**Informe Laboratorio 1**

**Samuel Freire 202111460 (K-medoids)**

**Juan Felipe Garcia 202014961 (K-means)**

**Lucciano Franco Márquez 202111458 (DbScan)**

**Informe Laboratorio 1**

1. **Entendimiento del problema**

Para realizar este análisis vamos a entrar a detallar los datos del archivo, se va a realizar un análisis exhaustivo para lograr entender de mejor manera su ordenamiento y su clasificación. Sin embargo, sería necesario recalcar que la población a analizar se consideró variada en cuanto a unas características, sin embargo, sería necesario para un futuro análisis tener mayor variedad, por ejemplo, a lo referente a idiomas o regios. Ahora bien, luego de haber aclarado esto vamos a entrar a detalle, en primer lugar, a el tipo de datos numérico. Dentro de los cuales vamos a detallar características que a primera vista no se encuentran, como detalles estadísticos. Para este análisis, vamos a tener en cuenta los siguientes puntos: cantidad de datos, media, desviación estándar, valor mínimo, 25% representativo, 50% representativo, 75% representativo, valor máximo. Luego de esto se va a realizar una cierta representación gráfica para entender de mejor manera este tipo de variables. Cabe aclarar que en el notebook adjuntado se pueden ver los algoritmos utilizados para realizar este tipo de análisis. Ahora bien, al hablar de la cantidad de datos entramos una similitud entre las columnas: index, isAdult, startYear, averageRating, ordering, numVotes, isOriginalTitle. Cabe aclarar que esta última es una columna que solo existe en el archivo de datos, mas no en el diccionario. Ahora bien, al hablar de la similitud de estos datos, está basada en cada uno hace referencia que se está contando la cantidad de registros que contienen este dato. Así que de cierta manera es de esperarse esta información. Luego, al hablar de la media, encontramos que para la columna de #, columna que tampoco está en el diccionario, es de 4043, es decir, es el dato promedio de esta columna. De la misma manera, para la columna index que tampoco está en el diccionario encontramos que su valor promedio es de 5440. Luego, para la columna isAdult, encontramos que su valor promedio es de 0, dado que esta columna solo contiene 1 o 0 si es o no de adulto. Luego, para la columna startYear, encontramos que su valor promedio es de 2013, lo que nos dice es que en promedio se estrenos películas este año. Por el mismo hilo, encontramos que el valor promedio para la columna AverageRating, es de 7.4, el cual resulta siendo la valoración promedio de todas las películas. Luego, para la columna de numVotes, la cual contiene el número de votos que se calcularon para la valoración, encontramos que su valor promedio es de 72828. Luego, para la columna de ordering, encontramos que si valor promedio es de 16.5. Por último, para la columna de isOriginalTitle, podemos decir que su promedio es de 0, ya que sucede algo similar que con la columna de si Adult solo contiene valores de 0 o 1. A partir de aka, como ya se mencionó no se van a mencionar los estadísticos de las columnas: isAdult, isOriginalTitle. Esto debido a que solo tomaron valores de 0, por lo que no tiene sentido mencionar que todos sus valores dieron 0, más adelante se va a mencionar esto como opción de tratamiento de datos. Ahora bien, luego de haber aclarado esto es necesario pasar a la desviación estándar de cada columna de cada columna. Cabe aclarar, que la desviación es cuan dispersa están los datos del promedio encontrado antes. Para el caso de la columna #, su desviación 2192.75. Luego, para la columna index es de 2752.82. Siguiendo por el mimo hilo, para la columna startYear es de 6.97. Luego, para la columna averageRating es de 2.77. Ahora bien, para la columna numVotes, es de 164023 Para finalizar, la desviación de la columna ordering es de 12.76. Ahora, se va a mencionar los valores mínimos de cada columna. Para el caso de la columna # y de index, su valor mínimo es de 1. Esto porque el índice 1 es el primero. Luego para la columna startYear es 1990, año en el cual la película más antigua fue estrenada. Ahora para la columna averageRating es de 6.5. Luego, la cantidad votos mínima fue de 10. Para finalizar, para la columna ordering fue de 1. Luego, de haber hablado de valores mínimos para necesario hablar de percentiles en esta muestra de datos. Ahora bien, para analizar este caso y la distribución o variedad de los datos se van a hablar del percentil, 25;50;75. Cabe aclarar que el valor que se va a mencionar en cada columna para cada percentil es el valor el cual representa que el número de percentil está por debajo de este. Esto quiere decir, que si el percentil 25 de la columna index es 36 quiere decir que el 25% de los datos está por debajo de este número. Luego, de haber aclarado esto es necesario mencionar los percentiles de cada columna. En primer lugar, para la columna # su percentil 25,50,75 son 2152, 4033 y 5914 respectivamente. Siguiendo por el mismo hilo los percentiles 25, 50, 75, de la columna index son 3073, 5421, 7641. Luego, para la columna startYear, sus percentiles 25,50,75 son 2008, 2015, 2019 respectivamente. Luego, los percentiles 25, 50, 75; para la columna averageRating, son 6.9, 7.3, 7.8 respectivamente. Luego, para la columna numVotes son 6265, 152790, 56419 respectivamente. Por último, para la columna sus respectivos percentiles son 6, 14, 24. Ahora para finalizar este análisis estadístico sería necesario mirar el valor máximo que alcanza cada columna. Para la columna # su valor máximo es 7849. Luego, para la columna index el valor máximo es 10274. Luego, para la columna startYear su valor máximo, o el valor más reciente de estreno es 2023. Siguiendo, por el mismo hilo, para el averageRating, o la valoración más alta es de 92. Luego, para el número de votos el valor máximo fue de 2197234. Por último, para la columna de ordernig el valor máximo fue de 119. Además, de estos datos analíticos, se puede hacer una comparativa entre las columnas numéricas. Después, de compararlas, se puede decir, que la columna numVotes el rango es muy grande su valor máximo a comparación del resto. Ahora bien, hablando de rangos también se puede decir que las columnas # e index, tiene un rango similar. También, podemos hacer análisis por tipo de número. En el cual, podemos ver proporciones donde el número de tipo float o inexacto es muy grande el rango también. Luego, al hablar del tipo int vemos que existe un rango similar entre las columnas index y #.

Ahora luego, de haber realizado el análisis estadístico o más bien, el análisis de las columnas de tipos numéricas es necesario realizar el análisis de las columnas categóricas. Antes de empezar a hablar y esto sería necesario recordar las columnas ctaegoricas que se encontraron. Estas columnas fueron: tconst, titleType, originalTitle, runtimeMinutes, región, language, types, attributes, main\_genre, secundary\_genre. Para realizar este análisis, se van a realizar diversos análisis de frecuencias, conteos, más frecuente. Para empezar este análisis, vamos a empezar hablando de los valores únicos que presenta cada categoría. En este caso, esta medida va a ayudar a identificar las diversas categorías, que se presentan dentro de cada columna. Antes de hablar de los valores únicos, sería necesario mencionar que además de esto se realizó un algoritmo de conteo. Sin embargo, se encontró que par la mayoría de las columnas categóricas, se encontró 7040, siendo solo diferente3s # e index, que fue 7041. Ahora bien, luego de haber aclarado esto, la cantidad de valores únicos para la columna titleType fueron 7. Siendo estas, movie, TvSeries, MOVIE, tvminiseries, MOvie Movie, tvMovie. Mas adelante, se adentra a detallar por qué existen errores o posibilidades de mejora en este punto. Para la columna de originalTitle se encontraron 7290 únicos. Luego, para el runTimeMinutes se encontraron 362 originales. Realmente, en esta última no sería necesario saber los únicos ya que pueden repetirse y no importa si existen o no únicos. Luego para la región se encontraron las únicas 4 que presenta el diccionario. Las cuales son US, IN, CA, CB. Luego para el lenguaje solo se encontró un valor único el cual es inglés. Luego para el tipo se encontraron 5 valores únicos. Para la columna de atributos se encontraron 1 atributo, que más adelante se detallar que no es de mucha importancia. Luego para genero principal o main\_genre se encontraron 22 géneros diferentes y para el género secundario o secundary\_genre se encontraron 26 diferentes. Ahora bien, luego de haber detalla esto, vamos a detallar algo un poco más específico y es el valor único que más se repite o que con mayor frecuencia apareció. Para el caso de la columna tconst, el valor más repetido fue tt0395843. Luego, para el tittleType el más repetido fue movie con 4673 apariciones. Para la columna de originalTitle el valor más repetido fue Home con 3 apariciones. Para la duración o runtimeMinutes el valor más repetido fue 60 minutos con 534 apariciones. Luego, la región que más apareció fue In con 5590 apariciones. Luego el lenguaje que más apareció fue ingles con 7470 apariciones. El type que más apareció fue imdbDisplay con 7395 apariciones. Luego, el valor más repetido en atributos fue \N con 7470 registros. Luego el género principal que más apareció fue Drama con 1767. Lo mismo que para genero secundario, sino que con 2428 apariciones.

Luego de haber realizado este análisis más detallado de ambos tipos de columnas o valores que se presentaron en el archivo podemos realizar, algo más abstracto y es realizar o definir relaciones que a futuro podemos usar, cambiar, o desechar, pero para tener presente proporciones que presentan estas columnas en el trabajo de estos datos. Una de las que se realizaron fue ver la frecuencia de cada longitud o duración de las películas donde, como ya se había dicho antes, la mayor fue la duración de 60 minutos. Sin embargo, podemos encontraron fenómenos interesantes como que la segunda duración más usada es de 30 minutos. Como se podría pensar una duración mayor sería mejor, sin embargo, los datos marcan que los estudios prefieren hacer películas más cortas. Otro análisis o relación que podríamos encontraron s la cantidad de votos de películas de acuerdo con la región. En este caso se realiza una proporción entre región y número de votos donde, por cada región se realiza su traza de votos. Donde podemos encontrar que la región In es una región que tiene a votar mucho junto con la región de CA, siendo estas las dos mayores, por lo que se podría usar como población o muestra en este estudio en caso de necesitarse.

1. **Calidad de los datos**

Luego de haber realizo un entendimiento de los datos, su contexto, entender que querían decir, que información encontramos, cuales fueron los datos más repetidos entre otros. Sería necesaria analizar si de toda esa información que se encontró es de calidad o es útil para el proyecto, en caso de no serlo que se podría modificar o mejorar para poder tener una etapa de procesamiento exitosa para el desarrollo del proyecto. De esta manera, se va a adentrar a análisis cuatro etapas o cuatro partes fundamentales de la calidad de los datos. Estas partes son: la completitud de los datos; si los datos estos completos si tiene nulo, etc… Luego, se va a analizar la unicidad de los datos, datos repetidos porque sucede esto entre otros. Luego, se va a analizar la consistencia de los datos; si aparecieron valores que no eran si aparecieron columnas que no eran entre otros. Y, por último, se va a analizar la validez de los datos, si aparecieron cosas que no son relevantes, si son necesarias para el estudio o si en su defecto no se requieren para realizar el análisis y de esta manera descartarlas para el procesamiento.

* 1. Completitud

Como se había dicho, se requiere empezar revisando si los datos entregados por MovieAlpes, están completos. En este punto es necesario analizar si existen valores nulos a lo largo de los registros. En este caso, vamos a realizar una evaluación por porcentajes para cada columna dad su totalidad de registros cuanto en porcentaje de su registro están nulos. Luego de haber realizado el conteo de registros que contaran con nulos o registros incompletos se encontró que las columnas numVotes, ordering, main\_genre, isOriginalTitle, attributes, types, language, región, secodary\_genre, averageRating, runtimeMinutes, startYear, e isAdult, todas cuentan con una falta o con un porcentaje de valores nulos del 0.0134% con respecto a su total. Se puede pensar que es un porcentaje muy bajo, que realmente representa que aun quitando esos datos se podría tener un buen panorama en cuanto a los resultados, dado que se está tomando una gran cantidad de registro completos para los estudios. Sin embargo, sería recomendable, que con la columna tconst, la cual es el identificador de la película se pudiera rellenar esa información faltante de los titulo que tiene espacios nulos en su registro. Dado que, si se tiene el identificador, se esperaría fuera más fácil la reconstrucción de la información. Por otro lado, las columnas, como se mencionó, index, originalTitle, titleType, tconst y # tiene el 100% de completitud. Por lo tanto, se puee decir que en el atribvuto de compleitutd el archivo de datos entregado es una muy buena base para el estudio o proyecto a realizar.

* 1. Unicidad

Para hablar de la unicidad, es necesario realizar un análisis sobre los datos en búsqueda registros totalmente repetidos o parcialmente repetidos. Esto con el fin de verificar si son necesarias mantener esas copias, o en su defecto es un error y se puede corregir para que los resultados de los estudios salgan con la mayor variedad sobre la población posible. Para esto vamos a remontarnos un poco a los cálculos de cuantas veces aparecían cada elemento en los datos. Al realizar una prueba de duplicidad total entre registros se encontrar que 6 registros de los datos estas completamente iguales. Entre estos registros encontramos CSI:NY, Tarnation, Les choses qu'on dit, les choses qu'on fait. Ahora bien, vamos a abría que adentrarse en fijarse en las filas en las que duplicados no deberían existir. En este caso, se consideran dos columnas donde los duplicados serian un error. La primera es tconst, la cual al seri el identificador de la película no tendría sentido que hubiera películas con duplicados. Sin embargo, se encontraron que existen 10 registros repetidos en esta columna. Luego, para la columna originalTittle, que también no deberían haber duplicados peor en menor medida, se encontró que existen 354 duplicados. Esto sin tomar en cuenta que pueden existir duplicados peor con un espacio de más o este tipo de errores de ortografía. Por lo que por el atributo de unicidad a pesar de que el porcentaje de erro no está grande si fuera necesario encontrar y corregir estos duplicados para que la muestra sea variada y no se repitan.

* 1. Consistencia

Para hablar de consistencia, ahora es necesario realizar un análisis más a fondo sobre cada columna de los datos. EN este caso, es necesario buscar cosas, que como su nombre lo indica, no sea consistentes con lo que sabemos o con lo que debería ser. Un ejemplo de esto puede ser que aparezca un dato que no tiene nada que ver con la columna o que aparezcan columnas que representan lo mismo. Luego, de dejar claro esto, vamos a revisar las inconsistencias que se encontraron. En primer lugar, en la columna de titleType s encontraron inconsistencias de tipo gramatical, si se quiere decir, dado que se encontraron 4 tipos diferentes de la palabra movie. Donde esta palabra debería estar toda unificada en la misma. Este tipo de errores sería fácil de corregir antes del procesamiento haciendo un reemplazo de estas fallas. Otra inconsistencia, de otro tipo, se encuentra en la columna averageRating, donde su valor máximo es 92. Sin embargo, cuando revisamos el rango de esta columna debería estar entre 0 y 10. Así que pueden existir varios de estos errores también que estén fuera del rango. Ahora bien, para agregar dos casos los cuales sucede que no están en el diccionario de la empresa. Esto pasa con las columnas #e index.

* 1. Validez

Para hablar de validez, es necesario ir a esas columnas o registro que se consideran que no aportan nada al análisis, es decir, que si se tiene en cuenta daría información irrelevante. Luego, de haber dejado claro esto es necesario resultar esas columnas que se consideran como invalidas o que no aportan información. En este caso se encontraron dos columnas específicas, las cuales deberían ser o reformuladas o sacas de los datos para, primer disminuir peso de archivo, para la rapidez del análisis y segundo para no perder espacio en información que no aporta. La primera columna, es la isAdult. En esta podemos ver que todos los valores son 0. Por lo tanto, esta columna sería ideal quitarla i lo que se quiere decir es que todas son para adultos o no, se sabría si ponerlo en cada registro. Luego, la otra columna en la que sucede algo similar es en language. La cual, todos sus valores son en, es decir, ingles. Por lo tanto, esta columna tambien seria susceptible de modificación o eliminación, ya que no aporta nada en los registros y se puede resumir en el contexto del proyecto. Algo similar sucede con la columna originalTitle. Tambien, sucede algo similar con la columna atriubutes. Son columnas que repiten el mismo valor en todos sus registros. Además, de estas columnas, se tiene que mencionar un caso que se mencionó antes que también se puede considerar de validez, y es el error de los números máximos de la columna averageRating

1. **Preparación de datos**
   1. La primera decisión tomar, como ya se había dicho antes, va a ser necesario desechar esas columnas que no aportan o tiene errores sobre ellas. En primer lugar, se va a desechar la columna attributes. Esto se debe a que tiene valores nulos al 100 por ciento, dado que no tiene nada en sus registros. Luego, la siguiente columna a desechar isOriginalTitle, sucede algo que se va a repetir a lo largo de varias columnas. En este caso, esta columna mantiene un valor de 0 a lo largo de todos sus registros. Caso similar sucede con la columna isAdult, que mantiene el mismo valor de 0 para todos sus registros. Luego, sucede algo similar con la columna language. Estas columnas que tiene valores similares se omiten dado que no aportan ninguna información relevante para realizar un análisis dado que no existe variedad en la muestra. Luego, se tiene tres casos para los cuales no se repiten sus valores, sin embargo, no aportan información para realizar los objetivos de este análisis. En primer lugar, está la columna #, la cual se espera mantenga el orden, además de existir en el diccionario de la empresa. Luego, un caso similar sucede con índex. Por último, tenemos el caso de la columna, ordering, dado que no aporta información que sepa. Además, de aportar orden que ya aportan las otras dos columnas mencionadas.
   2. Un segundo para la preparación de datos es necesario eliminar los porcentajes de duplicados. En este caso, retomamos los análisis realizados en la unicidad. En este caso se eliminarán las filas duplicadas, basándose en los valores duplicados en las columnas ttconst, la cual al ser un identificador no tiene sentido que se repitan para el análisis.
   3. Luego, en un tercer momento va a ser necesario quitar los valores nulos a lo largo los datos. Para este caso, vamos a ir definiendo ciertas columnas, las cuales van a ser necesarios para realizar el análisis. En este caso, la primera columna, sujeta a estos cambios, es que se van a eliminar las filas que contengan valores nulos de las columnas averageRating. Algo similar se va a realizar con la columna numVotes. Luego se va a realizar lo mismo con la columna newruntimeMinutes. En este caso, se quitaron estos valores nulos, ya que pueden llegar a afectar los algoritmos que se van a utilizar próximamente, los cuales son sujetos a valores numéricos y no categóricos.

Ahora bien, muchas de estas decisiones fueron sujetas a dos bases para el análisis. La primera de ellas va relacionada con los objetivos del negocio. Como se menciona, en el contexto se espera que este análisis ayude a realizar una estrategia de especialización del contenido. Por lo tanto, se consideró que las columnas, que se definieron o se dejaron en el análisis pueden llegar a ser importantes, o pueden llegar a portar información para realizar especialización del contenido. Luego, la segunda condición fue, los algoritmos que se van a utilizar. En este caso, fue necesario definir columnas, que fueran numéricos, o que fuera factible de convertir de categórica de tipo string, si se quiere decir, a categórica nominal. Esto a que los algoritmos que se van a utilizar que se van a mencionar y definir más adelante, son susceptibles a las columnas de tipo string. Por lo tanto, todas las decisiones de procesamiento de datos, va regidas por los algoritmos utilizar y van regidas al mismo tiempo por los objetivos de negocio.

1. **Modelamiento**
   1. **K-means**

El funcionamiento de K-means parte del hecho que es mejor usar variables numéricas en vez de variables categóricas para su uso, debido a esto en este caso se prefirió usar las columnas que tuvieran variables numéricas para llevarlo a cabo, aparte de esto K-means se basa en la distancia de un punto hasta un centroide entre más cerca este va a pertenecer a un grupo más oportunamente pero si está más lejos es un poco más complicado ubicar el punto a un grupo ya que lo que se realiza para encontrar la distancia es parecido a una norma para un vector. K-means es un algoritmo bastante útil pero que cuenta con una falencia bastante importante, esta falencia es el hecho de que al basarse en la distancia que tiene un dato hasta 1 centroide si este no tiene los datos normalizados en 1 misma escala entonces únicamente 1 de los datos va a ser significativo por ende hay que llevar a cabo 1 normalización para que todos los datos sean significativos en el modelo planteado, debido a esto nosotros llevamos a cabo la normalización con valores que no se parecían tanto lo que nos permitió establecer relaciones entre variables que a primera vista pareciera que no tuvieran relación alguna, la normalización de datos es un paso crítico en la preparación de datos antes de aplicar el algoritmo K-Means. En este contexto, la normalización se refiere a escalar todas las variables a una misma escala o rango, generalmente entre 0 y 1. Esto es fundamental porque K-Means se basa en medidas de distancia, y las diferencias en las escalas de las variables pueden distorsionar el proceso de agrupamiento lo que quiere decir que 1 variable seria la que dominaría el modelo lo que no nos es util. Ahora bien, para la elección del número de clústeres k se llevaron a dos pruebas las que se conocen como método del codo y el método de la silueta, la elección correcta del método del codo y la silueta es fundamental ya que entre mejor sea la elección de clúster el modelo va a tener un mejor planteamiento, este tema se explicara más adelante y nos permitirá ver la validez de nuestro modelo. Ahora bien, cabe aclarar que entre más veces se ejecutara el algoritmo el número de iteraciones mostraría alguna mejora con respecto a la iteración anterior pero tampoco era necesario llevar a cabo un número muy alto de iteraciones ya que hay un momento en el que el número de iteraciones no va a mostrar una mejora significativa.

* 1. **K-medoids**

Antes de ir a analizar o justificar el porqué de la elección de este algoritmo va a ser necesario realizar una definición corta de cómo funciona o que lo identifica como algoritmo de agrupación. En este caso, este algoritmo resulta en la idea de minimizar la distancia de su medio, y minimizar la suma de la distancia con los puntos aledaños a este o los puntos de su alrededor. Para este, punto se puede pensar que es muy similar a k-means, y realmente es que tiene el mismo sentido de separación, Sin embargo, en este caso, la base o los algoritmos matemáticos usados detrás son diferentes en ambos casos. Luego, otro aspecto importante sobre K-medodis, es que este escoge su centroide, si se quiere decir de alguna manera, no sobre el cálculo matemático, sino sobre uno ya existente. Esto quiere decir que K medoids, utiliza uno de los datos de los registros de los datos entrados como parámetros para realizar la elección de su centroide. Esto permite claramente tener una mejor definición, o una definición más real sobre los datos agrupados. Para este caso, el centro se va a definir como medoide. Este busca minimizar la distancia con este centroide al resto de sus centroides, o más bien busca reducir la distancia promedio con todos sus aledaños (scikit-learn-extra ,2023). Ahora bien, luego, de haber definido de manera corta este algoritmo. Es necesario describir por qué la elección de este algoritmo. En este caso, se escogió por su afinidad con k-menas. Esto quiere decir, que la mayoría de partes de los análisis eran muy similares, y se entendió que los análisis para ambos iban a ser mucho más entendibles, ya que se tenía un antecedente de entendimiento., Luego, se escogió por que mejora ciertas partes de los K-means.- Uno de estos valores, fue que los medoides, en este cao hacen parte del conjunto de datos. Esto, a pesar de ser una característica menor, puede ser mucho más fácil para el cliente entender que un centro de su grupo es un dato que debe conocer. Además, que es más amigable con los valores atípicos. Por lo tanto, su elección fue basad en estas ventajas que se mencionaron como lo pueden ser: su afinidad con k-measn, su capacidad de entendimiento y por su capacidad de manejo de valores atípicos (Rpubs, 2023). Ahora siguiendo con la justificación ahora va a ser necesario justificar la elección de los hiperparametros de este algoritmo. En este caso K-medoids, tiene dos hiperparametros. En este caso vemos k, el cual hace referencia al número de clústeres que se van a crear para agrupar los datos. El segundo de ellos es la métrica, la cual hace referencia a la métrica usada para definir la distancia entre los puntos y sus medios. Para el caso de K. se espera realizar un método de validación, Mas adelante, como el codo y la silueta para definir el nivel óptimo de numero de clústeres para que el análisis sea los entendible posible. Luego, para el caso de la métrica se va a definir como euclidiana. Lo que hace referencia a que se va a calcular la distancia a base de distancias euclidianas para lograr realizar los cálculos entre los puntos.

* 1. **DbScan**

DBSCAN es un algoritmo de clustering que se diferencia de K-means y K-medoids en su enfoque y flexibilidad. En lugar de depender de distancias euclidianas, DBSCAN se basa en la densidad de puntos en el espacio de características para identificar clústeres. Esto significa que DBSCAN puede descubrir clústeres con formas y tamaños irregulares, lo que es especialmente útil cuando los datos no siguen una estructura de clúster perfectamente circular o esférica. Este algoritmo es resistente a la influencia de valores atípicos, ya que los considera como puntos individuales o ruido en lugar de asignarlos a un clúster. Esto es muy importante ya que se trabaja con datos que pueden contener valores atípicos o errores. Además, el algoritmo es capaz de manejar conjuntos de datos con dimensiones mixtas, lo que lo convierte en una herramienta versátil para análisis multidimensionales. Una de las ventajas más notables de DBSCAN es que no es necesario especificar previamente el número de clústeres (k). En su lugar, DBSCAN detecta automáticamente el número de clústeres en función de la densidad de los datos, lo que lo hace ideal para conjuntos de datos donde el número de clústeres es desconocido. Además, DBSCAN es resistente a valores atípicos, lo que significa que puede identificar y aislar valores inusuales sin que afecten negativamente la formación de clústeres. Su capacidad para adaptarse a estructuras de datos de diferentes formas y dimensiones mixtas lo convierte en una elección sólida en una variedad de situaciones de análisis de datos. La elección de DBSCAN se basa en su capacidad para encontrar clústeres basados en densidad de manera robusta y su flexibilidad para adaptarse a la complejidad de los datos sin requerir una predefinición del número de clústeres.

1. **Validación**

Para esta sección se van a realizar múltiples pruebas sobre cada algoritmo mencionado anteriormente. Esto con el objetivo de probar o en su defecto validar, cuan buenos fueron los modelos realizados desde el punto de vista de parámetros o datos óptimos. E este caos se va a hablar únicamente de valores numéricos o parámetros de cada algoritmo que podrían mejorar.

* 1. **K-means**

En lo que respecta a las pruebas de validación que se realizaron para el algoritmo de k-means se pueden llevar a mucha información que nos puede ayudar a determinar lo bien o mal que esta el algoritmo. Para empezar, realizamos la prueba del codo la cual consiste en que se hace una gráfica con los números de clúster y la cantidad de distorsión que cada uno de estos, al inicio la distancia que existe de distorsión entre los primeros clústeres va a ser muy grande pero después va a haber un punto que esta cantidad reduzca demasiado y exista un cambio de pendiente bastante grande, cuando eso pase quiere decir que ese número de clústeres es candidato a ser el mejor, por lo que este número va a ser muy tomado en cuenta para realizar iteraciones del algoritmo en este caso 4, de igual modo los números anteriores no son desechados sino que también hay que ponerlos a prueba en este caso esos números son 2,3. Para dar un resumen del método del codo, esta es una técnica gráfica que te ayuda a identificar el número óptimo de clústeres en un algoritmo de agrupación al buscar el punto en el gráfico donde la curva de inercia muestra una disminución significativamente menor. La otra prueba que realizamos fue con el método de la silueta, El método de la silueta es una técnica de evaluación de la calidad de los clústeres en algoritmos de agrupación, como K-Means, que proporciona una medida de cuán bien están separados los clústeres. El método de la silueta coloca los clústeres en un rango de 0 a 1 lo ideal sería que estas siluetas estuvieran como mínimo en 0,5 para que fuera un modelo aceptable ya un modelo mejor sería que mínimo estuvieran en 0,7. Cabe aclarar que si existe una parte de la silueta que está en el rango negativo significa que hay valores de ese clúster que están más cerca de otro centroide del que fueron colocados originalmente. Como se ve en las imágenes se realizaron los clústeres de 2 a 3, cada uno de estos tiene una zona negativa que es debido a los outliers, estos datos no pueden ser eliminados ya que en la parte de la preparación de datos se decidió que era mejor dejarlos ya que eran significativos en el análisis y el modelo, debido a esto se ve la zona negativa en los 3 gráficos. Aparte de esto se ve que los mejores resultados obtenidos se ven en la gráfica con 3 y 4 clúster, por lo que la información que el método del codo proporciono era bastante apropiada y también los clústeres usados en anteriores iteraciones. En lo que respecta al modelo se puede decir que es bueno, es un modelo mejorable ya que se podría discutir una solución para los outliers y su silueta en la parte negativa, pero de igual modo lo consideramos un modelo valido para el análisis de los datos.

**Gráfico, Gráfico de líneas

Descripción generada automáticamente**

**Gráfico

Descripción generada automáticamente** Gráfico, Gráfico de embudo

Descripción generada automáticamenteGráfico, Gráfico de embudo

Descripción generada automáticamente

* 1. **K-medoids**
     1. **Validación cuantitativa**

Para este caso, se va a utilizar el método del codo para verificar la factibilidad el número de clústeres óptimos para mejorar el análisis. Luego, de haber realizado el análisis, el cual se encuentra en el notebook adjunto, se encontró que el numero apropiado de clústeres puede ser el 3 ya que aquí es cuando la distorsión empieza a tener un cambio más pequeño entre más se ajusten los clústeres. De igual modo también 2 y 4 clústeres son buenos indicativos para hacer análisis, pero el que es mejor candidato es 3 clústeres. Esto debido a que en la gráfica que se puede ver en el notebook, se encontró que un punto de inflexión que realiza que la gráfica tenga una tensión ya más recta, es el valor número 3 como se mencionó anteriormente por lo que se va a realizar una prueba con esta métrica. Luego de realizar esta prueba se encontró que claramente hubo una mejora en cuanto a la anterior grafica de formato película y numVotes, los clústeres en esta graficas se dividieron de manera más uniforme siendo el clúster 0 el que tiene los valores de bajo en ambos el clúster 3 tiene valores del medio y bajos, el clúster 1 tiene valores altos de averagerating y bajos de numVotes y el 2 que es el que tiene números altos de numVotes y de todo tipo de formatos de película. Esto demuestra que para estas variables que son las que tienen más relación tener 4 clústeres es idóneo ya que se cubre una gran cantidad de grupos y del mismo modo demuestra una normalización bastante efectiva.

* + 1. **Validación cualitativa**

Luego, de haber validado de manera cuantitativa el modelo, es necesario realizar un análisis cualitativo del mismo. En este caso, sé que concluyo que el modelo claramente se puede mejorar. Sin embargo, estas mejoras irían orientadas al caso final para que en el momento de presentar una oferta o estrategia tenga el mayor sustento posible. Ahora bien, para esta primera iteración se considera que el modelo permite al negocio encontrar solución rápidas o posibilidades de estrategias para la personalización del contenido. Además, se observa que la variable número de votos y la variable avreageRaitng es la variable que aporta más en las agrupaciones generadas, como se puede ver en la segunda agrupación. Sin embargo, estas variables por si solas, solo son útiles si se combinan con las otras variables. Por ejemplo, unir average Rating, numVotes y tipo de película. Podría ser una gran estrategia para lograr definir estrategias para producir más películas de ese estilo. Por lo tanto, sería necesario realizar una gran carga de estudios, para lograr genera estrategias a partir de las variables número de votos y average Rating. junto con otro tipo de variables categóricas que pueden ayudar a aumentar el objetivo del negocio de personalizar o genera más películas con un público objetivo definido y apoyar, la generación de la personalización que se busca en las películas. Sin embargo, el análisis realizo y los modelos, son suficientes para lograr identificar patrones en las películas de cierto tipo, o de cierto origen o de cierto género, se pueden llegar ligar con su grupo objetivo de personas, ayudando así a una estrategia de personalización mejor.

* 1. **DbScan**

Al realizar la verificación de validez del algoritmo, nos damos cuenta de que el algoritmo identifica los conjuntos de datos como se mencionó anteriormente. Se identificaron 3 clústeres y el ruido, esto gracias a que este va etiquetando en las iteraciones a los puntos y verifica si se cumple con los criterios de densidad. Analizando en profundidad, el modelo propuesto sugiere que este es eficiente a la hora de encontrar soluciones acordes a lo que se busca, que es, la personalización. Con ayuda de un histograma, podemos identificar tendencias en el comportamiento de los clientes y adaptar estrategias de marketing y promociones para satisfacer sus necesidades específicas. Además, podemos utilizar esta información para diseñar un diseño de tiendas y ubicación de productos más efectivo dentro del centro comercial. Este análisis de distribución de clústeres ayuda a comprender mejor a la clientela y a tomar decisiones más informadas para mejorar la experiencia del cliente y aumentar las ventas.

Por otro lado, cualitativamente, con ayuda del análisis por siluetas y los outliers que el algoritmo encontró, se ha permitido evaluar la calidad de la segmentación de los clientes en el caso de estudio. Los coeficientes de silueta indican que los clústeres son cohesivos y bien definidos, lo que significa que hemos tenido éxito en agrupar. También significa que el algoritmo tiene una buena segmentación a la hora de modificar sus parámetros. La coherencia interna es alta, esto se traduce a que los datos son parecidos en términos de las características que se utilizaron en el algoritmo. Por otro lado, a la hora de realizar una validación de calidad, nos damos cuenta de que efectivamente la elección de parámetros y la estructura de los clústeres es apropiada. Finalmente, lo anterior se concluye a que el modelo es bueno, se pueden tomar decisiones de negocio por ejemplo para el ruido y así alcanzar un buen nivel para el modelo de datos.

1. **Conclusiones**

a. En un primer momento, luego de haber realizado todos los análisis propuestos, se concluyó que uno de los mayores problemas o mas bien uno de los obstáculos para generar propuestas sólidas, fue que hubo falta de variedad en algunas variables. Esto ya que se sintió que, para poder generar estrategias de otros indoles o basadas en otras cosas, no había tanta variedad más allá de la obvia que fue las de géneros duración. Se concluye que la empresa debería mejorar en ese sentido la variedad de datos para lograr generar unos análisis más gruesos.

b. Usar la variable de averageRating junto con la variable de número de votos, acompañada de otras variables, puede generar estrategias interesantes para la empresa. Esto se debe a que realmente estas variables son las que ayudan a la empresa saber qué tipo de películas o géneros de estas, son las que están triunfando y en cuales habría que generar mayor publicidad o estrategias para lograr aumentarlas. Por lo que, sería necesario que, si en un futuro se sigue con el análisis, se enfoquen estas dos variables como pilares de este.

c. Como pilar en el número de clústeres para los tres algoritmos se encontró que un valor entre 2 y 4 clústeres es ideal para el análisis. Ahora bien, esta conclusión ya va un poco mas enfocad a la idea de futuros análisis basados en algoritmos de agrupamiento. Y es que luego, de haber evaluado los algoritmos con sus debidos métodos, se encontró que existe una fiabilidad que el número de clústeres sea cercan o este en el rango de [2,4]. Esto debido a que este rango genera una visualización muy clara de las tendencias sobre los gurpos de datos en los que se aplican.

d. Las variables types main\_genre y seondary\_gnere son variables que tiene la mayor variedad en cuanto a la posibilidad de análisis para personalización de contenido. Si bien, otras variables tienen a ser útiles como se había mencionado anteriormente, el caso de num votos o de averageRating. Las variables que se consideran pueden ayudar a visualizar la variedad sobre la población, y así mismo genera un estudio mucho mas completo, son estas que se mencionan al principio. Esto ya que, estas variables tienen una población muy variada y permiten ver el comportamiento de estas variables sobre más grupo s de datos que a comparación de las otras.

1. **Recomendaciones**
   1. Se recomienda realizar publicidad o generar películas de tipo drama, con una duración entre 100 y 120 minutos.​
   2. Realmente en este caso se recomienda que la duración, de sus películas principalmente vallan en un rango [75,150]​
   3. Luego, enfocarse en las películas de tipo imbdDisplay que son las de mayor aceptación según los datos.​
   4. Ahora, sería bueno hacer campañas para aumentar la visualización en los géneros de comedia y de aventura
   5. Se recomienda aumentar el contenido de los principales géneros, como drama, comedia y acción en formato de película.​
   6. Incluir más contenido audiovisual entre los 100 y 125 minutos, puesto que es lo que más MovieAlpes proyecta.​
   7. En general, MovieAlpes se desempeñade forma adecuada y mantiene calificaciones positivas, con el análisis mencionado previamente, se puede proporcionar recomendaciones más relevantes y personalizados
2. **Referencias**

* 1. *2. Clustering with KMedoids, CLARA and Common-nearest-neighbors — scikit-learn-extra 0.3.0 documentation*. (s/f). Readthedocs.Io. Recuperado el 3 de septiembre de 2023, de <https://scikit-learn-extra.readthedocs.io/en/stable/modules/cluster.html>
  2. *RPubs - Chapter 5. K.medoids*. (s/f). Rpubs.com. Recuperado el 3 de septiembre de 2023, de <https://rpubs.com/JoyP/1046516>
  3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
  4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
  5. [Machine Learning & Clustering: el algoritmo DBSCAN (datascientest.com)](https://datascientest.com/es/machine-learning-clustering-dbscan#:~:text=El%20DBSCAN%20es%20un%20algoritmo%20sencillo%20que%20define%20los%20cl%C3%BAsteres,%CE%B5%2Dvecindad%20de%20la%20observaci%C3%B3n.)
  6. [DBSCAN en Python: aprende cómo funciona - Ander Fernández (anderfernandez.com)](https://anderfernandez.com/blog/dbscan-python/)