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 Especifico

Por medio de nuestra base de datos centrada principalmente en contenido agregado en la plataforma de Netflix predecir el comportamiento de los datos que puedan llegar a agregarse a la plataforma en un futuro, destacando campos y hábitos como el de género mediante datos como el director, país, etc.

Objetivo General

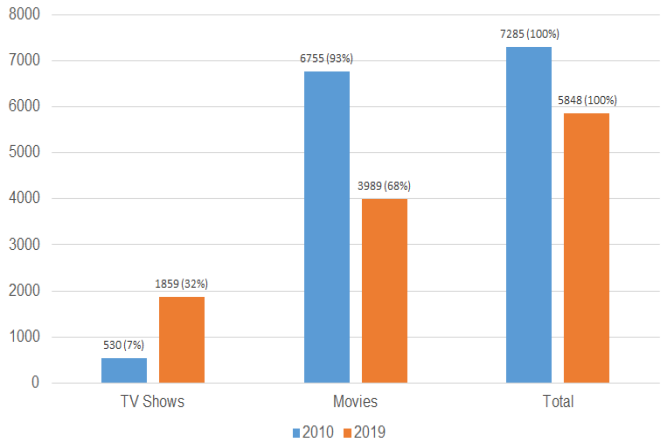
Obtener una idea clara y concisa del cambio a futuro que pueden tener las producciones encontradas en la plataforma de Netflix.

Palabras cláve

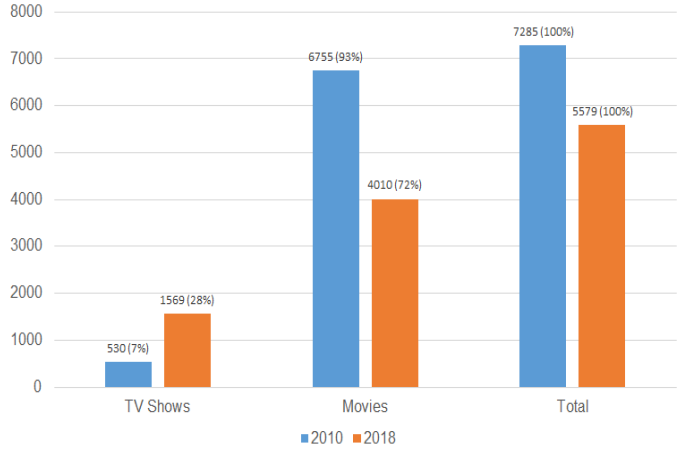
Netflix, Analisis de datos, problematica

1. **Recopilación de datos**
   Los datos que usaremos en este proyecto para su análisis sus 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 Analisis**
   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 permitan el poder predecir un género teniendo datos anteriormente cargados, aquí compilaremos y predeciremos tipos de datos como el director, país, actores, audiencia que tendrán lugar en un futuro cercano, estas predicciones se obtendrán por medio de la inteligencia artificial.

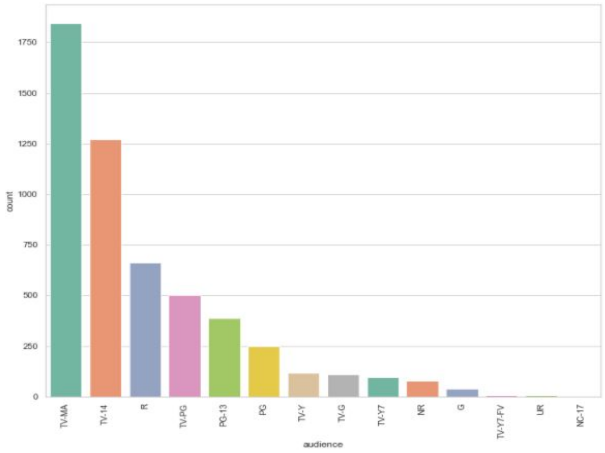
Así era el catálogo estadounidense de Netflix a principios de 2010.



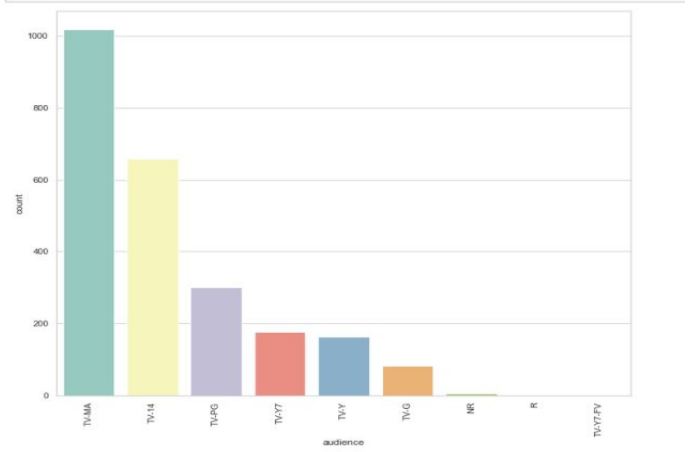
(a) Número de títulos en netflix en EE. UU. 2019



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



(c) Cantidad de películas según la audiencia



(d) Cantidad de series según audiencia

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.

Analisis de los datos entre audiencoa para peliculas 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)

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.

Categorias 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.

Los modelos de aprendizaje automático, y en concreto el aprendizaje por refuerzo, tienen una característica que los hace especialmente útiles para el mundo empresarial. “Es su flexibilidad y capacidad para adaptarse a los cambios en los datos a medida que ocurren en el sistema y aprender de las propias acciones del modelo. Ahí radica el aprendizaje y el impulso que faltaba en las técnicas anteriore.

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 de nuestras categorias tales como ”tipo”, ”pais”, rating”, .ºbejetivo de edad”, ”polaridad”, ”subjetividad”, previamente transoformadas en variables numericas en nuestro proceso de preprocesamiento, gracias a este logramaos tener una prediccion en números de los proximos titulos agregados en la plataforma, con resultados objetivos nuevamente dispersos y especificados en las categorias o caracteristicas previamente realizadas y proprocesadas, nuestro modelo fue confgurado y basado gracias al DecisionTreeClassifier con el fin de probar o testear el funcionamiento estable y preciso de nuestro modelo ademas de otorgarnos un 93,10 porciento de prediccion cercana del valor real predecido, de esta forma nuestro modelo aprenderá de los datos etiquetados, es decir, los datos del cual conoce la variable resultado, con todo esto nuestro modelo será capaz de hacer predicciones para ese resultado en nuevos conjuntos de datos. Nuestro proceso de machine learning buscamos predecir

las características de cada una de nuestras base de datos obteniendo un mejor estado de respuesta, ya que se busca una

predicción con un menor porcentaje de error, y así tener una predicción más acertada con respecto a estrenos de películas en Netflix en el futuro, es un proceso que se realiza para hacer un aprendizaje de datos para obtener y realizar un mejor estudio y funcionamiento del programa con respecto a la predicción y aprendizaje de datos

Evaluación experimental

Nuestra evaluacion experimental obtuvimos que tan preciso fue nuestro modelo, entendido como el porcentaje de valores clasificados correctamente con respecto al total de elementos., primero abalndo de nuestro modelo de arbol el cual nos arrojo un 82.02391667738202 porciento y para nuestro modelo de bosque un 82.61540439758261 porciento cuando usamos un metodo 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 description y le pedimos que para cada dato le haga la conversión y la agregue en el dataFrame, en este proceso de obtencion de sentimientos empleamos el uso de un world con el fin de desglozar las categorias que dariamos a nuestra base de datos y poder lograr una mayor concordancia acerca de los sentimientos aplicados a las peliculas que integran nuestra base de datos, despues de todo este procesologramos obtener un muy notable 93,10 porciento de prediccion cercana del valor real predecido.

Arreglo de datos

Hubo varias correciones en el mometno del procesamietno de datos pero uno de los apartados mas 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 dividia o categorizaba en unas 2000 o 3000 categorias para esto lo convertimos a una categoria general sin el uso de tantas con las que disponiamos, con esto obtuvimos unos datos mas ordenados y categorizados de una mejor forma, gracias a este cambio se logro de una forma muy eficeinte la clasificacion 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 3: https://www.claro.com.co/institucional/generos-cinematograficos/

Ya con esto, las principales categorías son:

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

- Reality show - Reality show En total 8 géneros, ahora es proceder de esto,

0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...

Conclusiones

Es interesante el saber cómo opera la base de datos de netflix, entender qué tipo de películas son las más relevantes a base de su categoría, en los datos cargados pudimos identificar los diferentes tipos de información en los que se pueden visualizar como el código de la película, el nombre de la película, el director por el cual fue dirigida, país de origen, duración, entre otros. Estos datos serán importantes en nuestro proyecto porque con estos datos proyectaremos y pronosticaremos sobre el próximo estreno de una película basándonos en su país, su género, sus actores y su tipo de audiencia.

Podemos concluir con respecto a la visualización de datos, encontramos un porcentaje alto de aceptación con respecto a los datos tomados por categorías para tener una mejor predicción y poder obtener una mejor respuesta.

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/