Informe Caso de Estudio – Aprendizaje No Supervisado

Integrantes:

Natalia Ximena Guzmán Zorrilla

Sebastián Cuartas Gil

Alexander Amaya León

Universidad de Antioquia, Facultad de Ingeniería

Analítica para la Toma de Decisiones [2502072]

Manuela Londoño Ocampo

Medellín, Colombia

29 de abril de 2024

Contenido

Planteamiento del problema

El objetivo principal de este proyecto es realizar un análisis de clustering sobre las valoraciones de atracciones turísticas en Europa, utilizando el conjunto de datos "Travel Review Ratings" obtenido de Google (más específicamente de Google Maps). Este conjunto de datos comprende opiniones de usuarios sobre 24 categorías de atracciones, desde iglesias hasta jardines, con valoraciones que van de 1 a 5.

Se busca identificar patrones y segmentar a los usuarios en grupos con preferencias similares en atracciones turísticas. Esto permitirá a los gestores turísticos comprender mejor las preferencias de los visitantes y personalizar las ofertas y servicios en función de estos segmentos.

Para lograr este objetivo, se aplicarán técnicas de clustering, incluyendo el algoritmo k-means y DBSCAN, con y sin reducción de dimensionalidad mediante PCA. La utilización de diferentes enfoques nos permitirá comparar y evaluar la eficacia de cada método en la segmentación de los usuarios.

Diseño de solución propuesto

1. Preprocesamiento de Datos:

Eliminar la variable 'User', ya que no aporta información para el clustering, realizar un análisis exploratorio de datos para identificar valores atípicos, si es necesario, escalar los datos para asegurar que todas las características tengan el mismo peso en el análisis.

2. Aplicación de Modelos de Clustering:

K-means sin PCA:

Seleccionar un número adecuado de clusters utilizando métodos como el codo o la silueta, aplicar el algoritmo k-means para agrupar a los usuarios en clusters según sus preferencias de atracciones turísticas, evaluar la calidad de los clusters utilizando métricas como la inercia o la silueta.

K-means con PCA:

Realizar una reducción de dimensionalidad utilizando PCA para reducir la cantidad de características, seleccionar el número de componentes principales que retengan la mayor cantidad de varianza explicada, aplicar k-means sobre los componentes principales obtenidos y evaluar los clusters resultantes.

DBSCAN sin PCA:

Configurar los parámetros de DBSCAN, como el radio y el número mínimo de puntos, aplicar DBSCAN para identificar grupos de usuarios basados en la densidad de las valoraciones, evaluar la calidad de los clusters y la capacidad de DBSCAN para identificar grupos de diferentes formas y tamaños.

Evaluación de Resultados:

Comparar los resultados de los diferentes enfoques de clustering (k-means sin PCA, k-means con PCA, DBSCAN), utilizar métricas como la homogeneidad, completitud y la medida F para evaluar la cohesión y separación de los clusters, interpretar y analizar los clusters obtenidos para comprender las preferencias de los usuarios y las características distintivas de cada grupo.

Presentación de Resultados:

Elaborar un informe detallado que incluya los pasos realizados, los resultados obtenidos y las conclusiones, visualizar los clusters utilizando gráficos como diagramas de dispersión o dendrogramas

para una comprensión más clara de los grupos identificados, proporcionar recomendaciones basadas en los hallazgos, tanto para la industria turística como para posibles investigaciones futuras.

Limpieza y transformación de los datos

Durante el proceso de limpieza y transformación de los datos, primero se procedió a cambiar el nombre de las variables originales, que estaban dadas como 'category 1' hasta 'category 24', por los nombres correspondientes como iglesias y jardines. Posteriormente, se transformaron todas las variables que estaban en formato object a tipo float para garantizar su compatibilidad con los algoritmos de clustering. Además, se eliminó la variable 'User', ya que no aportaba información relevante para el análisis de clustering.

Se identificaron y eliminaron tres valores duplicados que estaban presentes en el conjunto de datos. También se observó que existía una columna llamada 'Unnamed 25' que no contenía información relevante, por lo que fue eliminada del conjunto de datos. Se detectó que tres de las variables tenían un valor faltante cada una. Estos valores faltantes se llenaron utilizando el método de interpolación para mantener la integridad de los datos y evitar sesgos en el análisis.

Análisis exploratorio de los datos

Durante este análisis, se examina la distribución de las variables para detectar posibles valores atípicos y se investiga la correlación entre ellas para identificar patrones que puedan influir en los resultados del clustering. Además, se calculan estadísticas descriptivas para comprender la tendencia central y la dispersión de los datos, y se utilizan visualizaciones como histogramas, diagramas de dispersión y boxplots para explorar la distribución y las relaciones entre las variables. Se aborda la gestión de valores faltantes, y se exploran los datos en busca de posibles grupos o patrones que puedan guiar la elección del número de clusters. Finalmente, se considera la dimensionalidad de los datos para determinar si es necesario realizar una reducción de la misma. Este análisis proporciona una base sólida para la aplicación exitosa de técnicas de clustering.

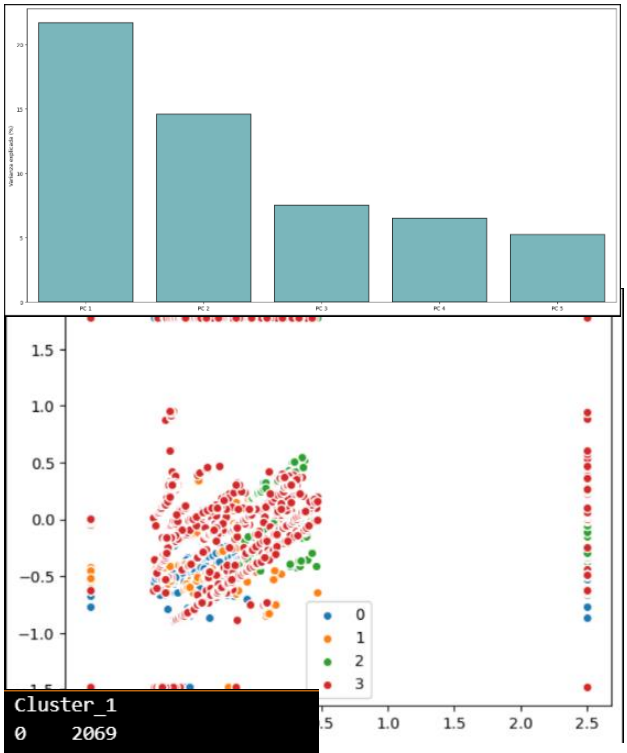
Selección de variables

Se eliminó la variable User ya que era categórica y no era muy representativa en nuestro caso, se conservaron el resto de variable, se aplicó reducción de dimensionalidad con el método de PCA.

Comparación y selección de técnicas para la aplicación del algoritmo asignado

K-means sin PCA:

Utilizamos la metodología del codo para visualizar el número de clúster, el cual confirmamos con el código para el número exacto de clúster, el cual obtuvimos un número de 4 clústeres, evaluamos su calidad mediante inercia que nos dio un valor de 90715.5395091095, Silhouette Score con un valor de 0.14182661012500977 y Calinski harabasz score con un valor de 805.9095783852654.




```
Cluster_1
0    2069
3    1835
1     860
2     692
Name: count, dtype: int64
```

La siguiente gráfica corresponde al agrupamiento de los clústeres para este caso:

se varió en múltiples ocasiones el número de k para determinar cuál era en óptimo, sin embargo, es como un valor de k = 4 que no se presentan mayores dificultades a la hora de generar el gráfico, sin embargo, a simple vista podríamos inferir un posible solapamiento de los clústeres.

A continuación, observamos la distribución de los datos en cada clúster:

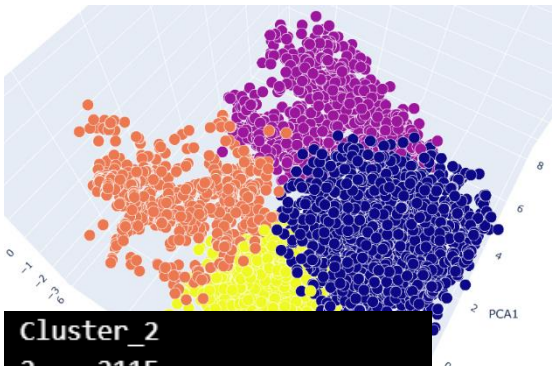
Distribución de los clústeres: Se puede ver la cantidad de puntos de datos asignados a cada clúster para entender su distribución, podemos evidenciar que el clúster 0 es el clúster dominante con la mayor cantidad de datos, sin embargo el clúster 3 se acerca al valor del clúster 0, por último el clúster 1 y 2 tienen una cantidad proporcional entre ellos teniendo 860 y 692 datos respectivamente.

K-means con PCA:

Aplicando el método de varianza explicada concluimos que con 3 componentes retenemos el 55% de varianza explicada para nuestra base de datos.

Utilizamos la metodología del codo para visualizar el número de clúster, el cual confirmamos con el código para el número exacto de clúster, el cual obtuvimos un número de 4 clústeres, evaluamos su calidad mediante inercia que nos dio un valor de 40114.48993834478, Silhouette Score con un valor de 0.2724516195951531 y Calinski harabasz score con un valor de 2221.507327024097.

La siguiente gráfica corresponde al agrupamiento de los clústeres para este caso:




```
Cluster_2
2    2115
0    1831
1     766
3     744
Name: count, dtype: int64
```

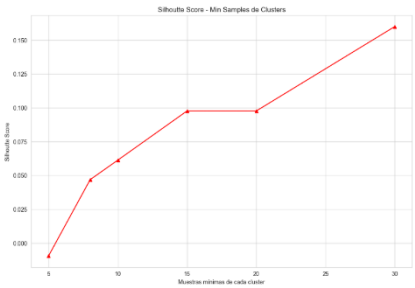
Podemos observar la agrupación de los clústeres con su respectivo color, podemos visualizar que no se evidencia solapamiento y la agrupación por clústeres está compacta o bien relacionada.

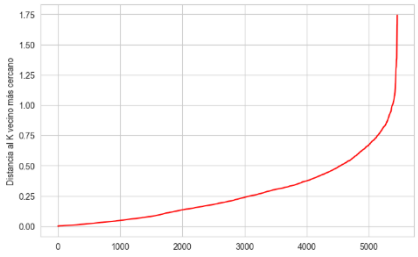
A continuación, observamos la distribución de los datos en cada clúster:

Distribución de los clústeres: Se puede ver la cantidad de puntos de datos asignados a cada clúster para entender su distribución, podemos evidenciar que el clúster 2 es el clúster dominante con la mayor cantidad de datos, sin embargo, el clúster 0 se acerca al valor del clúster 2, por último, el clúster 1 y 2 tienen una cantidad proporcional entre ellos teniendo 766 y 744 datos respectivamente.

DBSCAN sin PCA:

Primeramente, hallamos las distancias más cortas entre 2 puntos para saber cómo clasificarlos. Utilizamos este gráfico para hallar el punto de máxima curvatura de la línea, para poder hallar epsilon:

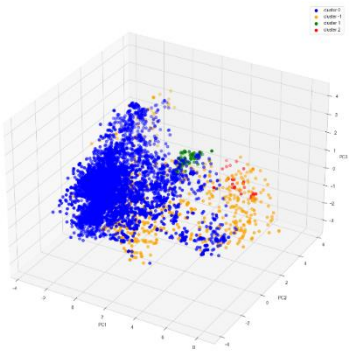




En este caso el punto de máxima curvatura fue 5403 y un epsilon de 3.322. Con estos parámetros sacamos el coeficiente de silueta y hallamos el número mínimo de puntos.

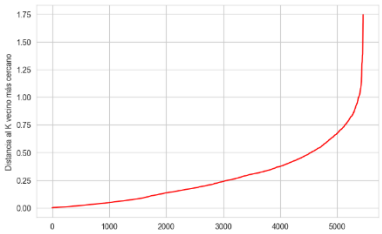
En la gráfica podemos observar que el número máximo sería 30, por lo cual usamos este parámetro para el algoritmo de DBSCAN. Finalmente nos dio como resultado: Silhouette Score: 0.15 y Calinski harabasz score: 161.13. Podemos analizar que son buenas métricas, pues el algortimo es capaz de separar los puntos y agruparlos en diferentes clusteres. En este caso nos dio como resultado los datos agrupados en 3 clusteres, mayor mente grupados en el cluster 0 y en números atípicos, distribuidos de la siguente manera,

Cluster_3	
0	4707
-1	635
1	84
2	30



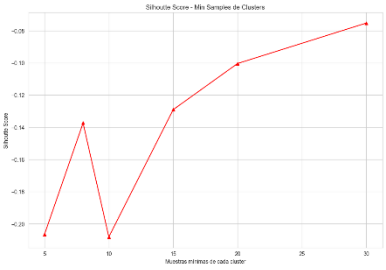
DBSCAN con PCA:

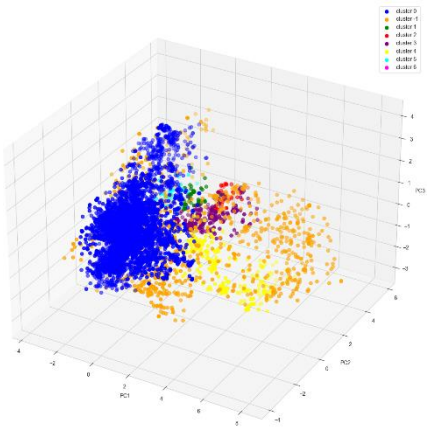
Primeramente, hallamos las distancias más cortas entre 2 puntos para saber cómo clasificarlos. Utilizamos este gráfico para hallar el punto de máxima curvatura de la línea, para poder hallar epsilon:



En este caso el punto de máxima curvatura fue 5408 y un epsilon de 1.07. Con estos parámetros sacamos el coeficiente de silueta y hallamos el número mínimo de puntos.

En la gráfica podemos observar que el número máximo sería 30, por lo cual usamos este parámetro para el algoritmo de DBSCAN. Finalmente nos dio como resultado: Silhouette Score: -0.07 y Calinski harabasz score: 290.09. Podemos analizar que el coeficiente de silueta no es tan buena métrica, ya que esta indica que el algoritmo con PCA no es capaz de separar los puntos en diferentes clusteres, o que estos se superponen por lo cual no es tan facil para el agrupar puntos teniendo en cuenta los parámetros dados anteriormente. En este caso nos dio como resultado los datos agrupados en 7 clusteres, mayor mente grupados en el cluster 0 y en números átipicos, distribuidos de la siguente manera:





Cluster_4	
0	3887
-1	891
4	274
3	215
1	57
2	52
6	43
5	37

Evaluación y análisis del modelo

K-means: En el K-means sin PCA evidenciamos una distribución de los datos no muy buena con un posible solapamiento y una inercia de 90715.5395091095, mientras que con PCA obtuvimos una distribución y una agrupación mejor de los datos, donde ya no se presentaba solapamiento, una inercia de 40114.48993834478, lo que nos puede concluir que los clúster son más compactos y cohesivos con PCA, en el Silhouette Score sin PCA obtuvimos un valor de 0.14182661012500977 y con PCA de 0.272451619595153, por lo cual evidenciamos una leve mejoría y para el Calinski harabasz score sin PCA nos dio un valor de 805.9095783852654 y con PCA un valor de 2221.507327024097, por lo cual, con PCA el cluster es más denso y mejor separado.

DBSCAN: La comparación de los resultados entre DBSCAN con y sin PCA revela diferencias significativas en la capacidad de agrupación de los datos. Sin la reducción de dimensionalidad proporcionada por PCA, el algoritmo DBSCAN logró una separación más clara entre clusters, evidenciada por métricas positivas como el coeficiente de silueta y el puntaje de Calinski-Harabasz. Esto sugiere una estructura más clara y distinta en los datos originales, lo que facilitó al algoritmo identificar y clasificar los puntos en grupos cohesivos de manera efectiva. Por otro lado, con la aplicación de PCA, se observa una pérdida de calidad en la agrupación, indicada por un coeficiente de silueta negativo y un puntaje de Calinski-Harabasz más alto. Esto sugiere que la reducción de dimensionalidad puede haber eliminado información crucial para la correcta separación de los clusters, resultando en una agrupación menos definida y posiblemente superposición entre los grupos.

En términos de impacto, la introducción de PCA pareció alterar la estructura subyacente de los datos al reducir su dimensionalidad, lo que a su vez afectó la capacidad del algoritmo DBSCAN para identificar patrones significativos. Aunque PCA puede ser útil para reducir el ruido y la redundancia en conjuntos de datos de alta dimensionalidad, su aplicación en este caso parece haber eliminado información valiosa

para la separación clara de clusters. Esto sugiere la importancia de evaluar cuidadosamente el impacto de técnicas de reducción de dimensionalidad como PCA en conjuntos de datos específicos, ya que pueden mejorar o deteriorar el rendimiento de algoritmos de clustering dependiendo de la estructura inherente de los datos.

Conclusiones y recomendaciones

- Al emplear K-means da mejores resultados al aplicar PCA
- La variable Users no fue representativa y fue la única que se eliminó
- En la metodología de clustering con DBSCAN arrojó mejores resultados sin PCA
- Atracciones preferidas en general:Las atracciones con las valoraciones más altas en general son los resorts, seguidos de museos, parques y teatros. Esto sugiere que los usuarios tienden a valorar positivamente las instalaciones de resort y las experiencias culturales como museos y teatros.
- Variedad de preferencias: Los usuarios muestran una variedad de preferencias en diferentes categorías de atracciones. Mientras que algunos clusters muestran una preferencia clara por ciertas atracciones, otros muestran una distribución más equilibrada de valoraciones entre diferentes tipos de atracciones.
- Diferencias entre clusters: Los clusters muestran diferencias significativas en las preferencias de atracciones. Por ejemplo, algunos clusters tienden a valorar más las atracciones naturales como playas y parques, mientras que otros pueden preferir actividades culturales como visitar museos o teatros.
- Atracciones menos valoradas: Las atracciones menos valoradas en general son los gimnasios y monumentos. Esto sugiere que estas atracciones pueden no ser tan populares entre los usuarios o pueden necesitar mejorar su calidad o servicios para atraer más visitantes.
- Importancia de la ubicación: Es importante considerar la ubicación geográfica de las atracciones y cómo esto puede influir en las preferencias de los usuarios. Por ejemplo, las playas pueden ser más valoradas en regiones costeras, mientras que los monumentos históricos pueden ser más relevantes en áreas con una rica herencia cultural.
- Oportunidades para mejorar: Los resultados del clustering pueden ayudar a los administradores de atracciones turísticas y a las autoridades locales a identificar áreas de oportunidad para mejorar la calidad y la oferta de atracciones en diferentes áreas geográficas. Por ejemplo, si un cluster muestra una baja valoración de ciertos tipos de atracciones, se pueden implementar estrategias para mejorarlas y atraer más visitantes.