

10 mo
CONGRESO NACIONAL

CoNaIISI

Congreso Nacional de Ingeniería Informática / Sistemas de Información

2022 Modalidad Híbrida

Facultad Regional Concepción del Uruguay

Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay
10mo. Congreso Nacional de Ingeniería Informática y Sistemas de Información /
compilación de Adrián Callejas ... [et al.]. - 1a ed. - Ciudad Autónoma de Buenos Aires
: Universidad Tecnológica Nacional, 2022.

Libro digital, PDF

Archivo Digital: descarga y online

ISBN 978-950-42-0218-9

I. Tecnología Informática. I. Callejas, Adrián, comp. II. Título.
CDD 004.071

ISBN 978-950-42-0218-9



Este obra está bajo una licencia de Creative Commons Reconocimiento-NoComercial 4.0 Internacional.



**10 mo. Congreso Nacional de
Ingeniería Informática / Sistemas de Información**

3 y 4 de Noviembre 2022

Facultad Regional Concepción del Uruguay

Análisis del contexto de Machine Learning en el enfoque de aplicaciones multimedia

Roxana Martínez, Rocío Vargas
Centro de Altos Estudios en Tecnología Informática (CAETI).
Universidad Abierta Interamericana,
Ciudad Autónoma de Buenos Aires, Argentina
{Roxana.Martinez, Rocio.Vargasnavarro}@uai.edu.ar

Resumen

El enfoque de Machine learning permite crear un valor agregado y ayuda a entender mejor los cambios en el comportamiento, las preferencias o bien las necesidades de una persona. A través de la utilización de los datos abiertos que son facilitados en distintos portales web, éstos pueden ser de gran utilidad para el proceso de análisis. Este trabajo se enfoca en un relevamiento del contexto de machine learning en aplicaciones multimedia identificando sus distintos tipos como así también trabajos relacionados para este contexto que, posteriormente, se estudian algunas de las falencias encontradas en los algoritmos utilizados en multimedia. Este trabajo, además, aporta un análisis de éstos a través de la utilización de un software en el que se realiza la comparativa de distintos algoritmos de machine learning en el contexto de aplicaciones multimedia detectando así, cuáles de estos podrían ser los más adecuados para este entorno. Para esto se presenta una comparativa en las que se cotejan los distintos valores arrojados y se enfoca en los resultados para la curva de rendimiento AUC-ROC, y el detalle de la matriz de confusión que validará los resultados demostrados en la comparación y la clasificación de precisión dentro de los valores: Accuracy, Precision, Recall y Area Under de ROC Curve (AUC).

Palabras Claves: Datos Abiertos, Machine Learning, Algoritmos para Multimedia, Over-the-top (OTT).

Introducción

En la actualidad la utilización de los datos es importante a la hora de realizar enfoques tanto de análisis, seguimiento, estadísticas y demás, por lo que éste, es cada vez más utilizado en distintas metodologías y áreas en general.

Desde el enfoque de datos abiertos públicos, se puede decir que, “mediante el acceso a los portales de datos

abiertos, se pueden extraer datasets de diferentes categorías de información. Algunos portales gubernamentales presentan APIs para obtener principalmente información en tiempo real, de transporte, tránsito, entre otros. Otros tipos de archivos públicos utilizados para la apertura suelen entregarse en formatos estandarizados como, por ejemplo: CSV, XLS, XLSX, XML, entre otros. La mayor diversidad de formatos ofrecidos facilitará que un tercero pueda elegir el formato conveniente para visualizar los datos o utilizarlos como entrada de otra herramienta o sistema” [1]. Cabe señalar que la cultura de datos abiertos trata de obtener un valor añadido de la información y de esta forma agregar un aporte a la comunidad social, esta “información no genera sólo su valor por estar reservada a unos pocos, sino que lo hace por su disponibilidad para ser interpretada y traducida por cualquier actor interesado en trabajar con ella” [2].

Desde una mirada hacia las aplicaciones multimedia, cada vez se encuentran más en pleno auge de utilización. Este concepto se centra en el uso combinado de diferentes medios de comunicación: texto, imagen, sonido, animación y video, básicamente, son los programas informáticos que utilizan de forma combinada sus objetivos en diferentes medios, y permiten la interacción con el usuario, éstas son las conocidas aplicaciones multimedia interactivas [3]. En base a esto existen dos características básicas para tener en cuenta en este tipo de aplicaciones, a saber:

a) **Multimedia:** Permite el uso de múltiples tipos de información (textos, gráficos, sonidos, animaciones, videos, etc.) integrados coherentemente;

b) **Hipertexto:** Permite la interactividad basada en los sistemas de hipertexto, que permiten decidir y seleccionar la tarea que deseamos realizar, rompiendo la estructura lineal de la información.

Este trabajo se basa en el aspecto de “multimedia” en el contexto de datos abiertos para el entorno de machine learning.

Machine Learning en multimedia

El mundo en multimedia está cambiando la metodología de visualización, la experiencia de los usuarios y el criterio de eje comercial en esta temática. “La llegada de Internet ha tenido un profundo impacto en el video y la forma en que nos relacionamos con él. La capacidad de entregar contenido a través de redes de telecomunicaciones cableadas e inalámbricas ha abierto la puerta a servicios en línea innovadores como YouTube [4], Vimeo [5] y Twitch [6], así como Netflix [7], Hulu [8], Amazon Video [9] y muchos otros” [10].

Para el usuario final, todo es transparente e invisible, pero con cada una de las tecnologías a utilizar, hay una amplia variedad de modelos de negocio que controlan el acceso al contenido. Existen algunos términos fundamentales a tener presente, en un relevamiento de estos aspectos [10], por lo que se resumen los siguientes:

- **VOD - Video on demand:** El término general para esta nueva generación de entretenimiento es obtener el contenido en cualquier momento, no cuando un emisor o bien propietario de contenidos, decide entregarlo a los usuarios.
- **OTT - Over-the-top:** Se refiere al material que se distribuye directamente a los espectadores a través de Internet. OTT es un subconjunto de VOD, que también incluye servicios de cable y satélite.
- **SVOD – subscription video on demand:** es similar a los paquetes de TV tradicionales, lo que permite a los usuarios consumir tanto contenido como deseen a una tarifa plana por mes. Los principales servicios incluyen Netflix [7], Amazon Prime Video [9] y Hulu [8], con nuevos servicios provenientes de Apple, HBO [11] y Disney+ [12]. Los proveedores de SVOD se enfrentan continuamente al desafío de retener a los consumidores, al proporcionar contenido nuevo exclusivo, esquemas de precios y demás.
- **TVOD – transactional video on demand:** TVOD es lo opuesto al video de suscripción, donde los consumidores compran contenido sobre una base de pago por visión. Hay dos subcategorías, conocidas como venta electrónica: a) EST (Electronic sell-through), donde se paga una vez para obtener acceso permanente a una parte de contenido; y b) DTR (download to rent), descarga para alquilar, donde los clientes acceden a una parte de contenido por un tiempo limitado por una tarifa menor. Los servicios de TVOD tienden a ofrecer lanzamientos más recientes, proporcionando a los titulares de derechos mayores ingresos y dando a los consumidores acceso oportuno a nuevos contenidos. Ejemplos de servicios TVOD incluyen: iTunes de Apple [13], Sky Box Office [14] y la tienda de videos de Amazon.

- **AVOD – advertising-based video on demand:** El Video bajo demanda basado en publicidad, a diferencia de los servicios SVOD y TVOD, AVOD es gratuito para los consumidores. Sin embargo, al igual que la televisión abierta, los consumidores tienen anuncios. Ejemplo de esto es: DailyMotion [15], YouTube [4] y 4OD [16], donde los ingresos publicitarios se utilizan para compensar los costos de producción y alojamiento.

A modo de resumen, se puede indicar que en los sistemas de transmisión de video como Subscription Video On-demand (SVOD) y Over-the-top (OTT), se encuentran algunas de las plataformas más populares a nivel internacional, como ser: Netflix [7], Disney+ [12] y Amazon Prime Video [9] como las más estudiadas por diversos artículos de investigación, y las que tienen como su principal objetivo de negocio, la reproducción de video para los usuarios.

Desde el punto de vista de los estudios de Machine Learning para este contexto, éstos se basan en base a las recomendaciones que les pueden dar a los usuarios para continuar viendo las opciones de reproducción de video o atraerlos a futuras posibles transmisiones que guarden en sus listas debido a las sugerencias que estás brindan a los usuarios.

Muchas campañas de marketing vienen dadas de las visualizaciones más populares de los usuarios de sus plataformas, lo más visto, lo más puntuado o recomendado. Estas campañas, por lo general, serán dirigidas a público muy similar de aquellos usuarios que se tomen los datos como origen.

Los datos abiertos en este contexto son sumamente importantes, por ejemplo, existen portales que disponibilizan datos, como ser “*Datos abiertos en AWS*” [17], que permite compartir cualquier volumen de datos con cuantas personas se desee. Cuando se comparten datos en AWS, cualquiera puede analizarlos y crear servicios sobre ellos, mediante un amplio rango de tecnologías informáticas y de análisis de datos. Compartir datos en la nube permite que los usuarios dediquen más tiempo al análisis de los datos y no a su obtención.

Uno de los puntos para tener en cuenta es *¿en qué se diferencia el aprendizaje automático en la nube?*, según [18], la combinación de la potencia de cómputo masiva, los lagos de datos, la seguridad, las capacidades de análisis y la habilidad para integrarse con los servicios en la nube está transformando el aprendizaje automático de una tecnología de nicho y experimental a un bloque de creación de negocios fundamental. Varias empresas, utilizan machine learning con el fin de preparar los datos para el análisis, y desarrollar modelos que permitan obtener beneficios para los usuarios finales, como ser: reconocimiento de voz, el análisis de imágenes y videos, el ofrecimiento de previsiones y recomendaciones, y muchas otras soluciones.

Otra de las preguntas para tener en cuenta es: *¿Cuáles son algunas de las ventajas que el aprendizaje automático*

en la nube aporta a los videos?, según [18], los proveedores de videos modernos tienen muchas preguntas en común:

- *¿Qué actores están en una escena?*
- *¿Cuándo se pronuncian determinadas palabras?*
- *¿Qué objetos hay en la pantalla?*

El aprendizaje automático para video en la nube ofrece una manera muy conveniente de responder a cada una de estas preguntas. A continuación, se incluyen algunos ejemplos extraídos de [18]:

- **Archivos de videos que admiten búsquedas:** Con los servicios de aprendizaje automático en la nube, los equipos de video pueden reducir sustancialmente el tiempo y los recursos invertidos en catalogar, buscar y crear recursos en sus archivos de video. Con la indexación de contenido impulsada por el aprendizaje automático y la generación de metadatos, se pueden habilitar varias aplicaciones con beneficios reales y significativos.
- **Subtítulos de video automáticos:** Los metadatos de los subtítulos son fundamentales para que los videos sean útiles y la audiencia pueda acceder a ellos. No obstante, puede ser costoso y lento transcribir recursos de video, y producir e integrar subtítulos precisos en los diferentes formatos que se requieren para asegurar el acceso en las distintas pantallas y dispositivos.
- **Generación de videoclip:** El proceso de generar y publicar videoclips ha requerido un flujo de trabajo manual para poder identificar el contenido relevante del video sin procesar, generar clips con tiempo codificado, luego trans codificar, empaquetar y distribuir esos clips para su publicación en las redes sociales. Las herramientas de aprendizaje automático pueden automatizar los pasos clave del proceso para ayudar a las emisoras a entregar clips de gran valor en las pantallas de los espectadores casi en tiempo real.
- **Personalización y monetización:** A fin de optimizar las oportunidades de ingresos para transmitir video, los proveedores de contenido deben equipar la infraestructura para poder suministrar publicidad que se oriente a los usuarios individuales y se entregue de modo que se adapte a cada espectador, por ejemplo, mediante anuncios personalizados.
- **Análisis y mediciones:** Los flujos de trabajo actuales tienen la capacidad de medir e informar cantidades aparentemente infinitas de información sobre las transmisiones en directo y de video bajo demanda, así como la infraestructura detrás de ellas. La aplicación de sistemas de aprendizaje automático ofrece a los proveedores de contenido nuevas maneras de optimizar cada aspecto del flujo de trabajo de video, incluidos el rendimiento

de ese flujo, el uso de los recursos de red, los resultados de monetización, y mucho más.

A continuación, se analiza la inclusión de la multimedia en el contexto de Machine Learning a través de un relevamiento de trabajos de investigación.

Trabajos Relacionados

Desde ya casi una década, las plataformas multimedia existentes intentan ajustar sus algoritmos con objetivos distintos ya sea por retención o captación de nuevos usuarios. Existen varias técnicas implementadas que se explicarán a lo largo de este trabajo. En primera instancia se debe mencionar al “Filtrado Colaborativo”, este método se basa en brindar recomendaciones de contenido basado en la información que los usuarios vieron en el pasado y les ha gustado. En algunos estudios realizados, se demuestra que este tipo de recomendaciones no eran útiles para predecir lo que los usuarios querían ver al día siguiente [19], de todas formas, se ha convertido en el modelo más frecuente y adoptado a comparación de otros modelos. También se estudiaron ciertos problemas conocidos como Scalability y Sparsity los cuales no han sido resueltos. Debido a esta limitación, la precisión en la recomendación puede llegar a ser baja [20]. De este modelo se desprenden dos tipos de filtrado colaborativos, “Basado en el Usuario” y “Basado en Memoria”.

En aspectos del filtrado “Basado en el Usuario”, éste compara similitudes del comportamiento del usuario. Por ejemplo: en ver los mismos contenidos, es decir, que se van a recomendar una cantidad similar de temas a un usuario con esas similares preferencias de otro usuario. Es importante aclarar que desde el punto de vista de machine learning y la utilización de algoritmos, para este tipo de filtrado, se utilizan algoritmos como Correlación de Pearson, Similitud coseno y K-nearest neighbors (KNN) de esta forma se pueden crear grupos con preferencias similares [20].

El segundo tipo de filtrado “Basado en Memoria” se utiliza para las predicciones, por ejemplo: como para conocer una futura baja del usuario en la plataforma o la caída de las ventas o suscripciones. De esta provienen tres problemas conocidos como Sparsity, debido a que no hay suficiente cantidad de datos disponibles, el siguiente es el Cold Start, en el cual no existen datos evaluados previamente que se puedan usar para la ejecución del modelo, y el tercer problema es Gray Sheep, en el cuál la muestra de los usuarios que van a ser evaluados y tienen similitudes entre ellos es muy pequeña. Para este tipo, se usan algoritmos tales como Clustering, Singular Value Decomposition (SVD), y Análisis de Componentes Principales (PCA) [20].

Para complementar los problemas de los modelos, se comenzaron a utilizar otras técnicas como el testeado de A/B test para comprender la psicología de sus usuarios y personalizar su experiencia como menciona en [19]. Por ejemplo, Netflix adopta estas pruebas (tests) para

ejecutarlos de forma online con el objetivo de medir cuál de las dos versiones tienen mejores rendimientos, por ejemplo, se toma una muestra de un grupo de control (A) y otro grupo (B) en donde se corre un prototipo para los dos casos y se observa el comportamiento de ambos grupos analizando los resultados y obteniendo información que llevará a decidir sobre los productos ofrecidos [21].

Otras de las técnicas como, “el Filtrado Basado en Contenido”, se utiliza donde se toma como referencia la descripción de los atributos que les han gustado a los usuarios y el detalle del elemento, por ejemplo: este puede ser un título de una película o una serie, género o tal vez información relevante para el análisis. Para este método se usan técnicas de aprendizaje automático como Clasificadores Bayesianos, Análisis de Cluster, Árboles de Decisión, and Redes Neuronales Artificiales, con el objetivo de estimar que tan probable es que ese contenido le guste al usuario en estudio [22]. Adicionalmente, otros trabajos proponen de forma práctica el uso de este método basándose en el algoritmo de distancias Euclidean [23].

Falencias detectadas

Debido a que ambos Filtrados Colaborativos y Basados en contenidos, han tenido problemáticas y limitaciones se genera el modelo Híbrido, combinando tanto el detalle del contenido como interacciones y comportamiento del usuario, de esta forma el modelo toma una robustez importante combinando los modelos. Otro de los modelos que pretende ayudar en esta problemática, es el denominado LightFM que es una implementación de varios algoritmos de recomendación y que permite incorporar metadatos de todos los elementos, tanto del usuario como del contenido, la información adicional que se suma en estos modelos podría ser edad, género, demografía, entre otros y debe ser utilizada en el entrenamiento [25]. A continuación, se observa la Figura 1, en la que se grafica a modo de resumen, los modelos explicados en la sección anterior, recopilando las ideas principales de [19] [23] [24]:

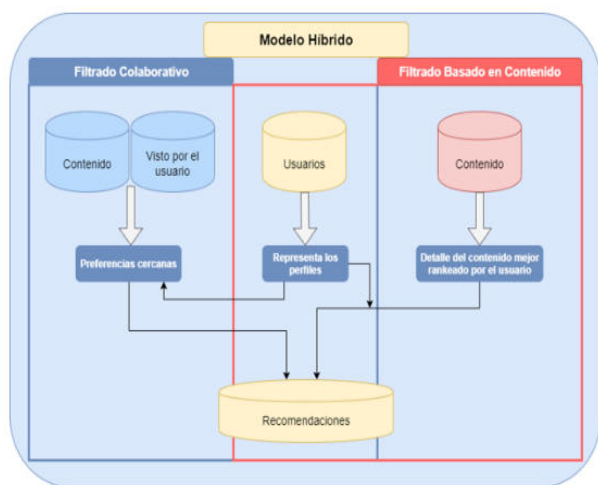


Figura 1. Resumen del modelo Híbrido sobre el filtrado colaborativo y basado en conocimiento.

En la Figura 1, se muestra el enfoque de las preferencias cercanas, tomando en cuenta el contenido y las vistas de usuarios, por otra parte, se muestra la representación e interacción de los perfiles de usuarios para ambos filtros. Finalmente, se observa el detalle del contenido mejor rankeado por el usuario desde la visión de un filtrado basado en contenidos.

En lo que respecta a los algoritmos utilizados de machine learning, se realiza una Tabla comparativa con dichos filtrados y los algoritmos utilizados en cada uno de estos casos, según lo relevado en los trabajos de investigación mencionados anteriormente, en la Tabla 1, se muestra la comparativa de éstos.

Tabla 1. Cuadro comparativo de algoritmos utilizados en cada método

Algoritmos utilizados a lo largo de los métodos		
Filtrado Colaborativo		Filtrado Basado en Contenido
Basado en Usuario	Basado en Memoria	
▪ Correlación de Pearson	▪ Clustering	▪ Clasificadores Bayesianos
▪ Similitud Coseno	▪ Singular Value Decomposition (SVD)	▪ Análisis de Cluster
▪ K-nearest neighbors (KNN)	▪ Análisis de Componentes Principales (PCA)	▪ Árboles de Decisión
		▪ Redes Neuronales Artificiales
		▪ Eclidean

Existen comparaciones realizadas de varios algoritmos de los mencionados para entender cuál es el peso de cada algoritmo, algunos trabajos comparan diversidad de algoritmos como se menciona en [25], donde se estudian J48-decision Tree, K-nearest neighbors (KNN), rule-based PART, y Suport Vector Machine (SVM); esto permite que se analice y se clasifiquen para el consumo para las OTT. Éstas se miden por la clasificación de precisión tales como: Recall, Precision y F-Measure.

Otros estudios realizan comparaciones con clasificadores de Precisión, Recall y Kappa [26], con el objetivo de comparar algoritmos combinados para datos de plataformas OTT.

En lo que respecta a las fuentes de datos utilizadas en este contexto, existen estudios como [27], en donde se profundiza el perfil del usuario y en el que se abordan datos obtenidos a través de encuestas personalizadas, de esta forma, se puede obtener datos más precisos del comportamiento de los usuarios tales como, “¿ha visto alguna transmisión de video on-demand las últimas 4 horas?”, “¿Cuántas o cuáles personas han visto series o películas a través de OTT?”, “¿Qué tipo de contenido ha visto?”, entre otras. Un detalle más amplio de estas

características es la obtención de un mayor análisis a la hora de hacer recomendaciones.

Un punto para destacar es la selección de datos y la combinación de ellos para los objetivos de análisis respondiendo a la pregunta de “¿Qué vamos a recomendar?”.

En otras investigaciones [28] emplean en sus análisis una gran gama de diversos datos, incluyendo la interacción de los usuarios como utilizar un “*Me Gusta*” o “*No Me Gusta*”, el texto del argumento de la película o serie, los actores involucrados en el reparto y los carteles de las portadas de las películas o series. En esta investigación [28], se destaca que pueden interpretar recomendaciones desde las particularidades de datos incluidos y mencionados anteriormente.

Propuesta

Como objetivo de esta sección, este trabajo de investigación presenta el análisis de la curva de rendimiento AUC-ROC para cruzar los algoritmos recomendados en los análisis previamente explicados, y la matriz de confusión que validará los resultados demostrados en la comparación y la clasificación de precisión será dentro de los valores de: Accuracy, Precision, Recall y Area Under de ROC Curve (AUC).

Los 5 algoritmos propuestos para el análisis son: K-NN, Naive Bayes, Decision Tree, Gradient Boosted Trees y Random Forest, a continuación, se describe el funcionamiento para cada caso:

- **K-NN** es un método de clasificación de agrupamiento. Los pasos para generar esta agrupación son: 1) confirmar a qué clase pertenecen los k vecinos, y 2) realizar la clasificación tomando el voto mayoritario basándose en el resultado. En el caso se parametrizar el valor “ k ” en 1, se asigna a la clase de vecino más cercano. Este algoritmo aporta sencillez con la asignación de sólo un parámetro, pero si la distribución es distorsionada debido a que los datos no pertenecen a la misma clase no se puede aplicar K-NN [29].
- **Naive Bayes** es otro típico método de clasificación que se basa en supuestos estadísticos del teorema de Bayes y del cuál, se parte de la suposición de que todas características de entradas son independientes. Debido a este supuesto, tiene un menor tiempo de aprendizaje, pero a su vez, podría llegar a no ser realista. La precisión puede disminuir al utilizar una muestra pequeña de datos [29].
- **Decision Tree y Gradient Boosted Trees** se basan en el modelo de árbol y se reutilizan para clasificar un conjunto de datos en varios subgrupos. Los nodos se despliegan desde la parte superior del árbol hacia los nodos inferiores demostrando las

clasificaciones correspondientes. Los algoritmos de árbol de decisión más típicos son ID3 y C4.5. [29]. En la propuesta se utiliza Gradient Boosted Trees para combinar la clasificación con la regresión, ambos son métodos de conjunto de aprendizaje progresivo que obtienen resultados predictivos a través de estimaciones mejoradas gradualmente [33].

- **Random Forest** crea o entrena un conjunto de árboles que se especifican en los parámetros iniciales. Cada nodo de un árbol representa una regla para un atributo específico, la regla va a clasificar los valores para separarlos en diferentes clases. Este algoritmo también puede utilizarse como método de regresión con el objetivo de reducir el error de la estimación. La diferencia con Decision Tree es que solo crea un árbol, donde todos los atributos están disponibles en cada nodo, la predicción en este caso puede ser más comprensible para su interpretación [33].

El uso de 5 algoritmos seleccionados proveerá indicadores para estimar principalmente como estos reaccionan ante un conjunto de datos real (dataset) y así, comprender mejor los posibles casos de estimaciones sobre las recomendaciones. Para esto, se han elegido estos algoritmos debido a los estudios previamente analizados y se determinó el “LABEL” con 2 categorizaciones con el objetivo de no fragmentar la muestra con una serie de clasificaciones para este primer análisis.

La propuesta es demostrar que, en un primer análisis comparativo de algoritmos estudiados, existen algoritmos con mejor performance para recomendar películas o programas de acuerdo con la historia obtenida de los usuarios que ya han vivido la experiencia de ver y realizar opiniones o bien puntuaciones dentro de la plataforma. Esto servirá para usuarios que actualmente ya se encuentran suscriptos y podría ser útil para personas que no se han suscripto, en las cuales poder focalizar la generación de nuevas campañas de Marketing que atraigan a un público dirigido con similitudes de los actuales usuarios.

a. Fuente de Datos (dataset)

Los datos utilizados en este trabajo para el primer análisis contienen 992 registros y 20 columnas, éstas incluyen los nombres: “*imdb_id*” (Internet Movie Database ID), “*title*” (título del programa/película), “*plot*” (trama del programa/película), “*type*” (tipo del programa/película), “*rated*” (calificación), “*year*” (año de lanzamiento), “*released_at*” (fecha de lanzamiento), “*added_at*” (fecha en la que se agregó el título a la plataforma), “*runtime*” (tiempo de reproducción), “*genre*” (género), “*director*” (director), “*writer*” (escritor), “*actors*” (actores), “*language*” (idioma disponible), “*country*” (país disponible), “*awards*” (Premios), “*metascore*” (valoración), “*imdb_rating*” (rating por vistas),

“*imdb_votes*” (votos de usuarios) y “*score*” (ponderación tomada para el análisis). El dataset llamado “*Disney Plus Movies and TV Shows*” [30] fue descargados desde el portal de Kaggle [31], y muestra el contenido de los listados de películas y programas de TV en Disney Plus.

El cálculo realizado para contemplar el “*score*” fue mediante la ponderación de los atributos “*metascore*” en el cual se tomó el promedio de la columna en el valor de 62, el “*imdb_rating*” en un valor medio de 6,5 y el “*imdb_votes*” en 58.700. Una vez obtenidos los valores ponderados se hizo un condicional en el cuál tanto el “*metascore*” como el “*imdb_rating*”, como valores contemplados para las vistas generadas por las películas o programas, debían superar el promedio para ser catalogados como “*Alto*”. Como segundo condicional se contempla la cantidad de votos del atributo “*imdb_votes*” en el cuál para ser catalogado como “*Alto*” debe tener un valor mayor a 58.700 independientemente de los ratings o “*metascore*”, este atributo fue considerado de suma importancia debido a que los usuarios colocan dicho voto.

En el estudio [32], se analiza este mismo conjunto de datos por grupos de edad de las películas que a los usuarios les gustaría ver, se exhibe que en Disney+ [12] tiene menos contenido respecto a otras plataformas y su distribución es mucho más fuerte en películas familiares, de aventuras y de animación.

b. Prototipo propuesto utilizado

Utilizando el software es RapidMiner Studio [33], inicialmente se hizo la limpieza correspondiente para los casos en los que no hubo: a) votos; b) ranking; y c) ni puntuación; por parte de los usuarios. Para este caso, se excluyeron 122 registros, con el fin de no alterar el promedio de los casos registrados. El siguiente paso fue utilizar la misma copia del conjunto de datos (dataset) para ejecutar cada uno de los algoritmos en donde se obtiene la Matriz de confusión de cada uno de éstos, y, por otra parte, las métricas referidas a las precisiones.

Como siguiente paso, se toma en cuenta que, al ejecutar el proceso en el software, también se ejecuta de manera simultánea la comparación ROC (Característica Operativa del Receptor) para obtener la comparativa de algoritmos, es decir, cotejar los 5 algoritmos en análisis. Finalmente, durante la ejecución de cada algoritmo también se agrega un vector de performance para evaluar la clasificación del campo Binominal Score seleccionado para el “*LABEL*”.

En la Figura 2, se muestra el proceso construido para el prototipo, donde se puede observar el dataset en el objeto “*Retrieve disney_plus_shows*”, el filtro utilizado para la limpieza con el nombre “*Filter Examples*”, el objeto “*Multiply*” que se encarga de distribuir una copia de datos a “*Compare ROCs*” y otra copia a la aplicación del algoritmo correspondiente dentro del operador de performance “*Cross Validation*”.

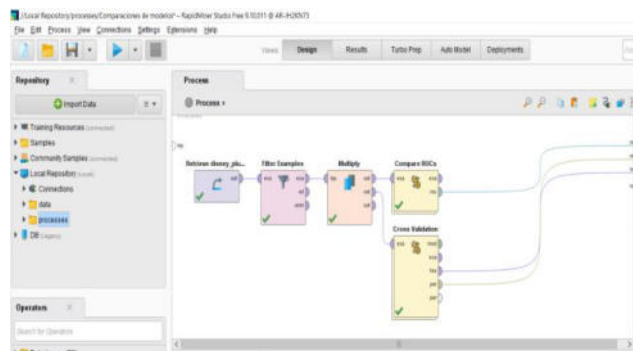


Figura 2. Prototipo del proceso desarrollado en RapidMiner Studio [32].

En la Figura 3, se observa el detalle del “*Compare ROCs*” dentro del proceso, donde se realiza la ejecución simultánea de los 5 algoritmos analizados.

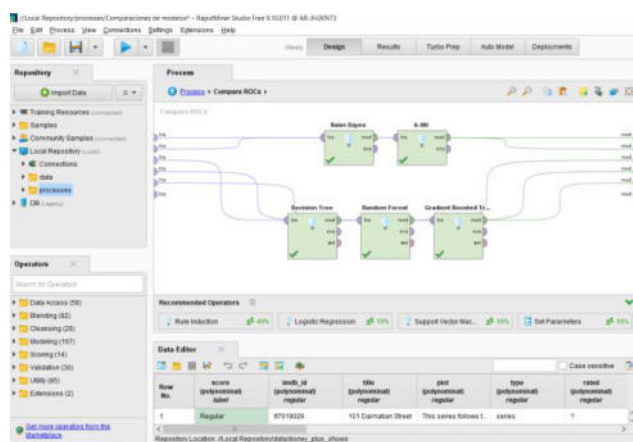


Figura 3. Proceso Compare ROCs desarrollado en RapidMiner Studio [32].

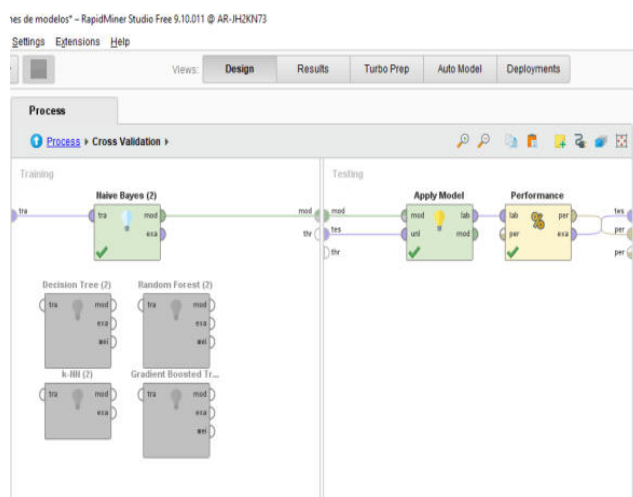


Figura 4. Proceso Cross Validation desarrollado en RapidMiner Studio [32].

En la Figura 4, dentro del operador “*Cross Validation*” del proceso, se pueden ver los 5 algoritmos que han sido procesados uno por vez para obtener los valores individuales de las predicciones. Se observa que, en la sección de “*Training*”, expuesta a la izquierda, se

encuentra habilitado “Naive Bayes”, en el momento de ejecutar el siguiente se habilita y deshabilita el que corresponda. En la sección “Testing” se aplica el modelo y se evalúa la performance.

A modo de resumen, los parámetros establecidos para cada algoritmo se indican a continuación:

- **Naive Bayes** por defecto se seleccionó “Laplace Correction” para evitar la probabilidad de obtener valores ceros.
- **Decision Tree** se parametrizó con “Maximal Depth” de 10 nodos y “Confidence” de 0,1.
- **Random Forest** se configuraron los parámetros de “Maximal Depth” de 10 nodos y un “Number of Trees” de 100.
- **Gradient Boosted Trees** fue establecido con “Maximal Depth” de 5 nodos y un “Number of Trees” de 50.
- **K-NN** se parametrizó con una “k” en 5 para obtener una clasificación dentro de 5 vecinos más cercanos.

c. Análisis de resultados

En esta sección se analizan los aspectos más determinantes en este estudio, por lo que se realiza la comparación de ROC (Característica Operativa del Receptor). En la Figura 5, se muestra la curva que revela la performance de los modelos de clasificación seleccionados. Por otra parte, el eje vertical (Y) indica la clasificación de Verdaderos Positivos y el eje horizontal (X) muestra la clasificación de Falsos Positivos. Si el umbral de la curva se reduce la clasificación de elementos positivos aumenta, tanto en Verdaderos Positivos como en Falsos Positivos.



Figura 5. Resumen del modelo Híbrido sobre el filtrado colaborativo y basado en conocimiento.

En la Figura 5, se muestra la comparación de los mejores algoritmos de machine learning que se ejecutaron con los datos de contenido de TV de Disney+ [12] de la sección anterior. Estos son: a) en primer lugar Random Forest”, b) luego “Decision Tree” y c) por último “Gradient Boosted Trees” con similitudes en su curva y en donde se puede apreciar que el umbral de la curva es menor que el resto de los algoritmos, por lo tanto, el Área Under de ROC Curve (AUR) aumenta, esto quiere decir que los algoritmos arrojarán mayor cantidad de casos positivos.

En contraparte, se observa que los algoritmos de Naive Bayes y K-NN presentan una curva ROC de menor precisión debido a que el umbral de la curva es mayor y por lo tanto el Área Under de ROC Curve (AUR) disminuye, esto permite afirmar que se predecirán menos casos positivos.

Para el análisis se toman los 3 mejores, con el fin de evaluar en la Matriz de Confusión y analizar cuales contenidos tienen una recomendación dentro de las clasificaciones “Alto” o “Regular”. En la Tabla 2, se explican los valores considerados para cada tipo (Regular y Alto):

Tabla 2. Cuadro comparativo de valores para el enfoque de clasificación Regular y Alto

	Clasificación Regular	Clasificación Alto
Pred. Regular	Verdaderos Positivos	Falsos Positivos
Pred. Alto	Falsos Negativos	Verdaderos Negativos

Las clasificaciones ayudan a medir la performance de los resultados de los algoritmos, en este caso, en 2 categorías. La Tabla 2 consiste en 4 celdas de las cuales cada una representa a una única combinación de predicción de valores.

Para la clasificación entre “Regular” y “Alto” si se toma la celda “Verdaderos Positivos” significa que la predicción se clasifica por su nivel de “sensibilidad”, lo que representa en este caso es que son datos predichos con una estimación altamente posible y bien calificada.

En el caso de “Verdaderos Negativos” la predicción fue “específica” y se puede predecir los casos que forman parte de la segunda clasificación, en este caso es “Alta”, esto no significa que sea buena o mala en términos de las “Recomendaciones” finales a los usuarios, simplemente se da la precisión “Verdadera” para los casos contrarios de la primera clasificación.

La clasificación “Falsos Positivos” predice valores “Positivos” en donde el valor real fue “Negativo”. Aquí el algoritmo no pudo predecir de forma correcta la clasificación “Alto” y la calificó de forma incorrecta en “Regular”.

En el último caso “*Falsos Negativos*” predice los casos en los que son “Negativos” pero el valor real fue “Positivo”, para estos casos tenía que haber sido una predicción “Regular” y la predijo como “Alto”.

En la Tabla 3, se demuestra un ejemplo de cada caso, los indicadores de Confianza se toman para el porcentual de estimación, cercano a 1 la estimación es más precisa y cercano a cero es menos precisa.

Tabla 3. Cuadro comparativo de valores reales que conforman la Matriz de Confusión

Resultados de la Matriz	Score	Predicción (score)	Confianza (Regular)	Confianza (Alto)
Verdadero Positivos	Regular	Regular	0,998	0,002
Verdaderos Negativos	Alto	Alto	0,004	0,996
Falso Positivo	Alto	Regular	0,913	0,087
Falso Negativo	Regular	Alto	0,266	0,734

En la Figura 6, la precisión en la predicción de valores fue la más alta de los algoritmos estudiados, brindando un 99,46%. Los “*Falsos Positivos*” (3 casos) son los que se predijeron como “Regulares” y debieron ser “Alto”, debido a esto la predicción “Regular” tiene una Precisión del 99,41%. Los casos actuales clasificados en “Alto” demuestran un Recall del 94%, por lo tanto, se observa que una gran proporción de las clases positivas predichas, siendo el 100% para la clasificación “Regular”.

accuracy: 99.46% +/- 0.87% (micro average: 99.46%)

	true Regular	true Alto	class precision
pred. Regular	504	3	99.41%
pred. Alto	0	47	100.00%
class recall	100.00%	94.00%	

Figura 6. Matriz de confusión de Random Forest.

En la Figura 7, se analiza que la precisión baja en un pequeño porcentaje a 98,74% en el cual se evalúan menores estimaciones tanto en Precisión como en Recall. Los casos “*Falsos Positivos*”, en este son 4 (cuatro) de los que fueron predichos como “Regular” y fueron incorrectamente categorizados, similar caso para los estimados como “Altos” y debieron ser “Regular”.

accuracy: 98.74% +/- 1.21% (micro average: 98.74%)

	true Regular	true Alto	class precision
pred. Regular	501	4	99.21%
pred. Alto	3	46	93.88%
class recall	99.40%	92.00%	

Figura 7. Resultados para la Matriz de confusión de Decision Tree.

El tercer caso, se muestra en la Figura 8, allí se observa que la precisión del 90,97% es la más baja comparada con los anteriores casos. Esto se debe a que no predice ningún caso como “Alto”, lo que indica que si tomamos este algoritmo para predecir tendríamos que mejorar la muestra, las variables consideradas, la estimación del “LABEL” o algún otro factor que pueda ser relevante.

accuracy: 90.97% +/- 0.08% (micro average: 90.97%)

	true Regular	true Alto	class precision
pred. Regular	504	50	90.97%
pred. Alto	0	0	0.00%
class recall	100.00%	0.00%	

Figura 8. Resultados de la Matriz de confusión de Gradient Boosted Trees.

Como comparativa final se muestran los indicadores para todos los algoritmos estudiados en este trabajo, a continuación, se muestra esto en la Tabla 4.

Tabla 4. Cuadro comparativo de algoritmos utilizados en cada método

Algoritmos	Accuracy	Clase positiva: Alto		
		Precision	Recall	AUC
Random Forest	99,46%	100%	94%	100%
Decision Tree	98,74%	93,88%	92%	92,40%
Gradient Boosted Trees	90,97%	0%	0%	50%
K-NN	93%	64,29%	72%	95,30%
Naive Bayes	84,63%	36,84%	98%	97,20%

Como conclusión sobre el análisis de los resultados arrojados por los algoritmos, a lo largo del prototipo estudiado, los primeros algoritmos, Random Forest y Decision Tree, son los mejores de los 5 comparados de acuerdo con los indicadores evaluados.

En la Tabla 4, para los casos K-NN y Naive Bayes, en los cuales la precisión (Accuracy) a pesar de ser alta, podemos observar que el indicador “Precisión” demuestra menor cantidad de casos predichos. En K-NN, el indicador “Recall” tiene una baja importante de clasificación de casos, lo que nos indica que no es capaz de estimar la clasificación de tipo “Alta”. También se evalúa que al disminuir el área AUC, se capta menor cantidad de valores positivos en los resultados del modelo. Se refleja la disminución del área en un 50% para Gradient Boosted Trees debido a que no es capaz de predecir “Falsos Negativos” y “Verdaderos Negativos”.

Es importante mencionar que, para este estudio, el puntaje correspondiente, analiza la recomendación de películas y programas, exceptuando las características individuales de los usuarios. No obstante, se examinó el rating generado por las vistas y la interacción de los usuarios con su voto.

Conclusiones

Machine learning es un enfoque que se encuentra en pleno auge y que cada vez son más los portales de datos abiertos que comparten subconjuntos de datos y disponibilidad a distintas personas su contenido de forma tal que puedan ser utilizados analizados en aspecto de mejoras cambios preferencias y nuevos enfoques que con la utilización de distintas herramientas tecnológicas permitan obtener un valor agregado a la comunidad.

Desde la perspectiva de las aplicaciones multimedia, este trabajo permitió conocer el contexto de los distintos tipos de aplicaciones como así también el entorno en el que se encuentra desde la perspectiva de machine learning. Es por ello por lo que se realizó una propuesta de forma tal que a través de la utilización de un software se generaron comparativas con los algoritmos más utilizados, según los trabajos relevados, en aplicaciones multimedia.

Existe una gran variedad de aplicaciones para herramientas de aprendizaje automático en la nube que pueden implementarse en los videos y que están en constante desarrollo y mejora. Es por ello por lo que, se consideran que los estudios realizados podrían incluir variedad de datos que no hemos utilizado en este primer análisis. Al seleccionar algoritmos utilizados en estudios previos, se observó que las comparativas finales, con un conjunto de datos de casi mil registros, funcionan de forma correcta.

Como objetivo futuro de este trabajo de investigación, se plantea la necesidad de estudiar una mayor muestra en cantidad de registros, con el fin de realizar pruebas aún más extensas con los mismos y/o nuevos algoritmos.

Cabe aclarar que este estudio, sólo se basó en algoritmos Supervisados y se enfocaron en la Clasificación desde el concepto de machine learning, por lo que, como líneas futuras, sería importante ampliar el radio de investigación con diversas pruebas, por ejemplo,

utilizando Algoritmos No Supervisados y estimar Clusters con el objetivo de encontrar preferencias más precisas sobre los usuarios y recomendar a grupos más fragmentados con características similares. Finalmente, de esta forma, se podría lograr reconocer preferencias personalizadas para usuarios suscriptos o no a las plataformas de transmisión de video.

Referencias

- [1] Martínez, R. (2022). Métricas de calidad para validar los conjuntos de datos abiertos públicos gubernamentales (Doctoral dissertation, Universidad Nacional de La Plata).
- [2] Magallón Rosa, R. (2017). Datos abiertos y acceso a la información pública en la reconstrucción de la historia digital. Disponible en: <https://revistas.ucm.es/index.php/HICS/article/view/57845>
- [3] Belloch, C. (2012). Aplicaciones multimedia. Comunicación interactiva. Disponible en: <https://repository.unad.edu.co/bitstream/handle/10596/5119/2/08062.pdf?sequence=1>
- [4] YouTube (2022). “YouTube”. Disponible en: <https://www.youtube.com/>
- [5] Vimeo (2022). “Host virtual events with Vimeo”. Disponible en: <https://vimeo.com/>
- [6] Twitch (2022). “Twitch.tv - Official Site”. Disponible en: <https://www.twitch.tv/>
- [7] Netflix (2022). “Películas y series ilimitadas y mucho más”. Disponible en: <https://www.netflix.com/ar/>
- [8] Hulu (2022). “Hulu: Stream TV and Movies Live and Online”. Disponible en: <https://www.hulu.com/welcome>
- [9] Prime Video. “Ver en Amazon.com - Prime Video”. Disponible en: https://www.primevideo.com/ref=atv_hm_offers_c_9zZ8D2_1_1?language=es_ES
- [10] Lydia Bird (2022). Imagen. “What are SVOD, TVOD, AVOD?”. Disponible en: <https://imagen.io/blog/what-are-svod-tvod-avod/>
- [11] HBO (2022). “HBO: Home to Groundbreaking Series, Movies, Comedies”. Disponible en: <https://www.hbo.com/>
- [12] Disney+ (2022). “Tus historias favoritas, todas en un mismo lugar”. Disponible en: <https://www.disneyplus.com/es-ar>
- [13] iTunes (2022). “iTunes”. Disponible en: <https://www.apple.com/la/itunes/>
- [14] Sky (2022). “Sky Sports Box Office hub”. Disponible en: <https://www.sky.com/boxoffice>
- [15] Dailymotion (2022). “Dailymotion, donde los videos importan”. Disponible en: <https://www.dailymotion.com/ar>
- [16] 4OD (2022). “All 4 | The on-demand channel from 4”. Disponible en: <https://www.channel4.com/>
- [17] AWS (2022). “Datos abiertos en AWS”. Disponible en: <https://aws.amazon.com/es/opendata/?wwps-cards.sort-by=item.additionalFields.sortDate&wwps-cards.sort-order=desc>

- [18] AWS (2022). “*Aprendizaje automático para aplicaciones multimedia*”. Disponible en: <https://aws.amazon.com/es/media/tech/machine-learning-for-media-applications/>
- [19] Song, M. (2021). A Comparative Study on Over-The-Tops, Netflix & Amazon Prime Video: Based on the Success Factors of Innovation. *International journal of advanced smart convergence*, 10(1), 62-74.
- [20] Ko, H., Lee, S., Park, Y., & Choi, A. (2022). A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1), 141.
- [21] Calderón, C. A., Holgado, P. S., & González, K. E. O. (2019). Las plataformas de entretenimiento on-demand: detrás del Machine Learning de Netflix, HBO y Spotify. In *La comunicación en el escenario digital: Actualidad, retos y perspectivas* (pp. 645-669). Pearson Educación de Perú.
- [22] Kashyap, A., Sunita, B., Srivastava, S., Aishwarya, P. H., & Shah, A. J. (2020). A movie recommender system: MOVREC using machine learning techniques. *International Journal of Engineering Science and Computing*, 10(6).
- [23] Reddy, S. R. S., Nalluri, S., Kuniseti, S., Ashok, S., & Venkatesh, B. (2019). Content-based movie recommendation system using genre correlation. In *Smart Intelligent Computing and Applications* (pp. 391-397). Springer, Singapore.
- [24] García Camargo, L. J., & Guerrero Ramos, G. A. (2020). Sistema de recomendación de contenido de la plataforma On Demand de Directv.
- [25] Choi, J., & Kim, Y. (2021). A heterogeneous learning framework for over-the-top consumer analysis reflecting the actual market environment. *Applied Sciences*, 11(11), 4783.
- [26] Rojas, J. S., Rendon, A., & Corrales, J. C. (2019). Consumption behavior analysis of over the top services: Incremental learning or traditional methods?. *IEEE Access*, 7, 136581-136591.
- [27] Kristoffersen, M. S., Shepstone, S. E., & Tan, Z. H. (2019). The importance of context when recommending TV content: Dataset and algorithms. *IEEE Transactions on Multimedia*, 22(6), 1531-1541.
- [28] Rychalska, B., Basaj, D., Dąbrowski, J., & Daniluk, M. (2020). I know why you like this movie: Interpretable Efficient Multimodal Recommender. *arXiv preprint arXiv:2006.09979*.
- [29] Choi, J., & Kim, Y. (2020). Time-aware learning framework for over-the-top consumer classification based on machine- and deep-learning capabilities. *Applied Sciences*, 10(23), 8476.
- [30] Kaggle (2022). “*Disney Plus Movies and TV Shows*”. Disponible en: <https://www.kaggle.com/datasets/unanimad/disney-plus-shows>
- [31] Kaggle (2022). “*Kaggle: Your Machine Learning and Data Science Community*”. Disponible en: <https://www.kaggle.com/>
- [32] Jeon, C., Ha, J. S., & Kim, Y. J. R. Data Analysis on OTT Platforms: Which Service Should I Choose?.
- [33] RapidMiner Studio (2022). “*Best Data Science Platform for Your Enterprise – RapidMiner*”. Disponible en: <https://rapidminer.com/platform/>