

Facultad de Estudios Superiores: Acatlán.

Diplomado en Ciencia de Datos

Proyecto:

*La magia del cine (o de los datos)*

Valencia Martínez Jesús Adolfo

La magia del cine (o de los datos)

Contexto

En el mundo, la información crece de manera constante y. con el tiempo, se han ido generando metodologías con el fin de convertir esta información en valiosos activos para su explotación. Hablando del arte, y más específicamente, en el llamado séptimo arte, este fenómeno no sería la excepción, la cantidad de producciones emerge y se va anexando a la base gigantesca de películas creadas por la humanidad.

El presente reporte expone la aplicación de diversas metodologías de aprendizaje de máquina para la obtención de información de valor haciendo uso de los datos generados por la industria del cine.

La fuente

Los datos de insumo corresponden a una extracción dentro de un sitio web llamado Kaggle. Estos datos se actualizan diariamente y detallan la información de las películas que se encuentran en la página de reseñas de internet llamada TMDb. Esta página cuenta no solo con la información de películas, también de series, capítulos de series y aunque la misma página ofrece una API para realizar extracción de información, la base que se encuentra en Kaggle contiene únicamente la información de películas junto con otra información relevante, por lo cual, se opta por elegir esta última base como insumo de trabajo.

Los datos

Los datos cuentan con la información de un total de 750,318 películas y 20 características:

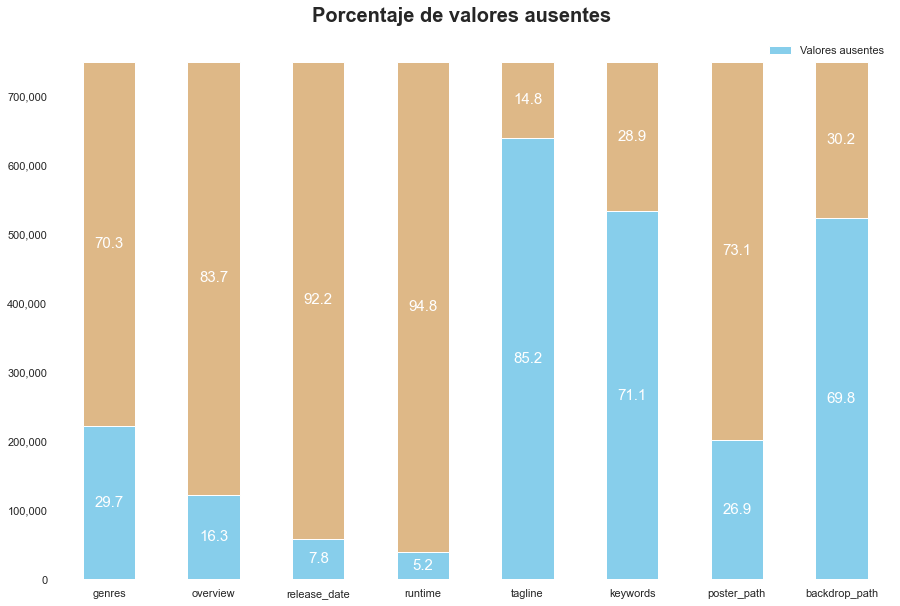
|  |  |
| --- | --- |
| Característica | Descripción |
| id | Identificador único en la página de TMDb |
| title | Nombre de la película |
| genres | Géneros relacionados |
| original\_language | Idioma original |
| overview | Resumen o sinapsis |
| popularity | Métrica de TMDb para medir la popularidad dentro del sitio web |
| production\_companies | Compañías productoras |
| release\_date | Fecha de estreno |
| budget | Presupuesto de producción |
| revenue | Ganancia reportada |
| runtime | Duración en minutos |
| status | Estado actual de la película (estrenada o no) |
| tagline | Breve frase alusiva a la trama de la película |
| vote\_average | Calificación promedio que los usuarios le dan a la película |
| vote\_count | Cantidad de calificaciones proporcionadas |
| credits | Nombre de los productores principales |
| keywords | Palabras clave sobre la película |
| poster\_path | URL para obtener el poster de la película |
| backdrop\_path | URL para obtener el backdrop de la película |
| recommendations | Recomendaciones a otras películas relacionadas |

De las características anteriores, hay algunas que, para propósitos de aplicación, no son relevantes en su uso, por lo que fueron descartadas. Las variables que se descartaron fueron:

|  |  |
| --- | --- |
| 1 | title |
| 2 | production\_companies |
| 3 | credits |
| 4 | recommendations |

Valores ausentes

Realizando la exploración de nuestros datos, se pudo observar que varias películas no cuentan con información de algunas características. Las características que no cuentan con algunos registros se presentan en el siguiente gráfico, donde, a su vez, se detalla el porcentaje de valores ausentes respecto al total de registros:



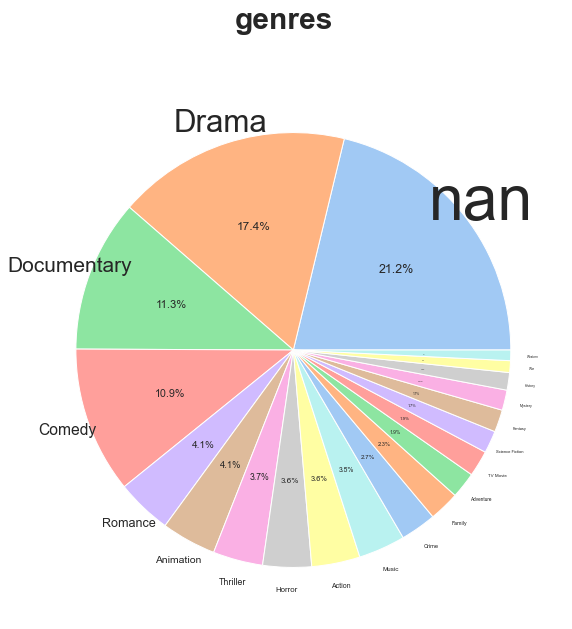
Del gráfico anterior, podemos observar que existen variables que cuentan con una gran cantidad de valores ausentes, sin embargo, el tratamiento que se optó por realizar deriva de la naturaleza de su uso. En consecuencia, el tratamiento de estos valores no se realiza de manera general dentro de la primera exploración. Posteriormente, se describe el tratamiento de los mismos en cada caso de uso.

Segmentación de variables

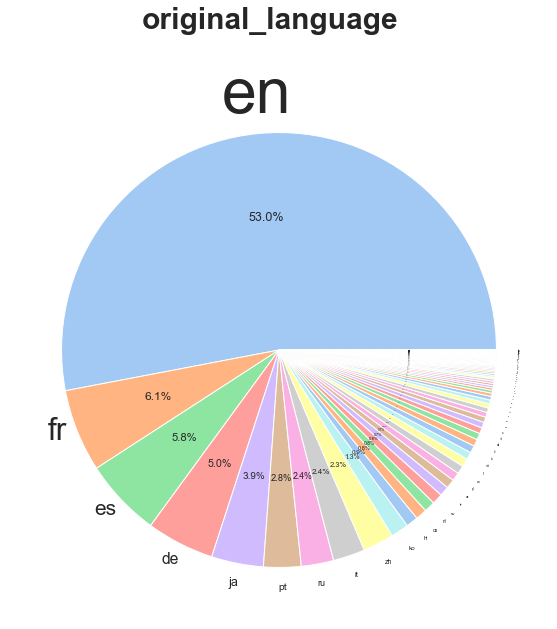
Una vez explicado el punto referente a los valores ausentes, las variables de nuestra base de datos se segmentaron de acuerdo a su tipo. La clasificación es la siguiente:

En los gráficos a continuación expuestos, se presenta un breve resumen de las características descritas

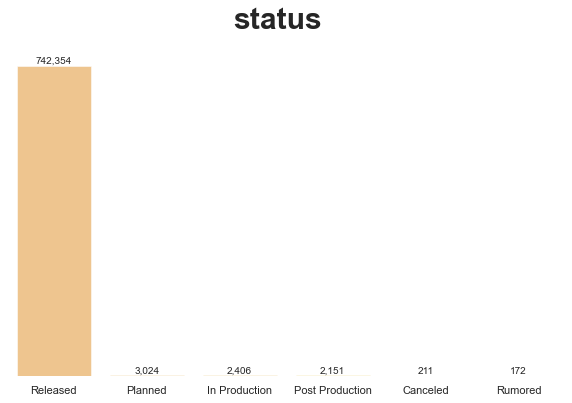
Variables categóricas



Analizando la gráfica de la derecha, la cantidad de valores ausentes juega un papel importante en la cantidad de géneros que se encuentran en nuestra información. Además, existen diversos géneros que si bien, existen en la base, su representación estadística es muy baja.



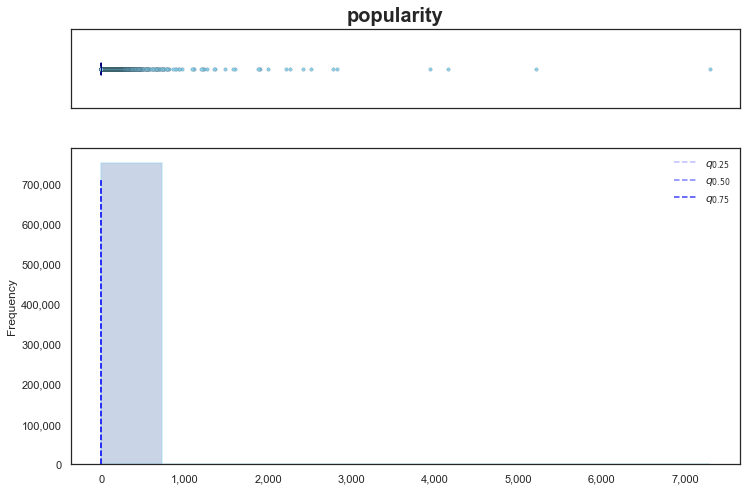
Para el caso del idioma original de las películas, aunque la variable no cuenta con valores ausentes, se puede apreciar la carga de información al idioma inglés, además de que varios idiomas no cuentan con una gran representación estadística.

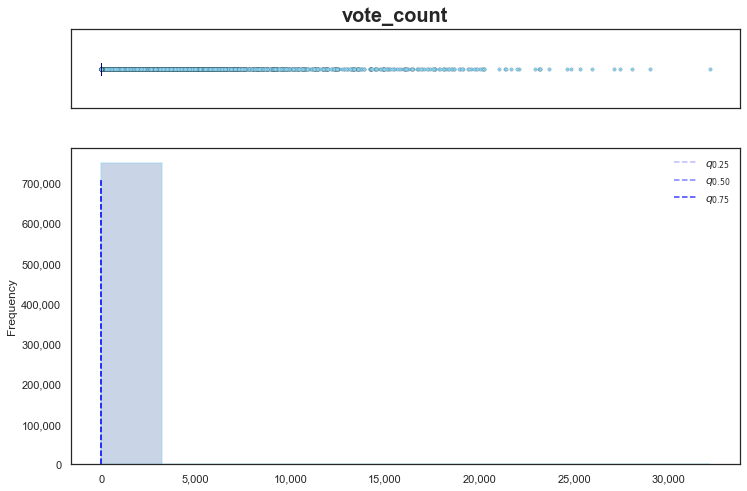
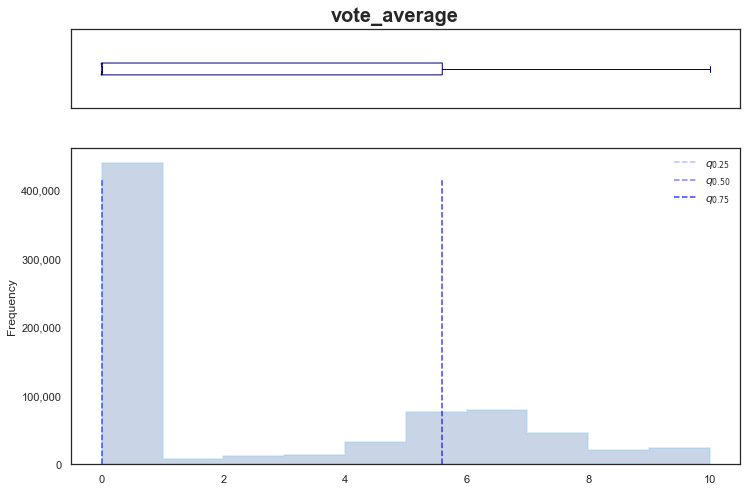


Finalmente, para la variable del estado de la película, es bastante evidente que la información prácticamente se basa en las películas que ya han sido estrenadas

Variables numéricas

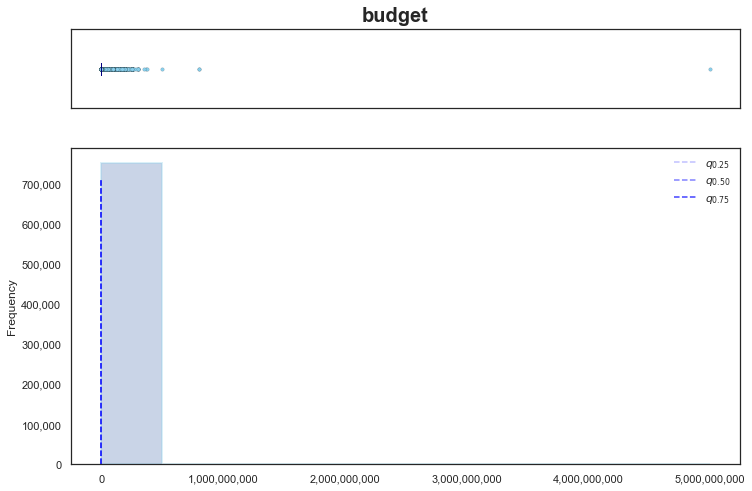
|  |
| --- |
| *Referentes a opiniones del usuario* |

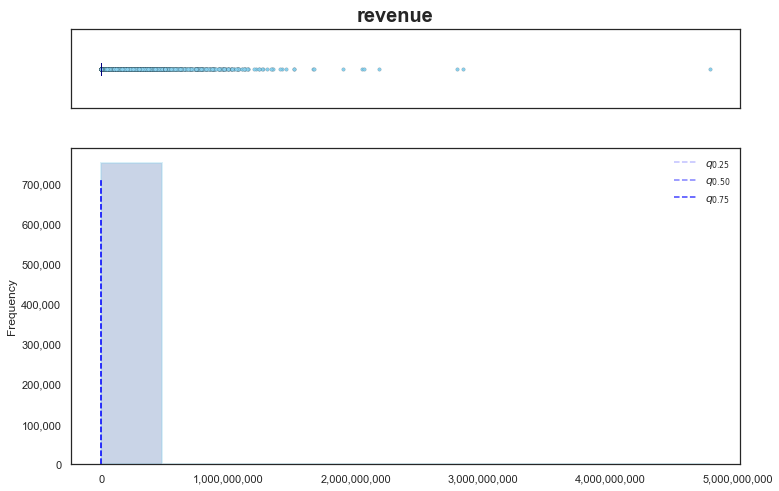




Las gráficas anteriores exhiben una carga de la distribución al valor 0. Esto es consecuencia a que existe una cantidad considerable de películas que, si bien, no cuentan con valores ausentes, los datos se registran como 0 y, por lo tanto, esto influye en la calificación promedio y en la métrica de popularidad.

|  |
| --- |
| *Referentes cifras monetarias* |





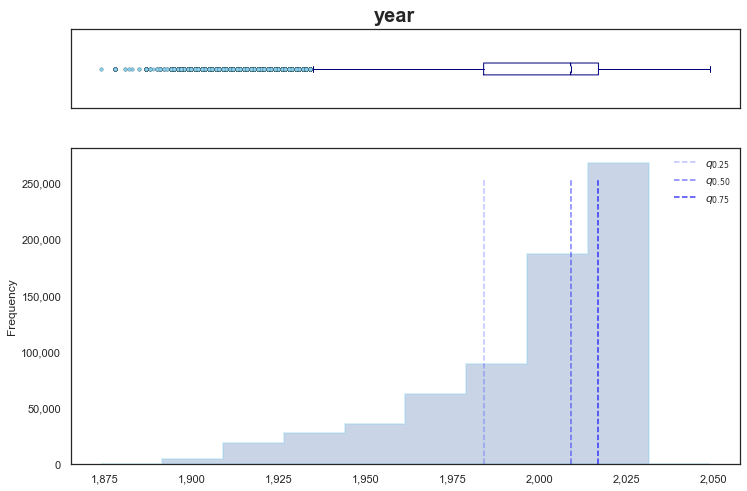
El comportamiento en estas variables es similar a las anteriores, si bien, no hay valores ausentes, su carga hacia el 0 se debe a que se registra la información como 0 al no contar con información

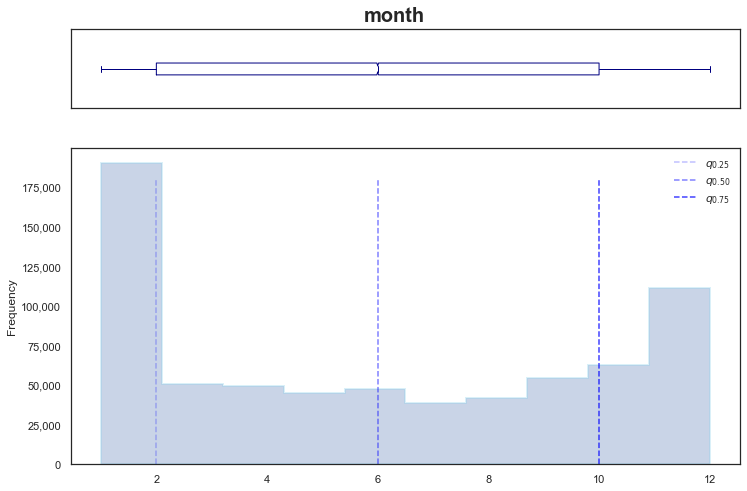
|  |
| --- |
| *Referentes información de la película* |

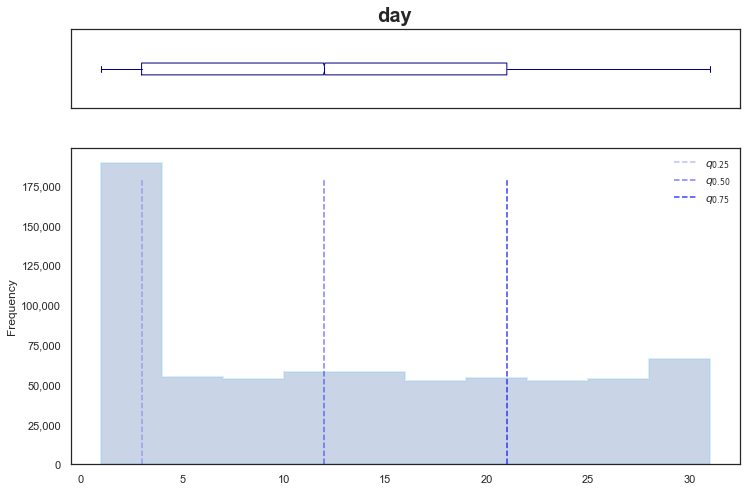


En este gráfico se puede apreciar la evidencia de un claro valor atípico o bien, una captura errónea que provoca un sesgo enorme en la distribución.

Las siguientes, son las distribuciones de la fecha de lanzamiento de las películas en la base:

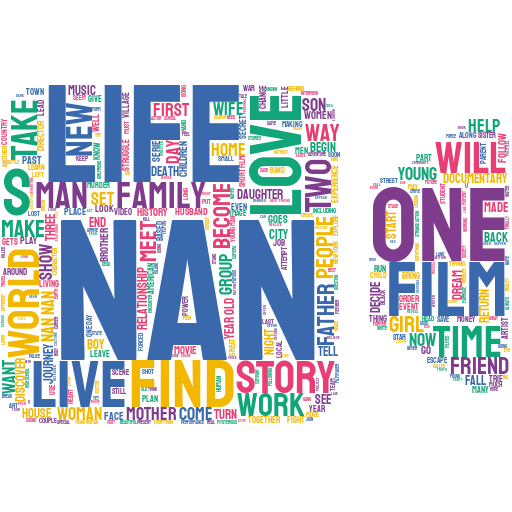


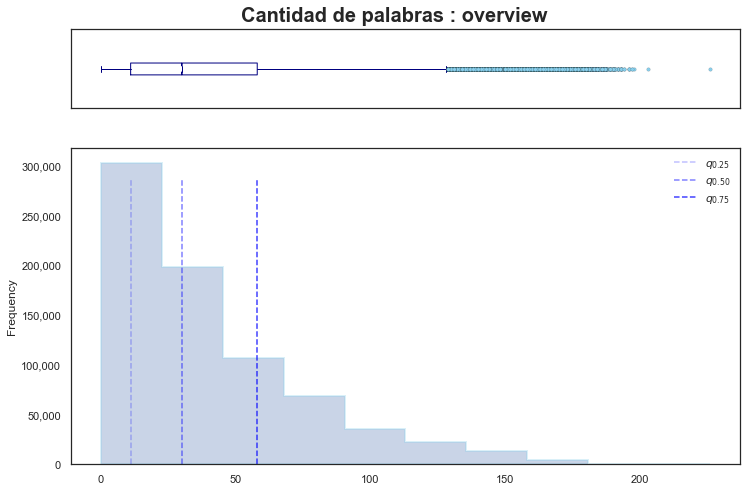




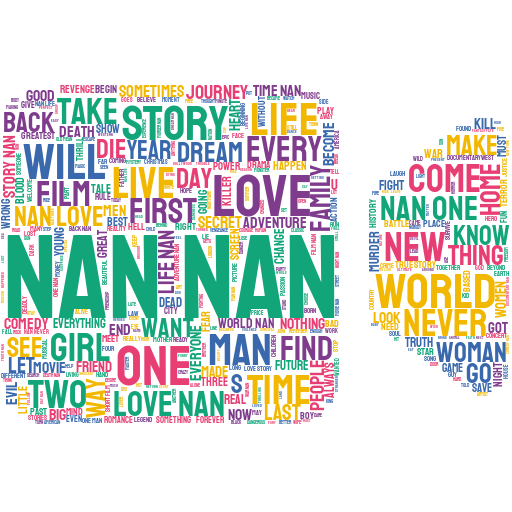
Variables de tipo carácter

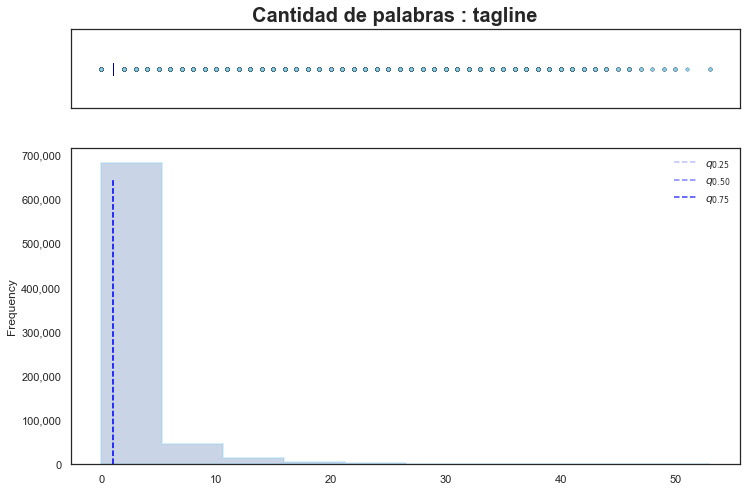
|  |
| --- |
| *overview* |



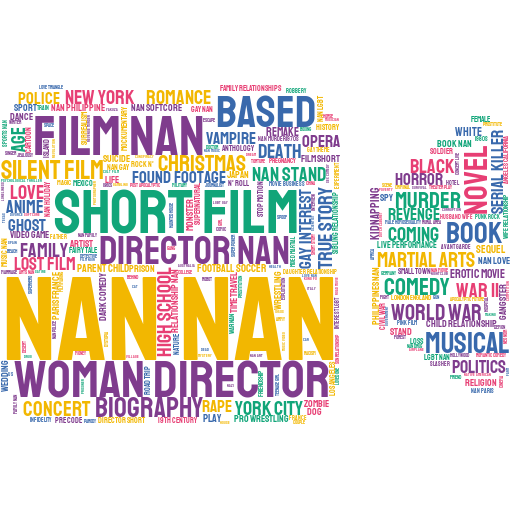


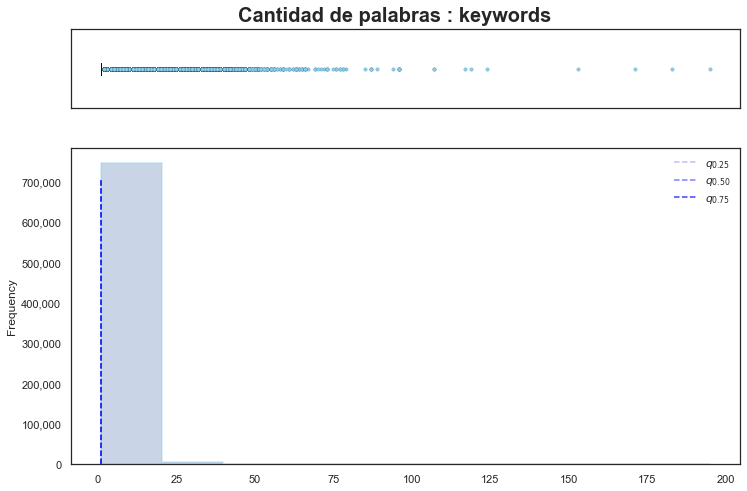
|  |
| --- |
| *tagline* |





|  |
| --- |
| *keywords* |





Finalmente, como conclusiones al análisis exploratorio, podemos mencionar que, aunque se cuenta con una gran cantidad de películas, muchas de ellas no cuentan con información relevante. Por esta razón, el uso de determinadas películas no será relevante para futuros análisis. A continuación, se presentan las metodologías aplicas con las consideraciones aplicadas para el correcto aprovechamiento de nuestros datos.

Modelaje supervisado

Objetivo

El incentivo principal del uso del modelaje supervisado se basa en la capacidad de predecir con ayuda de las variables que ya contamos. Por lo cual, el objetivo que se planteó y se trabajó para este apartado es el siguiente:

*“Predecir si una película es buena o mala”*

Este problema, al basarse en la predicción de alguna categoría, su naturaleza es de clasificación, por lo cual, se aplicaron diversos algoritmos de aprendizaje de máquina que atacan de manera eficaz nuestro objetivo:

Metodología

### Variable objetivo

Para realizar nuestro modelo de clasificación es necesario definir cuando una película es buena y cuando es mala. Por lo cual, el criterio que se tomó para esta segmentación se basa en la intuición natural sobre esta estratificación:

* **Una película es buena si sobrepasa el valor de 6.5 de calificación promedio.**

### Variables predictoras

Por otro lado, la elección de las variables que determinarían el que una película sea buena o mala obedece a la idea de poder ocupar variables que se encuentren disponibles incluso para una película recién estrenada. Por ejemplo, el ocupar como predictora la ganancia de una película no podría ser tan viable ya que es un dato que no necesariamente se puede obtener o bien, el dato no podría ser tan verídico.

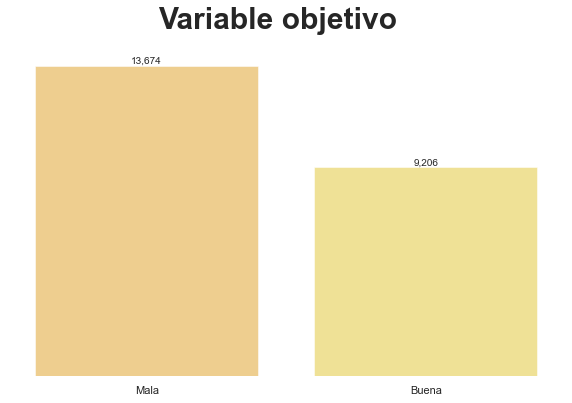
Teniendo en cuenta lo anterior, las variables que se consideran para la construcción del modelo son las siguientes:

Así, para tener insumos de valor en nuestros modelos, se procede a considerar los siguientes filtrados:

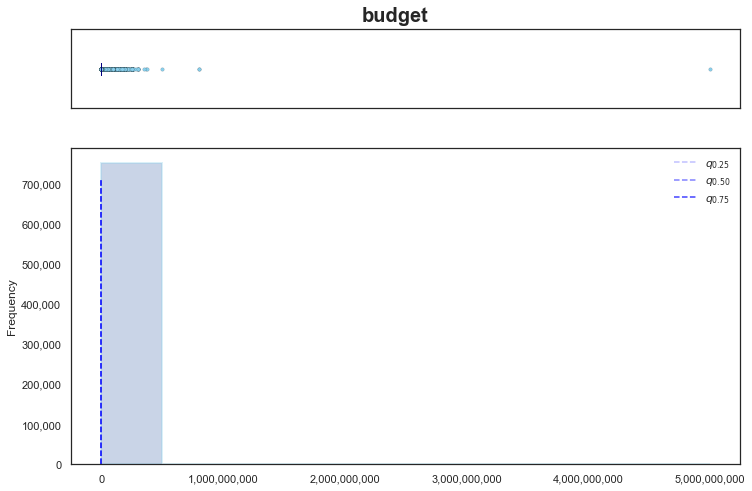
* Filtrado de películas con si cuentan con sinapsis
* Filtrado de películas que si cuentan con información del presupuesto
* Filtrado de películas que si cuentan con información de calificaciones
* Filtrado de películas que cuentan con información de géneros relacionados

**Una vez realizado el filtrado, se obtiene un base de datos con una cantidad total de 22,880 películas.**

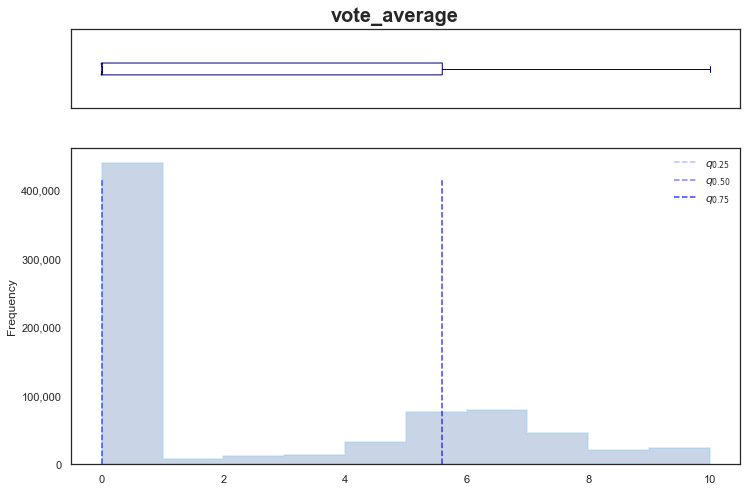
En el siguiente gráfico podemos ver la cantidad de películas, que, de acuerdo a nuestra segregación, se clasifican como una película buena o mala:

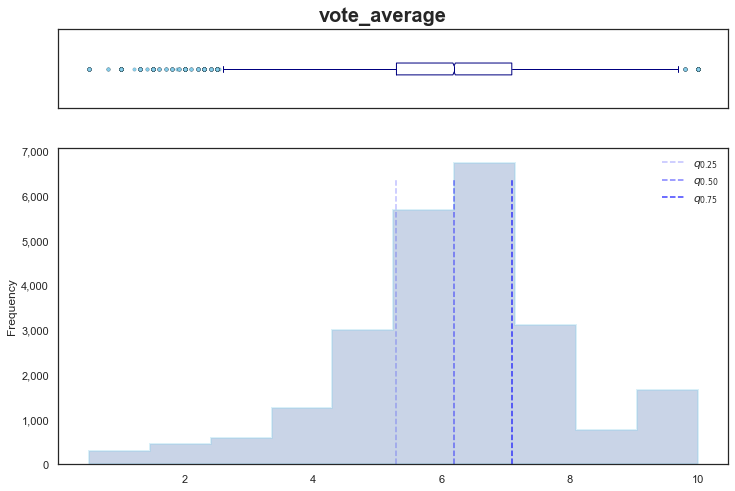


Adicional, se muestran los cambios significativos que sufren las distribuciones al realizar los filtrados.



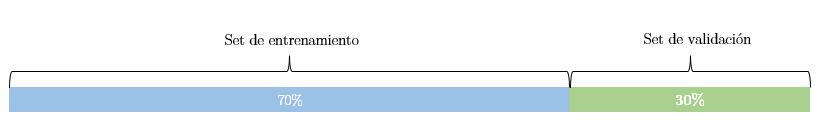






### División de los datos

Posterior, para poder entrenar nuestros modelos y generar métricas de desempeño, se realiza la división de nuestro set de datos en dos sets: el de entrenamiento y el de validación. Las proporciones se realizaron como sigue:

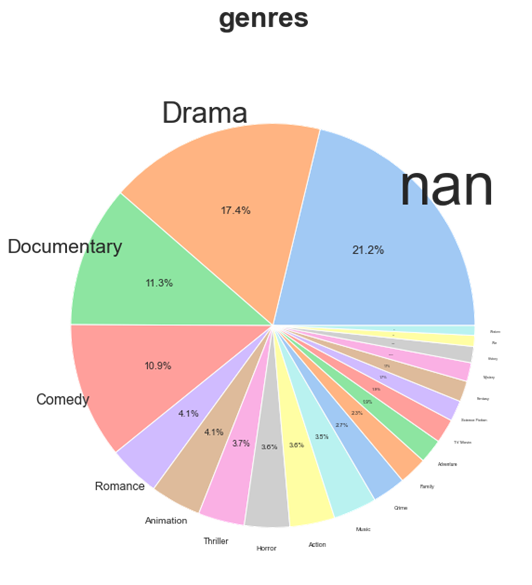


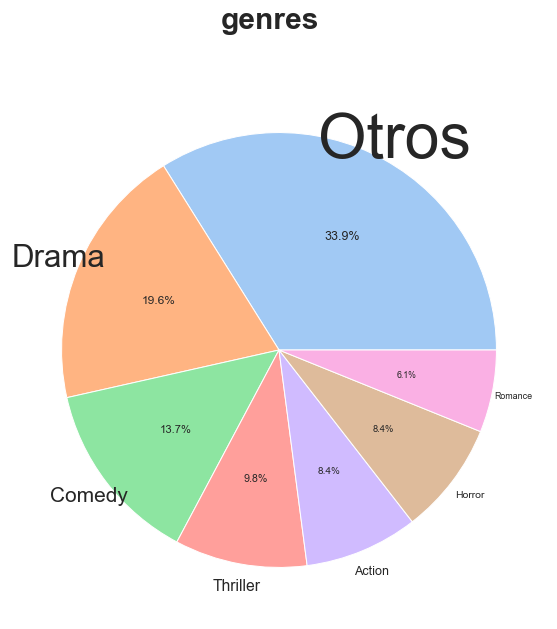
Finalmente, como consideraciones adicionales para poder realizar los entrenamientos se realizaron los siguientes pasos:

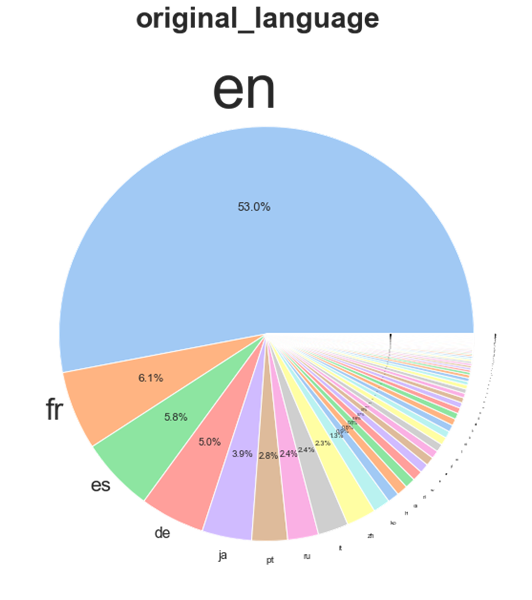
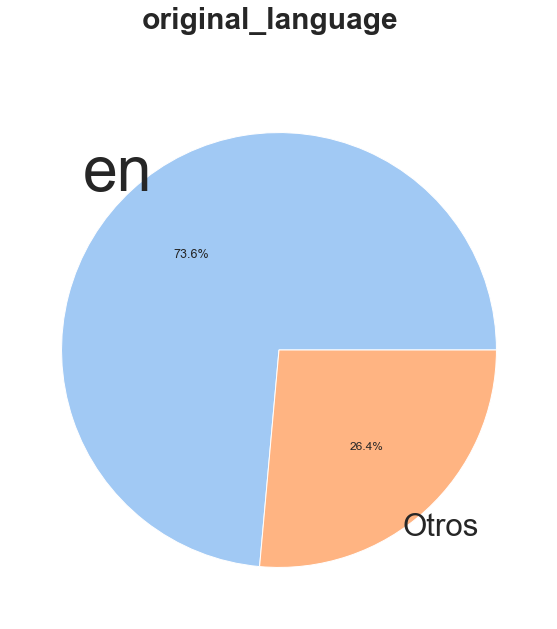
* Normalización de la variable de género
* One hot encoding de la variable de género
* Normalización de la variable de idioma
* One hot encoding de la variable de idioma
* Limpieza de las variables de tipo carácter
* Word embedding de las variables de tipo carácter
* Escalamiento de variables

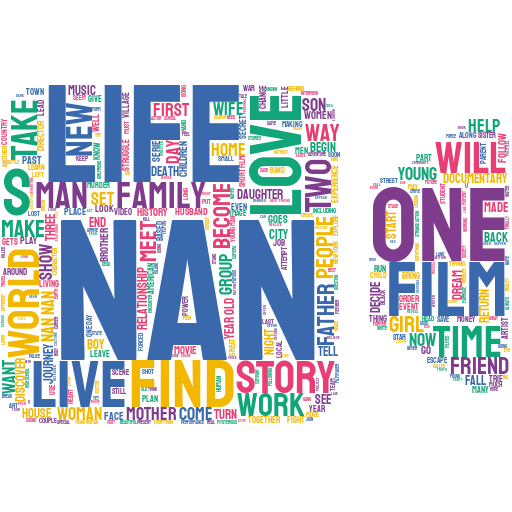
Cabe mencionar que estos últimos pasos se realizan dentro del set de entrenamiento, ya que se busca que estos pasos preserven la estructura general y se apliquen de manera independiente al set de validación. Así, en caso de requerir generar una predicción fuera de la muestra, los pipelines de datos pueden funcionar de manera correcta y se puede generar la nueva predicción.

Como consecuencia, en los siguientes gráficos podemos ver algunos cambios que se suscitan al realizar estos últimos pasos y los filtrados mencionados.











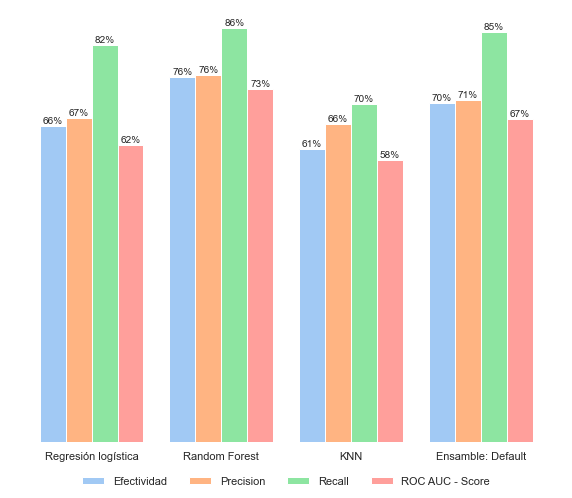
Resultados

Teniendo los datos necesarios de insumo, se procedió a realizar los entrenamientos. Se entrenaron dos modelos para cada algoritmo, uno correspondiente a un modelo con parámetros default y otro obtenido al realizar el tunning de parámetros, eligiendo aquellos parámetros que lograron el mejor desempeño en el conjunto de validación usando Cross-Validation. Adicional, se consideraron dos métodos diferentes de word embedding, el método TF-IDF y el método Doc2Vec. Por lo cual, se tienen los resultados de los entrenamientos ocupando cada tipo de word embedding.

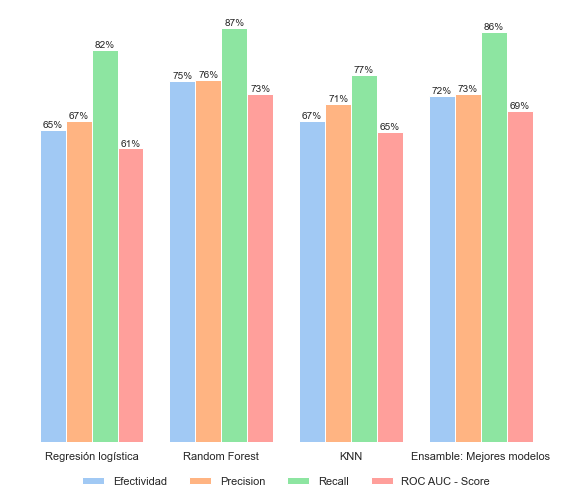
Es importante mencionar que se optó por elegir 100 features resultantes en cada word embedding y que las métricas expuestas fueron obtenidas evaluando los modelos en el conjunto de validación.

### Modelos ocupando TF-IDF.



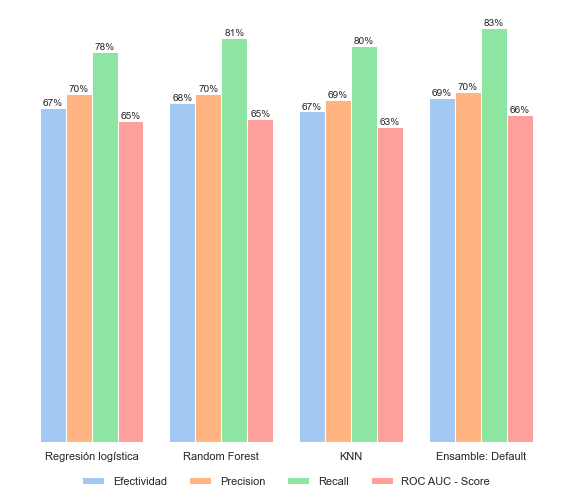




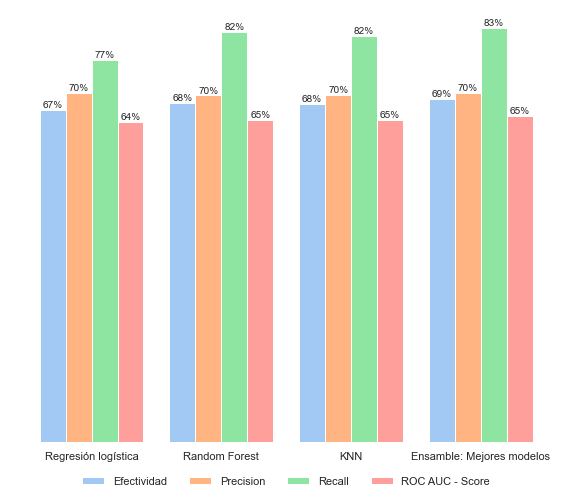


### Modelos ocupando Doc2Vec.







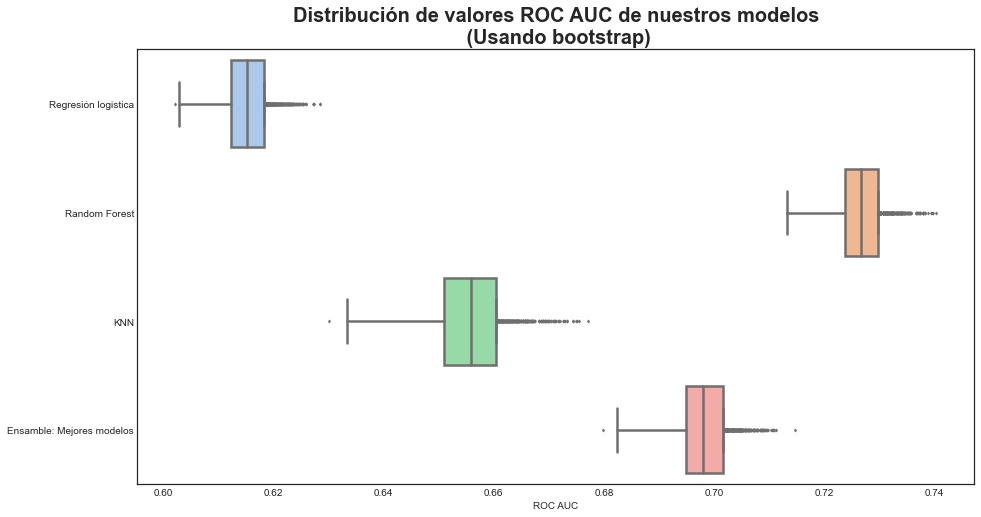


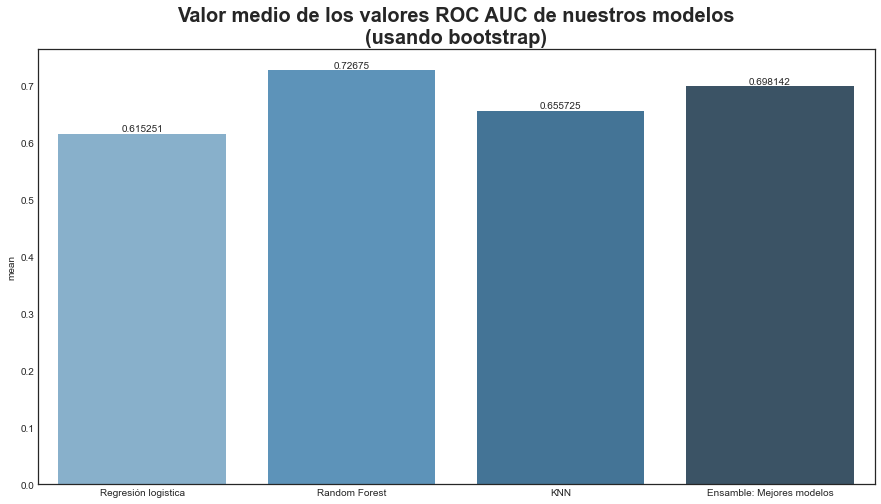
### Estabilidad

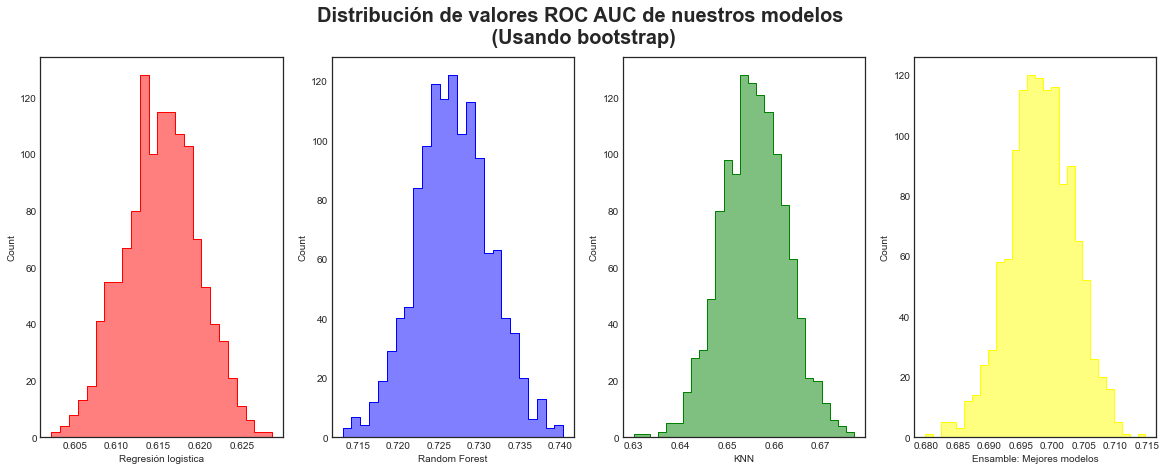
A manera de verificar que nuestros modelos estén brindando métricas confiables y poder generar una elección de modelo, se realizó un proceso de bootstrap con los modelos obtenidos de hacer el tunning de parámetros.

Para este proceso, solamente se hizo bootstrap con los mejores modelos usando el embedding de TF-IDF, ya que los modelos usando Doc2Vec aunado a que brindan peores resultados, el proceso de entrenamiento es considerablemente mayor al de TF-IDF, por lo cual, realizar bootstrap con este método de embedding resultaría en un tiempo excesivo de cómputo.

Los siguientes resultados recuperan el valor ROC AUC de cada modelo elegido realizando un total 1,250 iteraciones bootstrap que tomaron alrededor de 9 hrs de cómputo:







Por lo tanto, de las gráficas anteriores, podemos notar estabilidad en las métricas de validación y es viable elegir nuestro modelo final.

Elección de modelo

Las métricas obtenidas hacen evidente que los modelos ocupando Doc2Vec no brindan resultados que avalen la preferencia en su uso, por lo que, si nos enfocamos en los modelos usando TF-IDF, el modelo Random Forest demuestra tanto en métricas bootstrap como en entrenamientos sencillos, una eficacia notoria comparando contra los demás modelos.

Por lo tanto, la elección de nuestro modelo final es un Random Forest con los siguientes parámetros:

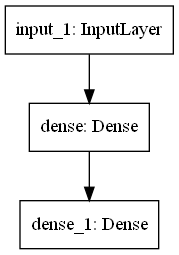
* max\_depth = 50
* min\_samples\_split = 4
* n\_estimators = 200

Haciendo uso de este modelo, podemos obtener una eficacia en nuestra clasificación de aproximadamente 76% y un valor de ROC AUC de aproximadamente 72%.

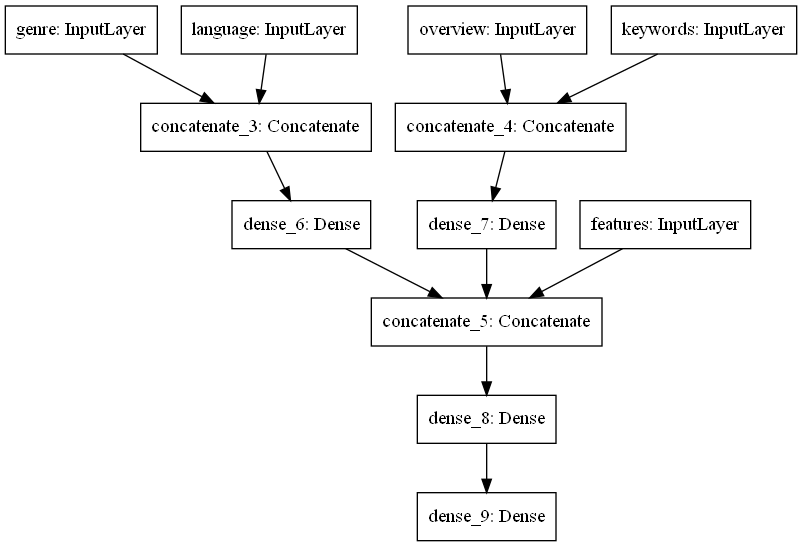
Redes neuronales

A manera comparativa, también se entrenaron tres tipos de redes neuronales. Todas se entrenaron con un early stopping de 50 épocas sobre la efectividad en el set de validación y las estructuras son las siguientes:

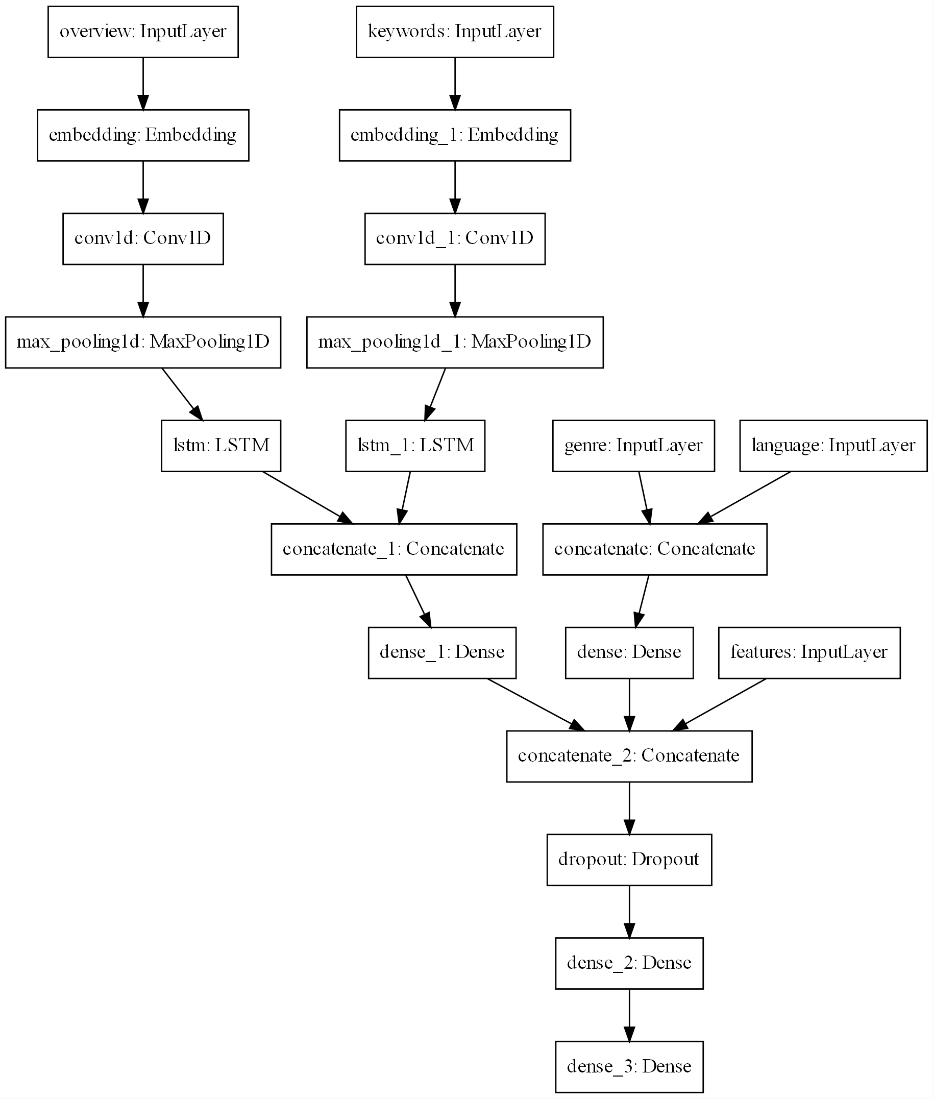
Estructura 1:



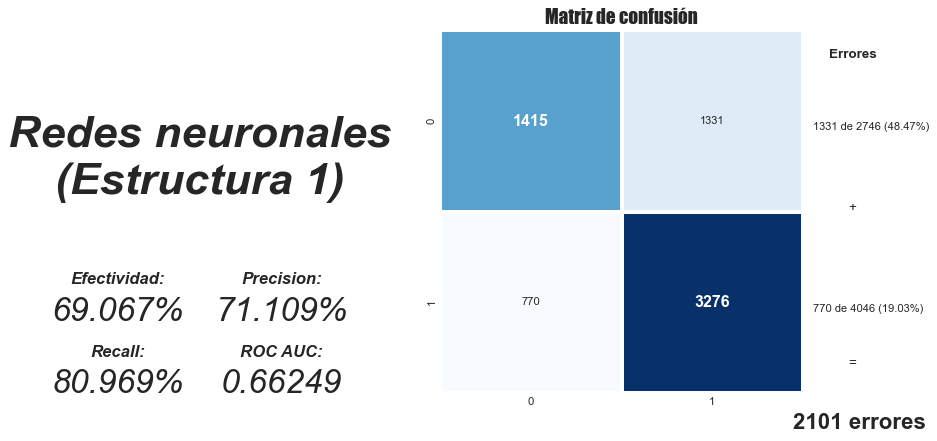
Estructura 2:

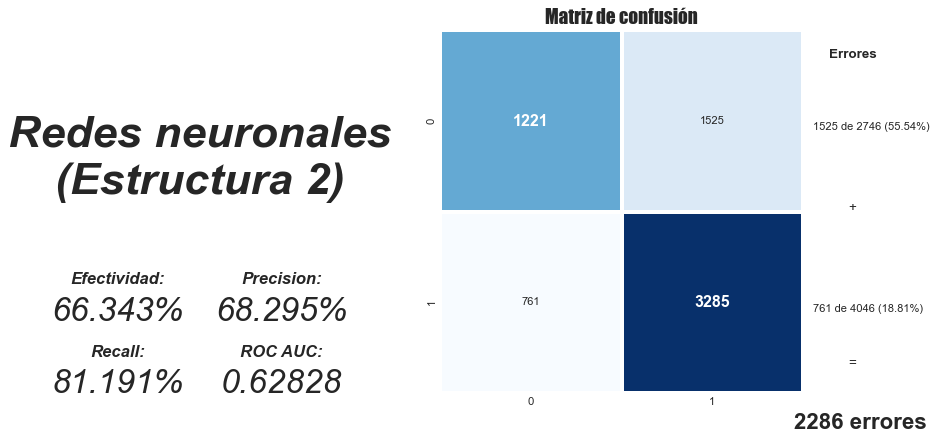


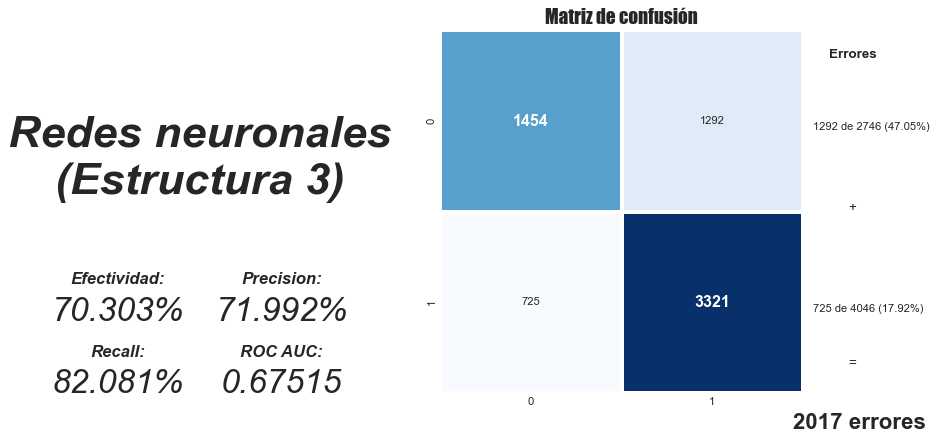
Estructura 3:



### Resultados







Conclusiones

Los resultados de las redes anterior entrenadas muestran un buen desempeño (sobre todo la última estructura) dando pauta a un modelo incluso mejor que algunos modelos que anteriormente habíamos expuesto, sin embargo, las métricas obtenidas aun no son mejores que el modelo Random Forest que se había elegido.

A pesar de esto, las redes neuronales brindan posibilidades de establecer un sin fin de parámetros que podrían incluso superar las métricas del modelo que elegimos, sin embargo, para esto, se requieren hacer muchísimas pruebas exhaustivas en busca de estos parámetros.

Por lo tanto, aunque el uso de redes neuronales puede resultar beneficioso para ciertos tipos de problemas, en este caso, su uso no es tan eficiente si comparamos contra otros modelos menos complejos.

Modelaje no supervisado

Objetivo

El uso de modelaje no supervisado tiene diversas aplicaciones para la predicción de valores; de manera muy general, pretende generar agrupaciones de la información. En particular, para nuestro estudio, su incentivo de uso en nuestro proyecto surge de la siguiente idea

*“Segmentar los usuarios con gustos por el cine”*

Metodología

Para lograr lo anterior, era necesario contar con una base de datos que proporcionara información de usuarios; nuestra base no contaba con dicha información. Así, realizando investigaciones, GroupLens cuenta con bases de usuarios de TMDb que proporcionan calificaciones a las películas del sitio.

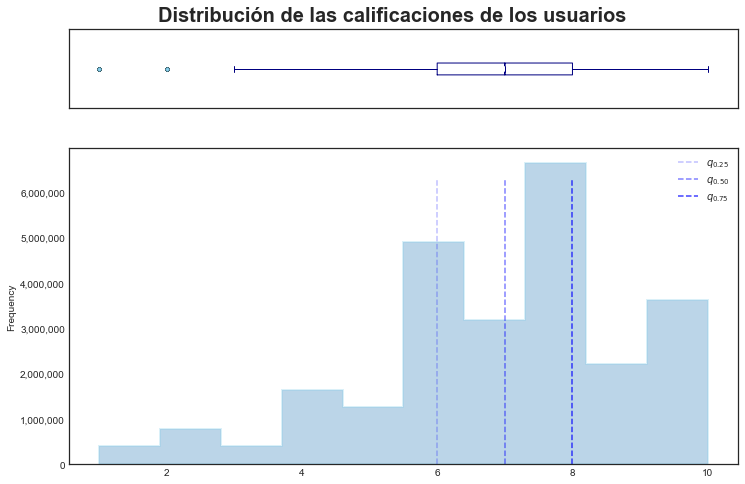
### Base de datos de calificaciones de usuarios

La información que contiene esta base de datos consolida la siguiente información:

En consecuencia, la base cuenta con:

* ID del usuario que califica
* ID de la película calificada
* Calificación brindada

El siguiente gráfico muestra la distribución de las calificaciones que los usuarios brindan



### Consideraciones

A pesar de ya contar con información de usuarios, nuestra base de películas contaba con información que no necesariamente sería relevante para nuestra segmentación de usuarios.

Como primera consideración, las variables categóricas no iban a figurar para el modelado. Posterior, al igual que en el modelaje supervisado, las variables elegidas obedecen a la idea de ser variables que se puedan obtener de una manera efectiva para todos los usuarios. En consecuencia, las variables conservadas para cada película fueron:

* Duración de película,
* Año de estreno
* Calificación promedio

Y, al igual que las variables categóricas, las variables de tipo carácter no fueron consideradas ya que, en el fondo, obedecen a la idea de generar categorías, en consecuencia, el entrenar modelos con estas variables no resultaría buenos modelos.

### Consolidación de la información

Teniendo en cuenta lo anterior, era necesario filtrar nuestra base de películas obedeciendo los criterios impuestos, por lo que se realizo:

* Filtrado de películas que tengan registrada duración
* Filtrado de películas que si cuenten con fecha de lanzamiento
* Filtrado de películas que tengan registrada una calificación promedio

Cabe mencionar que, aunque no figuran las variables categóricas no figuran en el modelado, estas se conversan para futuros análisis de perfilamiento.

Finalmente, teniendo las variables relevantes de nuestra base de películas, se procedió a realizar la consolidación de bases con ayuda de nuestra llave de unión, el ID de la película calificada.

### Generación de la TAD

Resultando un total de 11,839,138 calificaciones de usuarios al realizar el cruce, el siguiente paso para conseguir nuestro objetivo sería obtener la información a nivel usuario, sin embargo, hasta este punto la información se encuentra a nivel calificación a película, por lo cual se realizó la agrupación de información para cada usuario.

Como resultado, las variables finales de insumo de cada para generar nuestros modelos serían las siguientes:

**Finalmente, obteniendo la información de 162,541 usuarios.**

### Modelos a entrenar

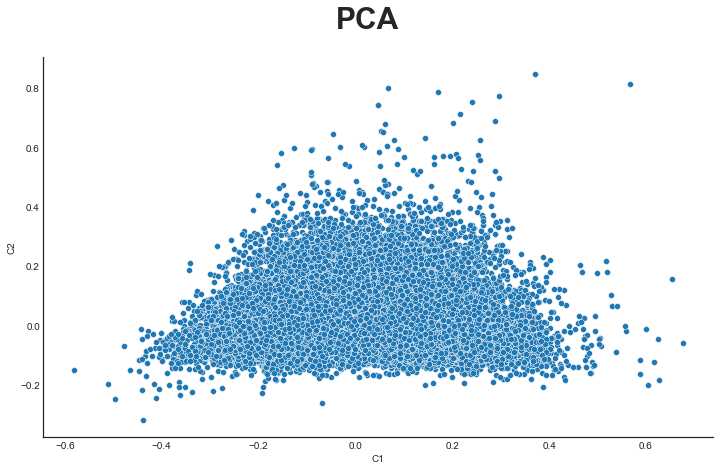
Los modelos a entrenar que nos ayudan a generar segmentaciones y lograr nuestro objetivo son los siguientes:

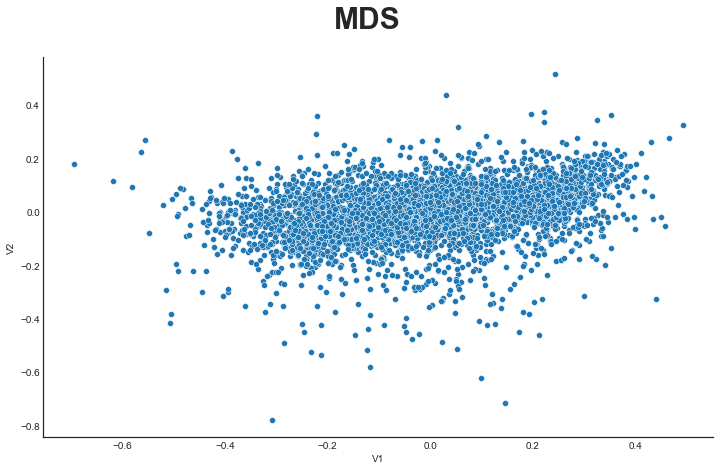
Resultados

A continuación, se presentan los resultados de los entrenamientos de los modelos anterior propuestos.

### Visualización en dimensión reducida

Primeramente, se realizó una visualización de nuestras características ocupando dos algoritmos de reducción de dimensión. De esta manera, con ayuda de esta visualización se pretendía llegar a inferir una evidente segmentación de información.





Como nota a esta última visualización, corresponde a una visualización de una muestra de solo 5,000 usuarios, ya que, al intentar ajustar el algoritmo con la totalidad de la información, este requería una cantidad enorme de recursos computacionales.

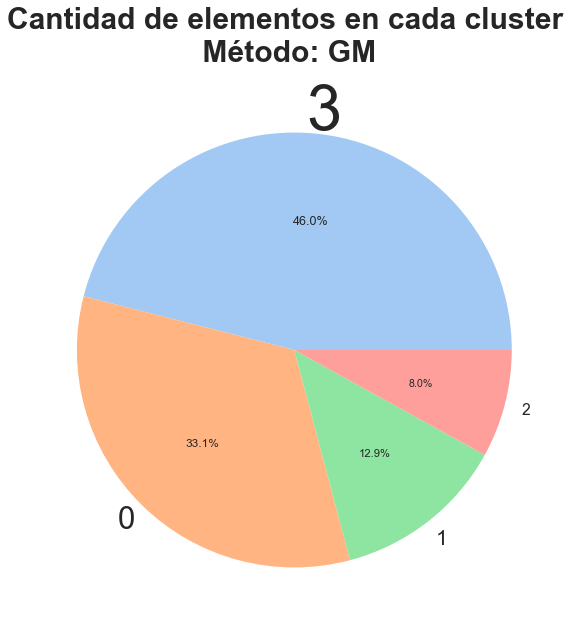
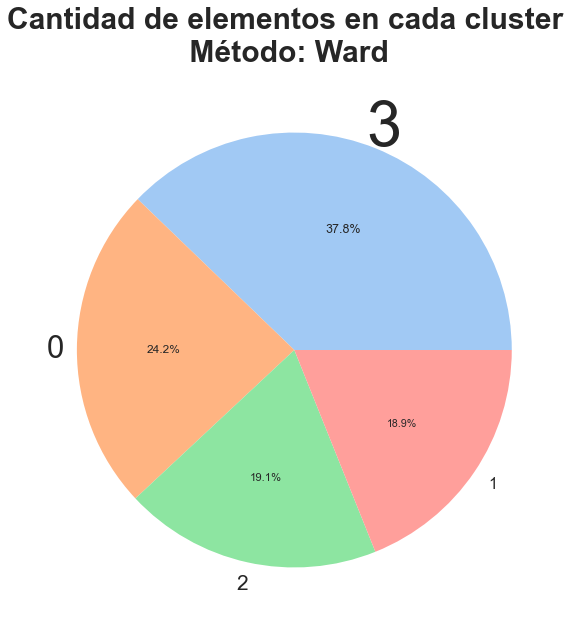
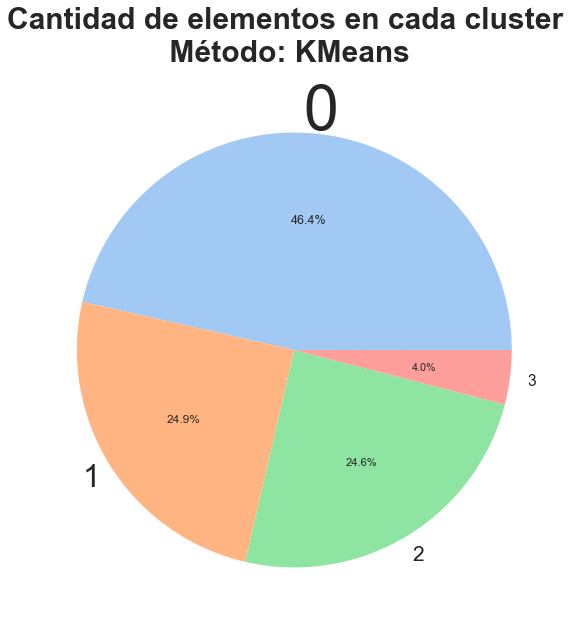
Así, de las visualizaciones en dimensión reducida, no marcan una pauta clara para la elección para una cantidad optima de segmentaciones.

### Modelos entrenados

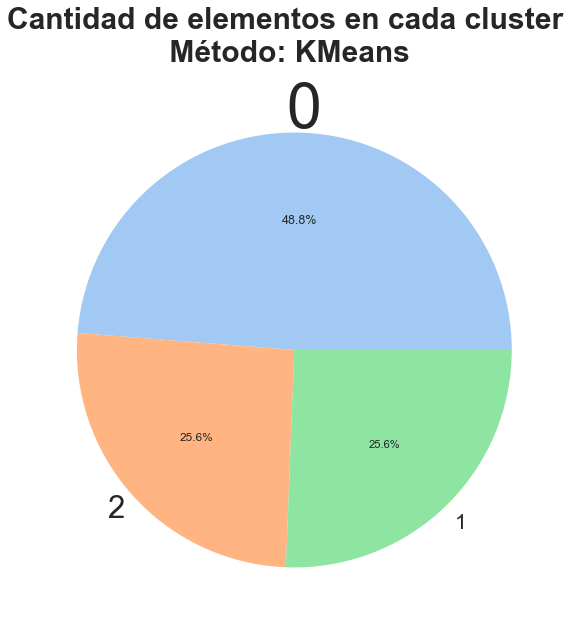
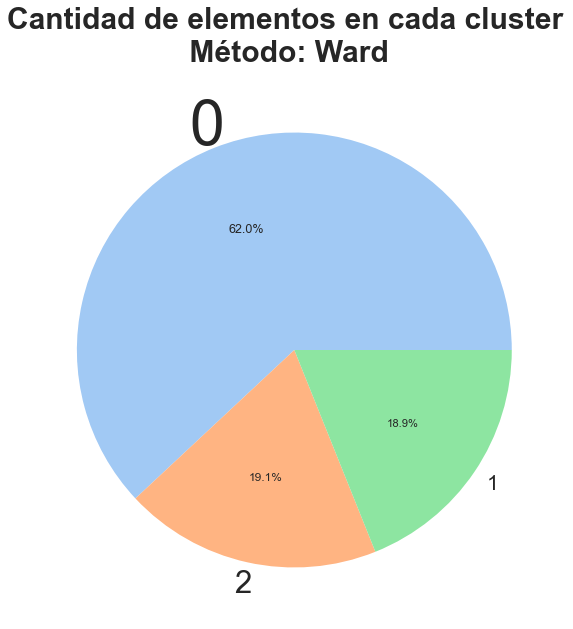
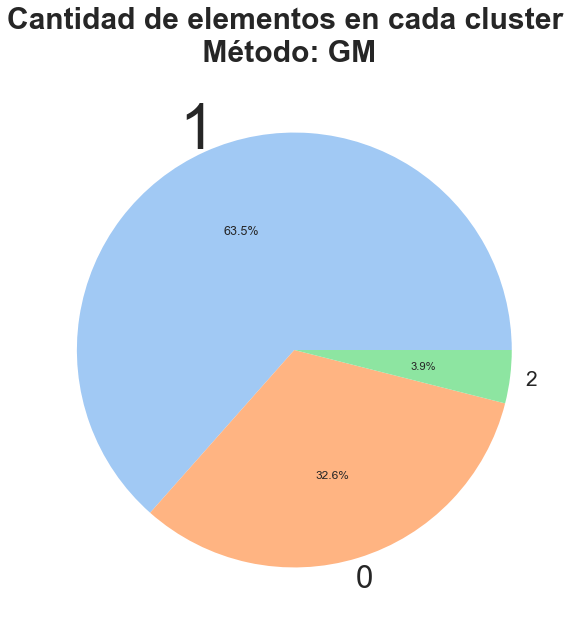
El modelo óptimo que se consideró fue aquel que presentara una estructura de segmentación adecuada y además tuviera un score de silueta lo más próximo a uno.

De esta manera, se realizaron diversos entrenamientos, probando cantidades diferentes del número de segmentaciones haciendo uso de los algoritmos de K-Means y Gaussian Mixture. El algoritmo agglomerative clustering presentaba fallos al requerir un tamaño de memoria excesivo, por lo que las métricas correspondientes a este algoritmo, fueron obtenidas entrenando los modelos con una muestra de 5,000 observaciones.

#### Ocupando 4 clústers



#### Ocupando 3 clústers



Elección de modelo

De los resultados anterior expuestos, podemos observar que el modelo con mejor score de silueta es el modelo K-Means con 4 clusters, sin embargo, si observamos la segmentación que realiza, el cuarto cluster prácticamente no cuenta con usuarios. Así, analizando el segundo modelo con mejor score de silueta también resulta ser un modelo entrenado con K-Means con la diferencia de estar ajustado con una cantidad total de 3 cluster pero en esta ocasión ahora este modelo cuenta con una clara segmentación de usuarios.

**Por lo tanto, el modelo que se elige para realizar la correcta segmentación de usuarios de TMDb es el modelo K – Means ajustado con un total de 3 cluster, logrando una eficacia en el score de silueta de 0.3476.**

Pruebas estadísticas.

A pesar ya haber elegido el modelo, es necesario verificar que el modelo elegido en efecto es confiable, por lo cual, se procede a realizar las siguientes pruebas estadísticas para poder validar que las variables usadas cuentan con suficiente evidencia estadística para determinar un buen modelo.

Las pruebas realizadas son las siguientes:

|  |  |
| --- | --- |
| Prueba | Uso |
| Kruskal - Wallis | Comparar variables entre los clusters |
| Kruskal - Wallis | Comparar variables de cada clúster contra la TAD |
| Tuckey | Comparar variables entre pares de clusters |

### Kruskal – Wallis (Entre clusters)



### Kruskal – Wallis (Contra la TAD)







### Tuckey



Conclusiones

La evidencia estadística nos brinda la posibilidad de concluir que el modelo elegido cuenta con un buen desempeño y que además las variables usadas para su entrenamiento han sido representativas.

**Finalmente, podemos afirmar que el mejor modelo para solventar nuestro objetivo de segmentación de usuarios es el modelo K – Means ajustado con 3 clusters.**

Perfilamiento

Elegido el modelo, a continuación, se presenta el análisis de los segmentos encontrados. Tres segmentos, cada uno con características diferentes.

Cabe mencionar que, para el presente perfilamiento, se opta por usar, además de las variables con las que se entreno el modelo, las variables categóricas. Estas variables consideran una ponderación por gusto, es decir, si el usuario proporciono una calificación mayor a 7 a la película, las variables categóricas correspondiente tomarán el valor de 1

A continuación, se presentan las estadísticas relevantes de los clústers:







De estas métricas, se propuso el siguiente perfilamiento para futuras predicciones y estrategias de mercado aplicadas a cada segmento:



Resumen

El mejor modelo para atacar el problema propuesto en la modelación supervisada:

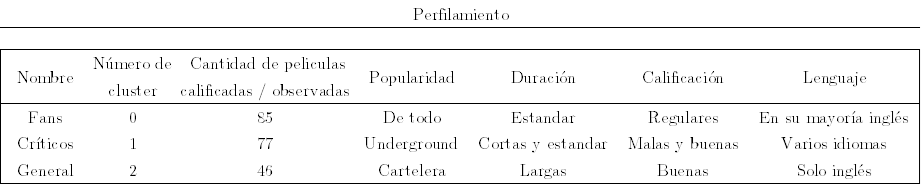
“Predecir si una película es buena o mala”

Es un Random Forest que brinda una eficacia promedio de 76%

El mejor modelo para atacar el problema propuesto en la modelación no supervisada:

“Segmentar los usuarios con gustos por el cine”

Es un modelo K – Means, ajustando 3 segmentaciones que resultan en el siguiente perfilamiento:



Comentarios finales

El uso de algoritmos de aprendizaje brinda posibilidades increíbles. Este proyecto es un ejemplo del aprovechamiento de estos algoritmos, logrando atacar de manera eficaz problemas que pueden surgir en el día a día.

Adicional, vale la pena mencionar que, aunque los modelos anteriores expuestos han tenido un buen desempeño, aún queda bastante trabajo por realizar en harás de encontrar un mejor desempeño y poder llevar estos modelos a un entorno más real y productivo.