Netflix Movies and TV Shows

1st Juan Pablo Ramirez Bernal
Facultad de Ciencias e Ingenieria
Universidad Jorge Tadeo Lozano
Bogota, Colombia
juanp.ramirezb@utadeo.edu.co

2nd Alex Lopez Gonzalez
Facultad de Ciencias e Ingenieria
Universidad Jorge Tadeo Lozano
Bogota, Colombia
alexi.lopezg@utadeo.edu.co

3rd Juan Camilo Jara
Facultad de Ciencias e Ingenieria
Universidad Jorge Tadeo Lozano
Bogota, Colombia
juanc.jarab@utadeo.edu.co

4nd John Sebastian Garcia Moreno
Facultad de Ciencias e Ingenieria
Universidad Jorge Tadeo Lozano
Bogota, Colombia
johns.garciam@utadeo.edu.co

Resumen

Netflix es un servicio de streaming por su suscripción que les permite a sus usuarios ver series y películas y se puede ver desde un dispositivo con conexión a internet, cualquier dispositivo que cuente con la aplicación de Netflix tendrá acceso a su contenido, surgió en 1997 como un video club online, que se limitaba a enviar DVD por correo a los clientes, esta plataforma tuvo un crecimiento bastante importante debido a la pandemia por lo que tuvo un crecimiento de 16 millones de suscriptores, siendo el mayor aumento en lo que va de su historia

Objetivo General

Se quiere obtener una predicción correcta con respecto a los géneros presentados en las bases de datos por medio del uso de machine learning.

Objetivo Especifico

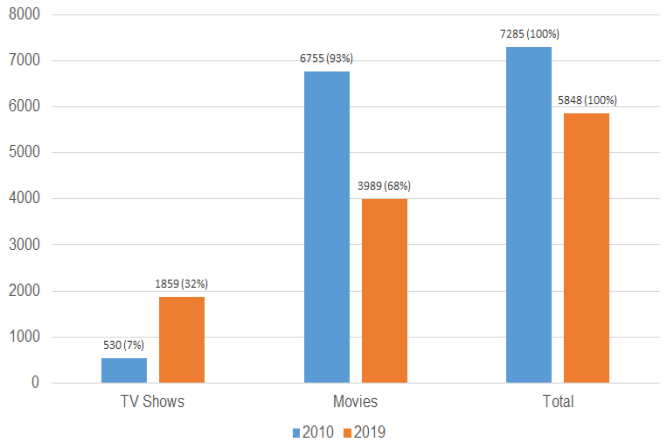
Por medio de nuestra base de datos centrada principalmente en contenido agregado en la plataforma de Netflix con el fin de generar una predicción correcta.

Modificar nuestra base de datos con el fin de que correspondan sus características en torno a las necesidades de aplicación en nuestro proyecto

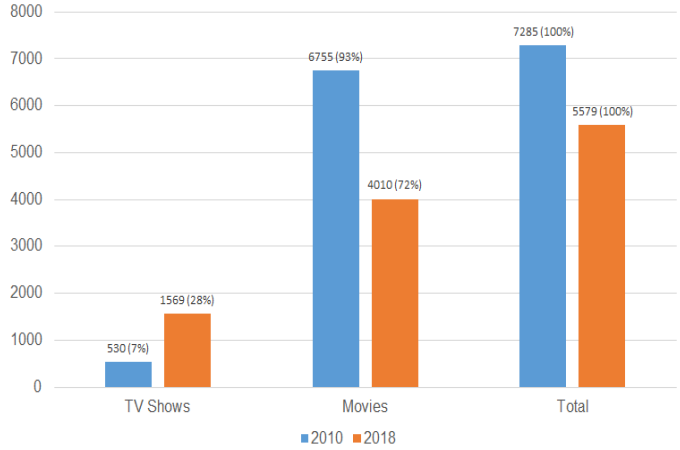
Palabras cláve

Netflix, análisis de datos, problemática

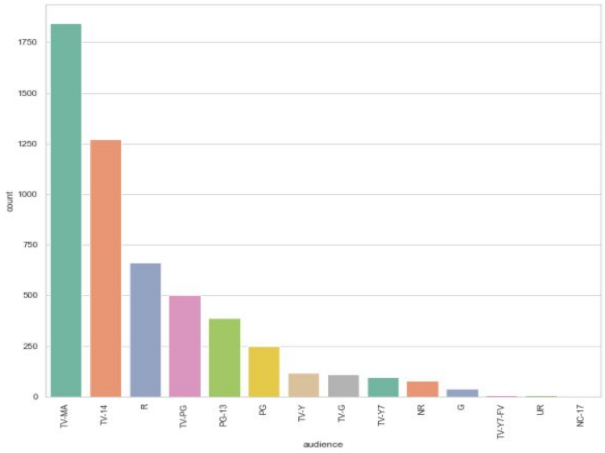
1. **Recopilación de datos**
   Los datos que usaremos en este proyecto para su análisis fueron suministrados por Kaggle, una plataforma importante para el estudio y aprendizaje de diversos temas, todos estos temas se centran en la ciencias de datos, el análisis predictivo y el machine learning, estos datos originalmente son de un motor de búsqueda de Netflix de terceros llamado Fixable, el autor Shivam Bansal se interesó debido a que la cantidad de programas de televisión de Netflix se ha triplicado desde 2010, su inspiración proviene del poder comprender que contenido se muestra en los diferentes países del mundo, también se interesó por identificar el contenido similar haciendo coincidencias basadas en texto.
   En estos datos encontraremos una serie de lista de películas en donde encontramos un código único para cada película, también vemos el identificador en donde sabemos si es un programa o una película, el director, el país donde fue creada la película, entre otros.
2. **Tema de análisis**
   En el presente proyecto se busca analizar y responder preguntas por medio de la búsqueda de datos, aquí tendremos que verificar diferentes componentes y generar algoritmos los cuales nos den una predicción.



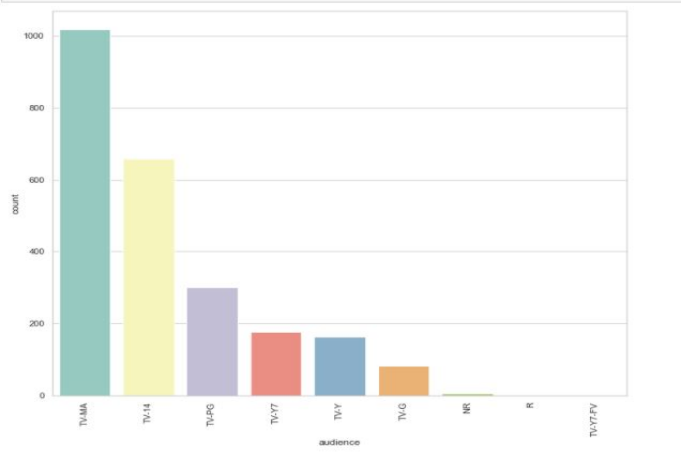
(a) Número de títulos en netflix en EE. UU. 2019



(b) Número de títulos en Netflix en EE. UU. 2018



(c) Cantidad de películas según la audiencia



(d) Cantidad de series según audiencia

Así era el catálogo estadounidense de Netflix a principios de 2010.
Análisis de los gráficos A y B

Se encontraron un total de 5386 películas y 2416 TV Shows esto gracias a que Netflix mantiene en constante cambio y evolución su plataforma entre catálogo de películas y series, también tener en consideración que el mayor número de contenido son parte de TV-MA la cual es una categoría impulsada o dirigida para un público de edades avanzadas, a pesar de que Netflix destaca bastante en el público joven y hasta llegar al punto de tener la opción de crear un perfil exclusivamente con contenido dirigido para los más pequeños de la casa por lo que sorprende que solo 247 títulos son de clasificación PG, también el contenido TV-14 que es la segunda calificación con más contenido en esta base de datos con un total de 1931 títulos entre TV Shows y películas.

Análisis de los datos entre audiencia para películas y audiencias en series C y D

- Tenemos más variedad de audiencias en películas que en series.
- Se puede observar que hay mayor cantidad tanto de series como de películas para la audiencia apta para mayores (TV-MA).
- Se puede observar que no hay series clasificadas con visualización restringida.
- La cantidad de películas disponibles para la audiencia en la que se recomienda encarecidamente la supervisión de los padres (TV-14) es aproximadamente el doble de las series disponibles para esta misma.
- Se produce aproximadamente la misma cantidad de series como de películas aptas para todo público (TV-G)
- Hay un 40 por ciento más de películas que de series para la audiencia que se recomienda supervisión de los padres (TV-PG)

análisis de datos

En nuestra evaluación experimental empleamos varias formas de entender nuestros datos una de esta fue el uso de graficas las cuales entre más grande se muestra la palabra quiere decir que más veces se recurre en los datos dada la función wordcloud.

MACHINE LEARNING

El machine learning es una disciplina del campo de la inteligencia artificial que, a través de algoritmos, dota a los ordenadores la capacidad de identificar patrones en datos masivos y elaborar predicciones, en machine learning tenemos tres categorías de algoritmos. Una habilidad esencial para hacer sistemas que no solo sean inteligentes, sino autónomos y capaces de identificar patrones en los datos para convertirlos en predicciones. Esta tecnología está presente actualmente en un sinfín de aplicaciones, como las recomendaciones de Netflix y Spotify, las respuestas inteligentes de Gmail o el habla natural de Alexa y Siri.

categorías de Machine learning

- Aprendizaje Supervisado: estos algoritmos cuentan con un aprendizaje previo basado en un sistema de etiquetas asociadas a unos datos que les permiten tomar decisiones o hacer predicciones. Un ejemplo es un detector de spam que etiqueta un e-mail como spam o no dependiendo de los patrones que ha aprendido del histórico de correos (remitente, relación texto/imágenes, palabras clave en el asunto, etc.).
- Aprendizaje no supervisado: estos algoritmos no cuentan con un conocimiento previo. Se enfrentan al caos de datos con el objetivo de encontrar patrones que permitan organizarlos de alguna manera. Por ejemplo, en el campo del marketing se utilizan para extraer patrones de datos masivos provenientes de las redes sociales y crear campañas de publicidad altamente segmentadas
- Aprendizaje por refuerzo: su objetivo es que un algoritmo aprenda a partir de la propia experiencia. Esto es, que sea capaz de tomar la mejor decisión ante diferentes situaciones de acuerdo con un proceso de prueba y error en el que se recompensan las decisiones correctas. En la actualidad se está utilizando para posibilitar el reconocimiento facial, hacer diagnósticos médicos o clasificar secuencias de ADN.

Machine learning en NETFLIX

Para nuestro proyecto usaremos el aprendizaje supervisado, debido a que debemos entrenar nuestro modelo asignándole etiquetas que le permitan tomar decisiones o hacer las predicciones para dar resultados más precisos, estos datos de entrenamiento consiste en pares de objetos, el objetivo que de nuestro aprendizaje supervisado es que nuestro modelo sea capaz de predecir con el uso de nuestras categorías tales como tipo, país, rating, objetivo de edad, polaridad, subjetividad, Gracias al Label Encoder logramos interpretar las cadenas como datos numéricos en nuestro

preprocesamiento, además haciendo uso de varios métodos como el DecisionTreeClassifier, KNeighborsClassifier, RandomForestClassifier los cuales son ampliamente explicados más adelante en este documento.

Evaluación experimental

En este aspecto convertimos nuestras películas y series en valores numéricos para una mejor interpretación de estos, esto también lo justificamos en el hecho de ser nuestros datos objetivo. Este procedimiento también lo realizamos a nuestros datos de país, Rating, categorías, y finalmente con Obj ages haciendo uso del Label Encoder y del algoritmo de sentimientos de Twitter.

Para este punto debimos centrarnos en un aspecto más complejo pues se tuvo que realizar un análisis de sentimientos para la descripción de cada una de las películas.

Como inicio para nuestras predicciones (Cadenas de string) se muestra en la imagen siguiente, en el cual se creó una función que permitirá la visualización de los datos.

	k_type	k_obj_ages	k_country	k_rating	polarity	subjectivity
0	1	3	39	8	0.050000	0.562500
1	0	3	308	8	-0.425000	0.600000
2	0	5	379	5	-0.475000	0.575000
3	0	0	549	4	0.100000	0.950000
4	0	3	549	4	0.900000	1.000000
...	...	...	...	...	...	...
7782	0	3	429	8	-0.329167	0.233333
7783	0	3	229	6	-0.325000	0.425000
7784	0	2	549	8	-0.333333	0.333333
7785	1	7	12	9	0.055556	0.333333
7786	0	2	473	8	0.120000	0.440000

7777 rows × 6 columns

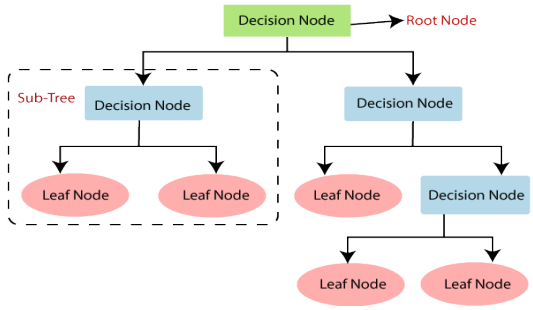


Figura 2: Imagen de referencia de un árbol de decisión.

Con el uso de nuestro decisionTreeClassifier evaluando a en primera instancia nuestros datos de entrenamiento y de testeo obtuvimos un gran resultado con un elevado porcentaje de validacion.

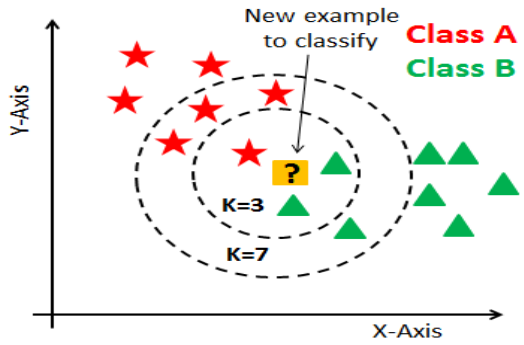
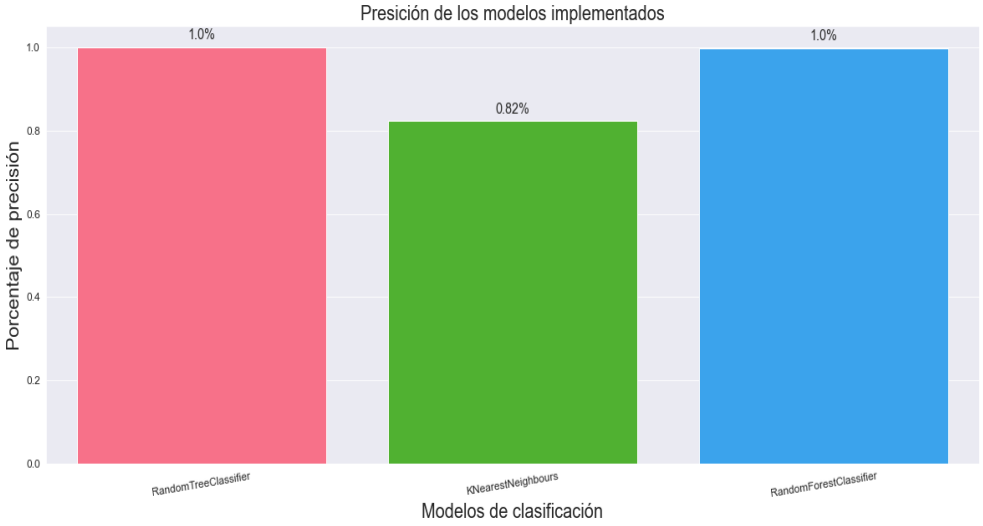


Figura 3: Imagen de referencia del modelo de K vecinos clasificador

Bajo la evaluación del modelo de K vecinos consideramos que tuvimos una evaluación bajo pues la tasa de éxito de los vecinos más cercanos fue de un 82.31 por ciento la cual consideramos baja por lo que se decidió continuar.

Como consideramos que el porcentaje de eficacia con nuestro modelo de K vecinos es bastante bajo decidimos hacer el uso de un modelo considerablemente más eficaz y eficiente, usamos el Bosque Aleatorio clasificador, con nuestro Random Forest Classifier logramos un 99.7000856898029 por ciento en torno a nuestro puntaje de precisión por lo que consideramos que fue muy acertado todo nuestro proceso de desarrollo.

En esta imagen podemos evidenciar como se comparan la precisión de los modelos implementados en nuestro trabajo de izquierda a derecha el DecisionTreeClassifier, K NearestNeighbours, y Random Forest Classifier con eficacias bastante altas exceptuando el modelo de K vecinos por lo que nos quedamos muy satisfechos con el rendimientos de estos.



Nuestra evaluación experimental obtuvimos que tan preciso fue nuestro modelo, entendido como el porcentaje de valores clasificados correctamente con respecto al total de elementos, primero hablando de nuestro modelo de árbol el cual fue muy eficaz a la hora de ser evaluado para nuestro modelo de K vecinos tuvimos un pequeño disgusto pues apenas logramos un 82 por ciento de validación y finalmente el modelo de bosque muy alto porcentaje de eficiencia cuando usamos un método de clasificación sea como en este caso RandomForestClassifier, con base a lo anterior el (Autor en Towards Data Science, 2021) entendemos que el método calcula la puntuación de precisión de forma predeterminada el método de puntuación no necesita las predicciones reales, pero estas predicciones fueron muy precisas gracias a un paso anterior realizamos una operación mucho más compleja, debemos de realizar un análisis de sentimientos para la descripción de cada película, para esto creamos y asignamos la variable el valor de descripción y le pedimos que para cada dato le haga la conversión y la agregue en el dataFrame, en este proceso de obtención de sentimientos empleamos el uso de un world con el fin de desglosar las categorías que damos a nuestra base de datos y poder lograr una mayor concordancia acerca de los sentimientos aplicados a las películas que integran nuestra base de datos.

Problemas de desarrollo

Hubo varias correcciones en el momento del procesamiento de datos pero uno de los apartados más importantes fue con el uso de un Word en el cual con el fin de no dejar los datos dispersos, esto porque nuestra bases de datos se dividía o categorizaba en unas 2000 o 3000 categorías para esto lo convertimos a una categoría general sin el uso de tantas con las que disponemos, con esto obtuvimos unos datos más ordenados y categorizados de una mejor forma, gracias a este cambio se logró de una forma muy eficiente la clasificación con estos datos.

Como se vio, el problema de esta base de datos es que posee categorías con sub categorías, o séase, nos provee datos demasiado específicos, un ejemplo (drama, pero este puede dividirse en drama acción o drama suspenso, sigue siendo drama pero mucho más específico), de cierto modo los datos que son específicos son buenos, pero para un método de predicción no, el no podrá diferenciar cual dato es igual a otro, pues estos amplían la definición de drama, ojo, solo nos hemos centrado en drama, pero con todas los tipos de genero está haciendo lo mismo. Por ello, nos baremos en un artículo realizado por RED+ Noticias y provisto por Claro.com, este nos da los principales géneros cinematográficos.

Figura 4: https://www.claro.com.co/institucional/generos-cinematograficos/

Ya con esto, las principales categorías son:

- Ciencia ficción-Sci-Fi
- Acción - Aventura - Action and Adventure
- Comedia - Comedy
- Fantasía - Fantasy
- Drama - Dramas
- Musical - Musical
- Romántico - Romantic
- Horror - Horror
- Documentales - Documentary
- Reality show - Reality show

Figura 5: Visualización de los géneros antes de su corrección

Conclusiones

Podemos concluir con respecto a la visualización de datos, encontramos un porcentaje alto con el uso de los métodos de DecisionTreeClassifier, K vecinos y de RandomForestClassifier los cuales fueron muy efectivos a la hora del desarrollo de nuestro trabajo.

Se llego a la comprensión de que no todas las bases de datos permiten, Se llego a la comprensión de que no todas las bases de datos permiten no todas las bases están preparadas para los métodos de machine learning, justificándose en la falta de relación entre ellos.

Se entiende que se debe realizar un buen preprocesamiento de datos con el fin de obtener resultados correctos y precisos a la hora de utilizar métodos de machine learning.

Repositorio GIT

https://github.com/BJUANR/Proyecto

Referencias

[1] ¿Qué es Netflix? - https://help.netflix.com/es/node/412

[2] Netflix Museum - https://flixable.com/netflix-museum/

[3] Claro Generos Cinematograficos - https://www.claro.com.co/institucional/generos-cinematograficos/