UNIVERSIDAD TECNOLÓGICA METROPOLITANA
UTEM del Estado de Chile

Avance proyecto de minería de datos: “Análisis de recomendaciones de películas basado en perfiles de usuarios”

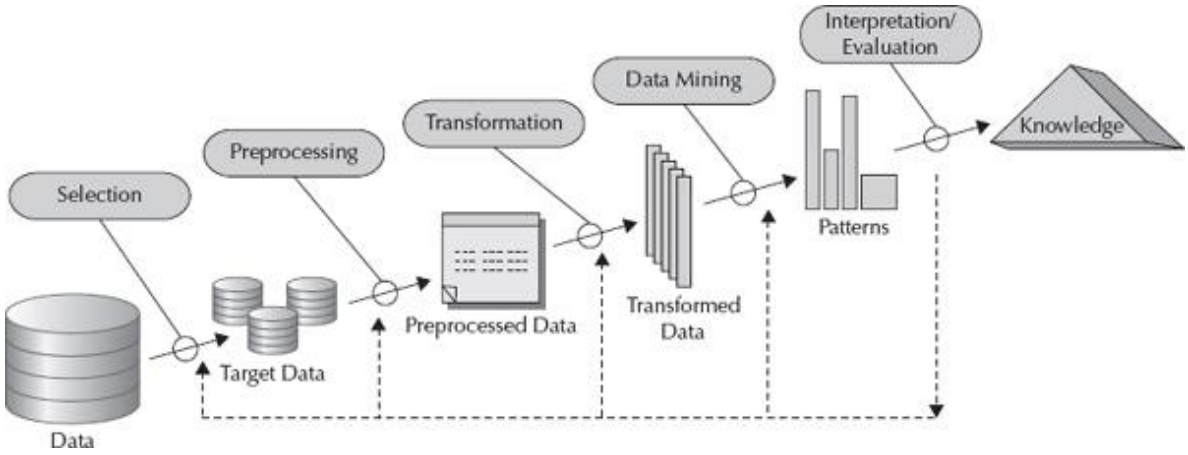
Profesor: Pablo Figueroa Plaza
Asignatura: Minería de datos
Estudiante: Francisco Sánchez Fuentes
Fecha: 28/11/2017

Indicé

Marco Teórico

Para obtener información de un conjunto de datos es necesario seguir una metodología que esté vigente y cumpla con el objetivo de brindar al equipo de trabajo una serie de pasos que concluyan en conocimiento, este conocimiento particularmente contesta una serie de preguntas que son realizadas en el análisis de un problema y luego se concretan en la definición del problema.

Para el desarrollo de este proyecto se utilizará la metodología KDD. La metodología KDD se puede definir como "el proceso no trivial de identificar patrones válidos, novedosos y potencialmente útiles y en última instancia comprensible a partir de los datos". KDD también supone la convergencia de distintas disciplinas de investigación, podemos nombrar algunas tales como el aprendizaje automático, estadística, inteligencia artificial, técnicas de visualización de datos, sistemas para el apoyo a la toma de decisión (DSS).



Etapas del KDD

1) Selección de datos: Consiste en buscar el objetivo y las herramientas del proceso de minería, identificando los datos que han ser extraídos, buscando los atributos apropiados de entrada y la información de salida para representar la tarea. Esto quiere decir, primero se debe tener en cuenta lo que se saber lo que se quiere obtener y cuáles son los datos que nos facilitarán esa información para poder llegar a nuestra meta, antes de comenzar el proceso en tal.

2) Limpieza de datos: En este paso se limpian los datos sucios, incluyendo los datos incompletos (donde hay atributos o valores de atributos perdidos), el ruido (valores incorrectos o inesperados) y datos inconsistentes (conteniendo valores y atributos con nombres diferentes). Los datos sucios en algunos casos deben ser eliminados ya que pueden contribuir a un análisis inexacto y resultados incorrectos.

3) Integración de datos: Combina datos de múltiples procedencias incluyendo múltiples bases de datos, que podrían tener diferentes contenidos y formatos.

4) Transformación de datos: Consisten principalmente en modificaciones sintácticas llevadas a cabo sobre datos sin que supongan un cambio para la técnica de minería aplicada. Las transformaciones discretas de los datos[HLT99] tienen la ventaja de que mejoran la comprensión de las reglas descubiertas al transformar los datos de bajo nivel en datos de alto nivel y también reduce Significativamente el tiempo de ejecución del algoritmo de búsqueda. Su principal

Desventaja es que se puede reducir la exactitud del conocimiento descubierto, debido a que puede causar la perdida de alguna información. Existen diferentes métodos de transformación de variables continuas a discretas que se pueden agrupar según distintas aproximaciones: métodos locales (realizan la transformación discreta en una región del espacio de las instancias, por ejemplo, utilizando un subconjunto de las instancias), métodos globales (utilizan el espacio de las instancias), métodos supervisados (utilizan la información de la clave (valor del atributo objetivo).

5) Reducción de datos: Reducir el tamaño de los datos, encontrando las características

Más significativas dependiendo del objetivo del proceso. Se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas, o para encontrar otras representaciones de los datos.

reducción de dimensiones (la extracción irrelevante y débil de atributo), compresión de datos (reemplazando valores de datos con datos alternativos codificados),

reducción de tamaño (reemplazando valores de datos con representación alternativa más pequeña), una generalización de datos (reemplazando valores de datos de niveles conceptuales bajos con niveles conceptuales más altos), etc.

6) Minería de datos: Consiste en la búsqueda de los patrones de interés que pueden

Expresarse como un modelo o simplemente que expresen dependencia de los datos. Se tiene que especificar un criterio de preferencia para seleccionar un modelo de un conjunto de posibles modelos. También se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está determinado en el algoritmo de minería).

7) Evaluación de los patrones: Se identifican verdaderamente patrones interesantes que representan conocimiento usando diferentes técnicas incluyendo análisis estadísticos y lenguajes de consultas.

8) Interpretación de los patrones: Consiste en entender los resultados del análisis y sus implicaciones y puede llevar a regresar a algunos de los pasos anteriores.

Definición del problema

El problema que se presenta con este conjunto de datos es determinar cuándo a un usuario le gustara la película que se recomienda. Si bien cada persona tiene sus propios gustos estos se pueden relacionar con los de otros usuarios en base a una similitud de gustos los cuales puedan compartir. Sin embargo, como la calificación en base a gustos es subjetiva, se debe estar consiente respecto a los resultados que pueda entregar el programa.

En profundidad el problema radica en la identificación de perfiles de usuario basado en gustos por ciertas películas, la calificación de una película se encuentra medido por la variable *Rating* donde el usuario clasifica en un rango definido que tan buena encuentra que es la película, este procedimiento se reitera las veces necesarias por los distintos usuarios que quiera clasificar la misma película u otra.

Los perfiles de usuario y sus recomendaciones por película están sujetos a *factores subjetivos* ya que cada recomendación depende de una opinión personal del usuario. Otro factor que puede incidir en la calificación son los factores sociales como Revistas, Redes sociales y Blog entre otros.

Objetivo general

> *Identificar perfiles de usuario y utilizar un modelo que entregue como respuesta predicción de recomendaciones de pelicular para usuario con un perfil similar.*

Objetivos específicos

- Consolidar un grupo de datos que sirva para el estudio de la calificación basado en información de la película e integración con la información del usuario.
- Encontrar un modelo al cual se pueda adaptar para obtener recomendaciones de películas basado en los datos de prueba.
- Definir el modelo apropiado junto con la solución del problema.

Solución Propuesta

Para encontrar una solución al problema de recomendaciones primero de debe entender la lógica del negocio.

El negocio tiene por objetivo brindar a sus usuarios un servicio para ver películas en línea, basado en el gusto de la persona y un criterio personal o quizás compartido calificara la película. Esta calificación queda registrada en el sistema con su respectiva fecha realizando la unión entre cliente y película, además de guardar la calificación. Por otro lado, tenemos la información afiliada por cada película donde sabemos que cada película tiene un identificador, titulo, genero, lenguaje de origen y otras variables. Es importante entender que un cliente al momento de calificar una película *generara una historia* (Un registro histórico) que marcara los gustos de dicha persona.

Hay varios métodos para poder realizar correlaciones entre las variables y encontrar que usuarios tienen los mismos gustos. Además, se tiene la posibilidad de realizar agrupaciones entre los distintos géneros y ver que usuario comparten mismos grupos, por ende, tenemos la posibilidad de agrupar variables, también se suma la posibilidad de realizar un seguimiento a las películas y en segundo lugar a los usuarios privilegiando grupos de películas antes que gustos de usuario e inclusive relacionar por el rating antes que usuario o películas.

Descripción de los datos seleccionados

Recomendación de películas

- **ACTOR_ID:** Identificador numérico del actor en la base de datos.
- **ACTOR_GENERO:** número del genero a cuál pertenece el actor. Este número está en un rango de 0, 1 y 2 donde significan “No definido”, “Femenino”, “Masculino” respectivamente.
- **DIRECTOR_ID:** Identificador numérico del director en la base de datos.
- **DIRECTOR_POPULARIDAD:** indicador de popularidad del director de la película recopilado desde sitio web dedicado a criticas de cine.
- **PRESUPUESTO:** Monto invertido para la producción de la película.
- **INGRESOS:** registro de monto de las ganancias obtenidas por la película.
- **ID_PELICULA:** Identificador numérico de la película en la base de datos.
- **ID_USUARIO:** Identificador numérico del usuario en la base de datos.
- **DIRECTOR_GENERO:** Número del genero al cual pertenece el director de la película, análogamente con el ACTOR_GENERO este se clasifica en 0, 1 y 2.
- **TIEMPO:** Corresponde al tiempo de duración de la película.
- **VOTO_PROMEDIO:** Voto promedio corresponde a un indicador evaluado entre el 1 al 7 para indicar el agrado de la película según el público.
- **POPULARIDAD_DETALLE:** Popularidad detalle corresponde a un indicador de evaluación similar el RATING, pero este campo posee un valor distinto en varias situaciones al RATING por lo cual se conservó para el estudio.
- **VOTO_CONTADOR:** Este es un campo numérico que cuenta cuantos votos se realizaron en el sitio web para calificar la película.
- **GENERO:** Corresponde a un campo numérico que clasifica a la película dentro de una categoría.
- **ANIO:** Corresponde al año de publicación de la película.
- **POPULARIDAD:** Este indicador establece una relación entre los usuarios y la película demostrando que tan popular es dentro del sitio web.
- **RATING_PROMEDIO:** Corresponde a la clasificación promedio que fue otorgada por los usuarios.
- **RANKING:** Es en realidad el RATING entregada por cada usuario, al momento de transcribir este variable se nombró mal en el archivo de datos y debe ser corregida por “RATING”, cabe destacar que esta variable corresponde a una calificación única por cada usuario y de esta misma se calculó el rating promedio.

Descripción de las variables Input

Cabe destacar que en este punto se realizó a lo menos una reducción de dimensión la cual esta detallada en la sección de **análisis de los datos**.

- **DIRECTOR_GENERO**
- **TIEMPO**
- **VOTO_PROMEDIO**
- **POPULARIDAD_DETALLE**
- **VOTO_CONTADOR**
- **GENERO**
- **ANIO**
- **POPULARIDAD**
- **RATING_PROMEDIO**
- **RANKING**

Descripción de las variables output

- **PREDICCION:** Esta variable de salida corresponde al Rating proporcionado por el modelo que dicta basado en el perfil del usuario si la película le gustara o no. La salida puede estar en el mismo rango que se clasifican las películas y su exactitud es dependiente del modelo.

  PREDICCION: [1;5]

- **ID_PELICULA:** Se está evaluando la posibilidad de agregar una segunda salida que esté relacionada con la predicción la cual puede entregar una colección de identificadores de películas recomendadas para el usuario basado en otros usuario con comportamiento similar.

  ARREGLO_ID_PELICULAS: [ID_PELICULA_1, ID_PELICULA_2, …, ID_PELICULA_N]

Análisis de los datos

Para el análisis de los datos se trabajó con dos conjuntos, el primero describe la película junto con su información asociada como el actor principal, el director de la película, el año de estreno y el tiempo de duración entre otros detalles. En cuanto al segundo conjunto, se trabajó con la clasificación de los usuarios mezclado con la información de las películas. El motivo de diferenciar estos dos conjuntos es porque al evaluar una película por separado obtenemos información relacionada únicamente de las películas y cuando cruzamos los datos con la clasificación del usuario obtenemos un comportamiento asociado a la reproducción de la película junto con la opinión (clasificación) del usuario. A continuación se realizara un análisis a ambos conjunto.

Conjunto N°1: Películas y su información

Para este conjunto de datos utilizamos el programa SPSS IBM el cual nos proporciona los siguientes resultados en el análisis de reducción de dimensiones.

Prueba de KMO y Barlett

En esta prueba determinamos si es factible continuar con el análisis de reducción de dimensiones. Como el KMO es mayor a 0.5 podemos continuar con el análisis. Respecto a la prueba de esfericidad de Barlett nos dice que no es significativa la hipótesis nula de variables iniciales no correlacionadas, por lo tanto, es menor que 0.05 lo cual es adecuado para continuar con el análisis factorial.

Prueba de KMO y Bartlett

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,773
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	33855,211
	gl	120
	Sig.	,000

Comunalidades

Esta prueba nos indica que tan explicadas se encuentran las variables en la solución factorial, como podemos apreciar los indicadores la mayoría se encuentra sobre el 0.5 y la variable generada a partir de la variable de estudio, o sea, RATING_PROMEDIO tiene un 0.658 de explicación lo cual no es el mejor indicador, pero se puede continuar con el estudio. Respecto al ID_PELICULA resulta curioso que una variable generada durante la ejecución del almacenamiento de una película (es una variable que incrementa su valor en cada inserción) se encuentre tan bien explicada.

Comunalidades

	Inicial	Extracción
ACTOR_ID	1,000	,566
ACTOR_GENERO	1,000	,479
DIRECTOR_GENERO	1,000	,516
DIRECTOR_ID	1,000	,580
DIRECTOR_POPULARIDAD	1,000	,514
TIEMPO	1,000	,413
VOTO_PROMEDIO	1,000	,611
PRESUPUESTO	1,000	,593
POPULARIDAD_DETALLE	1,000	,728
VOTO_CONTADOR	1,000	,786
INGRESOS	1,000	,681
GENERO	1,000	,268
ANIO	1,000	,587
ID_PELICULA	1,000	,954
POPULARIDAD	1,000	,770
RATING_PROMEDIO	1,000	,658

Método de extracción: análisis de componentes principales.

Varianza total explicada

En los datos analizados obtener un 60% de varianza explicada. Si se desea obtener una mejor explicación es necesario reducir algunos factores. Tener un 60% de varianza explicada no es un indicador favorable, pero permite continuar con el estudio.

Varianza total explicada

	Autovalores iniciales			Sumas de extracción de cargas al cuadrado			Sumas de rotación de cargas al cuadrado		
Componente	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	4,201	26,256	26,256	4,201	26,256	26,256	3,701	23,133	23,133
2	1,847	11,546	37,802	1,847	11,546	37,802	1,949	12,183	35,316
3	1,510	9,440	47,242	1,510	9,440	47,242	1,668	10,425	45,741
4	1,142	7,139	54,380	1,142	7,139	54,380	1,377	8,607	54,348
5	1,002	6,262	60,643	1,002	6,262	60,643	1,007	6,295	60,643
6	,952	5,952	66,595						
7	,935	5,844	72,439						
8	,778	4,861	77,301						
9	,708	4,428	81,728						
10	,668	4,172	85,901						
11	,569	3,555	89,456						
12	,558	3,485	92,941						
13	,454	2,838	95,779						
14	,417	2,605	98,384						
15	,154	,963	99,347						
16	,104	,653	100,000						

Método de extracción: análisis de componentes principales.

Matriz de componente rotado

Esta prueba permite ver la relación entre las variables según una división de dimensiones. Respecto al RATING_PROMEDIO se pueden concluir dos cosas, primero que la relación en esa dimensión es muy débil ya que solo tiene un 0.276 de explicación y en segundo lugar se tiene que las variables directamente relacionadas con este indicador son:

- DIRECTOR_POPULARIDAD (0.444)
- PRESUPUESTO (0.731)
- POPULARIDAD_DETALLE (0.795)
- VOTO_CONTADOR (0.876)
- INGRESOS (0.825)
- ANIO (0.252)
- POPULARIDAD (0.820)

Si bien, ANIO tiene una relación muy baja en la dimensión evaluada de todas formar será incluida para el entrenamiento.

Matriz de componente rotado[a]

	Componente				
	1	2	3	4	5
ACTOR_ID			-,309	,637	
ACTOR_GENERO			,275	-,601	
DIRECTOR_GENERO			,710		
DIRECTOR_ID			-,732		
DIRECTOR_POPULARIDAD	,444		,541		
TIEMPO		,381		,447	
VOTO_PROMEDIO		,749			
PRESUPUESTO	,731				
POPULARIDAD_DETALLE	,795	,251			
VOTO_CONTADOR	,876				
INGRESOS	,825				
GENERO				,473	
ANIO	,252	-,651		,271	
ID_PELICULA					,975
POPULARIDAD	,820				
RATING_PROMEDIO	,276	,744			

Método de extracción: análisis de componentes principales.
Método de rotación: Varimax con normalización Kaiser.

a. La rotación ha convergido en 8 iteraciones.

Conjunto N°2: Películas con su información y calificación del usuario

Para este estudio por capacidad de computo se realizaron tres pruebas a tres subconjuntos de datos pertenecientes al conjunto 2. Para facilitar el entendimiento se compararán las tres pruebas en los análisis correspondientes.

Comunalidades

Respecto a las comunalidades de las tres pruebas podemos notar que hay un alto índice de explicación en el análisis factorial, pero la variable de interés que es el RANKING tiene una baja explicación en dos de las tres pruebas. Este resultado no favorece mucho al estudio ya que dentro de cada componente la variable implicada en el estudio se ve débilmente explicada, sin embargo la variable derivada RATING_PROMEDIO si tiene una buena explicación dentro de las componentes.

	Prueba 1	Prueba 2	Prueba 3
ACTOR_ID	0,576	0,567	0,751
ACTOR_GENERO	0,688	0,654	0,788
DIRECTOR_GENERO	0,507	0,654	0,822
DIRECTOR_ID	0,562	0,428	0,573
DIRECTOR_POPULARIDAD	0,795	0,876	0,786
TIEMPO	0,656	0,702	0,521
VOTO_PROMEDIO	0,833	0,897	0,765
PRESUPUESTO	0,871	0,903	0,744
POPULARIDAD_DETALLE	0,812	0,751	0,916
VOTO_CONTADOR	0,887	0,735	0,750
ID_DETALLE	0,874	0,902	0,650
INGRESOS	0,814	0,76	0,882
GENERO	0,248	0,594	0,950
ANIO	0,718	0,759	0,882
ID_PELICULA	0,742	0,778	0,784
ID_EXTERNO	0,874	0,902	0,877
POPULARIDAD	0,901	0,811	0,633
RATING_PROMEDIO	0,83	0,748	0,807
ID_USUARIO	0,997	1	0,965
RANKING	0,292	0,101	0,443

Varianza explicada

La varianza explicada es buena dentro de las tres pruebas que se realizaron lo cual permite continuar con el estudio, sin embargo, para aumentar el porcentaje de explicación se deben eliminar factores y como se vio anteriormente en las comunalidades **la variable RANKING es un factor candidato a ser removido para mejorar la varianza total.**

	Prueba 1	Prueba 2	Prueba 3
Varianza Total explicada	72%	72%	76%

Matriz de componente rotado

Respecto a la matriz de componente se muestra la siguiente tabla resumen.

Tabla de resumen para matriz de componente

	Prueba 1	Prueba 2	Prueba 3
DIRECTOR_GENERO		0,267	
VOTO_PROMEDIO	0,828	0,889	0,335
TIEMPO		0,762	
POPULARIDAD_DETALLE	0,54	0,485	
VOTO_CONTADOR	0,694	0,643	
ID_DETALLE	-0,292		
GENERO	-0,318	0,368	
ANIO		-0,572	
ID_EXTERNO	-0,292		
POPULARIDAD	0,564	0,685	
RATING_PROMEDIO	0,894	0,803	0,652
RANKING	0,515	0,303	0,659
DIRECTOR_ID			0,293
ID_PELICULA			0,274

En esta tabla podemos apreciar que hay dos variables que son transversales en las tres pruebas, la primera corresponde al VOTO_PROMEDIO que es el número de votaciones que recibe una película y en segunda instancia tenemos el RATING_PROMEDIO, como podemos verificar en la comunalidad el RATING_PROMEDIO tiene un alto índice de explicación dentro de los datos lo cual lo hace un candidato importante para el modelo final.

Prueba 1

Matriz de componente rotado[a]

	Componente					
	1	2	3	4	5	6
ACTOR_ID			,710			
ACTOR_GENERO	-,592		-,515			
DIRECTOR_GENERO				,651		
DIRECTOR_ID					-,695	
DIRECTOR_POPULARIDAD				,264	,763	
TIEMPO	,370				,681	
VOTO_PROMEDIO	-,325	,828				
PRESUPUESTO	,861			,285		
POPULARIDAD_DETALLE		,540	-,381	,578		
VOTO_CONTADOR	-,395	,694		,446		
ID_DETALLE		-,292	,878			
INGRESOS	,849					
GENERO	-,339	-,318				
ANIO	,468			,651		
ID_PELICULA	-,823					
ID_EXTERNO		-,292	,878			
POPULARIDAD		,564	-,339	,638		
RATING_PROMEDIO		,894				
ID_USUARIO						,998
RANKING		,515				

Método de extracción: análisis de componentes principales.
Método de rotación: Varimax con normalización Kaiser.

a. La rotación ha convergido en 9 iteraciones.

En esta prueba la variable RANKING tiene un 0.515 de explicación en la componente 2, acompañando a esta variable se encuentran las siguientes:

- VOTO_PROMEDIO (0,828)
- POPULARIDAD_DETALLE (0,540)
- VOTO_CONTADOR (0,694)
- ID_DETALLE (-0,292)
- GENERO (-0,318)
- ID_EXTERNO (-0,292)
- POPULARIDAD (0,564)
- RATING_PROMEDIO (0,894)
- RANKING (0,515)

Matriz de componente rotado[a]

	Componente					
	1	2	3	4	5	6
ACTOR_ID		,722				
ACTOR_GENERO		-,280		-,723		
DIRECTOR_GENERO	,267				,744	
DIRECTOR_ID		,435		-,301	,306	
DIRECTOR_POPULARIDAD			,286	,882		
TIEMPO	,762			,263		
VOTO_PROMEDIO	,889					
PRESUPUESTO			,908			
POPULARIDAD_DETALLE	,485	-,332	,414		,441	
VOTO_CONTADOR	,643		,406			
ID_DETALLE		,922				
INGRESOS			,827			
GENERO	,368		,360	-,569		
ANIO	-,572		,494	,422		
ID_PELICULA					,821	
ID_EXTERNO		,922				
POPULARIDAD	,685	-,340	,389		,272	
RATING_PROMEDIO	,803					
ID_USUARIO						1,000
RANKING	,309					

Método de extracción: análisis de componentes principales.
Método de rotación: Varimax con normalización Kaiser.

a. La rotación ha convergido en 14 iteraciones.

En esta prueba la variable RANKING tiene un 0.309 de explicación en la componente 1, acompañando a esta variable se encuentran las siguientes:

acompañando a esta variable se encuentran las siguientes:

- DIRECTOR_GENERO (0,267)
- TIEMPO (0,762)
- VOTO_PROMEDIO (0,889)
- POPULARIDAD_DETALLE (0,485)
- VOTO_CONTADOR (0,643)
- GENERO (0,368)
- ANIO (-0,572)
- POPULARIDAD (0,685)
- RATING_PROMEDIO (0,803)
- RANKING (0,309)

Prueba 3

Matriz de componente rotado[a]

	Componente						
	1	2	3	4	5	6	7
ACTOR_ID	,821						
ACTOR_GENERO			,818		-,262		
DIRECTOR_GENERO			,792		,403		
DIRECTOR_ID	,558	,308				,293	
DIRECTOR_POPULARIDAD					,842		
TIEMPO		-,542		,424			
VOTO_PROMEDIO	-,427	-,490		,410		,335	
PRESUPUESTO		,658	,313		,354		
POPULARIDAD_DETALLE				,897			
VOTO_CONTADOR		,680	,400				
ID_PELICULA			,709			,274	
ID_EXTERNO	,862	-,271					
POPULARIDAD				,951			
ID_DETALLE	,862	-,271					
INGRESOS		,598	,497	-,374			
GENERO					,902		
ANIO		,629	-,408				
RATING_PROMEDIO		-,475			,300	,652	
ID_USUARIO							,981
RANKING						,659	

Método de extracción: análisis de componentes principales.
Método de rotación: Varimax con normalización Kaiser.

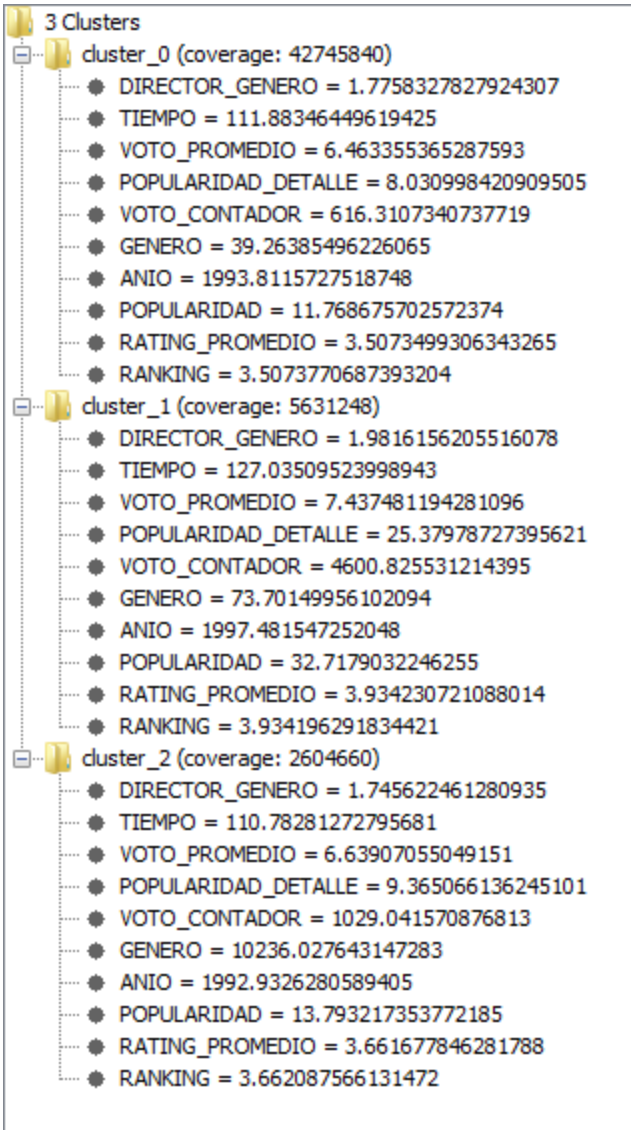
a. La rotación ha convergido en 14 iteraciones.

En esta prueba la variable RANKING tiene un 0.659 de explicación en la componente 6, acompañando a esta variable se encuentran las siguientes:

- DIRECTOR_ID (0,293)
- VOTO_PROMEDIO (0,335)
- ID_PELICULA (0,274)
- RATING_PROMEDIO (0,652)
- RANKING (0,659)

k-means

Se realizó una prueba de k-means aplicando 3 cluster con 10 iteraciones, como resultado se obtuvieron los siguientes centroides:



Modelo aplicado a datos

En esta etapa se utilizarán dos modelos, árbol de decisiones y random forest. Antes de aplicar los modelos se trabajará con los datos para conocer los detalles de estos.

Conociendo los datos

Primero cargamos los datos desde un archivo csv a un DataFrame llamado **myData**, posteriormente se con el método *shape* obtenemos las filas y columnas de los datos trabajados.

```
myData.shape
```

Como resultado obtenemos un total de 20.000.000 de filas de registros junto con 12 columnas de datos que fueron los seleccionados por el análisis factorial.

```
(20000000, 12)
```

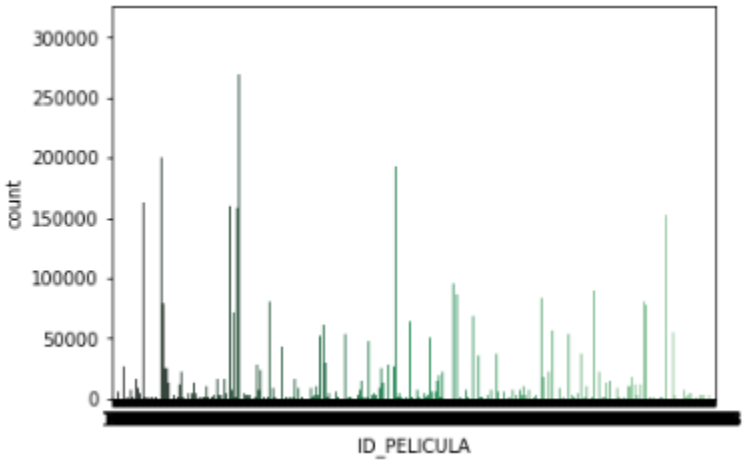
Ahora, agrupamos los datos por el identificador de las películas (ID_PELICULA) para obtener las frecuencias de sus calificaciones y además el número de películas que estamos operando.

```
myData.ID_PELICULA.value_counts()
```

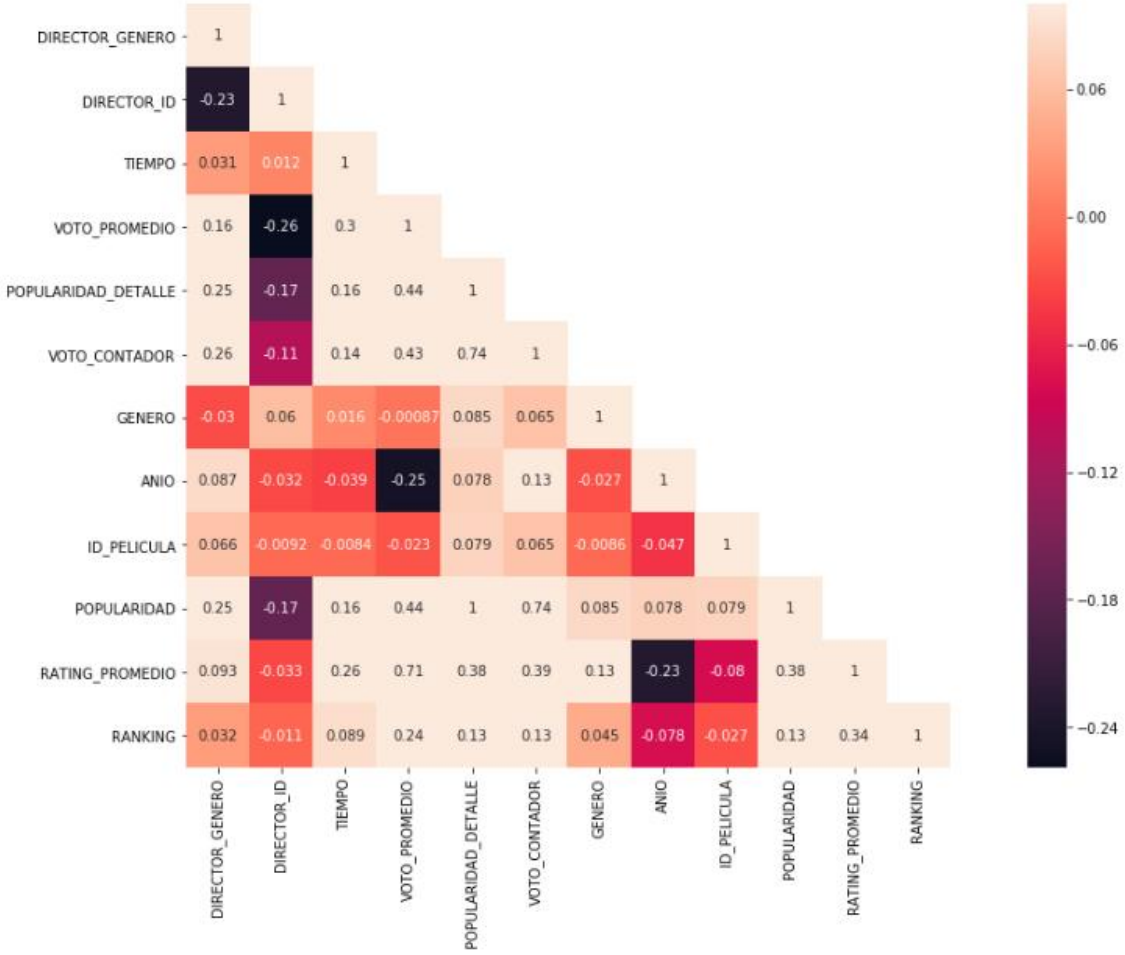
Registrando un total de 1.577 películas

```
Name: ID_PELICULA, Length: 1577, dtype: int64
```

Posteriormente generamos una gráfica donde se muestran las frecuencias de las clasificaciones por película



Como se puede apreciar en el gráfico, hay una minoría de películas que presenta una alta frecuencia de calificaciones. Ahora veremos la correlación entre las variables.



Observando nuestro variable target que es RANKING vemos que no tiene una relación fuerte respecto a las demás variables, la más cercana es RATING_PROMEDIO que como ya mencionamos anteriormente resulta ser una variable generada desde la variable target. En este grafico podemos concluir que hay dos variables “representativas” que serían VOTO_PROMEDIO y RATING_PROMEDIO, agregando dos excepciones se puede considerar POPULATIDAD_DETALLE y VOTO_CONTADOR.

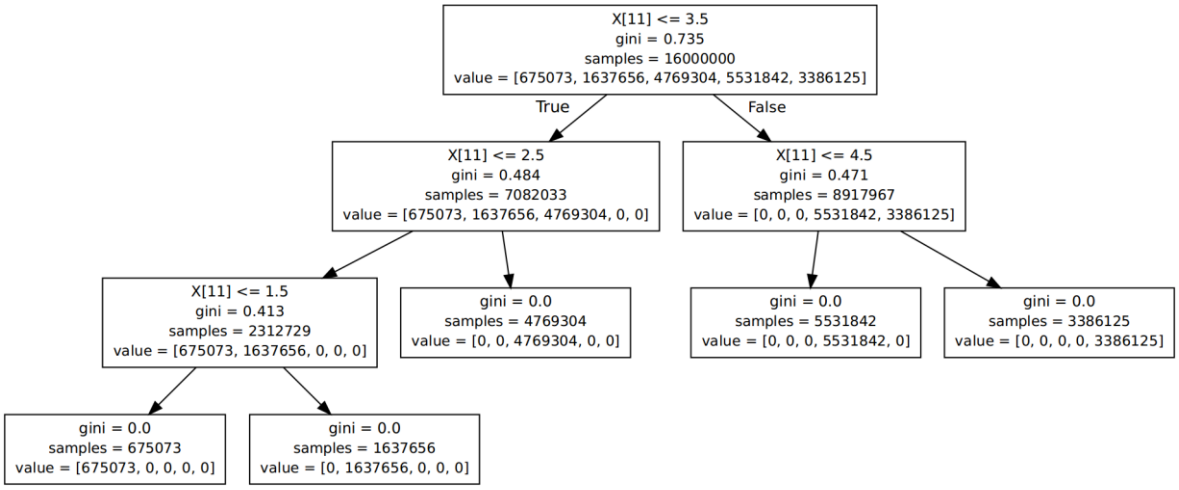
Árbol de decisiones

Para realizar el árbol de decisiones primero se necesita limpiar los datos utilizando la función "dropna" la cual omite los campos vacíos, posteriormente generamos dos variables, primer "predictors" quien almacena el conjunto de datos a predecir y segundo la variable "targets", con dos conjuntos de datos seleccionados preparamos la división de datos entre entrenamiento y prueba. Luego de haber separado los conjuntos de datos se instancia el árbol como objeto y se entregan las variables para ajustar el árbol. Finalmente se entrena y predice obteniendo la matriz de confusión la cual evalúa el acierto de la clasificación además del puntaje de acierto.

```
In [15]: # Obtenemos la matriz de confusion
         sklearn.metrics.confusion_matrix(tar_test, predictions)

Out[15]: array([[ 168607,       0,       0,       0,       0],
                [      0,  409857,       0,       0,       0],
                [      0,       0, 1192535,       0,       0],
                [      0,       0,       0, 1381728,       0],
                [      0,       0,       0,       0,  847273]])

In [16]: # Medida de acierto
         sklearn.metrics.accuracy_score(tar_test, predictions)

Out[16]: 1.0
```

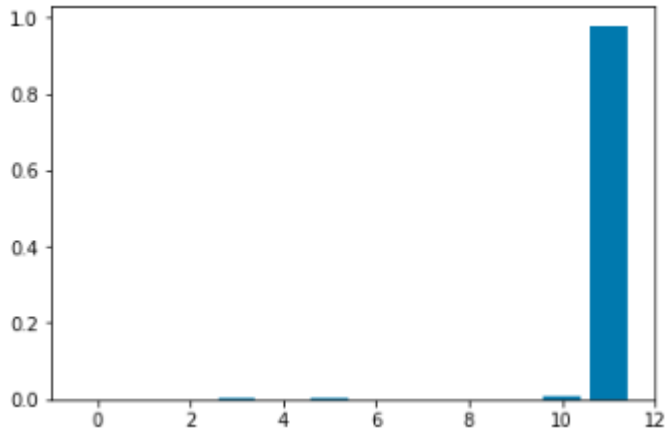
Respecto a la medida de acierto indica que tiene un 100% de acierto lo cual puede indicar una buena predicción o sobre entrenamiento.

Árbol de decisión generado:



Random forest

Utilizando este modelo se sigue la mayoría de los procesos de limpieza y procesamiento de datos para el Árbol de decisiones. En lo que respecta al modelo también resulta un puntaje del 100% de acierto lo cual también puede significar un buen modelo entrenado o sobre entrenamiento. Este modelo entrega dentro de dos resultados una gráfica con la variable y su importancia dentro del modelo.



Esto describe que la variable RATING_PROMEDIO presenta una importancia destacable dentro del modelo.

Conclusión

Posterior al análisis factorial se realizó el filtro en las variables a utilizar, además de volver a verificar la correlación de las variables se procede a probar dos modelos, el primero es el árbol de decisiones donde genero un gráfico para la toma de decisiones y además presento una buena puntuación de acierto para ambos modelos. Respecto al Random Forest se obtuvo un gráfico de relevancia de variables la cual es RATING_PROMEDIO, cabe recalcar que esta variable es un derivado de la variable objetivo.