Módulo 2

Sebastián Buitrago Gómez

Sebastián Ciro

Juan Esteban Marulanda Ayala

Docente:

Juan Camilo España Lopera

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería industrial

Aplicaciones de la Analítica

Medellín

2023-1

a.) Diseño de la solución propuesto

Problema de negocio.

La plataforma online de películas no cuenta con un sistema de recomendación para los usuarios. La ausencia de un sistema de recomendación no deja conocer tendencias en películas, no se genera un conocimiento segmentado de los clientes o individual y no permite crear proyecciones y estrategias de servicio en la plataforma; lo anterior posiblemente implique una reducción significativa en los cineastas que visitan la plataforma. El objetivo es crear modelos de recomendación que permita fidelizar a los usuarios y brindar una experiencia más personalizada y amena; lo que se traduce como más estabilidad del usuario dentro de la plataforma y más visitas a las películas dentro de ella.

Problema analítico

Realizar una separación de las soluciones que se pueden dar, una parte a partir del análisis exploratorio en donde se realiza una entrega de informes de recomendación por popularidad o ranking, agrupaciones por géneros, por películas e incluso horas; posteriormente se hace una valoración de modelos basados en contenido, inicialmente a partir de una sola película vista. Se utilizan dos métodos, una funciona a partir de las correlaciones de la misma película y una clusterización con un modelo no supervisado de aprendizaje automático, *KNN*.
El segundo modelo basado en contenido se utiliza para todas las películas vistas por el usuario knetworks centroide, en donde se obtiene una lista de películas recomendadas. Finalmente se entregan un par de modelos de aprendizaje supervisado, los filtros colaborativos, el primero basado en usuarios, que mide cuales usuarios se parecen según las películas bien calificadas; y segundo basado en ítems. Ver el gráfico 1.

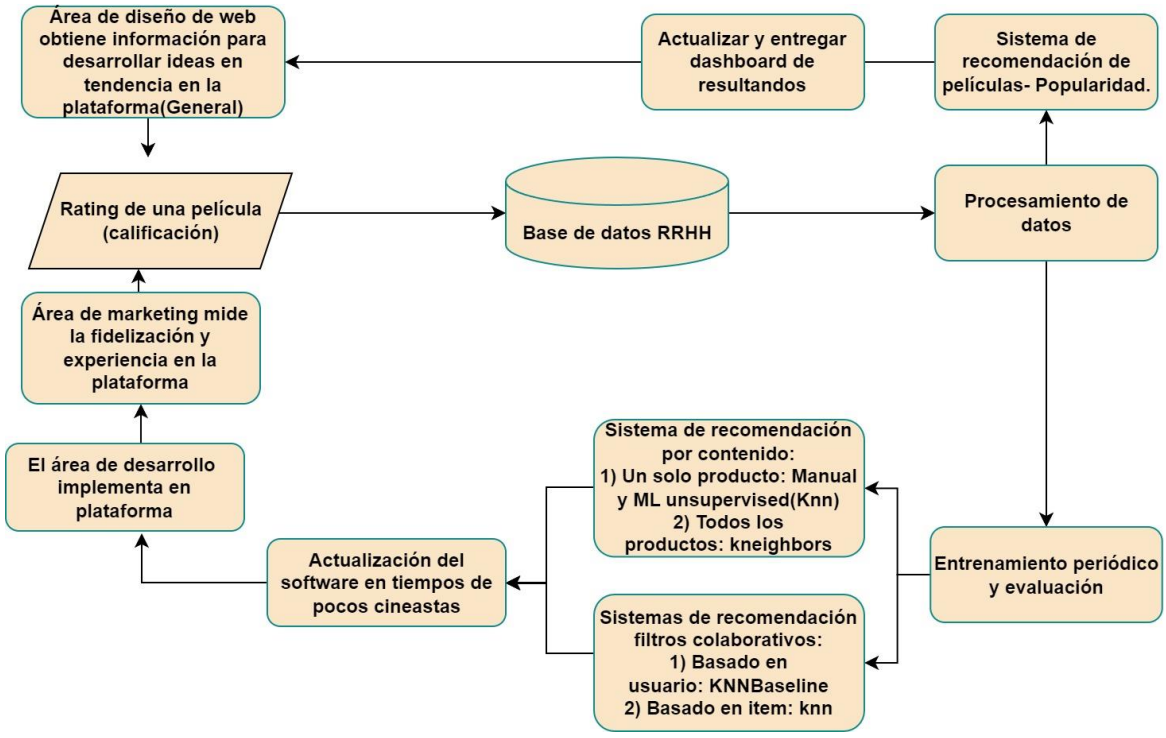


Gráfico 1/ Diseño de la solución

Limpieza y transformación. Para la limpieza de la información se utilizó lenguaje SQL buscando una descripción inicial de las tablas y la separación de las variables, se encuentran 9742 películas dentro del catálogo de la plataforma, una lista de 100.836 calificaciones de 610 usuarios que cuentan en promedio con poco más de 80 calificaciones. En términos de géneros, el drama y la comedia son los que más aceptación tienen, con más del doble respecto a las demás categorías. Se hace una separación de la variable género para obtener información más segmentada y organizada sobre las particularidades de cada película, se convierte la variable "tiempstap" en tipo dato fecha para un manejo más uniforme. Se integran las tablas al realizar el escalado de la variable año, al hacer la transformación numérica de las variables categóricas y la separación de los géneros como dummies.

Análisis exploratorio. Dentro del análisis exploratorio se utiliza también una extensión SQL para realizar algunas consultas referentes a la información de las tablas; se encuentra información pertinente al comportamiento de los datos de forma general, un resumen se encuentra a continuación en donde se observa un promedio del ranking de la plataforma, 3,9, las visualizaciones por año, mensuales, y los comportamientos del tiempo que muestra la tendencia de las visualizaciones a lo largo del día, por género.

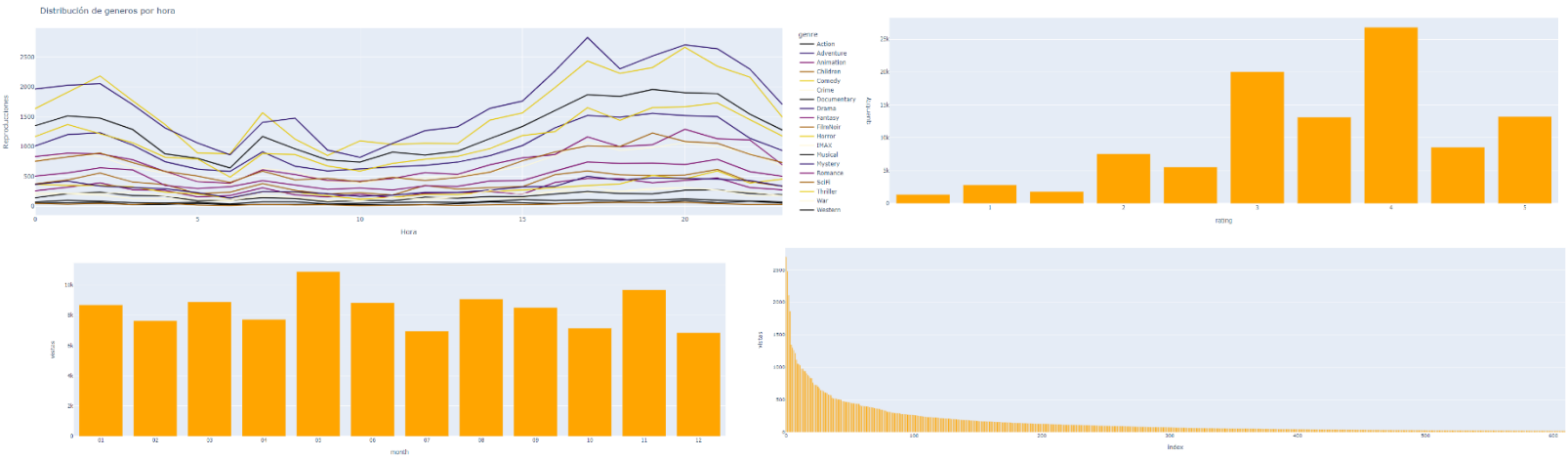
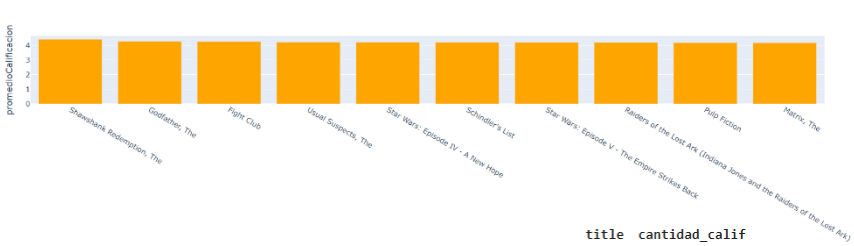
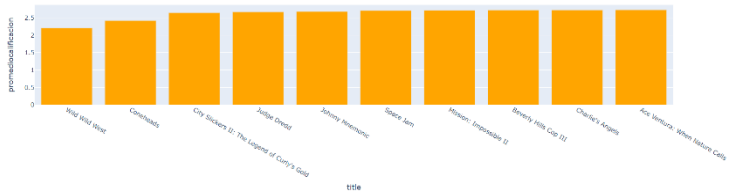


Imagen1/ Resumen exploratorio

En términos de popularidad, se encuentra que hay un nicho particular de películas con un número considerable de calificaciones. Considerando que las calificaciones son un elemento que habla en gran medida sobre la visualización e intensión de películas, en términos globales, Forrest Gump es una de las películas más calificadas, mientras que Toy Story es aquella que cuenta con mayor diversificación en términos de género, es una película versátil y que puede recomendarse a un amplio segmento. En términos de las mejor calificadas, hay una lista bastante amplia respecto a una calificación mayor a 4, las mejores 10 se encuentran en la figura de promedio de calificación, se encuentra que Sueño de Fuga, El Padrino y El Club de la Pelea son aquellas con la mejor puntuación y se convierten en un buen grupo de recomendación. Igual de importante considerar aquellas que poca calificación generan ya que una mala puntuación, en términos globales, puede tratarse como una mala experiencia del usuario dentro de la plataforma.





	title	cantidad_calif
0	Forrest Gump	329
1	Shawshank Redemption, The	317
2	Pulp Fiction	307
3	Silence of the Lambs, The	279
4	Matrix, The	278
5	Star Wars: Episode IV - A New Hope	251
6	Jurassic Park	238
7	Braveheart	237
8	Terminator 2: Judgment Day	224

	title	total_genres
0	Toy Story	5
1	Jumanji	3
2	Grumpier Old Men	2
3	Waiting to Exhale	3
4	Father of the Bride Part II	1

Recomendación basada en contenido: Una sola película

1) Método manual

movie: Toy Story 3

	correlación	title
1712	1.000000	Toy Story 3
1780	0.872776	Shrek Forever After
2931	0.872195	Ant Bully, The
3226	0.820863	Robots
286	0.798880	Moana
555	0.798764	The Good Dinosaur
1146	0.798508	Turbo
1265	0.798369	Rise of the Guardians
1309	0.798369	Madagascar 3: Europe's Most Wanted
1584	0.798221	Cars 2
1521	0.798221	Puss in Boots

movie: Resident Evil: Retribution

	correlación	title
1270	1.000000	Resident Evil: Retribution
1314	0.830378	Prometheus
1842	0.767027	Resident Evil: Afterlife
2626	0.765826	I Am Legend
331	0.726304	The Purge: Election Year
351	0.726304	Kill Command
508	0.726028	Wyrmwood
674	0.726028	The Gracefield Incident
747	0.725742	Edge of Tomorrow
869	0.725742	The Amazing Spider-Man 2
1045	0.725447	Gravity

2) KNN

movie_name: Predator 2

```
['Predator 2',
 'Captain America: Civil War',
 'Insurgent',
 'Garm Wars: The Last Druid',
 'Universal Soldier: Day of Reckoning',
 'Chronicle',
 'Total Recall',
 'Lockout',
 'Predators',
 'Repo Men',
 'Death Race 2',
 'X-Men Origins: Wolverine',
 'Gamer',
 'Surrogates',
 'Next',
 'X-Men: The Last Stand',
 'Déjà Vu',
 'Island, The',
 'Chronicles of Riddick, The',
 'Paycheck']
```

ie_name: 3 Ninjas Knuckle Up

```
Ninjas Knuckle Up',
ghty Morphin Power Rangers: The Movie',
manji: Welcome to the Jungle',
ngled Ever After',
y Next Door, The',
rate Kid, The',
y Kids 3-D: Game Over',
Ninjas: High Noon On Mega Mountain',
rbo: A Power Rangers Movie',
Ninjas Kick Back',
xt Karate Kid, The',
. Nanny',
Ninjas',
enage Mutant Ninja Turtles II: The Secret of the Ooze',
credibles 2',
lo: A Star Wars Story',
ptain Underpants: The First Epic Movie',
leficent',
ave',
Force']
```

Recomendación basada en contenido: Todas las películas

user_id: 408

	title
3521	SpaceCamp (1986)
6319	Man of the Year (2006)
6825	Mutant Chronicles (2008)
6189	9/11 (2002)
8194	Pacific Rim (2013)
4702	Presumed Innocent (1990)
1082	La Cérémonie (1995)
3888	Crocodile Hunter: Collision Course, The (2002)
17	Four Rooms (1995)
115	Up Close and Personal (1996)
4563	Brief History of Time, A (1991)

Filtros colaborativos: Basado en usuarios.

Elección del modelo: teniendo en cuenta el comportamiento de los modelos a partir de la validación de sus errores, se encuentra que el modelo KNNBaseline, el cual calcula el rating ponderado por la distancia con usuarios/películas, ofrece un error absoluto medio (MAE) y un error cuadrático medio(RMSE) menor a los demás, sin dejar de mencionar que es el segundo mejor eficiente en términos computacionales.

	MAE	RMSE	fit_time	test_time
knns.KNNBaseline	0.668472	0.874577	0.291683	3.257002
knns.KNNWithMeans	0.684570	0.895827	0.265191	2.483213
knns.KNNWithZScore	0.680765	0.897480	0.314616	2.190660
knns.KNNBasic	0.725864	0.947464	0.189939	2.459405

```
0.8892021265903682
```

EL código utilizado con la función GRIDSearchCV nos ayudó a encontrar cuáles son los mejores hiperparámetros de nuestro modelo evaluado en base al KNNBaseline utilizando la métrica MAE y RMSE mencionadas anteriormente. La función realiza una búsqueda sobre unos valores estándar que puedan ser útiles para la evaluación de los hiperparámetros que serán utilizados dentro del “param_grid”, para así encontrar el conjunto óptimo de hiperparémetros que minimizan la métrica de evaluación, en este caso el MAE y el RMSE, ahora con ayuda del atributo “best_params” nos muestra la mejor combinación para minimizar las métricas de evaluación.

En nuestro caso la selección que se realizó basados en los resultados obtenidos fue la siguiente: para el parámetro “name” seleccionamos “msd”, para el parámetro “min_support” que equivale a la cantidad de vecinos más cercanos con los que se realizará la predicción se seleccionó 5, y por último para “user_based” se seleccionó “TRUE” tomando la decisión de entrenar el modelo basado en usuarios.

Predicción: Se realiza una predicción (recomendación) de 20 películas para cada usuario, en este caso para el usuario 3; Se obtiene una lista con las películas para las que, según el modelo, hay aceptación de tal usuario:

	index	iid	est	title
0	24295	3404	4.830946	Titanic (1953)
1	26906	5490	4.790349	The Big Bus (1976)
2	27120	132333	4.790349	Seve (2014)
3	28492	25947	4.717819	Unfaithfully Yours (1948)
4	23185	85	4.581362	Angels and Insects (1995)
5	28148	3379	4.535998	On the Beach (1959)
6	21099	96004	4.535998	Dragon Ball Z: The History of Trunks (Doragon ...
7	22600	67618	4.485777	Strictly Sexual (2008)
8	27820	107780	4.363643	Cats (1998)
9	24388	25906	4.359663	Mr. Skeffington (1944)
10	24419	77846	4.359663	12 Angry Men (1997)
11	24427	93008	4.359663	Very Potter Sequel, A (2010)
12	27431	6201	4.359202	Lady Jane (1986)
13	23621	5034	4.359202	Truly, Madly, Deeply (1991)
14	27375	4495	4.359202	Crossing Delancey (1988)
15	26910	5915	4.357007	Victory (a.k.a. Escape to Victory) (1981)
16	25659	50610	4.355591	Beer League (2006)
17	25655	7926	4.355591	High and Low (Tengoku to jigoku) (1963)
18	25862	4180	4.334896	Reform School Girls (1986)
19	21941	26326	4.319728	Holy Mountain, The (Montaña sagrada, La) (1973)

Despliegue: Teniendo en cuenta las cualidades del planteamiento del problema, se pretende entregar experiencia y diferenciación en la plataforma, se desea generar impacto visual por las mismas cualidades del servicio, por tal motivo se considera al área de Desarrollo y Marketing como clientes elementales de la solución.

Se obtendrán calificaciones diarias de los usuarios, en cuánto a popularidad, se llevará la información a un archivo csv que actualizará un dashboard de los rankings relacionados con los géneros, los títulos, las películas más vistas y mejores puntuadas. Se realizarán envíos diariamente a Desarrollo Web, ellos generarán actualizaciones del diseño visual de la plataforma con el área de Marketing para gestionar lo relacionado con la imagen y publicidad de la plataforma.

Al generar la calificación, también se ejecutará un código de limpieza y trasformación, llevará los datos a un esquema standart y entrenará un modelo Knn antes cargado en formato .pkl, finalmente se hará la predicción de las películas a cada cineasta, el modelo realizará la búsqueda de las películas dentro de otra data en la que reposa cada película con su respectiva imagen de portada. Buscará 4 recomendaciones y mostrará en la pantalla cada una de ellas con el mensaje de “Te podría interesar” (este aspecto concierne en gran medida a Desarrollo Web).