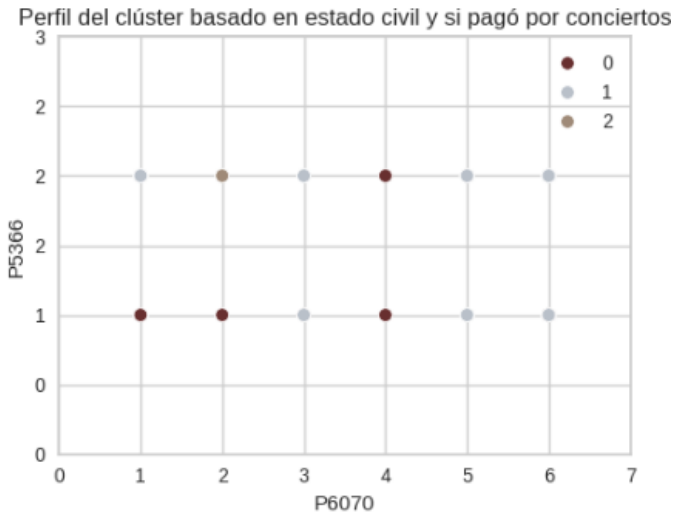
Gráfica 7. Distribución de los clústeres en los ingresos de los encuestados por edades

En la siguiente gráfica se puede establecer que el clúster 0 está más presente en las edades más jóvenes que invierten una cantidad menor en conciertos contrario al clúster 1 que muestra mayor gasto en edades más adultas. En cuanto al clúster 2, muestra un gasto muy bajo en todas las edades.



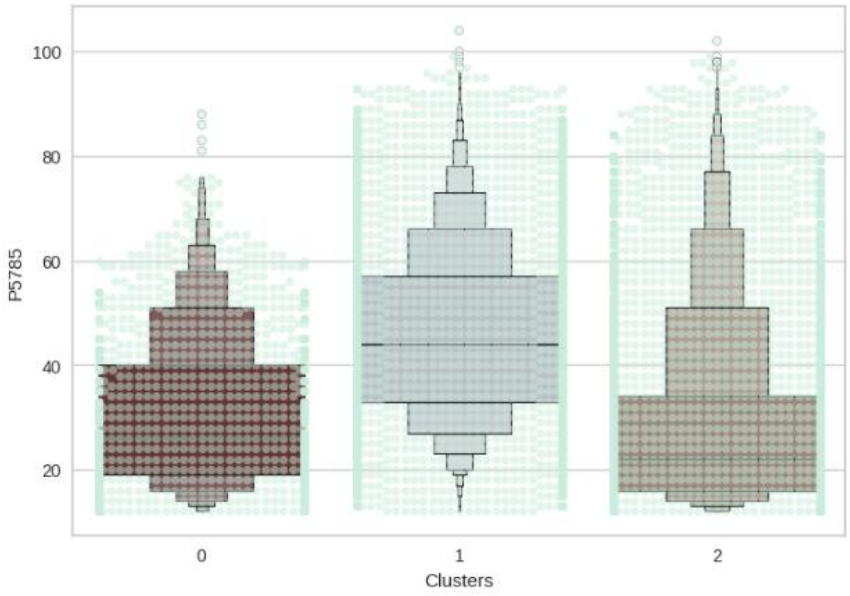
Gráfica 8. Distribución de los clústeres en los datos de los gastos en tiques para conciertos por edades.

Se analizó también la asistencia de los encuestados a conciertos dependiendo de su estado civil. La gráfica muestra el clúster 1 presente en la asistencia a conciertos por parte del grupo 1, 2 y 4 que corresponden a No casados con pareja viviendo menos de dos años, No casado con pareja viviendo más de dos años y personas separadas-divorciadas respectivamente. Por otra parte, el clúster 2 representa los grupos 3, 5 y 6 que corresponden a Casados, Viudos y Solteros respectivamente.



Gráfica 9. Distribución de los clústeres en los datos de pago en tiques para conciertos por estado civil.

A continuación, en el modelo, se grafican los boxplots para analizar la distribución de los clústeres a lo largo de las edades para saber en qué rango de los datos predomina cada grupo y determina la media de los datos en cada conglomerado.

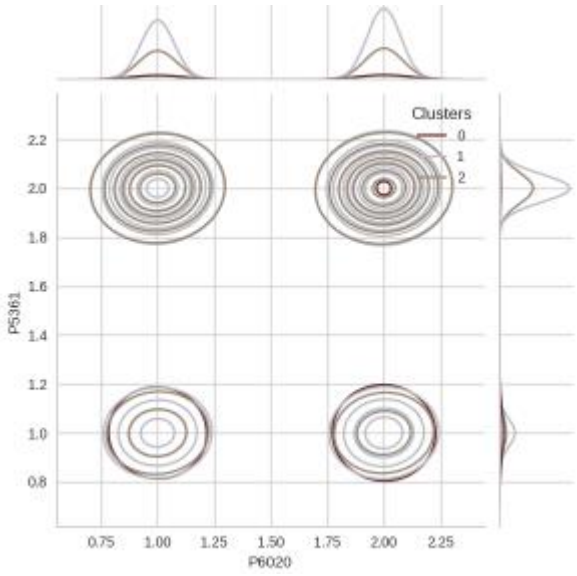


Gráfica 10. Distribución de los datos en los clústeres por edades.

3.5 Despliegue y Resultados

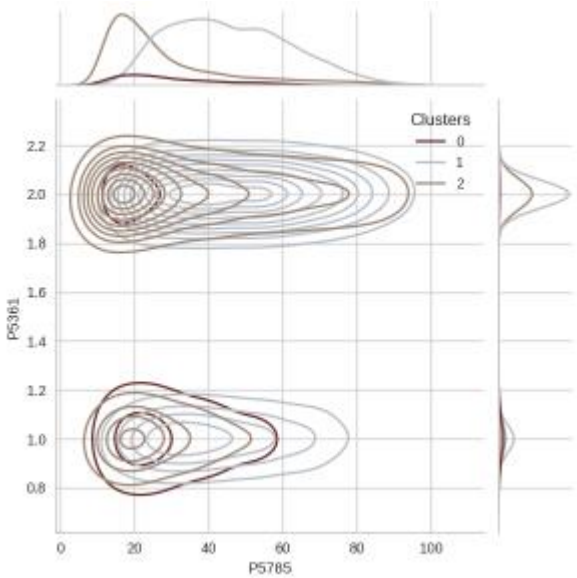
Como última instancia se procedió a aplicar los clústeres a diferentes datos de la encuesta cultural relacionándolos con las características demográficas para encontrar patrones y para realizar la clasificación de las audiencias correspondientes a los 3 clústeres.

Se analizó la asistencia de los encuestados a conciertos, recitales y/o presentaciones de música en espacios abiertos o cerrados en vivo. Esta pregunta se analizó para diferentes características demográficas de los colombianos encuestados.



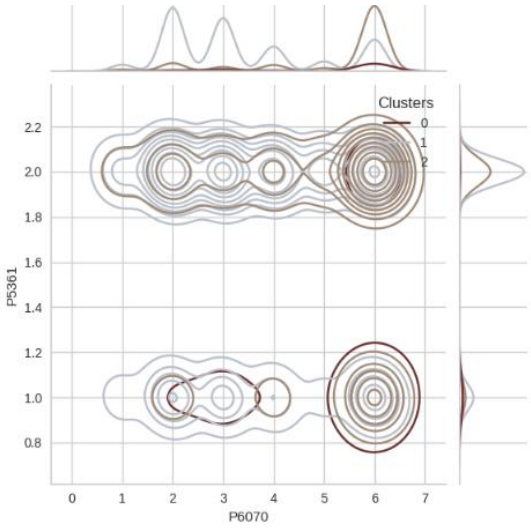
Gráfica 11. Distribución de los clústeres en la respuesta a la pregunta 5361 desglosada por la característica demográfica de género de los encuestados.

En primera instancia se analizó la asistencia o no a conciertos y espectáculos clasificando las respuestas entre mujeres y hombres. La anterior figura representa en su eje x los géneros Masculino (1.0) y femenino (2.0). En el eje Y se muestran las opciones verdadero (1.0) y Falso (2.0) a la pregunta con el código P5361. Según como se muestra, el clúster 2 corresponde a mujeres que en su mayoría asisten a conciertos, mientras que en los grupos 0 y 1 se encuentran los hombres asistentes a este tipo de eventos.



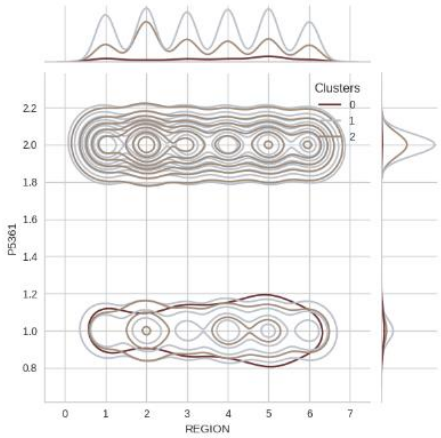
Gráfica 12. Distribución de los clústeres en la respuesta a la pregunta 5361 desglosada por la característica demográfica de las edades de los encuestados.

En la anterior gráfica se analizó la pregunta de la asistencia o no a conciertos, recitales y/o presentaciones de música para diferentes edades. El clúster 0 muestra predominancia en la asistencia a conciertos entre las edades de 15 y 59 años. El clúster 1 muestra un rango más amplio de edades que irían entre los 18 y los 79 años. Por su parte, el clúster 2 está entre los rangos de los 15 a los 50 años de quienes asisten a este tipo de eventos.



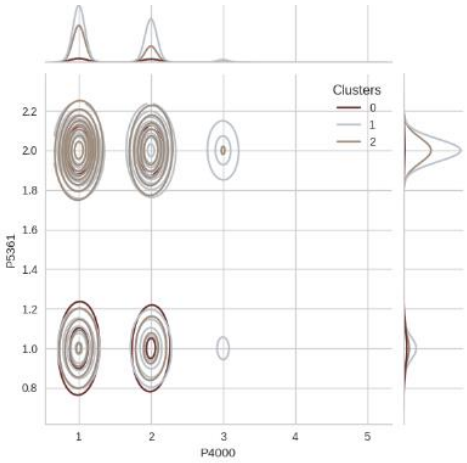
Gráfica 13. Distribución de los clústeres en la respuesta a la pregunta 5361 desglosada por la característica demográfica de estado civil de los encuestados.

En la anterior gráfica se muestra la asistencia a conciertos y presentaciones de música de acuerdo con el estado civil de los encuestados. Se puede evidenciar que hay una concentración de los clústeres en la opción 6 lo cual indica que los solteros son quienes más asisten a conciertos.



Gráfica 14. Distribución de los clústeres en la respuesta a la pregunta 5361 desglosada por las regiones donde se encontraban los encuestados.

A la pregunta desglosada por Región se puede observar que los clústeres se distribuyen a lo largo de todas las regiones determinando que los encuestados son de variadas partes del país especialmente hablando del clúster 0. Por su parte los encuestados del clúster 1 pertenece principalmente a las regiones del caribe, la parte Oriental del país, Central, pacífica y Amazonía. Los asistentes a conciertos del clúster 2 pertenecen a las regiones Caribe, Central y Pacífica principalmente.



Gráfica 15. Distribución de los clústeres en la respuesta a la pregunta 5361 desglosada por el tipo de vivienda de los encuestados.

Los asistentes a conciertos y presentaciones de música que pertenecen a los 3 clústeres viven en Casas y apartamentos y para el clúster 1 una menor proporción de encuestados viven en un Cuarto. Esto muestra la proporción de encuestados que viven en las cabeceras urbanas del país.

4. CONCLUSIONES

Al realizar el despliegue del modelo, se evidenció que las características de los asistentes a los conciertos y pertenecientes a los diferentes grupos son similares, mostrando una mayor concentración de las respuestas principalmente en los clústeres 0 y 1 especialmente en el clúster 1, haciendo este su mayor aparición en los grupos de audiencias clasificados tanto por ingresos agrupados por edades como por el gasto en boletos para conciertos y presentaciones musicales. Por su parte, el clúster 2 realizó una menor presencia siendo el grupo de audiencias conformado por mujeres principalmente de edades entre 15 a 50 quienes menos gastaron en conciertos en el 2019.

En esta investigación se pretendía clasificar los diferentes asistentes a conciertos y espectáculos en función de sus ingresos, características demográficas y gastos monetarios en este tipo de eventos porque se quería determinar qué grupo era más predominante y qué características cumplen especialmente aquellos que tiene ingresos promedio con el fin de incentivar su asistencia a eventos culturales y para ayudar a los artistas a ofrecer sus contenidos al público más acorde según sus preferencias. Podría concluirse que mediante la técnica utilizada esto se cumplió a cabalidad.

Por otra parte, se realizó un modelo de clusterización porque era el más acertado para cumplir con los objetivos de la presente investigación. También se pretendió llenar un vacío en la investigación de nichos clasificados para el sector de entretenimiento en Colombia ya que se encontraron pocos proyectos relacionados hechos en el país y se considera este es otro de los objetivos logrados que ayudará como base para futuras investigaciones.

REFERENCIAS

[1] M. Čertický, M. Čertický, P. Sinčák, G. Magyar, J. Vaščák, and F. Cavallo, "Psychophysiological indicators for modeling user experience in interactive digital entertainment," *Sensors (Switzerland)*, vol. 19, no. 5, Mar. 2019, doi: 10.3390/s19050989.

[2] R. Agarwal, M. P. L V N P, and S. Yunus Sait, "Video Classification into Academic and Entertainment using Subtitles," 2021.

[3] A. J. C. Trappey, C. V Trappey, A.-C. Chang, F.-L. Lee, H.-C. Hsieh, and M.-H. Chao, "PROMOTING AND POSITIONING ENTERTAINMENT ARTISTS USING CLUSTERING AND CLASSIFICATION APPROACHES," 2012.

[4] X. Yang and Q. Ge, "A Concert-planning Tool for Independent Musicians by Machine Learning Models," Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.11200

[5] "Universidad Jorge Tadeo Lozano Facultad de Ciencias Naturales e Ingeniería." 14

[6] M. Kim, S. Lim, C. Jang, and J. Song, "A study on entertainment TV show ratings and the number of episodes prediction," *The Korean Journal of Applied Statistics*, no. 6, pp. 809–825, 2017, doi: 10.5351/KJAS.2017.30.6.809.

[7] F. Terroso-Saenz, J. Soto, and A. Muñoz, "Evolution of global music trends: An exploratory and predictive approach based on Spotify data," Entertain Comput, vol. 44, p. 100536, 2023, doi: https://doi.org/10.1016/j.entcom.2022.100536.

[8] A. Bhave, H. Kulkarni, V. Biramane, and P. Kosamkar, "Role of different factors in predicting movie success," in 2015 International Conference on Pervasive Computing: Advance Communication Technology and Application for Society, ICPC 2015, Institute of Electrical and Electronics Engineers Inc., Apr. 2015. doi: 10.1109/PERVASIVE.2015.7087152.

[9] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," Expert Syst Appl, vol. 30, no. 2, pp. 243–254, Feb. 2006, doi: 10.1016/j.eswa.2005.07.018.

[10] K. Grolinger, A. L'Heureux, M. A. M. Capretz, and L. Seewald, "Energy Forecasting for Event Venues: Big Data and Prediction Accuracy," *Energy Build*, vol. 112, pp. 222–233, 2016, doi: https://doi.org/10.1016/j.enbuild.2015.12.010.

[11] A. Emerson *et al.*, "Early Prediction of Visitor Engagement in Science Museums with Multimodal Learning Analytics," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, in ICMI '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 107–116. doi: 10.1145/3382507.3418890.

[12] S. Jiang, J. Ferreira, and M. C. González, "Clustering daily patterns of human activities in the city," Data Min Knowl Discov, vol. 25, no. 3, pp. 478–510, 2012, doi: 10.1007/s10618-012-0264-z.

[13] A. P. Kirilenko, S. O. Stepchenkova, and J. M. Hernandez, "Comparative clustering of destination attractions for different origin markets with network and spatial analyses of online reviews," *Tour Manag*, vol. 72, pp. 400–410, 2019, doi: https://doi.org/10.1016/j.tourman.2019.01.001.

[14] A. Drachen, R. Sifa, C. Bauckhage, and C. Thurau, "Guns, swords and data: Clustering of player behavior in computer games in the wild," in *2012 IEEE Conference on Computational Intelligence and Games (CIG)*, 2012, pp. 163–170. doi: 10.1109/CIG.2012.6374152.

[15] M.-C. Alarcón-del-Amo, C. Lorenzo-Romero, and M.-Á. Gómez-Borja, "Classifying and Profiling Social Networking Site Users: A Latent Segmentation Approach," *Cyberpsychol Behav Soc Netw*, vol. 14, no. 9, pp. 547–553, Feb. 2011, doi: 10.1089/cyber.2010.0346.

[16] S. Amaro, P. Duarte, and C. Henriques, “Travelers’ use of social media: A clustering approach,” *Ann Tour Res*, vol. 59, pp. 1–15, 2016, doi: https://doi.org/10.1016/j.annals.2016.03.007.

[17] K. Dela Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, “Topical Clustering of Tweets,” 2011.

[18] T. Althoff, D. Borth, J. Hees, and A. Dengel, “Analysis and forecasting of trending topics in online media streams,” in *Proceedings of the 21st ACM International Conference on Multimedia*, in MM ’13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 907–916. doi: 10.1145/2502081.2502117.

[19] K. Kim and E. D. Tucker, “Assessing and segmenting entertainment quality variables and satisfaction of live event attendees: A cluster analysis examination,” *Journal of Convention & Event Tourism*, vol. 17, no. 2, pp. 112–128, Apr. 2016, doi: 10.1080/15470148.2015.1101035.

[20] C. Jurowski and A. Z. Reich, “An Explanation and Illustration of Cluster Analysis for Identifying Hospitality Market Segments,” *Journal of Hospitality & Tourism Research*, vol. 24, no. 1, pp. 67–91, 2000, doi: 10.1177/109634800002400105.

[21] A. Cardona, “Tuboleta revela datos del negocio del ‘ticketing’ en Colombia: así son los gastos en eventos y entretenimiento,” Jun. 12, 2024. Accessed: Aug. 18, 2024. [Online]. Available: https://www.valoraanalitik.com/tuboleta-datos-ticketing-conciertos-eventos-en-colombia/