Predicción y clasificación de conductas juveniles

David Santiago López, Isabella Martínez.

Universidad del Rosario
Escuela de Ingeniería, Ciencia y Tecnología

November 28, 2020

1 Resumen ejecutivo

Este proyecto pretende estudiar los gustos de los jóvenes eslovacos del año 2013 con respecto a la música, las películas y los hobbies, para esto, en primer lugar, se utilizará el método de aprendizaje no supervisado *k-medoids* para realizar agrupamiento de los usuarios con base en sus gustos y así, poder predecir los gustos de otros usuarios. En segundo lugar, se utilizará PCA y FA para no solo realizar una reducción de dimensionalidad sino también para ver la estructura de la varianza de los datos. Por último, se realizará un análisis de correspondencia para evaluar ciertas hipótesis sobre algunos estereotipos de la sociedad y además, encontrar otros patrones en los gustos de cada uno de los aspectos mencionados anteriormente y el género de los jóvenes eslovacos.

2 Introducción y descripción del problema

En la era de las redes sociales masivas se ha vuelto muy popular por parte de estas el uso de algoritmos de predicción de comportamiento con el fin de mostrarle a los usuarios publicidad efectiva. ¿Cómo es que Netflix o Spotify saben qué recomendarte? ¿Cómo es que Google te muestra la publicidad de algo que justo necesitabas comprar y en oferta?

Todas estas estrategias de marketing se basan en algoritmos que analizan los datos que tan amablemente las personas proporcionan como pago por utilizar una plataforma de forma gratuita, y de esta manera se vuelve tan lucrativo aunque no estés pagando por usarlo.

Es conveniente saber a qué "tribu social" pertenece alguien con el fin de recomendarle amigos, o poder predecir sus hábitos de consumo dependiendo de sus intereses para mostrarle ofertas.

3 Datos a usar

Se utilizará una encuesta realizada en 2013 a jóvenes eslovacos entre 15 y 30 años [2], fue realizada tanto en formato físico como digital. Se les preguntó sobre preferencias musicales, preferencias de películas, intereses y hobbies, fobias, hábitos de salud, rasgos de personalidad, puntos de vista sobre la vida, opiniones, hábitos de consumo, y demografía.

Así pues el dataset cuenta con 1010 personas entrevistadas y 150 variables, de las cuales 139 son ratings y 11 son categóricas. En la mayoría de las preguntas se les pidió a los entrevistados un número del 1 (Muy en desacuerdo) al 5 (Totalmente de acuerdo) con respecto a determinadas afirmaciones (por ejemplo, "Disfruto escuchar música").

4 Análisis de datos

4.1 Clustering y Clasificación

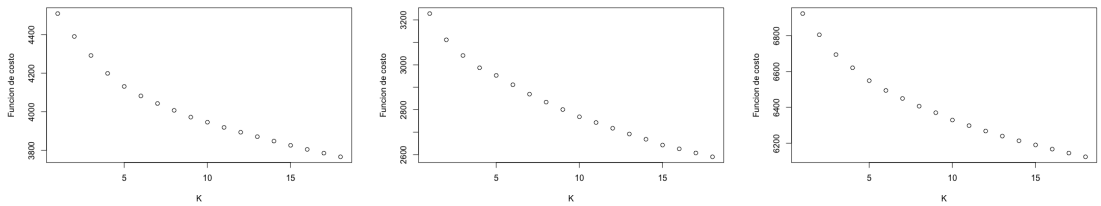
Esta sección pretende analizar los resultados obtenidos al realizar clustering y clasificación, técnicas vistas a lo largo del curso. Antes de empezar, se le pretende recordar al lector que clustering es una técnica de aprendizaje no supervisado la cual nos permite agrupar las observaciones en clusters que son generados a partir de la información obtenida por medio de las variables predictoras. Clasificación, a diferencia de clustering, es un a técnica de aprendizaje supervisado, la cual, como su nombre lo indica, pretende clasificar observaciones en 2 o más poblaciones (clases).

Ahora bien, con respecto al proyecto, se decidió dividir cada uno de los dataframes (musica, películas y hobbies) en 3 partes: training (70 %), validation (20 %) y simulation (10 %), con el fin de realizar clustering con los datos de training y validation (entrenar el modelo) para identificar grupos de usuarios que tuviesen gustos similares en cada uno de estos aspectos y así generar etiquetas a cada una de las observaciones (etiquetas que están basadas en los clusters formados), posteriormente crear un modelo de clasificación (LDA, QDA), el cual es entrenado con solamente los datos de training y evaluado, por medio del error aparente (APER) que está basado en la matriz de confusión, con los datos de validation. Finalmente, se realizó una simulación de como se comportaría el modelo de clasificación con datos que no conociera y no tuviese previamente etiquetas, para esto se utilizaron los datos de simulation.

Cabe mencionar que en un inicio se intentó realizar dicho agrupamiento por medio del método *K-means* y cluster jerárquico, sin embargo, debido a que los datos presentaban *outliers* y además, las variables a trabajar eran categóricas, estos no se comportaron del todo bien, por lo tanto, se decidió buscar un método un poco mas robusto; *K-medoids* es un método de clustering muy parecido a *K-means* (minimiza distancias entre observaciones), sin embargo, este asigna como

centroide la mediana y no la media del grupo, además, a diferencia de *K-means* es capaz de recibir diversas métricas de distancia (no solo la euclidiana). Así mismo, el modelo de clasificación utilizado, fue LDA (Análisis de discriminante lineal) Y QDA (Análisis de discriminate cuadrático).

Una pregunta muy frecuente cuando se quiere realizar clustering y que no es tan fácil de resolver es cuántos clusters se quieren formar con los datos, muchas veces es difícil definir el K para aplicar algún método de clustering, en especial cuando los datos están en una dimensionalidad mayor a 3 y no se puede tener una visualización pura de estos. En este caso, cada observación estaba en $\mathbf{R}^{18}$ (musica), $\mathbf{R}^{11}$ (películas) y $\mathbf{R}^{32}$ (hobbies) y aunque se intentaron diferentes técnicas de reducción de dimensionalidad no fue posible diferenciar visualmente estos clusters por lo que se decidió utilizar el criterio del codo para definir el K. Las diferentes gráficas obtenidas se muestran a continuación.



(a) Musica (b) Películas (c) Hobbies

Figure 1: Criterio del codo para la selección del K para el método K-medoids

Por lo tanto, a partir del criterio del codo, se decidió utilizar $K = 8$ para la musica, $K = 6$ para las películas y $K = 8$ para los hobbies y se realizo el respectivo clustering.

Ahora bien, visualizar estos clusters es de suma importancia para darse una idea de cómo fueron agrupados los datos (gusto de los usuarios por música, películas y hobbies) y asimismo la estructura que tienen. Como se menciono anteriormente, los datos correspondientes a la música están en $\mathbf{R}^{18}$, a las películas en $\mathbf{R}^{11}$ y a los hobbies en $\mathbf{R}^{32}$, por lo cual se decidió utilizar el método TSNE, el cual hace una muy buena reducción de dimensionalidad para visualizarlos (solo sirve para visualizar). Los resultados con un análisis de cada uno de estos se muestran a continuación.

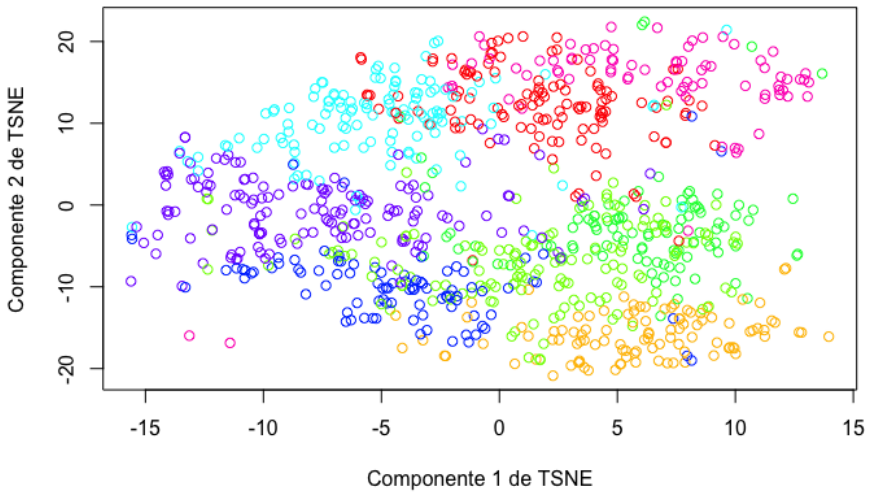


Figure 2: Proyección de los datos referentes a los gustos por la musica en R^2

En la gráfica anterior se pueden identificar los clusters formados, en donde cada color corresponde a un cluster. Podemos observar que algunos clusters están bien definidos mientras otros se solapan entre sí. Ahora bien, es de suma importancia tener en cuenta que esto es una aproximación de los datos en R^2 y es probable que al realizar la proyección mucha información se pierda. Por lo que, aunque en la gráfica mostrada anteriormente no se vea tan clara la división entre cada cluster esto no indica que la agrupación no sea adecuada.

Teniendo en cuenta lo anterior, se decidió evaluar este modelo por medio de su silueta, la cual se muestra a continuación.

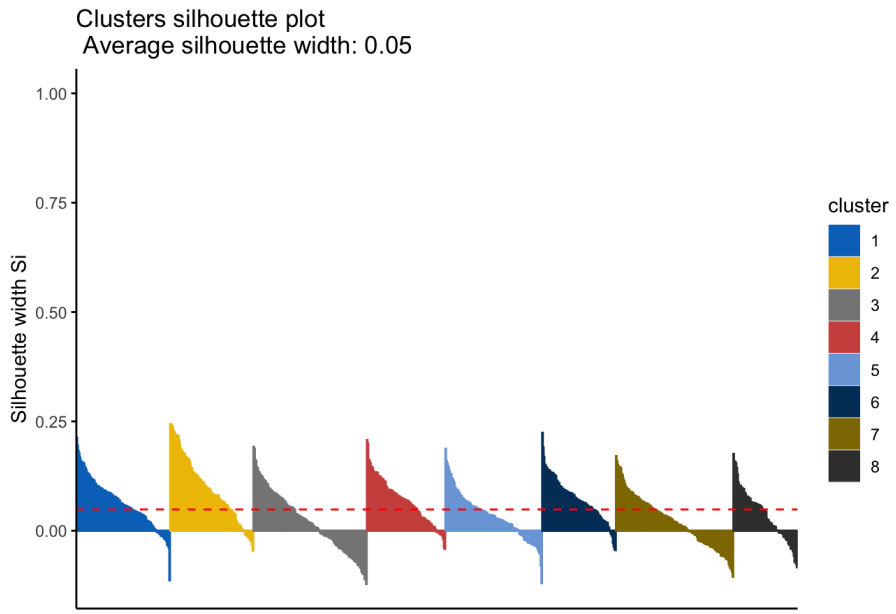
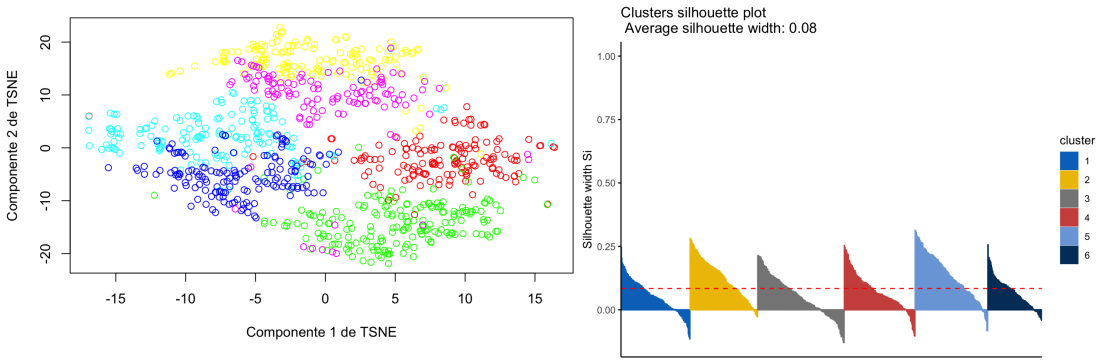


Figure 3: Silueta del modelo para la musica

En este caso, se puede observar que el modelo se comporta relativamente bien, puesto que aunque en la silueta se observan colas negativas, estas son de muy pocas observaciones y a su vez, no alcanzan valores tan negativos, en cambio, en la mayoría de los datos de cada cluster los valores de la silueta superan el promedio, lo cual nos indica que esas observaciones están más cerca a las observaciones de su cluster que a la de otros clusters.

Ahora bien, con respecto a las películas, la proyección de los datos en R^2 y la silueta del modelo se muestra a continuación.



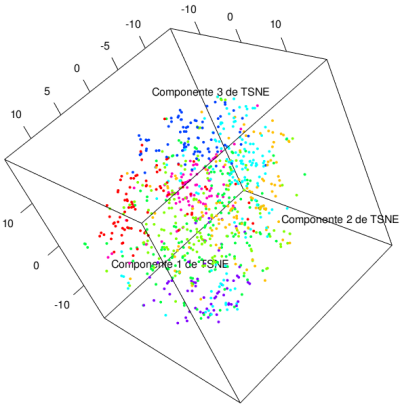
(a) Proyección de los datos referentes a los gustos por las películas en R^2

(b) Silueta del modelo para las películas

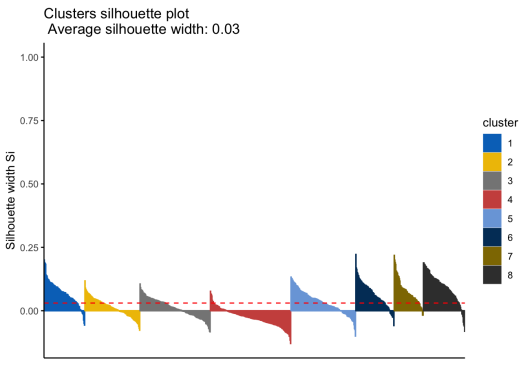
Figure 4: Análisis de clustering para películas

Podemos ver en este caso que los clusters se identifican mucho mejor en comparación con los de los gustos por la música y aunque algunos clusters están solapados, estas observaciones tienden a ser insignificantes comparados con la gran mayoría. De igual forma, en la silueta se puede ver que el modelo tiene un comportamiento bastante bueno, pues aunque se presentan colas negativas estas se presentan en muy pocas observaciones. Teniendo en cuenta esto, podemos concluir que la estructura de los datos de las películas es más separable en grupos con pequeña varianza entre las observaciones dentro de ellos pero con gran varianza entre las observaciones fuera de ellos, lo cual hace que un algoritmo de aprendizaje no supervisado como clustering tenga mejores resultados.

Por último, se muestran los resultados obtenidos para los hobbies. En este caso, estos datos pertenecen a R^{32}, esta es una dimensión considerablemente grande comparada con la de las películas y la música. Esto causa que una reducción a dos dimensiones no sea suficiente para interpretar las observaciones, por lo tanto, para este conjunto de datos se decidió realizar una visualización en 3D.



(a) Proyección de los datos referentes a los gustos por los hobbies en R^3



(b) Silueta del modelo para los hobbies

Figure 5: Análisis de clustering para los hobbies

Se puede observar que en este caso muchos de los clusters están solapados en la gráfica en 3D, además, en este caso la silueta del modelo presenta más colas negativas que en los anteriores, esto nos indica que existen observaciones de algunos clusters que están mas cerca a las observaciones de otros clusters distintos al suyo. Por lo tanto, podríamos decir que *k-medoids* no logra agrupar del todo bien las observaciones en el caso de los hobbies.

Ahora bien, basados con el agrupamiento realizado anteriormente, se fijaron las etiquetas para cada una de las observaciones para así crear modelos LDA y QDA con los datos de training y validarlo con los datos de validation, con el fin de evaluar como se comportarían los modelos

si se conocieran las etiquetas. Cabe mencionar que esta practica esta sujeta a la suposición de que *k-medoids* logró agrupar de forma correcta las observaciones, cosa que como se explicó anteriormente no sucedió en todos los casos, por lo tanto, es importante tener en cuenta que lo que se hizo fue una simulación con dicha suposición, lo cual no nos permite analizar el comportamiento real de los modelos. Los errores de clasificación obtenidos se muestran en la siguiente tabla.

Modelo	**Error Musica**	**Error Películas**	**Error Hobbies**
LDA	17%	10%	23%
QDA	33%	21%	43%

Table 1: Análisis Factorial de Música

Por lo tanto, en todos los dataframes trabajados se comportó mejor LDA.

4.2 Técnicas de Reducción de Dimensionalidad

En esta sección se buscará aplicar técnicas de reducción de dimensionalidad a los datasets considerados previamente, pues todos contienen más de 10 variables. La reducción de dimensionalidad es la transformación de datos de un espacio de alta dimensión en un espacio de baja dimensión, de modo que dicha representación preserva algunas propiedades significativas de los datos originales. En el presente proyecto se utilizó Análisis de Componentes Principales (PCA) y Análisis de Factores (FA), y como se pudo ver previamente se aplicó TSNE para propósitos de visualización.

Es importante recalcar que aunque PCA y FA tienen el mismo objetivo, sus enfoques adoptados son diferentes. El Análisis Factorial está diseñado con el objetivo de identificar ciertos *factores no observables* de las variables observadas, mientras que en el mejor de los casos PCA proporciona una aproximación a los factores requeridos.

4.2.1 Análisis de Componentes Principales

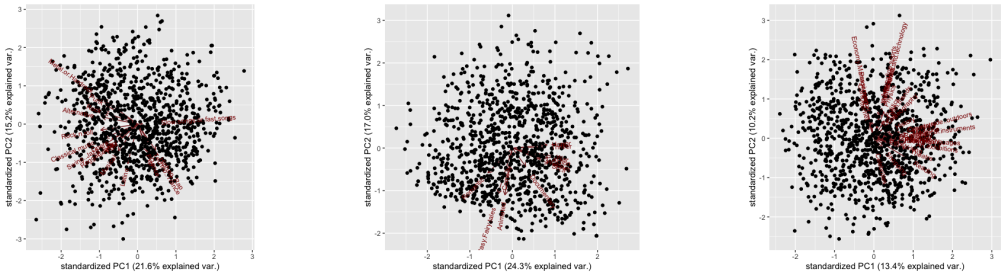
Siendo la principal técnica lineal para la reducción de dimensionalidad lo que hace es un mapeo lineal de los datos a un espacio de menor dimensión de manera que se maximiza la varianza de estos en la representación de baja dimensión.

Para el dataset de música (que contaba con 18 variables) se encontró que las 10 primeras componentes principales representaban un 80% de proporción acumulada de varianza total, lo cuál permitiría reducir la dimensionalidad a más o menos la mitad. Sin embargo se puede argüir que con un 80% se sigue perdiendo un buen porcentaje de información con esta transformación. En el caso de requerir un porcentaje del 90% se necesitarían ya 14 componentes principales. En ese orden de ideas no tendría sentido aplicar PCA para reducir solo 4 dimensiones.

De esta misma manera para el dataset de películas, que contaba con 11 variables, se encontró que se necesitan 7 componentes principales para superar el 80% de proporción acumulada de varianza total y 9 para superar el 90%. Nuevamente, solo se podrían reducir 5 o 3 dimensiones.

Finalmente, para el dataset más grande de todos, el de hobbies (con 32 variables) se necesitan al menos 18 componentes principales para superar el 80% de proporción acumulada de varianza total y 24 para superar el 90%. En este caso se conseguiría reducir 14 o 8 dimensiones, lo cuál puede parecer una cantidad considerable pero para propósitos de visualización sigue sin ser muy útil.

Ahora, como se puede ver en la Fig. 6 es imposible visualizar de buena manera los datos en dos dimensiones utilizando sus respectivas primeras dos componentes principales. De hecho, en la sección anterior se obtuvo una mejor visualización utilizando el algoritmo de TSNE, que en este caso es mejor debido a la no linealidad de los datos considerados.



(a) PCA para música (b) PCA para películas (c) PCA para hobbies

Figure 6: Representación de los datos en dos dimensiones utilizando PCA

4.2.2 Análisis de Factores

El Análisis de Factores se usa para describir la variabilidad entre las variables correlacionadas observadas en términos de un número potencialmente menor de variables *no observadas* llamadas factores. La teoría detrás de los métodos analíticos de factores es que la información obtenida sobre las interdependencias entre las variables observadas se puede utilizar más adelante para reducir el número de variables en un conjunto de datos.

En el presente proyecto se realizó a cabo un análisis de factores mediante MLE con rotación `varimax` con el fin identificar estos de forma más directa.

Para el caso del dataset de música se determinaron 11 factores que representan un total del 60% de varianza acumulada. Además, en cuanto a la prueba de hipótesis respecto a si 11 factores son suficientes se obtuvo un valor p de 0.348, por lo que no se puede rechazar la hipótesis nula de que en efecto, son suficientes. Así pues, se realizó la siguiente interpretación de los factores obtenidos:

Factor	Significado	Loadings Predominantes
Factor 1	Música Pesada	Rock, Punk y Metal
Factor 2	Música Elegante	Música Clásica, Musicales y Ópera
Factor 3	Música Bailable	Dance y Techno
Factor 4	Música Rápida	Canciones rápidas o lentas
Factor 5	Música Popular	Pop
Factor 6	Música Acústica	Folk y Country
Factor 7	Música Urbana	Reggae & Ska y Hip-Hop & Rap.
Factor 8	Música Rock	Rock & Roll
Factor 9	Música Latina	Latino
Factor 10	Música Alternativa	Alternative
Factor 11	Música Jazz	Swing & Jazz

Table 2: Análisis Factorial de Música

En cuanto a las películas se determinaron 6 factores que representan un total del 54% de varianza acumulada. Similarmente al caso anterior, se obtuvo un valor p de 0.335 por lo que no se puede rechazar la hipótesis nula de que son suficientes para representar los datos.

Factor	Significado	Loadings Predominantes
Factor 1	Películas de Fantasía	Fantasía, Cuentos de Hadas y Animación
Factor 2	Películas Maduras	Guerra, Documental y Occidental
Factor 3	Películas de Horror	Horror
Factor 4	Películas de Misterio	Thriller
Factor 5	Películas Emocionantes	Acción y Ciencia Ficción
Factor 6	Películas Románticas	Comedia y Romance

Table 3: Análisis Factorial de Películas

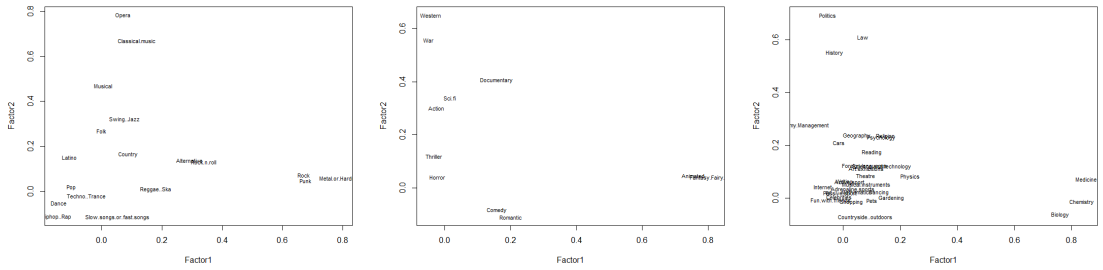
Finalmente, para los hobbies se determinaron 18 factores que representan un total del 57% de varianza acumulada, con un valor p de 0.46.

Factor	Significado	Loadings Predominantes
Factor 1	Ciencias Naturales	Biología, Química y Medicina
Factor 2	Ciencias Políticas	Historia, Política y Leyes
Factor 3	Ciencias Exactas	Matemáticas y Física
Factor 4	Tecnología	PC e Internet
Factor 5	Farándula	Celebridades y Shopping
Factor 6	Arte	Teatro y Exhibiciones de Arte

Factor 7	Adrenalina	Deportes Extremos y Deportes Activos
Factor 8	Creatividad	Instrumentos Musicales y Escritura
Factor 9	Baile	Bailar
Factor 10	Ciencia y Tecnología	Ciencia y Tecnología
Factor 11	Mundo	Geografía
Factor 12	Cultura	Lectura y Lenguajes Extranjeros
Factor 13	Conexión con la naturaleza	Religión y Campos Abiertos
Factor 14	Economía	Manejo de la Economía
Factor 15	Mente	Psicología
Factor 16	Plantas y Animales	Jardinería y Mascotas
Factor 17	Emociones leves	Deportes pasivos y Diversión con los amigos
Factor 18	Química	Química

Table 4: Análisis Factorial de Hobbies

A continuación se muestran los resultados obtenidos al realizar una proyección de los datos en los primeros dos factores:



(a) FA para música (b) FA para películas (c) FA para hobbies

Figure 7: Representación de las variables en dos dimensiones utilizando FA

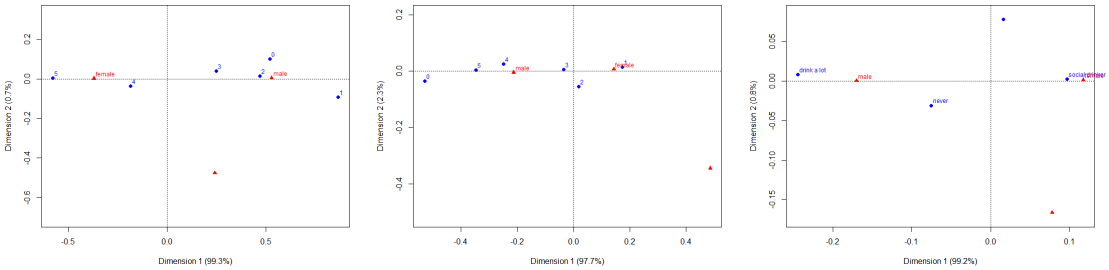
4.3 Análisis de Correspondencia

El análisis de correspondencia (CA) es una técnica descriptiva multivariada estadística que conceptualmente es similar al análisis de componentes principales, pero se aplica a datos categóricos en lugar de continuos. De manera similar a PCA, proporciona un medio para mostrar o resumir un conjunto de datos en forma gráfica bidimensional.

En este orden de ideas se utilizará como un factor determinante para este análisis el género de los usuarios, que cuenta con tres niveles: masculino, femenino, y NaN. Teniendo en cuenta que los valores NaN solo representan el 0.6% de los datos se decidió trabajar con todo el conjunto de

datos (pues no es una porción significativa).

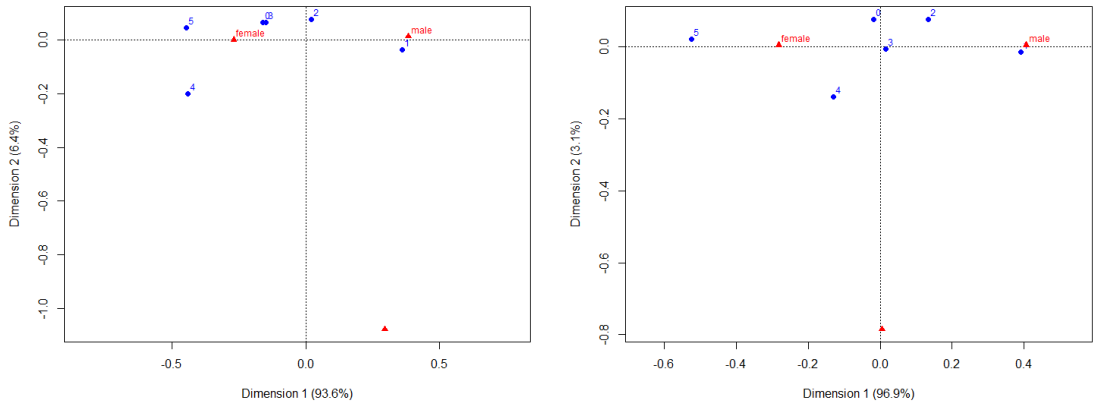
Empezamos considerando el gusto por las películas románticas con el género del usuario en la Fig. 8a. Como lo sugiere la intuición, las mujeres están más relacionadas con calificaciones altas de 4 y 5, en contraste con los hombres. En la Fig. 8b se consideró de forma similar el gusto por las matemáticas, que refleja un problema actual que poseen las carreras STEM: la falta de mujeres en las mismas. Vemos que las mujeres están fuertemente asociadas con el poco gusto por las matemáticas, mientras los hombres están más cerca de los valores altos. En la Fig. 8c se consideraron los hábitos de consumo del alcohol. En este caso los hombres están más cercanos a beber más, mientras que las mujeres tienden a beber más en eventos sociales. Además, el no beber nunca, parece estar asociado de la misma forma a los dos géneros.



(a) CA Romance - Género (b) CA Matemáticas - Género (c) CA Alcohol - Género

Figure 8: Análisis de Correspondencia

En la Fig. 9a vemos (nuevamente, en concordancia con los estereotipos) que los hombres están muy cerca a las menores calificaciones con respecto a si disfrutan bailar, mientras que las mujeres están más cercanas a los valores altos. Finalmente en la Fig. 9b vemos que las mujeres presentan un mayor miedo a las arañas que los hombres, los cuales están muy cercanos al mínimo de 1.



(a) CA Bailar - Género

(b) CA Miedo a las arañas - Género

Figure 9: Análisis de Correspondencia

5 Conclusiones

1. Basados en los tres tipos de dataframes trabajados a lo largo del proyecto, se pudo ver que *k-medoids* se comportó bastante bien en los gustos de los usuarios por las películas y la música, permitiendo agrupar a los usuarios que tienen gustos similares en estos aspectos. Una posible aplicación de esto es en los sistemas de recomendación de películas o de música, puesto que teniendo esta los usuarios agrupados, a un usuario se le pueden realizar recomendaciones basadas en las opiniones de los usuarios de su mismo cluster. Es importante tener en cuenta que para dichos sistemas de recomendación existen otras técnicas que no necesariamente utilizan clustering, sin embargo, es una forma de hacerlo. Por otro lado, con respecto a los hobbies, el método de *k-medoids* no resultó muy eficiente, puesto que como se explicó anteriormente, por medio de la silueta del modelo se identificó que habían una gran cantidad de observaciones de un cluster que estaban más cerca a observaciones de otros clusters.

   Ahora bien, con respecto a los modelos de clasificación LDA y QDA, se pudo observar que en los tres dataframes trabajados (música, películas y hobbies) LDA se comportó mejor. La métrica usada para evaluar dicho comportamiento fue el error de clasificación aparente. Cabe mencionar que dicho proceso de clasificación se realizó asumiendo que las etiquetas generadas por *k-medoids* eran acertadas para la muestra, cosa que no necesariamente es cierta.

2. Para los datasets considerados se pudo observar que realizar reducción de dimensionalidad utilizando análisis de componentes principales no resultó eficiente, debido a que no se logró una proporción de varianza acumulada significativa en pocas componentes princi-

pales. Esto se debe a la no linealidad de los datos con los que se trabajó. Debido a este problema se evidenció que el algoritmo de TSNE realizó un trabajo excelente de reducción de dimensionalidad para la visualización apropiada de los clusters planteados, ya que este algoritmo tiene en cuenta la no linealidad de los datos.

3. En cuanto al análisis de factores, se encontraron variables no observadas (los denominados factores) muy interesantes que consiguen una especie de agrupación de variables similares (por ejemplo, nótese el primer factor para el dataset de hobbies). Además, siendo los valores p relativamente altos, no se puede rechazar la hipótesis de que los factores planteados sean suficientes para describir los datos.

4. El análisis de correspondencia resultó muy útil como técnica descriptiva para identificar patrones en los datos y confirmar estereotipos comportamentales debido a la gran cantidad de variables categóricas en el dataset considerado. Se recalcan conclusiones planteadas en el proyecto como el hecho de que las mujeres parecen disfrutar más de las películas románticas que los hombres, o que los hombres están más interesados en las matemáticas que el otro género.

References

[1] Johnson & Wichern. *Applied Multivariate Statistical Analysis, 6th Ed.* Pearson, 2007.

[2] Young People Survey

[3] Las librerías utilizadas fueron: `MASS`, `rgl`, `ClusterR`, `factoextra`, `gtools`, `Rtsne`, `ggbiplot` y `ca`.