

Tesis de Maestría

**Análisis Predictivo de soporte a la toma de decisiones en la industria del entretenimiento**

Lic. Marcelo Fabian Magarzo Velasco

**FACULTAD DE INGENIERÍA**

**Maestría en Tecnología de la Información**

Director de Tesis

Lic. Daniel Tokman

Buenos Aires, Argentina

Junio 2020

**RESUMEN DE LA TESIS**

Aún no se pueden capturar datos procedentes del futuro, pero ya es posible predecir lo que sucederá en base a datos del pasado; este proceso se denomina análisis predictivo. A diferencia de las prácticas tradicionales de inteligencia de negocios, que presentan una naturaleza más retrospectiva, el enfoque del análisis predictivo se centra en ayudar a las empresas a desarrollar inteligencia procesable sobre la base de los datos históricos.

El presente documento busca mediante un modelo predictivo analizar los datos actuales e históricos para predecir el éxito o fracaso, en función a distintas variables y aristas, de una nueva producción en el mundo de la industria del entretenimiento.

De esta manera, los ejecutivos en el negocio podrán tomar una decisión acertada sobre cómo publicitar, cómo distribuir e incluso no llegar a producir o estrenar en caso de no obtener una predicción favorable o deseada.

**TABLA DE CONTENIDOS**

[1. Introducción 7](#_Toc37432664)

[2. Enfoque teórico 11](#_Toc37432665)

[3. Impacto del análisis predictivo para la industria del entretenimiento 20](#_Toc37432666)

[4. Técnicas predictivas 31](#_Toc37432667)

[5. Propuesta de modelo predictivo 44](#_Toc37432668)

[6. Experimentos 76](#_Toc37432669)

[7. Resultados 79](#_Toc37432670)

[8. Perspectivas de crecimiento 86](#_Toc37432671)

[9. Conclusiones 87](#_Toc37432672)

[10. Bibliografía 89](#_Toc37432673)

**LISTA DE FIGURAS**

[Figura 1: Proceso del análisis predictivo 12](#_Toc42802421)

[Figura 2: Proceso de descubrimiento del conocimiento (KDD) 14](#_Toc42802422)

[Figura 3: Principales tipos de aprendizaje en la minería de datos 15](#_Toc42802423)

[Figura 4: Principales usos del análisis predictivo en las industrias 18](#_Toc42802424)

[Figura 5: Áreas de impacto del análisis predictivo en la industria del entretenimiento 20](#_Toc42802425)

[Figura 6: Categorías e industrias del entretenimiento que utilizan análisis predictivo 22](#_Toc42802426)

[Figura 7: Ingresos de Netflix, T1 2011 a T4 2019, millones de USD [25] 26](#_Toc42802427)

[Figura 8: Árbol de decisión simple 33](#_Toc42802428)

[Figura 9: Clasificaciones correctas / instancias que alcanzan cada nodo 35](#_Toc42802429)

[Figura 10: Simple Perceptrón Multi-Capa 37](#_Toc42802430)

[Figura 11: Cruce del Algoritmo Genético 41](#_Toc42802431)

[Figura 12: Cruce de la Programación Genética 42](#_Toc42802432)

[Figura 13: Proceso de recopilación de datos de las categorías identificadas 45](#_Toc42802433)

[Figura 14: Proceso de análisis de datos y selección de variables 48](#_Toc42802434)

[Figura 15: Proceso de limpieza y tratamiento de datos con Tableau 49](#_Toc42802435)

[Figura 16: Limpieza del CDE en Tableau Prep 51](#_Toc42802436)

[Figura 17: Unión del CDE en Tableau Prep 57](#_Toc42802437)

[Figura 18: Agregación del CDE en Tableau Prep 59](#_Toc42802438)

[Figura 19: Organización del CDE en Tableau Prep 61](#_Toc42802439)

[Figura 20: Unión final del CDE en Tableau Prep 62](#_Toc42802440)

[Figura 21: Salida resultante del CDE en Tableau Prep 63](#_Toc42802441)

[Figura 22: Limpieza y organización del CDV en Tableau Prep 64](#_Toc42802442)

[Figura 23: Proceso general de modelado predictivo 67](#_Toc42802443)

[Figura 24: Modelo predictivo - Regresión Lineal 71](#_Toc42802444)

[Figura 25: Modelo predictivo - Regresión de Árbol de Decisión Potenciado 72](#_Toc42802445)

[Figura 26: Modelo predictivo - Regresión de Redes Neuronales 73](#_Toc42802446)

[Figura 27: Experimento predictivo – Regresión de Árbol de Decisión Potenciado con CDV 76](#_Toc42802447)

[Figura 28: Predicción de Puntaje en el CDV 77](#_Toc42802448)

[Figura 29: Predicción de Puntaje en el CDV filtrado por país 78](#_Toc42802449)

[Figura 30: Multibox Plot - Relación entre variables Nombre y Puntaje 79](#_Toc42802450)

[Figura 31: Multibox Plot - Relación entre variables Categoría y Puntaje 80](#_Toc42802451)

[Figura 32: Multibox Plot - Relación entre variables Fecha Lanzamiento y Puntaje 81](#_Toc42802452)

[Figura 33: Multibox Plot - Relación entre variables Género y Puntaje 81](#_Toc42802453)

[Figura 34: Multibox Plot - Relación entre variables País y Puntaje 82](#_Toc42802454)

[Figura 35: Gráfico de dispersión - Relación entre variables Ingresos y Puntaje 83](#_Toc42802455)

[Figura 36: Multibox Plot - Relación entre variables Canal y Puntaje 84](#_Toc42802456)

[Figura 37: Modelo predictivo de soporte a la toma de decisiones en la industria del entretenimiento 84](#_Toc42802457)

[Figura 38: Prueba de web service - Escenario de predicción N°1 85](#_Toc42802458)

[Figura 39: Prueba de web service - Escenario de predicción N°2 85](#_Toc42802459)

**LISTA DE TABLAS**

[Tabla 1: Precios y características de Tableau 49](#_Toc42802406)

[Tabla 2: Proceso de limpieza del CDE 55](#_Toc42802407)

[Tabla 3: Proceso de limpieza y organización del CDV 65](#_Toc42802408)

[Tabla 4: Precios y características de Machine Learning Studio 66](#_Toc42802409)

[Tabla 5: Comparativa de estimación de rendimiento de los distintos modelos predictivos 75](#_Toc42802410)

**LISTA DE FÓRMULAS**

[Fórmula 1: Regresión lineal 31](#_Toc42802416)

[Fórmula 2: Pureza ganada para separación del AD 33](#_Toc42802417)

[Fórmula 3: Estimación de probabilidad Laplace 35](#_Toc42802418)

[Fórmula 4: Función logística unipolar 38](#_Toc42802419)

[Fórmula 5: Función sigmoidea bipolar 38](#_Toc42802420)

# INTRODUCCIÓN

Los críticos tradicionales de Hollywood, como Scott Feinberg de The Hollywood Reporter, confiaban en que sus años de experiencia y relaciones dentro de la industria se traducirían en mejores perspectivas. “Sé que no hay una manera científica de predecir los Oscar”, dijo Feinberg al Wall Street Journal. Él cree que la ciencia de la información no puede sustituir la “inteligencia clandestina” de los 6.000 votantes de la Academia, a quienes señala que tienen preferencias caprichosas y subjetivas. Sin embargo, desde las elecciones de los votantes hasta la consultoría corporativa, los científicos de los datos aprecian el potencial del análisis predictivo para mejorar los pronósticos y ofrecer información inteligente sobre industrias y empresas. Los Oscar fueron otro símbolo de las aplicaciones de la ciencia de datos para la industria de entretenimiento y medios de comunicación. La toma de decisiones corporativas en todas las industrias puede beneficiarse de la ciencia de datos. [[1](#Gol13)]

Está en la naturaleza humana querer saber y predecir qué deparará el futuro. El análisis predictivo trata la predicción de futuros eventos basados en datos históricos previamente observados mediante la aplicación de métodos sofisticados como el aprendizaje automático. Los datos históricos se recopilan y transforman utilizando diversas técnicas como el filtrado, correlación de datos, etc. [[2](#Mis12)]

Para comprender los puntos básicos del análisis predictivo no es necesario ser un científico de datos, simplemente hay que conocer las tres claves de toda predicción:

* **Los datos:** para poder hacer pronósticos sobre el comportamiento de los clientes, los atributos que debería tener un producto/ servicio o las estrategias que llevarán a cabo empresas de la competencia es necesario contar con información fiable y consistente. La calidad de los datos es incluso más importante que su volumen, en su aplicación a la analítica predictiva. No obstante, cuanto mayor es el volumen de información disponible, mayor profundidad se alcanzará en la interpretación.
* **Las estadísticas:** los modelos predictivos se estructuran en torno a las diferentes variables y correlaciones. Aplicando algoritmos y ecuaciones los analistas pueden obtener resultados expresados en términos de probabilidad, que explican lo que sucederá en el futuro, con un margen de error mínimo.
* **Los supuestos:** la actualización es esencial y determina la precisión de los modelos. Los supuestos pueden cambiar con el tiempo, a medida que lo hacen los hábitos de los clientes, las circunstancias de mercado o incluso los propios procesos internos de la empresa. Si no se realiza un seguimiento adecuado que compruebe que estos supuestos no se encuentran defectuosos u obsoletos, el modelo predictivo no será válido y la toma de decisiones se verá afectada. [[3](#LOG)]

Además de entender los tres pilares de la analítica predictiva, se debe recopilar información acerca de:

* Las fuentes de los datos empleados para el análisis.
* La relevancia de las métricas extraídas del estudio de los datos.
* La existencia de valores atípicos en los conjuntos de datos y la medida en que pueden afectar a los resultados.
* Los supuestos detrás del análisis y las condiciones que pueden afectarlos.

Sin embargo, hay que tener en cuenta el dinamismo actual y asumir que, tan pronto como un modelo es creado, se empieza a degradar ya sea por la entrada de nuevos competidores en el mercado, cambios en los precios o fluctuaciones de divisas, entre otros. Muchos de estos factores escapan al control de la organización que sólo puede responder de una forma: monitorizando el rendimiento del modelo y teniendo preparado un plan para actualizarlo o reemplazarlo, en función de las necesidades. [[4](#Dys14)]

La forma en que los modelos predictivos producen valor en una organización es simple en concepto; hacen posible tomar decisiones correctas, rápidamente y con menos gasto. Pueden brindar apoyo para decisiones humanas, haciéndolas más eficientes y efectivas, o en algunos casos, pueden ser utilizados para automatizar todo un proceso de toma de decisiones.

El objetivo general de la presente tesis es diseñar un modelo de análisis predictivo para la toma acertada de decisión de una nueva producción, el canal de distribución y la ventana de distribución para cada uno de los canales definidos.

Como objetivos específicos se presentan los siguientes:

* Recolectar e identificar posibles variables de análisis con datos actuales e históricos de medios de entretenimiento en distintas fuentes de datos.
* Recolectar datos de las páginas y/o grupos de medios de entretenimiento en las redes sociales para realizar un Sentiment Analysis e identificar las variables relevantes a tomar en cuenta para el modelo propuesto.
* Diseñar un modelo de análisis utilizando técnicas predictivas y aprendizaje automático para obtener resultados que den soporte a la toma de decisiones en función a distintas variables.
* Validar el modelo predictivo propuesto analizando un caso de estudio con datos de la industria cinematográfica en Argentina.

El negocio del entretenimiento no es un entretenimiento. Se trata de negocios, que abarca desde finanzas, mercadeo hasta distribución, todo conducido a escala global. El éxito en este entorno empresarial cada vez más complejo, a menudo requiere más que grandes artistas y contenido deseable. Requiere de ciencia: la ciencia del análisis predictivo. [[5](#Del)]

# ENFOQUE TEÓRICO

El análisis predictivo es el uso de datos, algoritmos estadísticos y técnicas de aprendizaje automático para identificar la probabilidad de resultados futuros basados ​​en datos históricos. Los modelos predictivos utilizan la información conocida para desarrollar un modelo que se puede utilizar para predecir valores. El modelado da como resultado predicciones que representan valores de probabilidad para datos nuevos o diferentes.

El análisis predictivo es utilizado para predecir tendencias, mejorar el rendimiento, impulsar la toma de decisiones y predecir comportamientos. [[6](#SBa17)]

El proceso de extraer conocimiento a partir de grandes volúmenes de datos ha sido reconocido por muchos investigadores como un tópico de investigación clave en los sistemas de bases de datos, y por muchas compañías como una importante área y una oportunidad para obtener mayores ganancias. [[7](#TIM13)] Fayyad, Piatetsky-Shapiro & Smyth lo definen como el proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos. [[8](#Usa96)]

El análisis predictivo consiste en una variedad de técnicas estadísticas de modelado, aprendizaje automático y minería de datos que analiza hechos actuales e históricos para predecir futuros eventos. Los modelos predictivos explotan patrones encontrados en datos transaccionales para identificar riesgos y oportunidades. Los modelos capturan relaciones a través de muchos hechos que permiten la evaluación del riesgo o potencial asociado a un conjunto particular de condiciones. La secuencia de pasos que involucra el análisis predictivo se detalla en la figura 1.



Figura 1: Proceso del análisis predictivo

* **Captura de requerimientos:** Para desarrollar un modelo predictivo, los objetivos de la predicción deben estar claros. Se deben obtener los requerimientos del usuario/cliente y cómo se verá beneficiado con las predicciones obtenidas.
* **Recopilación de datos:** Se debe recopilar el conjunto de datos necesarios para desarrollar el modelo predictivo. Dichos datos pueden pertenecer a diferentes fuentes y estar de forma estructurada y no estructurada.
* **Análisis y limpieza de datos:** Se procede a analizar los datos obtenidos. Los datos no estructurados se convierten en una forma estructurada. Una vez que los datos completos están estructuradamente disponibles, se prueba su calidad. Hay posibilidades de que existan datos erróneos en el conjunto de datos principal o que falten valores en relación con los atributos, todos estos casos deben corregirse. La efectividad del modelo predictivo depende totalmente de la calidad de los datos.
* **Selección de técnica predictiva:** Todos los modelos de análisis predictivo se basan en técnicas estadísticas y/o de aprendizaje automático. Se debe seleccionar aquella técnica que se considere apropiada para el objetivo de la predicción.
* **Modelado predictivo:** Se desarrolla un modelo predictivo en base a la técnica seleccionada. Después del desarrollo, se prueba en el conjunto de datos de entrenamiento que forma parte del conjunto de datos recopilado principal para verificar la validez del modelo. Una vez finalizada la validación, el modelo puede hacer predicciones precisas sobre los nuevos datos ingresados ​​como entrada al sistema.
* **Estimación y rendimiento de predicciones:** Se monitorea constantemente el rendimiento del modelo para asegurar que los resultados sean correctos y las predicciones sean precisas. [[9](#Vai)]

Un concepto importante dentro del análisis predictivo es la minería de datos. Los términos minería de datos y extracción de datos a menudo se confunden el uno con el otro; pero hay una diferencia significativa. La extracción de datos implica obtener datos de una fuente y cargarlos en una base de datos específica o data warehouse. Por otro lado, la minería de datos es la extracción de información predictiva oculta en grandes bases de datos o almacenes de datos.

El término minería de datos predictivos se aplica generalmente para identificar proyectos de minería de datos con el objetivo de identificar un modelo o un conjunto de modelos que se pueden usar para predecir algunas respuestas de interés. [[10](#Deb)]

En la figura 2, se observa el proceso de Descubrimiento del Conocimiento (KDD) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan significado a los patrones encontrados. [[11](#Vio04)]

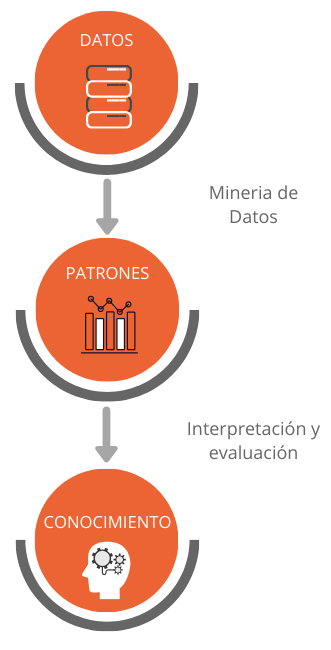


Figura 2: Proceso de descubrimiento del conocimiento (KDD)

Las herramientas y técnicas de minería de datos construyen un modelo de análisis predictivo. El primer paso consiste en extraer los datos por acceso a bases de datos masivas. Los datos son procesados con la ayuda de algoritmos avanzados para encontrar patrones ocultos e información predictiva.

El principal objetivo de la minería de datos es construir un modelo que se puede usar para predecir la ocurrencia de un evento. Los constructores del modelo extraerán el conocimiento de los datos históricos y se representarán de tal forma que el modelo resultante pueda ser aplicado a nuevas situaciones. El proceso de análisis de datos extrae información útil sobre la que se aplica una o más técnicas de minería de datos para descubrir previamente patrones desconocidos dentro de los datos, o encontrar tendencias en los datos que pueden luego ser utilizados para predecir comportamientos futuros.

La minería de datos se puede dividir en dos tipos de aprendizaje: supervisado (predictivo) y no supervisado (descriptivo). En el aprendizaje supervisado, los datos se modelan a partir de “datos entrenados” para encontrar patrones que se pueden utilizar para predecir un valor al ingresar ciertos parámetros. El aprendizaje supervisado es el proceso de crear modelos predictivos usando datos históricos que contienen los resultados que se tratan de predecir. Los tipos de datos determinan si este modelo se logrará utilizando un algoritmo de Regresión o Clasificación. El aprendizaje no supervisado se refiere al problema de tratar de encontrar una estructura oculta en datos sin etiquetar. A diferencia del aprendizaje supervisado, los algoritmos que se utilizan para el aprendizaje no supervisado no “aprenden” de datos históricos con valores conocidos. Las técnicas estándares no supervisadas incluyen agrupaciones, caracterización, minería de reglas de asociación y técnicas de cambios y desviaciones (anomalías). [[12](#Pra15)]



Figura 3: Principales tipos de aprendizaje en la minería de datos

Los modelos de decisión describen la relación entre todos los elementos de una decisión: los datos conocidos (incluidos los resultados de modelos predictivos), la decisión y los pronósticos de los resultados de decisión, con el fin de predecir los resultados de las decisiones que involucran muchas variables. Estos modelos pueden ser utilizados en optimización, maximizando ciertos resultados mientras otros son minimizados.

El núcleo del análisis predictivo se basa en la captura de relaciones entre las variables explicativas y las variables pronosticadas de ocurrencias pasadas, siendo así explotadas para predecir futuros resultados. Sin embargo, es importante tener en cuenta, que la precisión y la facilidad de uso de los resultados dependerán en gran medida del nivel de análisis de datos y la calidad de los supuestos. [[13](#Mus)]

Debido al auge e influencia de las redes sociales en los últimos años, ha surgido una nueva arista dentro de la minería de datos, conocida como minería de opinión que tiene como objetivo brindar la información y herramientas necesarias para realizar un análisis de sentimiento. Todos los días, los usuarios y clientes participan en conversaciones en línea sobre productos y servicios en redes sociales como Facebook, Twitter, Instagram y ofrecen comentarios valiosos para las organizaciones. [[14](#Sit10)]

El análisis de las redes sociales ofrece a las industrias la capacidad de identificar patrones en la confianza del usuario y medir su efectividad de marketing. Por muchas razones es esencial identificar la información subjetiva del contenido existente en el mundo digital. En primer lugar, permite interactuar de forma rápida y simultánea con usuarios, clientes, empleados o público en general, especialmente con las generaciones más jóvenes. En segundo lugar, representa una forma de bajo costo con la que las organizaciones pueden construir su marca y comunicarse con diversos interesados dentro y fuera de la misma. En tercer lugar, brinda la oportunidad de aprender de la información en función de la connotación positiva o negativa del lenguaje ocupado en la misma.

Cada día, los usuarios generan miles de mensajes y ofrecen información valiosa en foros públicos sin costo alguno. Los profesionales de marketing reciben toda la información y los detalles que necesitan para que puedan tener una imagen completa de los hábitos de consumo, las opiniones y las tendencias. Desde el momento en que estos datos se combinan con una estrategia específica y métricas de rendimiento, se puede obtener un amplio conocimiento que puede ayudar a tomar decisiones de negocios. [[15](#Kra08)]

La recopilación de datos masivos sobre el comportamiento social es una oportunidad única para observar y estudiar los fenómenos sociales. Las consultas de los motores de búsqueda o las publicaciones en redes sociales se han utilizado para pronosticar elecciones, tendencias, comportamiento del mercado, etc. Un claro ejemplo dentro de la industria del entretenimiento, son los programas de televisión, donde los datos de las plataformas sociales se están convirtiendo cada vez más en una parte importante del programa. Este hecho amplifica la importancia de los indicadores que pueden extraerse de estos medios para monitorear varias cosas. Además, el mayor uso de teléfonos inteligentes y dispositivos móviles crea información geolocalizada sobre la actividad de redes sociales que se explora. La información geográfica es un ingrediente clave para lograr poder predictivo. El entorno empresarial de hoy exige la existencia de la capacidad de predecir con precisión el resultado de eventos futuros. Por ejemplo, las predicciones de ventas y el comportamiento futuro de los consumidores finales pueden ayudar realmente a las industrias a reconsiderar su planificación de producción y la investigación. Lo primero que se debe hacer es identificar las fuentes de información de las distintas redes sociales, que incluyen opiniones relevantes para el conjunto de eventos y extraer la información adecuada. En un segundo paso, se aplica un método para procesar contenido, por ejemplo, análisis de sentimiento para los comentarios y opiniones. Los encargados de marketing evalúan el contenido que se genera todos los días y actualizan en consecuencia las acciones. Luego envían ofertas de compra o venta de productos relacionados con un indicador específico u ofertas para los productos o servicios que consideran más exitosos. [[16](#Tho16)]

Hoy en día la importancia del análisis predictivo no es un misterio debido a que muchas empresas y organizaciones están recurriendo a la ciencia de datos y análisis de negocio para ayudar a resolver problemas, dar soporte a la toma de decisiones y descubrir nuevas oportunidades. En la figura 4, se muestran los usos más importantes en las distintas industrias:



Figura 4: Principales usos del análisis predictivo en las industrias

* **Mejora de operaciones:** Muchas industrias utilizan modelos predictivos para pronosticar el inventario y administrar los recursos. Por ejemplo en los hoteles se intenta predecir el número de huéspedes para una noche determinada para maximizar la ocupación y aumentar los ingresos. El análisis predictivo permite a las organizaciones funcionar de manera más eficiente.
* **Reducción de riesgo:** Los datos de los que aprenden los modelos predictivos incluyen los ejemplos negativos y también los positivos, tanto los éxitos como los “errores” inevitables del día a día de una industria. Las evaluaciones de crédito se utilizan para valorar la probabilidad de un cliente de realizar compras predeterminadas y son un ejemplo muy conocido de analítica predictiva.
* **Optimización de marketing:** El análisis predictivo se utiliza para determinar respuestas o compras de clientes, además de promocionar oportunidades de ventas cruzadas. Los modelos predictivos ayudan a las industrias a atraer, retener e incrementar el número de sus clientes más rentables.
* **Detección de fraude:** La combinación de múltiples métodos analíticos puede mejorar la detección de patrones y prevenir la conducta criminal. A medida que la ciberseguridad se convierte en una preocupación cada vez mayor, la analítica conductual de alto desempeño examina todas las acciones en una red en tiempo real para detectar anormalidades que puedan indicar fraude, vulnerabilidades desde el día cero y amenazas persistentes. [[17](#SAS)]

# IMPACTO DEL ANÁLISIS PREDICTIVO PARA LA INDUSTRIA DEL ENTRETENIMIENTO

Mediante la implementación de una solución de análisis predictivo es posible obtener un conocimiento de gran valor que, de otro modo, se hubiese perdido o ni siquiera hubiera existido. [[18](#Way07)]

Las decisiones empresariales en el entretenimiento se han guiado tradicionalmente por experiencia e intuición. Pero en el mundo complejo de hoy, los desafíos son demasiado complicados, y lo que está en juego es demasiado alto, para confiar únicamente en corazonadas y sentimientos viscerales. Mientras que muchas empresas de medios y entretenimiento recopilan datos en la sociedad, pocas tienen la capacidad de comprender cómo los consumidores toman decisiones, cómo identificar de manera proactiva nuevas oportunidades de mercado y cómo desarrollar estrategias de acción para capitalizar las oportunidades.

A medida que proliferan las redes sociales, el desarrollo de productos tradicionales y el marketing los enfoques deben evolucionar. El impacto que se obtiene al aplicar el análisis predictivo en la industria del entretenimiento se puede categorizar en 6 principales áreas:

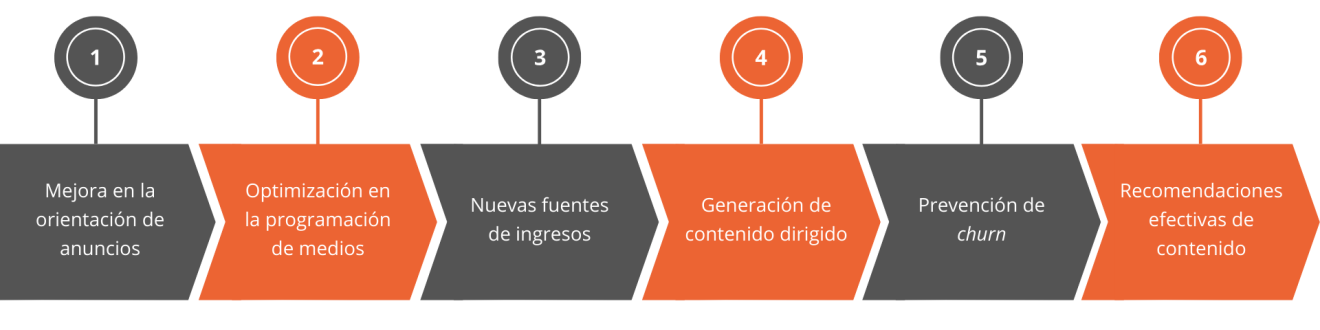


Figura 5: Áreas de impacto del análisis predictivo en la industria del entretenimiento

* **Mejora en la orientación de anuncios**: aprovechar el poder del análisis predictivo significa una orientación fácil y rápida para que las personas adecuadas vean exactamente los anuncios correctos en los que es más probable que se interesen, lo que se traduce en un mayor retorno de la inversión (ROI).
* **Optimización en la programación de medios**: para una programación optimizada, incluso las fuentes de datos externas inesperadas pueden ser útiles. Como el clima, por ejemplo, una empresa puede ajustar las transmisiones de programación de medios en función de una audiencia más cautiva en un día lluvioso.
* **Nuevas fuentes de ingresos**: en el mercado actual, ser un activo valioso es poder identificar fuentes de ingresos adicionales e innovadoras, muy aparte de la publicidad y las asociaciones tradicionales.
* **Generación de contenido dirigido**: usar la ciencia de los datos y el aprendizaje automático para dictar qué mostrar a los usuarios, incluso si no saben que es lo que quieren, es una forma revolucionaria de impulsar la estrategia.
* **Prevención de *churn***: saber qué clientes tienen intención de abandonar y ser capaces de dirigirse específicamente a los que probablemente regresarían con ofertas y marketing personalizado. Esto es fundamental para el éxito.
* **Recomendaciones efectivas de contenido**: por supuesto, no hay mejor manera de aumentar la participación de los usuarios que proporcionar recomendaciones de alta calidad. Pero un buen motor de recomendación es más fácil de decir que de hacer y también implica usar diferentes tipos de datos y aplicar diferentes estrategias dependiendo del tipo de contenido y la composición de la base de usuarios. [[19](#Dan17)]

En un futuro no muy lejano, la ciencia de los datos será adaptable e integral para manifestaciones críticas en el sector de medios y entretenimiento. Para cine, música, medios digitales y televisión, la composición de big data y análisis predictivo ofrecerán implementaciones que generen mayor valor agregado en la industria.

En la figura 6, se detallan las categorías y algunas industrias del entretenimiento que utilizan el análisis predictivo como herramienta de transformación:



Figura 6: Categorías e industrias del entretenimiento que utilizan análisis predictivo

1. **Música**

El cambio a un modelo de negocio basado en datos frente a los servicios de suscripción de música representa una de las mayores transformaciones de la industria que la música ha presenciado en muchos años, volcando por completo el modelo de ingresos de las compañías de música. El crecimiento de los servicios de streaming como Spotify y Apple ha acelerado fuertemente ese cambio. La mayoría de los ejecutivos ahora entienden el valor de analizar el comportamiento del usuario y adaptar estrategias basadas en esas métricas. [[20](#DAT20)]

**Caso Spotify:** En 2008, las principales discográficas accedieron a participar en un experimento: primero EMI y Sony, seguidas por Warner Music y Universal, acordaron ceder su enorme catálogo sonoro a una pequeña *startup* sueca. A sus 25 años, el cofundador y actual CEO de la empresa escandinava, Daniel Ek, planeaba revolucionar la industria musical y borrar la sombra de la piratería con un servicio de música digital sin precedentes. Su nombre es una combinación de las palabras inglesas *Spot* e *Identify* porque «Spotify te ayuda a encontrar e identificar tus canciones favoritas, tanto las que has olvidado como las que todavía no conoces», anunció la compañía durante su lanzamiento. Hace una década que Ek y su colega Martin Lorentzon fundaron la empresa Spotify AB en Estocolmo. El servicio de música en streaming ha saludado en este tiempo a 75 millones de usuarios y ha conseguido atraer a 20 millones de suscriptores.

La herramienta ha madurado desde la primeriza versión beta, lanzada en 2007. A finales de aquel año, la novedad era compartir listas de canciones. Hoy, gracias a una maraña de algoritmos cada vez más complejos, no solo ofrece recomendaciones personalizadas, sino que es capaz de descubrir artistas emergentes y predecir los hits que encabezarán las listas de éxitos. La base de usuarios de Spotify solo constituye una pequeña parte de todas las personas que escuchan música globalmente, pero es suficientemente grande para que sus recomendaciones tengan un efecto medible, sostiene Sander Dieleman, un ingeniero que ha desarrollado modelos de recomendación musical basados en *deep learning* para la empresa sueca.

Los investigadores de Spotify consiguen prever los ganadores de los premios Grammy basándose en el big data proporcionado por sus usuarios (aproximadamente 1,5 terabytes diarios de información): hábitos, suscripciones a listas o a un artista, popularidad de los grupos, canciones y álbumes. Su objetivo es comprobar si los galardonados reflejan las rutinas y gustos musicales exhibidos en la plataforma. En el año 2015, sus herramientas de análisis pronosticaron los ganadores de las cuatro categorías más importantes. Sam Smith, vencedor en las secciones de mejores canción y álbum del año, aparecía en segundo lugar en los rankings de Spotify, aunque lo señalaron acertadamente como mejor artista revelación. [[21](#Luc16)]

1. **Cine y televisión**

En comparación con la industria de la música, las compañías de cine y televisión aún no han pasado por su gran transición. Algo de esto ya está sucediendo con los servicios de streaming y suscripción, pero la mayoría de los estudios de cine todavía operan principalmente en el modelo tradicional. Los servicios de streaming están a la vanguardia de la revolución de la inteligencia artificial y la ciencia de datos. Las compañías de producción, incluidas Amazon, Hulu y Netflix, analizan patrones en big data para determinar los tipos de contenido que crean y hacer recomendaciones de visualización personalizadas. De esta manera, la ciencia de datos puede ayudar al arte de producir y comercializar entretenimiento a niveles nunca antes vistos. [[20](#DAT20)]

**Caso Netflix:** Los comienzos de Netflix no están en Internet, ya que inició sus actividades en 1997, desde California, enviando DVD’s físicos por correo común. La idea era competir con los videoclubes con un servicio que nunca cobrara multa por no devolver a tiempo las películas alquiladas. Netflix, en cambio, permitía quedarse con los DVD’s tanto tiempo como uno quisiera, a cambio de pagar 15,95 dólares al mes por suscripción. [[22](#Lau16)] Desde entonces, Netflix se convirtió en el servicio de entretenimiento por streaming líder en el mundo. Presente en más de 190 países, sus 183 millones de suscriptores de pago pueden ver series, documentales y largometrajes de una amplia variedad de géneros y en diversos idiomas. Los suscriptores de Netflix pueden ver las series y películas que quieran, cuando quieran, donde quieran, y en cualquier pantalla conectada a Internet.

En 2011, Netflix tomó una decisión muy importante basándose en la analítica digital: compró, por delante de AMC o HBO, los derechos para la versión estadounidense de *House of Cards*. El precio de esta operación fue de más de 100 millones de dólares. Netflix estaba seguro del éxito de *House of Cards* incluso antes de la emisión del primer episodio ya que la decisión de producir la serie, así como cada momento y detalle de la propia grabación y del guion, estuvo basada en los resultados del análisis previo de los gustos de los usuarios:

* La mayoría habían visto *Social Network*, dirigida por David Fincher.
* La versión británica de *House of Cards* había sido bien valorada.
* Quienes habían visto la versión británica de la serie también habían visto películas protagonizadas por Kevin Spacey y/o dirigidas por Fincher.
* La combinación de estos factores, sumados a la popularidad de las tramas políticas parecía esconder la fórmula secreta del éxito pero aún quedaban un par de elementos vitales a tener en cuenta.

Uno de los factores más determinantes fue la innovadora estrategia de distribución y comunicación. Gracias al análisis de datos, Netflix pudo saber que a la gente le gustaba hacer maratones de series los fines de semana por lo que optó por la estrategia pionera de poner a disposición de los usuarios todos los capítulos de la primera temporada. La plataforma incluso conocía, según los gustos de los usuarios, a que personajes había que “matar”. De esta manera, el big data influyó en el desenlace de los episodios finales con el fin de captar la máxima atención posible hasta el último momento. [[23](#Key17)]

Con el tiempo, Netflix ha implementado varios algoritmos y mecanismos que hacen uso de los datos que registran los suscriptores y generan ideas críticas que ayudan a dirigir a la empresa en la dirección correcta.

Los ingresos de Netflix para el cuarto trimestre de 2019 fueron de 5,46 mil millones de dólares. Esto representa un aumento del 31% año con año. Los ingresos trimestrales de Netflix han aumentado más de ocho veces desde 2011 y se ha más que duplicado desde 2017. [[24](#Man20)]

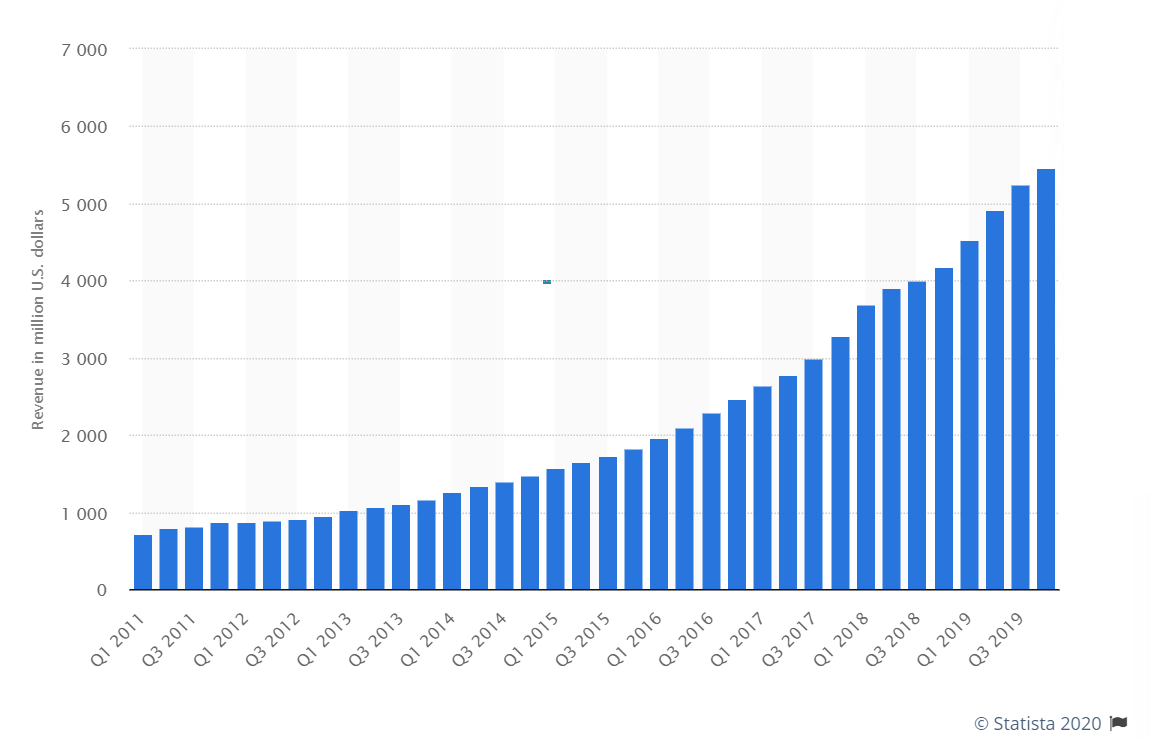


Figura 7: Ingresos de Netflix, T1 2011 a T4 2019, millones de USD [[25](#Sta20)]

1. **Videojuegos**

Los datos siempre han jugado un papel importante en los medios y el entretenimiento. Esto es especialmente cierto para la industria del juego, donde es esencial generar conocimientos sofisticados de los clientes para comprender las percepciones de los jugadores sobre los juegos y desarrollar estrategias de productos altamente personalizadas y efectivas. Millones de personas están jugando videojuegos en cualquier momento en todo el mundo. Los jugadores interactúan constantemente con los juegos mismos, con los creadores e influenciadores del juego, y entre ellos, todo lo cual genera una cantidad significativa de datos sociales y de clientes valiosos. Ser capaz de transformar de manera efectiva todos estos datos en ideas empresariales accionables es clave para los desarrolladores de videojuegos. [[20](#DAT20)]

**Caso Zynga:** Dado que los juegos han puesto un pie en las redes sociales, obtienen algo muy importante, grandes cantidades de datos. A medida que compañías como Zynga han creado juegos como *Zynga Poker*, *Words with friends* y *Farmville* en vivo en las redes sociales, han generado muchas conexiones de usuarios. Aquí es donde surge una perspectiva interesante sobre cómo estas empresas pueden hacer uso de los datos para proporcionar a los jugadores una distracción compulsiva y novedosa para mantenerse por delante de la competencia. *Farmville*, que es una simulación para cultivar una granja, ofrece continuamente a los jugadores bonos en el juego, como ganado adicional, por cantidades pequeñas de dinero en efectivo. La visión más inteligente de Zynga fue darse cuenta de la importancia de ofrecer a sus usuarios lo que realmente quieren. Por ejemplo, en las versiones anteriores, los animales solían estar solo en el fondo. Pero cuando los datos revelaron que eran muy populares entre los jugadores, se les dio papeles prominentes en las últimas versiones del juego.

La forma en que Zynga utilizó los datos podría calificarse como un buen ejemplo de cómo el big data puede convertirse en un pionero de la industria del juego. Utilizando el concepto de big data, estos juegos se habían vuelto extremadamente populares y llevaron a la compañía a la posición más alta en la industria del juego. [[26](#Ana16)]

1. **Noticias**

Las crecientes preocupaciones sobre la rentabilidad a largo plazo de los medios de comunicación impresos y digitales han dejado a las organizaciones de noticias buscando respuestas. Ahora, más que nunca antes, los editores necesitan reinventar sus modelos de manera que les permita llegar a su público objetivo de manera más inteligente y eficiente. Además, necesitan soluciones sostenibles para monetizar su contenido. Hoy en día hay más fuentes de datos de clientes y herramientas de análisis disponibles para editores en línea que nunca antes en la industria de los medios. Estas fuentes únicas proporcionan a las empresas grandes cantidades de datos, lo que les permite tener información precisa y completa sobre su base de clientes, su demografía, comportamiento y preferencias. [[20](#DAT20)]

**Caso New York Times:** En 2015, en una presentación de la conferencia *Predictive Analytics World* en Boston, el científico principal de datos del New York Times, Christopher Wiggins, habló sobre cómo él y su equipo utilizan algoritmos de análisis predictivo para realizar mejoras en el contenido como el análisis de embudo para ver cómo las personas se convierten en suscriptores y cómo influir más para hacerlo. También utilizan el procesamiento del lenguaje natural para comprender los temas de contenido que generan la mayor participación del lector, de modo que los equipos de marketing puedan saber qué tipos de artículos promocionar.[[27](#EdB15)]

En 2020, Wiggins se presentó en el *Jersey City branch* del Colegio de Informática Ying Wu donde citó el viejo chiste entre analistas de datos; de que si torturas los números lo suficiente, confesarán cualquier cosa. En el Times, el trabajo de Wiggins se usa para tres tipos de problemas: descriptivo, predictivo y decisivo, que son términos de aprendizaje no supervisado, supervisado y de refuerzo. Los problemas en los que trabaja, a menudo utilizando su herramienta favorita, que es una biblioteca de código abierto llamada *Scikit-learn* para el lenguaje de programación Python, incluyen un estudio sobre la detección de los comportamientos de los lectores que conducirían a suscripciones canceladas y qué temas de noticias evocan qué tipo de sentimientos. [[28](#Eva20)]

Hay tantos casos en la industria del entretenimiento donde todo parecía ser perfecto para una futura producción, donde se invierten recursos financieros y creativos, pero finalmente, termina en un fracaso absoluto. A continuación se detallan algunos casos:

* **Anuncios**: Uno de los mayores fracasos de 2018 fue la película de Walt Disney “*Solo: A Star Wars Story*”. Con un presupuesto de 250 millones de dólares en la taquilla mundial, la película obtuvo solo 213 millones. Se dice que es la razón principal fue el marketing: el *teaser* salió demasiado tarde y la campaña publicitaria ganó plena capacidad solo un mes antes del estreno de la película. [[29](#Ole19)]
* **Casting**: Realizar el remake de “*Carrie”* en 2013, fue un esfuerzo inútil. Después de todo, el original de Brian De Palma es un clásico, y es difícil superar la perfección. Evidentemente, el público y los críticos estuvieron de acuerdo, ya que la película tuvo una carrera decepcionante en la taquilla nacional y obtuvo un índice de aprobación del 44% en *Rotten Tomatoes* [[30](#Rot)]. Y aunque ciertamente no se puede atribuir toda la culpa a los hombros de la estrella, Chloë Grace Moretz, si se puede cuestionar qué podría haber pasado si MGM hubiera elegido otra actriz para realizar el papel protagónico. Desafortunadamente, Moretz es demasiado bonita para el papel de Carrie; como lo expresó Jay Bauman de *Red Letter Media*, “se parece al tipo de chica que estaría molestando a Carrie en una de estas películas”. [[31](#Ama17)]
* **Contenido**: En 2014, Fox canceló su ambiciosa serie “*Utopia*” , dos meses después de lo que se suponía era una crónica de 12 meses de los esfuerzos de un grupo de extraños para crear una nueva sociedad mientras vivían en un rancho de forma aislada. El programa se basaba en un exitoso formato holandés: *Talpa Media* del productor John de Mol. Pero la versión estadounidense estaba claramente fuera de sintonía con los gustos contemporáneos y requería un gran compromiso por parte de los espectadores en un momento de mucha competencia televisiva. A medida que el precio del espectáculo se disparó a un precio de 50 millones de dólares, algunos ejecutivos de Fox comenzaron a preguntarse si la serie valía la pena y la respuesta fue bastante clara. [[32](#Tim)]

# TÉCNICAS PREDICTIVAS

Las técnicas predictivas y los algoritmos de aprendizaje se utilizan en muchos dominios, y distintas métricas de rendimiento son apropiadas para cada dominio. Las diferentes métricas de rendimiento miden diferentes intercambios en las predicciones hechas por un clasificador, y es posible que las técnicas y métodos de aprendizaje funcionen bien en una métrica, pero no sean óptimos en otras. [[30](#Ric06)]

Debido a esto, es importante analizar las distintas técnicas predictivas para identificar la adecuada para un determinado dominio.

1. **Regresión Lineal**

La regresión lineal es una técnica simple adecuada para la predicción numérica que es utilizada frecuentemente en aplicaciones estadísticas. [[9](#Vai)] La idea es encontrar la cantidad de medida de cada atributo a1, a2,…,ak dentro de un conjunto de datos que contribuyen al objetivo del valor x. A cada atributo se le asigna un factor w1 y un factor extra que es utilizado para constituir el nivel base del atributo predicho.

Fórmula 1: Regresión lineal

El objetivo de la regresión lineal es encontrar pesos óptimos para las instancias de entrenamiento, minimizando el error entre los valores reales y predichos. La regresión lineal es bastante intuitiva y fácil de entender pero la desventaja es que maneja pobremente los atributos no numéricos y no puede manejar problemas no lineales más complejos. [[31](#RIK)]

1. **Árbol de Decisiones**

Los árboles de decisión (AD) son los modelos de aprendizaje automático más adecuados para las tareas de clasificación pero también sólo aplicables para problemas de regresión [[32](#Yan15)]. En la rama de minería de datos, los AD se han vuelto muy populares, ya que son relativamente rápidos de entrenar y producen modelos transparentes. El AD puede ser visto como una serie de preguntas, organizado en una estructura de árbol que lo lleva a un conjunto de clases predefinidas. Más a menudo un AD consta de nodos que contienen una predicción o una condición con dos nodos hijos con las respuestas correspondientes (si/no).

Las condiciones se eligen para dividir el conjunto de datos utilizado para el entrenamiento, en conjuntos de registros más pequeños y puros, es decir, conjunto de instancias que están dominadas por una sola clase objetivo. Cuando es imposible encontrar una condición en un nodo que haga que los conjuntos resultantes sean más puros, se marca como un nodo hoja y se etiqueta para predecir la clase con mayor número de instancias que llegan al nodo.

Cuando es utilizado para clasificación, el nodo raíz es primeramente evaluado seguido del nodo derecho o izquierdo dependiendo el resultado de la condición. Este proceso se repite para cada nodo resultante hasta llegar al nodo hoja donde se encuentra la predicción. Un camino de la raíz a la hoja puede ser visto como una regla que consiste en una condicional simple (if/else). [[31](#RIK)]

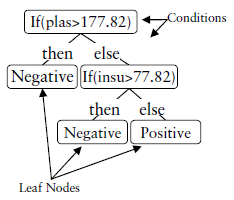


Figura 8: Árbol de decisión simple

**Creación**

Un árbol de decisión se construye para minimizar el error de clasificación de un conjunto de datos en entrenamiento. La creación del árbol se hace recursivamente separando los datos en variables independientes. Cada posible separación es evaluada calculando la pureza ganada que resultará si fuese utilizada para dividir los datos D en nuevos sub datos S = {D1, D2,…, Ds}. La pureza ganada es la diferencia de pureza entre el dato original y los sub datos, como se define en la fórmula 2 donde P(D1) es la proporción de D que está situado en Di. La separación resultante con la pureza ganada más alta es seleccionada y el procedimiento se repite recursivamente por cada sub dato en esa separación.

Fórmula 2: Pureza ganada para separación del AD

**Reducción**

Cuando un AD está completamente crecido, es optimizado en el conjunto entrenado que a menudo conduce a un ajuste excesivo del modelo y existe un alto error de generalización. El método de reducción analiza el árbol de decisión para hacerlo más general, quitando ramas débiles. Al remover partes del árbol, claramente habrá un decrecimiento en la aproximación entrenada del árbol, pero la idea es que esto ayude al AD a ser más general y obtener un mejor rendimiento con los nuevos datos que ingresen.

Hay dos enfoques para la reducción; pre reducción que trata de detener el crecimiento del árbol antes de que existan ramas débiles y post reducción donde se crea un árbol completamente crecido y luego es reducido. En este contexto se puede observar que la post reducción tiene algunas ventajas favorables como por ejemplo una serie de condiciones que pueden ser poderosas aun cuando todas sean ramas débiles unitariamente. [[31](#RIK)]

En general, los algoritmos de post reducción generan a los candidatos del subárbol que son evaluados en los datos entrenados o un nuevo conjunto de validaciones que contienen instancias no vistas previamente. Cómo son exactamente creados estos subárboles difieren de los algoritmos, pero todos aplican *reemplazo de subárbol* y/o *crecimiento de subárbol* de alguna manera. El reemplazo de subárbol comienza en las hojas del árbol y reemplaza los subárboles seleccionados con una sola hoja. El crecimiento del árbol mueve un subárbol a la posición más alta de su rama, eliminando nodos intermedios. Durante la reducción, un gran número de subárboles candidatos pueden ser creados y el árbol con el mejor rendimiento en la validación de datos será seleccionado como el árbol final.

**Estimación de probabilidad**

Aunque el funcionamiento normal de un árbol de decisión es predecir una etiqueta de clase basada en un vector de entrada, también puede ser utilizada para producir probabilidades de membresía de clase; en ese caso son referidos como Estimación de Probabilidad de Árboles (EPAs). La manera más fácil de obtener la probabilidad de una clase es usar la proporción de instancias entrenadas correspondientes a una clase específica en cada hoja. En la Figura 9, seis instancias entrenadas alcanzan la hoja derecha más baja, y cuatro de esas pertenecen a la clase *Positiva*. La probabilidad de clase asignada para la clase *Positiva*, en la hoja, se convertirá en 4/6 = 0.67. Consecuentemente, una futura prueba de instancia clasificada por la hoja será clasificada como clase *Positiva*, con una probabilidad estimada de *0,67*.

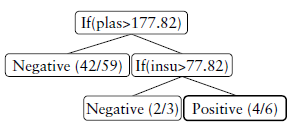


Figura 9: Clasificaciones correctas / instancias que alcanzan cada nodo

Normalmente, las frecuencias relativas no son utilizadas directamente para estimaciones de probabilidad, ya que no consideran el número de instancias entrenadas que respaldan la clasificación. Es por eso que *la estimación de* *Laplace* es comúnmente utilizada para producir estimación de probabilidad basada en el respaldo. La fórmula 3 muestra cómo la estimación de probabilidad p es calculada utilizando Laplace. N es el número total de instancias, C es el número de clases y k es el número de instancias entrenadas respaldando la predicción de la clase A.

Fórmula 3: Estimación de probabilidad Laplace

En la Figura 9, la estimación de la probabilidad para el nodo derecho más bajo será calculado como 4/6 = 0.67 sin Laplace y ((4 + 1) / (6 + 2)) = 0.63 utilizando Laplace. Se debe notar que la estimación de Laplace introduce una probabilidad uniforme previa para cada clase; es decir, antes de que alguna instancia alcance una hoja (*k = N = 0*), la probabilidad para cada clase es *1/C*.

1. **Redes Neuronales Artificiales**

Las Redes Neuronales Artificiales (RNA) es una técnica de aprendizaje automático basado en el funcionamiento del cerebro humano. Las RNA son extremadamente poderosas en el sentido de que son aproximadores funcionales universales, es decir, pueden aproximar cualquier función hacia cualquier medida deseada. Los Perceptrones Multi-Capas (PMC) son unos de los tipos más comunes de RNAs y han sido utilizados para resolver una amplia variedad de problemas y son frecuentemente usados para tareas de clasificación y regresión debido a su inherente capacidad de asignación de salida de entrada arbitraria. [[31](#RIK)]

**Fundamentos de una Red Neuronal**

Los Perceptrones Multi-Capas (PMC) consisten en un conjunto de unidades (neuronas) que se ordenan en capas y se conectan por enlaces de pesos que pasan señales a través de las unidades. Las unidades en la primera capa, la capa de entrada, están conectadas a sus variables de entradas y reciben señales de entrada correspondientemente. Cuando una señal es enviada a través de un enlace, ésta es multiplicada por el peso del enlace el cual decide el impacto de la unidad recibida. Todas las señales entrantes son resumidas y una activación es aplicada para calcular la respuesta de salida. Las unidades en la capa de entrada están conectadas a una capa oculta, que puede estar conectada a otra capa oculta o a la capa de salida. [[33](#Han)]

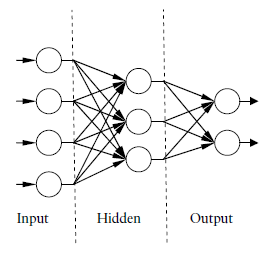


Figura 10: Simple Perceptrón Multi-Capa

* **Capas ocultas y unidades ocultas:** El número de capas ocultas y nodos ocultos en una RNA es una decisión crucial de diseño ya que permiten a la red detectar las características y capturar el patrón en los datos realizando mapeos no lineales complicados entre las variables de entrada y salida. Una capa oculta es suficiente para una RNA para aproximar cualquier función continua y obtener el grado de exactitud deseado. Dos capas ocultas, a veces pueden ser preferibles ya que pueden lograr el mismo resultado con menos unidades ocultas y aproximar funciones discontinuas. Una red con pocas unidades será más rápida de entrenar y en principio, generalizar mejores pronósticos.
* **Capa de salida y unidades de salida:** Seleccionar el número de unidades de salida es otro factor importante para el rendimiento de la red. Para problemas de regresión, una unidad de salida es suficiente pero, las tareas de clasificación son a menudo diseñadas con más de una clase, donde cada una es representada por una unidad de salida. La clasificación es hecha de acuerdo a la clase asociada para la unidad con más alta señal de salida.
* **Funciones de activación:** Las funciones de activación determinan la relación que existe entre las entradas y las salidas de un nodo y una red. En general, la función de activación introduce un grado de no linealidad que es valioso para la mayoría de aplicaciones de RNA. La función de activación más común son las funciones sigmoideas: la *función logística unipolar* (4) y la *función sigmoidea bipolar* (5).

Fórmula 4: Función logística unipolar

Fórmula 5: Función sigmoidea bipolar

Diferentes funciones de activación pueden ser usadas en una red pero usualmente sólo un tipo es utilizado en cada capa.

**Entrenamiento de una red neuronal**

El entrenamiento de una RNA es básicamente el mapeo entre las variables de entrada y salida de los datos. El entrenamiento de una RNA es un proceso de optimización de los pesos conectados. El algoritmo comúnmente más utilizado es la *propagación hacia atrás* (*back propagation*) que es un método gradiente en descenso. Se debe establecer un tamaño de paso denominado *tasa de aprendizaje* y un parámetro de ímpetu antes de que pueda comenzar el entrenamiento. Estos parámetros juegan un papel crítico en el rendimiento de la red final, pero no hay manera de calcular la velocidad de aprendizaje e impulso óptimo para un problema dado. A menudo se realizan varios experimentos con diferentes ajustes antes de seleccionar los valores finales. [[31](#RIK)]

1. **Algoritmos Evolutivos**

Los algoritmos evolutivos (AE) son una colección de algoritmos y técnicas inspirados en los principios de Darwin de selección natural y supervivencia del más apto. Los AE son esencialmente métodos de optimización basados en el proceso natural que ocurre con frecuencia en la naturaleza cuando una especie se adapta a los cambios de su entorno.

La idea principal en los AE es resolver un problema utilizando una *población de individuos en competencia*. El problema define las constricciones de un *entorno* en el que los individuos (*soluciones candidatas*) van a competir. Si no existe un conocimiento previo sobre la naturaleza del entorno, la población inicial se crea de forma aleatoria, ya que esto asegurará un rendimiento promedio aproximado de la población en general. Luego, se evalúa a cada individuo y se le asigna una puntuación (*aptitud física*) que se correlaciona con la forma en que se adapta al entorno. Finalmente, los individuos se seleccionan para formar parte de un *grupo de apareamiento* en el que estarán expuestos a operaciones genéticas para crear descendencia para una nueva generación más fuerte. Hay varias estrategias diferentes sobre cómo se seleccionan los individuos para formar parte del grupo de apareamiento, dos de los más comunes son la *ruleta* y la *selección de torneos*.

En la selección de la ruleta, cada individuo es seleccionado con una probabilidad proporcional a su forma física. Esto puede ser representado por una rueda de la ruleta donde cada individuo obtiene una ranura donde el tamaño de la ranura corresponde a su aptitud relativa. La estructura de la rueda permanece igual para cada selección, lo que permite seleccionar un individuo varias veces (*re-selección*).

La selección del torneo se realiza seleccionando *k* individuos al azar y luego seleccionando el que tiene el mejor estado físico. Una *k* mayor pondrá un mayor peso en el valor de la condición física, mientras que una *k* ignorará la condición física y elegirá a los individuos al azar.

Al crear una nueva población, hay tres operaciones genéticas fundamentales; *reproducción*, *cruce* y *mutación*. La reproducción hace una copia de un individuo; el cruce divide a cada uno de los padres en dos partes y los combina en dos nuevos individuos. La mutación hace un cambio aleatorio a un individuo. La mayoría de los AE se realizan en un modo llamado *estado estacionario* donde las operaciones genéticas se aplican a individuos en el grupo de apareamiento hasta que la nueva población tenga el mismo tamaño que la población actual. Las nuevas generaciones evolucionan continuamente de esta manera hasta que se alcanza un cierto número de generaciones o hasta que se encuentra una solución lo suficientemente buena. [[31](#RIK)]

**Algoritmos Genéticos**

En los algoritmos genéticos, un individuo o un cromosoma están representados por una cadena de bits. Normalmente, el tamaño de la cadena se fija para facilitar operaciones simples de cruce. La población inicial se crea asignando aleatoriamente 1 o 0 a cada bit de cada individuo. Una población puede contener cientos o incluso miles de individuos según el problema y el costo computacional de la evaluación de un individuo. Para evaluar a un individuo, la cadena de bits se traduce primero en una solución candidata y luego se evalúa mediante una función de aptitud. En los AG se realiza un cruce en dos padres que se seleccionan al azar del grupo de apareamiento. A continuación, ambos padres se dividen en dos partes en el mismo punto de cruce decidido al azar. Finalmente, se crean dos nuevos descendientes combinando una parte de cada padre de manera ordenada que preserva la colocación de cada parte.

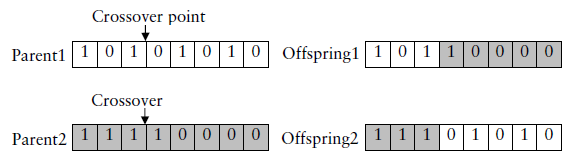


Figura 11: Cruce del Algoritmo Genético

La mutación se realiza simplemente cambiando un bit aleatorio y la reproducción hace una copia exacta del individuo seleccionado.

**Programación Genética**

En la programación genética, un individuo o programa se representa mediante una estructura de árbol similar a un árbol de decisión, que contiene *funciones* y *terminales* predefinidas. Sin embargo, es importante que la función y los conjuntos de terminales se elijan teniendo en cuenta el problema, ya que la PG solo puede encontrar soluciones que contengan estos elementos. La selección de las funciones y terminales incorrectas puede resultar en un espacio de búsqueda que no incluye la solución óptima, mientras que los elementos superfluos darán lugar a un ritmo de búsqueda innecesariamente grande.

Para problemas de minería de datos, donde un aspecto importante es la comprensibilidad, los lenguajes de representación simples como *Boolean* o *If/Else* son los más utilizados. Para las reglas booleanas, el conjunto de funciones F consta de *and*, *or* y los operadores relacionales *mayor*, *menor* e *igual*, mientras que el conjunto terminal T se define por los atributos del conjunto de datos actual y las constantes aleatorias.

Hay tres estrategias principales sobre cómo se crea la población inicial en la PG. *Full* crea árboles seleccionando funciones al azar hasta que se alcanza una cierta profundidad, cuando se seleccionan terminales para todas las funciones. Por lo tanto, un árbol creado con *Full* siempre está equilibrado, es decir, todos las terminales se definen a la misma profundidad. *Grow* selecciona aleatoriamente funciones y terminales y, por lo tanto, puede crear árboles desequilibrados. La última estrategia *Ramped half and half* crea la mitad de la población usando *Grow* y la mitad usando *Full* mientras que al mismo tiempo varía la profundidad máxima. *Ramped half and half* es la estrategia de creación más utilizada, ya que presenta a la población más diversa.

En la PG, la operación de cruce se realiza seleccionando al azar un punto de cruce en cada árbol y recombinando las partes en dos nuevos individuos como se ve en la figura 12. El punto de cruce puede ser diferente en cada árbol, pero los árboles resultantes tienen que ser dos programas legales. Si se eligen dos puntos incompatibles, los programas resultantes se descartan y el procedimiento se realiza nuevamente.

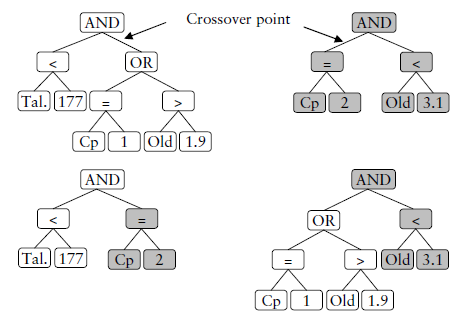


Figura 12: Cruce de la Programación Genética

La mutación se realiza seleccionando al azar un punto de mutación y eliminando todo lo que se encuentra debajo de este punto. A continuación, la estrategia de creación de *Grow* se aplica para crear una nueva parte del programa a partir del mismo punto de mutación. La reproducción se realiza de la misma manera que en los AG, es decir, simplemente copiando el programa en la siguiente generación. [[31](#RIK)]

# PROPUESTA DE MODELO PREDICTIVO

En términos generales, se divide el proceso de modelado y análisis predictivo en las siguientes etapas: recopilación de datos, análisis de datos, limpieza y tratamiento de datos, modelado predictivo y estimación de rendimiento.

1. **Recopilación de datos**

La búsqueda y selección de datos contempla las siguientes categorías dentro de la industria del entretenimiento:

1. Cine
2. Televisión
3. Música
4. Videojuegos

El conjunto de datos para entrenar el modelo predictivo de las categorías identificadas se obtuvieron de la plataforma web *Kaggle* [[34](#Kag)]: una comunidad en línea que permite a los usuarios encontrar y publicar conjuntos de datos, explorar y construir modelos en un entorno de ciencia de datos, trabajar con otros ingenieros expertos en aprendizaje automático y científicos de datos. [[35](#Kag19)]

Los datos utilizados para validar el modelo predictivo resultante fue proveído por INCAA (Instituto Nacional de Cine y Artes Audiovisuales). [[36](#Ins)]

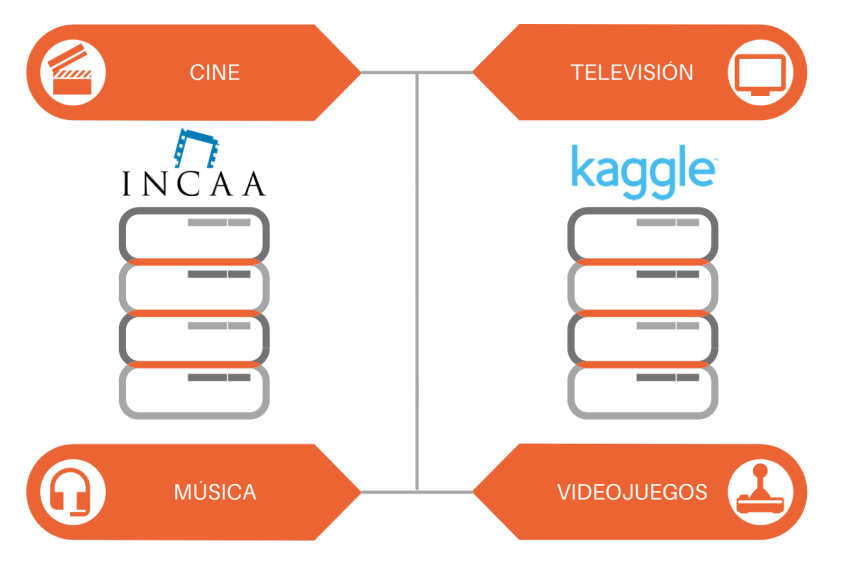


Figura 13: Proceso de recopilación de datos de las categorías identificadas

A continuación se detallan las descripciones de los datos recolectados:

1. **The Movie Database 5000 Movie Dataset:** contiene datos sobre el elenco, el equipo, la fecha de lanzamiento, la popularidad, el idioma original, el presupuesto y los ingresos de miles de películas. Este conjunto de datos se generó a partir de la *API de The Movie Database* [[37](#The)]: que proporciona acceso a datos sobre muchas películas, actores, actrices y programas de televisión. [[38](#Kag17)]
2. **1000 Netflix Shows:** en los últimos 5-10 años, Netflix ha capturado una gran población de espectadores. Este conjunto de datos contiene información sobre la clasificación, fecha de lanzamiento, puntaje y tamaño de calificación del usuario de mil películas y programas de Netflix. [[39](#Kag171)]
3. **Movie Ratings Dataset:** este conjunto de datos contiene registros de calificaciones IMDB de películas lanzadas desde 2000 al 2017. [[40](#Kag196)]
4. **Feature Film/TV Series:** son 100 películas y series ordenadas por popularidad, obtenidas de IMDB. El conjunto de datos cuenta con información del director, los actores y actrices, el género, la calificación, la cantidad de estrellas y los votos de la película. [[41](#Kag18)]
5. **Music Artists Popularity:** el conjunto de datos consta de más de 1.4 millones de artistas musicales presentes en la base de datos de *MusicBrainz*: sus nombres, etiquetas y popularidad, basados ​​en datos recogidos de *last.fm* [[42](#las)]: una plataforma online y además un sistema de recomendación de música que construye perfiles y estadísticas sobre gustos musicales, basándose en los datos enviados por los usuarios registrados. [[43](#Kag1911)]
6. **Trending YouTube Video Statistics:** para determinar los videos más populares del año, YouTube utiliza una combinación de factores que incluyen la medición de las interacciones de los usuarios (número de vistas, compartidos, comentarios y me gusta). Este conjunto de datos incluye varios meses de datos en videos de YouTube con tendencias diarias. Se incluyen datos para las regiones de EE. UU., GB, DE, CA, FR, RU, MX, KR, JP y IN (Estados Unidos., Gran Bretaña, Alemania, Canadá Francia, Rusia, México, Corea del Sur, Japón e India, respectivamente), con hasta 200 videos de tendencias listados por día. Los datos se recopilaron utilizando la *API de YouTube*. [[44](#You)] [[45](#Kag1933)]
7. **Video Games Sales Dataset:** venta de videojuegos en la plataforma Playstation 4 de Sony y XboxOne de Microsoft en el periodo 2016 y 2017. Además cuenta con un conjunto de datos de ventas de otras plataformas como Wii, Nintendo, etc. [[46](#Kag1967)]
8. **Histórico de películas INCAA:** el área de estadísticas de INCAA, facilitó la disponibilidad de los siguientes datos: Título, fecha de estreno, país/es de origen, género/s, espectadores de películas en un periodo entre 2013 a 2018.
9. **Análisis de datos**

El análisis de datos está alineado con el objetivo general definido en el presente documento, es por ello que la finalidad de la evaluación de los datos recopilados tiene que contemplar todas las variables que orienten a la toma de decisiones de una nueva producción dentro de la industria del entretenimiento.

Luego de ordenar, identificar, comparar y filtrar los distintos conjuntos de datos obtenidos, se definen las siguientes variables unificadas que ayudarán al entrenamiento del modelo predictivo:

1. **ID:** Cadena que representa el número identificador. Formato donde *x* representa el número secuencial del conjunto de datos origen.
2. **Nombre:** Cadena que contiene el título de la producción o el nombre del grupo/artista en caso de corresponder.
3. **Categoría:** Cadena que selecciona el tipo de producción. Tiene 4 valores posibles: *Cine, Televisión, Música* y *Videojuegos.*
4. **Fecha Lanzamiento:** Formato día, mes y año. En caso de tener sólo el año como dato, se asume el primer día y mes del año. Es posible tener valores nulos.
5. **Género:** Cadenas que representan los géneros que pueden existir dentro de las distintas categorías. En caso de no existir el valor se utilizará el texto *ND (No Definido)*.
6. **País:** Cadena con el nombre del país productor o de origen. En caso de no existir el valor se utilizará el texto *ND (No Definido)*.
7. **Ingresos:** Monto expresado en dólares equivalente a los ingresos generados por la producción. Se utilizará el número cero para aquellos registros sin la información necesaria.
8. **Puntaje:** Rango de números entre 0 y 10 para representar el puntaje otorgado a dicha producción o grupo/artista.
9. **Canal de distribución:** Cadena que selecciona el tipo de distribución. Tiene 5 valores posibles: *Sala de cines, Venta CD/DVD/Blu-ray, Alquiler digital, Streaming, Televisión por cable/abierta*.

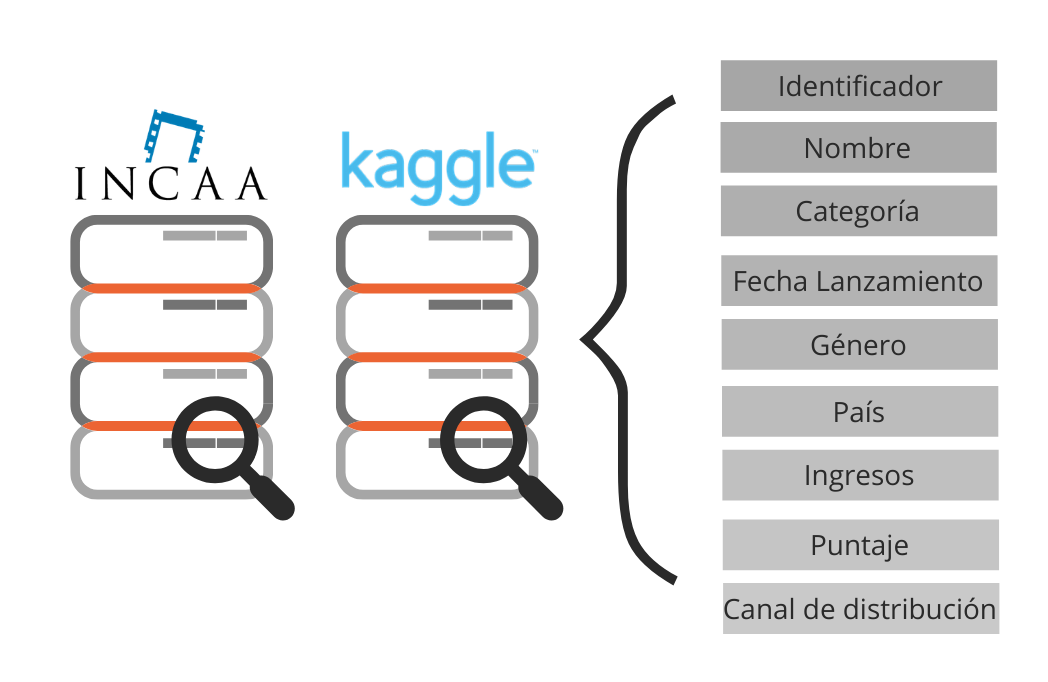
****

Figura 14: Proceso de análisis de datos y selección de variables

1. **Limpieza y tratamiento de datos**

Para la de manipulación de datos se utiliza la herramienta *Tableau Prep Builder 2019.1* para crear los distintos flujos de datos. Proporciona una forma visual y directa de combinar, dar forma y limpiar los datos. [[47](#TAB)]

|  |  |  |
| --- | --- | --- |
|  | **Prueba** | **Individual** |
| Precio | Licencia gratuita por un año para estudiantes/profesores. | USD 70 por usuario por mes facturado anualmente. |
| Características | Tableau Desktop y Tableau Prep Builder. | Tableau Desktop, Tableau Prep Builder y una licencia Creator para Tableau Server o Tableau Online. |

Tabla 1: Precios y características de Tableau

A continuación se detallan los distintos pasos que se empelaron para limpiar y organizar los conjuntos de datos acorde a las variables definidas, donde **CDE** representa el Conjunto de Datos para Entrenar y **CDV** representa el Conjunto de Datos para Validar el modelo predictivo:

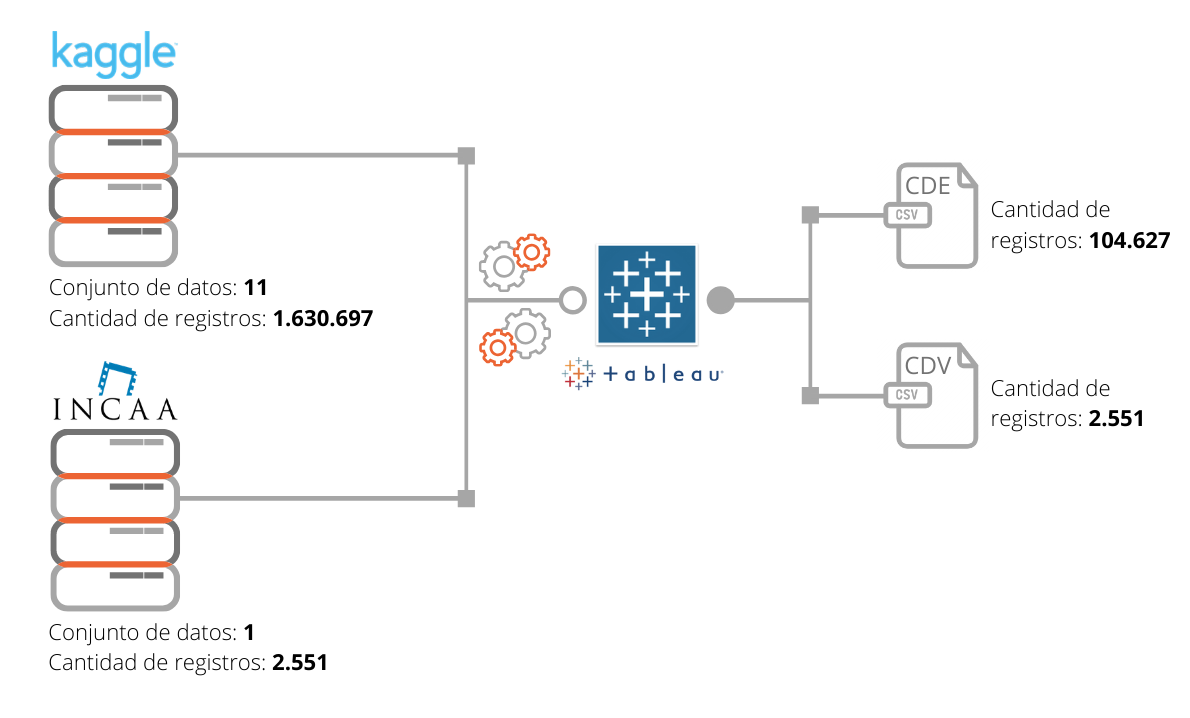


Figura 15: Proceso de limpieza y tratamiento de datos con Tableau

**Paso 1: Limpieza del CDE**

Las conexiones que se manejaron para aplicar la limpieza y organización del CDE fueron las siguientes:

1. tmdb\_5000\_movies.csv (The Movie Database 5000 Movie Dataset).
2. Netflix Shows.csv (1000 Netflix Shows).
3. movie\_ratings.csv (Movie Ratings Dataset).
4. top100popmovts\_dataframe.csv (Feature Film/TV Series).
5. artists.csv (Music Artists Popularity).
6. USvideos.csv (Trending YouTube Video Statistics – United States).
7. MXvideos.csv (Trending YouTube Video Statistics - México).
8. CAvideos.csv (Trending YouTube Video Statistics – Canadá).
9. PS4\_GamesSales.csv (Video Games Sales Dataset – PS4).
10. XboxOne\_GameSales.csv (Video Games Sales Dataset – Xbox One).
11. Video\_Games\_Sales\_as\_at\_22\_Dec\_2016.csv (Video Games Sales Dataset – Otras consolas).

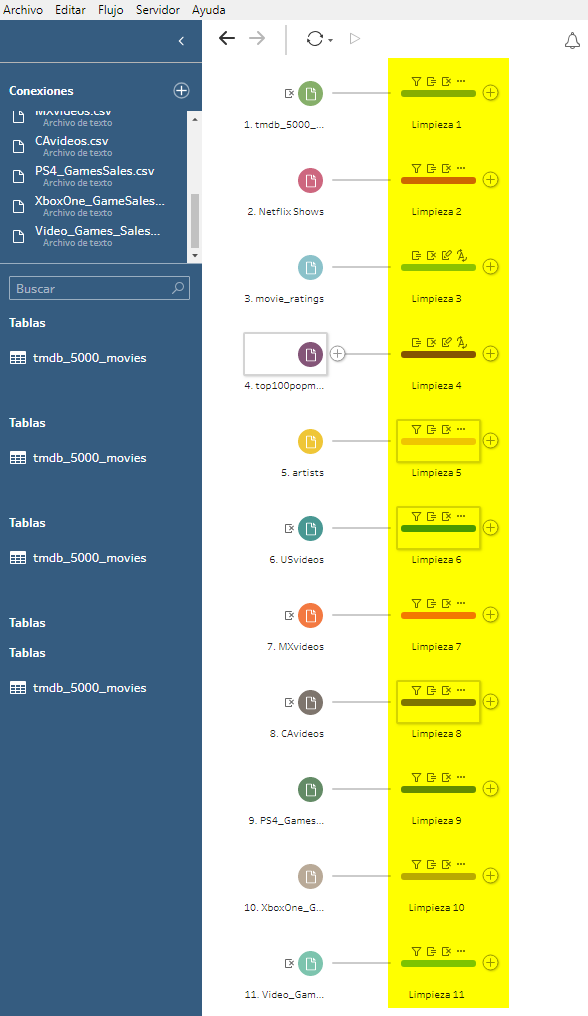
****

Figura 16: Limpieza del CDE en Tableau Prep

|  |  |
| --- | --- |
|  | **Paso 1** |
| Limpieza 1  tmdb\_5000\_movies.csv | **ID:** DS01…  **Nombre:** Se obtiene del campo *original\_tittle*, se eliminan los espacios innecesarios.  **Categoría:** Cine.  **Fecha Lanzamiento:** Se obtiene del campo *release\_date*, se filtran los elementos no nulos.  **Género:** Se extrae el primer valor del campo JSON *genres*, se filtran valores no nulos y se convierten a mayúscula.  **País:** Se extrae el primer valor del campo JSON *production\_countries*, se filtran valores no nulos y se convierten a mayúscula.  **Ingresos:** Se obtiene del campo *revenue*, se convierten los valores a formato decimal.  **Puntaje:** Se obtiene del campo *vote\_average*, se filtran rangos entre 0 a 10. |
| Limpieza 2  Netflix Shows.csv | **ID:** DS02…  **Nombre:** Se obtiene del campo *tittle*, se eliminan los espacios innecesarios.  **Categoría:** Televisión.  **Fecha Lanzamiento:** Se obtiene del campo *release\_year*, se convierte el valor a formato fecha.  **Género:** TV Serie.  **País:** ND (No Definido).  **Ingresos:** 0.  **Puntaje:** Se obtiene de dividir entre 10 el campo *user rating score.* En caso de ser nulo se establece el valor 0. |
| Limpieza 3  movie\_ratings.csv | **ID:** DS03…  **Nombre:** Se obtiene del campo *movie*, se eliminan los espacios innecesarios.  **Categoría:** Cine.  **Fecha Lanzamiento:** Se obtiene del campo *year*, se convierte el valor a formato fecha.  **Género:** ND (No Definido).  **País:** ND (No Definido).  **Ingresos:** 0.  **Puntaje:** Se obtiene del campo *imdb.* |
| Limpieza 4  top100popmovts\_dataframe.csv | **ID:** DS04…  **Nombre:** Se obtiene del campo *movieTittle*, se eliminan los espacios innecesarios.  **Categoría:** Cine/Televisión. (Se hace la separación en función a la fecha, es decir, si tiene un rango de fechas es una serie.  **Fecha Lanzamiento:** Se obtiene del campo *movieDate*, se convierte el valor a formato fecha.  **Género:** Se extrae el primer valor del campo *movieGenre*, se eliminan los espacios innecesarios y se convierten a mayúscula.  **País:** ND (No Definido).  **Ingresos:** Se calcula a partir del campo *movieGross,* en caso de ser nulo devuelve 0, en caso de ser de categoría Televisión o película del año 2019 se multiplica por 1.000.000.Se realizó este ajuste para obtener resultados más reales.  **Puntaje:** Se obtiene del campo *movieRating.* En caso de ser nulo se establece el valor 0. |
| Limpieza 5  artists.csv | **ID:** DS05…  **Nombre:** Se obtiene del campo *artist\_mb*, filtran valores no nulos y que no contengan caracteres especiales al inicio. Además se filtra que no sea un artista ambiguo (según información que proveen los datos).  **Categoría:** Música.  **Fecha Lanzamiento:** nulo.  **Género:** Se extraen los valores del campo *tags\_mb*, se filtran valores que no contengan caracteres especiales al inicio y se convierten a mayúscula.  **País:** Se obtiene del campo *country\_mb*, se filtran valores no nulos y se convierten en mayúscula.  **Ingresos:** 0.  **Puntaje:** Se obtiene de multiplicar por 10 y dividir entre 5.500.000 (cantidad máxima de oyentes) el campo *listeners\_lastfm.* Se redondea el valor a dos decimales. |
| Limpieza 6  USvideos.csv | **ID:** DS06…  **Nombre:** Se obtiene del campo *title*, se eliminan caracteres especiales y se filtran aquellos registros que contengan las siguientes palabras: Trailer, Movie, Music, Videoclip, TV, Serie, Videogame.  **Categoría:** Cine/Música/Televisión/Videojuegos. Dependen del título que contengan.  **Fecha Lanzamiento:** Se obtiene del campo *publish\_time*, se filtran los elementos no nulos.  **Género:** ND (No Definido).  **País:** United States of America.  **Ingresos:** 0.  **Puntaje:** Se multiplica el campo *likes* por 10 y se divide entre la suma entre *likes* y *dislikes.* Se redondea el valor a dos decimales. |
| Limpieza 7  MXvideos.csv | **ID:** DS07…  **Nombre:** Se obtiene del campo *title*, se eliminan caracteres especiales y se filtran aquellos registros que contengan las siguientes palabras: Trailer, Movie, Music, Videoclip, TV, Serie, Videogame.  **Categoría:** Cine/Música/Televisión/Videojuegos. Dependen del título que contengan.  **Fecha Lanzamiento:** Se obtiene del campo *publish\_time*, se filtran los elementos no nulos.  **Género:** ND (No Definido).  **País:** México.  **Ingresos:** 0.  **Puntaje:** Se multiplica el campo *likes* por 10 y se divide entre la suma entre *likes* y *dislikes.* Se redondea el valor a dos decimales. |
| Limpieza 8  CAvideos.csv | **ID:** DS08…  **Nombre:** Se obtiene del campo *title*, se eliminan caracteres especiales y se filtran aquellos registros que contengan las siguientes palabras: Trailer, Movie, Music, Videoclip, TV, Serie, Videogame.  **Categoría:** Cine/Música/Televisión/Videojuegos. Dependen del título que contengan.  **Fecha Lanzamiento:** Se obtiene del campo *publish\_time*, se filtran los elementos no nulos.  **Género:** ND (No Definido).  **País:** Canadá.  **Ingresos:** 0.  **Puntaje:** Se multiplica el campo *likes* por 10 y se divide entre la suma entre *likes* y *dislikes.* Se redondea el valor a dos decimales. |
| Limpieza 9  PS4\_GamesSales.csv | **ID:** DS09…  **Nombre:** Se obtiene del campo *Game*, se eliminan los espacios innecesarios.  **Categoría:** Videojuegos.  **Fecha Lanzamiento:** Se obtiene del campo *Year*, se convierte el valor a formato fecha.  **Género:** Se obtiene del campo *Genre*, y se convierten los valores a mayúscula.  **País:** ND (No Definido).  **Ingresos:** Se obtiene de multiplicar el campo *Global* por 1.000.000.  **Puntaje:** 0. |
| Limpieza 10 XboxOne\_GameSales.csv | **ID:** DS10…  **Nombre:** Se obtiene del campo *Game*, se eliminan los espacios innecesarios.  **Categoría:** Videojuegos.  **Fecha Lanzamiento:** Se obtiene del campo *Year*, se convierte el valor a formato fecha.  **Género:** Se obtiene del campo *Genre*, y se convierten los valores a mayúscula.  **País:** ND (No Definido).  **Ingresos:** Se obtiene de multiplicar el campo *Global* por 1.000.000.  **Puntaje:** 0. |
| Limpieza 11  Video\_Games\_Sales\_as\_at\_22\_Dec\_2016.csv | **ID:** DS11…  **Nombre:** Se obtiene del campo *Name*, se filtran valores no nulos y que no contengan caracteres especiales al inicio.  **Categoría:** Videojuegos.  **Fecha Lanzamiento:** Se obtiene del campo *Year\_of\_Release*, se convierte el valor a formato fecha.  **Género:** Se obtiene del campo *Genre*, y se convierten los valores a mayúscula.  **País:** ND (No Definido).  **Ingresos:** Se obtiene de multiplicar el campo *Global\_Sales* por 1.000.000.  **Puntaje:** Se obtiene del campo *User\_Score*, y si es valor nulo se establece el valor 0. |

Tabla 2: Proceso de limpieza del CDE

**Paso 2: Unión del CDE**

Una vez estandarizados los distintos conjuntos de datos, con las variables definidas en el análisis, se procede a unir aquellos pasos de limpieza que resulten relacionados entre sí, ya sea por categoría o similitud de datos.

El resultado son 3 uniones de los siguientes pasos:

* **Unión 1:** Es el resultado de una simple unión de los pasos de Limpieza 1, 2, 3 y 4 que contienen datos correspondientes a películas y series pertenecientes a las categorías Cine y Televisión respectivamente.
* **Unión 2:** Es el resultado de una simple unión de los pasos de Limpieza 6, 7 y 8 que contienen datos correspondientes a las visualizaciones de videos de YouTube en los países Estados Unidos, Canadá y México dentro de las categorías Cine, Música, Televisión y Videojuegos.
* **Unión 3:** Es el resultado de una simple unión de los pasos de Limpieza 9, 10 y 11 que contienen información de las ventas y popularidad de videojuegos de distintas compañías.

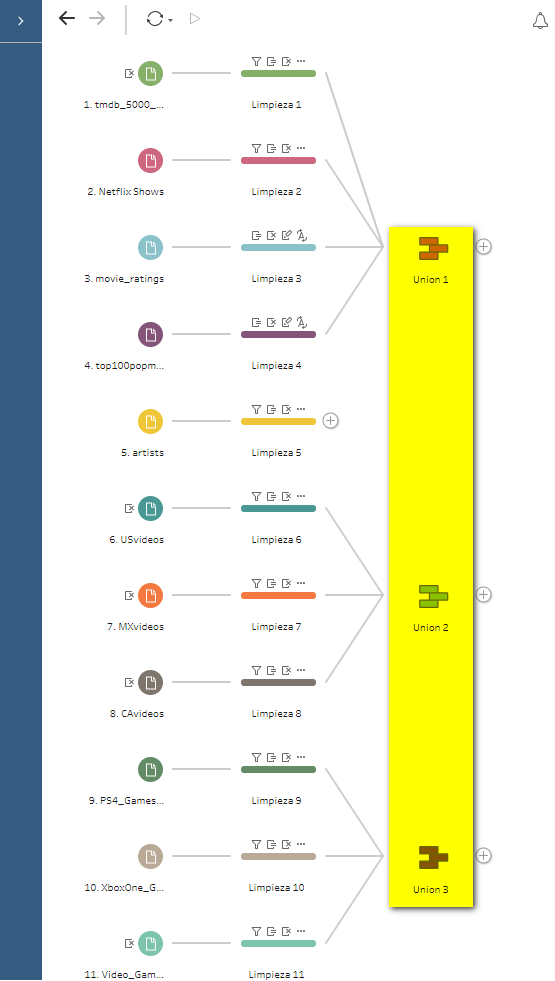
****

Figura 17: Unión del CDE en Tableau Prep

**Paso 3: Agregación del CDE**

Para ajustar la granularidad de los datos, se utiliza la opción de agregación con el fin de crear un paso que agrupe los datos. Los datos se agregarán o se agruparán en función del tipo de datos:

* **Agregación 1:**
  + *Campo agrupado:* Nombre (título de la película o serie).
  + *Campos agregados por valor máximo:* Categoría, Fecha Lanzamiento, Género, País, Ingresos, Puntaje.
* **Agregación 2:** 
  + *Campos agrupados:* Nombre (título del video), País.
  + *Campos agregados por valor máximo:* Categoría, Fecha Lanzamiento, Género, País, Ingresos.
  + *Campo agregado por la media:* Puntaje.
* **Agregación 3:**
  + *Campo agrupado:* Nombre (título del videojuego).
  + *Campos agregados por valor máximo:* Categoría, Fecha Lanzamiento, Género, País, Ingresos, Puntaje.

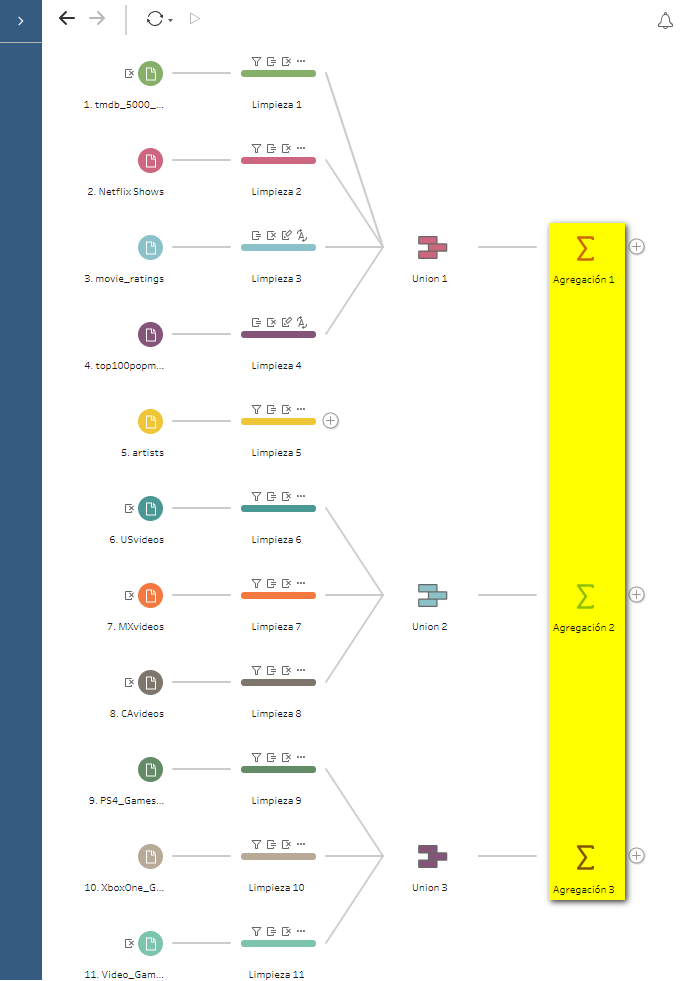
****

Figura 18: Agregación del CDE en Tableau Prep

**Paso 4: Organización del CDE**

Se afinan los últimos detalles de los resultados de agregación para generar calidad de los datos generados. A este paso se denomina Organización y el objetivo es generar la última variable faltante identificada en el análisis de datos: El canal.

El canal de distribución está representado por una letra: **“C”** (Sala de cines), **“V” (**Venta CD/DVD/Blu-ray), **“A”** (Alquiler digital), **“S”** (Streaming), **“T”** (Televisión por cable/abierta).

* **Organización 1:**
  + *Canal:* Si la Categoría es igual a “Cine” y los Ingresos son mayores a cero entonces el valor es “C”. Sino si la Categoría es “Televisión” y el Género es distinto de “ND” entonces el valor es “S”. Caso contrario el valor es “T”.
* **Organización 2:** 
  + *Canal:* El valor es “S”.
  + *País:* Se reemplaza el valor del país Estados Unidos (en inglés) para que coincida con el resto de conjunto de datos.
* **Organización 3:**
  + *Canal:* El valor es “S”.
  + *Nombre:* Se eliminan todos los caracteres de puntuación del texto.
* **Organización 4:**
  + *Canal:* El valor es “V”.
  + *Nombre:* Se eliminan todos los caracteres de puntuación del texto.

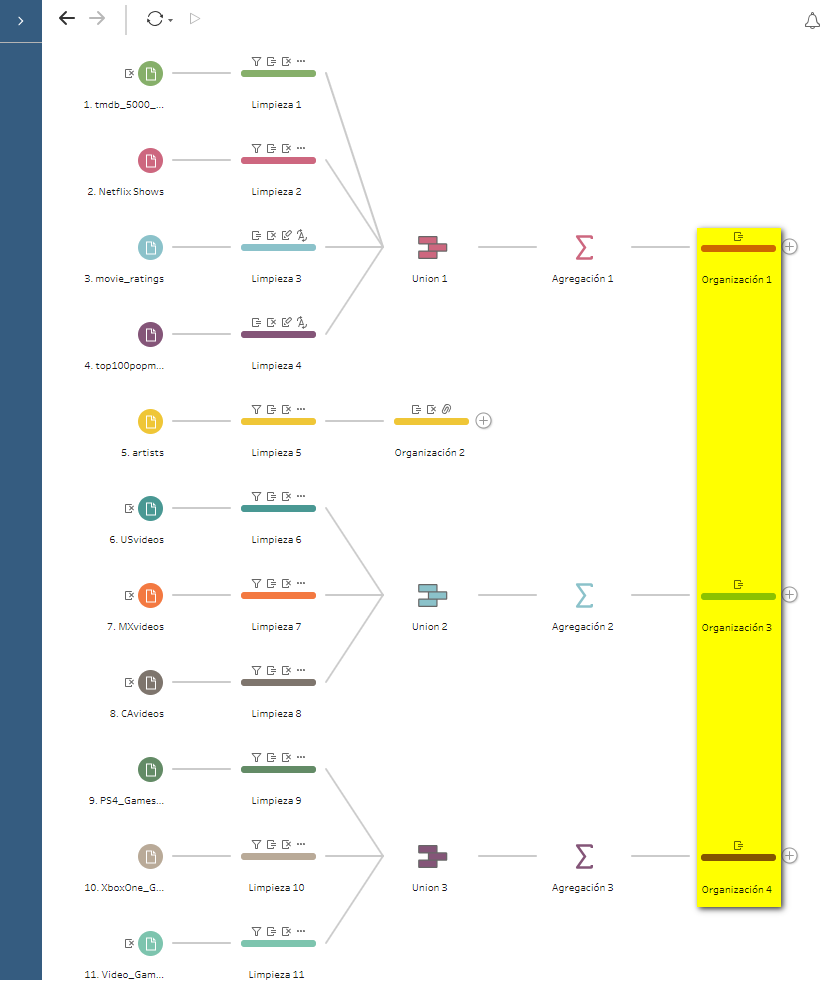
****

Figura 19: Organización del CDE en Tableau Prep

**Paso 5: Unión final del CDE**

Una vez se tiene organizados todos los conjuntos de datos, se procede a realizar la unión final, es decir, generar en un solo paso la combinación de los datos de todas las categorías.

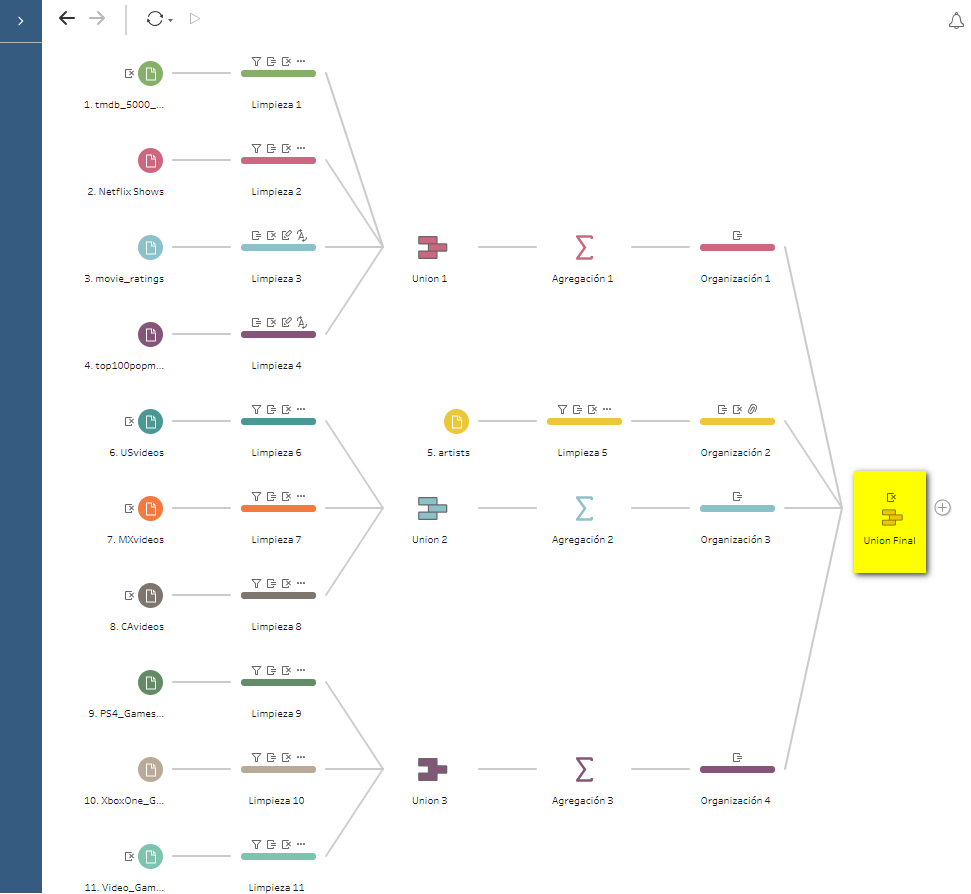
****

Figura 20: Unión final del CDE en Tableau Prep

Finalmente, los datos están preparados para generar el archivo de salida y ser posteriormente analizados. Para ello, se añade un paso de Salida, donde se especifica el nombre y la ubicación para guardar el archivo: **CDE.csv.**

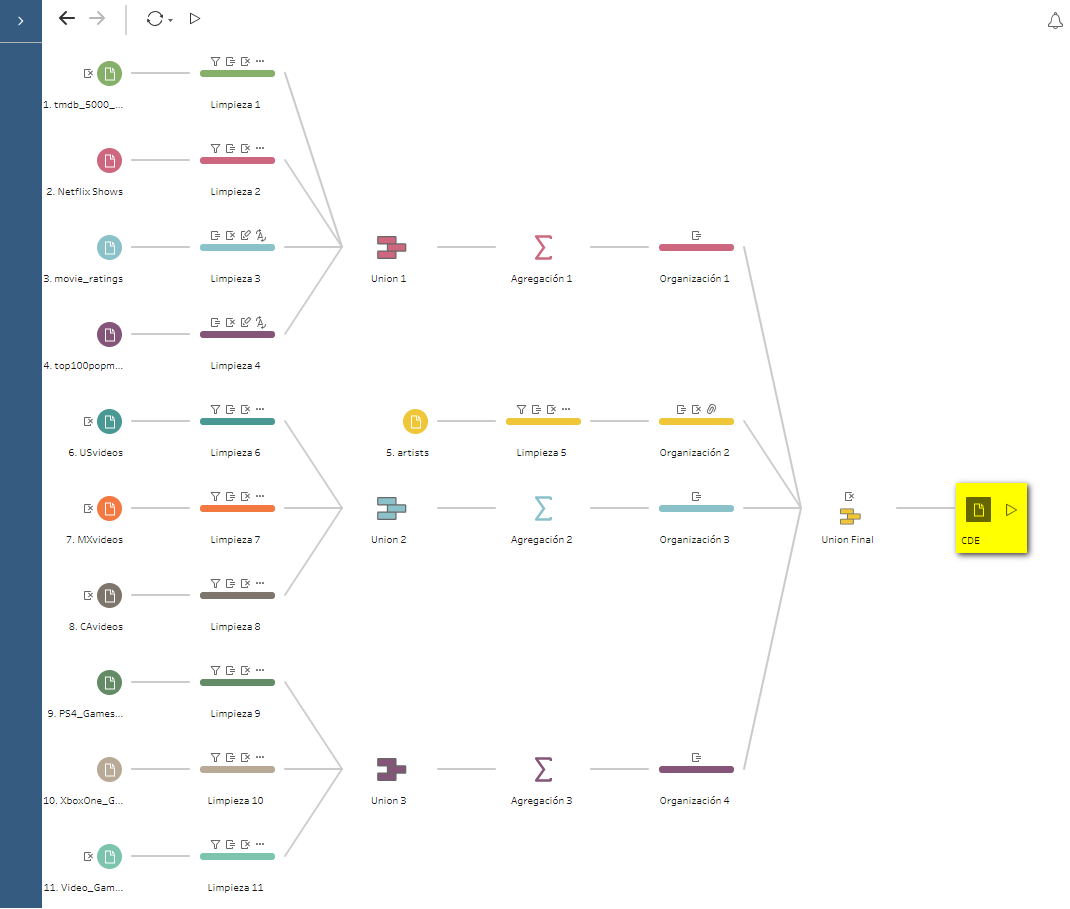


Figura 21: Salida resultante del CDE en Tableau Prep

**Paso 6: Limpieza y organización del CDV**

La conexión que se maneja para aplicar la limpieza y organización del CDV es un archivo excel proveído por INCAA:

1. INCAA Dataset.xslx.

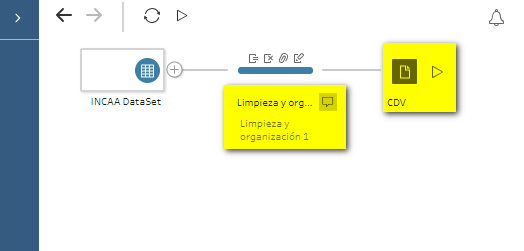


Figura 22: Limpieza y organización del CDV en Tableau Prep

|  |  |
| --- | --- |
|  | Paso 2 |
| Limpieza 1  INCAA Dataset.xslx | **ID:** DS12…  **Nombre:** Se obtiene del campo *título*, se eliminan los espacios innecesarios.  **Categoría:** Cine.  **Fecha Lanzamiento:** Se obtiene del campo *fecha\_estreno.*  **Género:** Se obtiene del campo *género* y se eliminan los espacios innecesarios. Los géneros se encuentran descritos en español, por ende se procede a reemplazar los valores por los géneros detectados en los otros conjuntos de datos que están detallados en inglés.  **País:** Se obtiene del campo *origen\_pelicula,*  donde existen dos posibilidades: si el valor es *“Nac”* el país es Argentina, si el valor es *“Ext”* el país no está definido (ND).  **Ingresos (Valor supuesto):** Para el cálculo de ingresos por película se tomaron en cuenta los precios de entradas en el rango de años desde 2013 hasta 2018. Según los datos de fiscalización de INCAA [[48](#INC)], se tomaron los siguientes precios para sacar un promedio de cada año:   * **2013:** $35,25 * **2014:** $44,14 * **2015:** $59,18 * **2016:** $77,33 * **2017:** $94,49 * **2018:** $116,14   Además, para obtener el valor equivalente en dólares, se obtuvo la cotización del dólar del último día del año para el mismo rango según datos históricos del Banco Nación [[49](#Cot)]:   * **2013:** $6,48 * **2014:** $8,48 * **2015:** $12,90 * **2016:** $15,82 * **2017:** $18,48 * **2018:** $36,96   Con estos datos, se realiza el respectivo cálculo utilizando el campo *entradas\_total.*  **Puntaje (Valor supuesto):** Se obtiene de multiplicar por 10 y dividir entre 4.972.400 (cantidad máxima de entradas vendidas) el campo *entradas\_total*. Se redondea el valor a dos decimales.  **Canal:** El valor es “C” debido a que son datos informativos de películas estrenadas en salas de cine. |

Tabla 3: Proceso de limpieza y organización del CDV

1. **Modelado predictivo**

En esta instancia se procede a construir el modelo predictivo utilizando como referencia las distintas variables identificadas como relevantes en el conjunto de datos y aplicando diferentes algoritmos de predicción para determinar el modelo más conveniente en la hipótesis planteada.

Para generar el modelo predictivo se utiliza la herramienta de Microsoft en entorno Azure denominada *Machine Learning Studio* [[50](#Azu)] que de manera muy visual permite desarrollar, probar y trabajar modelos de aprendizaje automatizado de forma práctica.

|  |  |  |
| --- | --- | --- |
|  | **Gratis** | **Estándar** |
| Precio | Gratis | USD 9,99 por espacio de trabajo de estudio ML por mes. |
| Número máximo de módulos por experimento. | 100 | Ilimitado |
| Duración máxima del experimento | 1 hora por experimento | Hasta 7 días por experimento con un máximo de 24 horas por módulo. |
| Max espacio de almacenamiento | 10 GB | Ilimitado |

Tabla 4: Precios y características de Machine Learning Studio

Los algoritmos utilizados para entrenar los modelos predictivos de prueba que nos brinda la herramienta de Microsoft son los siguientes:

* **Linear Regression (Regresión Lineal):** La regresión lineal intenta establecer una relación lineal entre una o más variables independientes y un resultado numérico o la variable dependiente.
* **Boosted Decision Tree Regression (Regresión de Árbol de Decisión Potenciado):** *Potenciación* significa que cada árbol depende de árboles anteriores. El algoritmo aprende ajustando el valor residual de los árboles que le preceden. Por lo tanto, la potenciación de un conjunto de árboles de decisión tiende a mejorar la precisión.
* **Neural Network Regression (Regresión de Redes Neuronales):** Aunque las redes neuronales son ampliamente conocidas por su uso en problemas complejos de aprendizaje profundo y modelado, como el reconocimiento de imágenes, se adaptan fácilmente a los problemas de regresión. Cualquier clase de modelo estadístico puede denominarse red neuronal si utiliza pesos adaptables y puede aproximar funciones no lineales de sus entradas. Por lo tanto, la regresión de red neuronal es adecuada en los problemas para los que un modelo de regresión más tradicional no puede encontrar una solución.

Las variables que se buscan predecir dentro del conjunto de datos, son las siguientes:

* **Puntaje:** Esta variable da una idea general de la valoración de la nueva producción entre los críticos y el público.
* **Ingresos:** Esta variable espera obtener un aproximado de ganancias que generará la nueva producción.

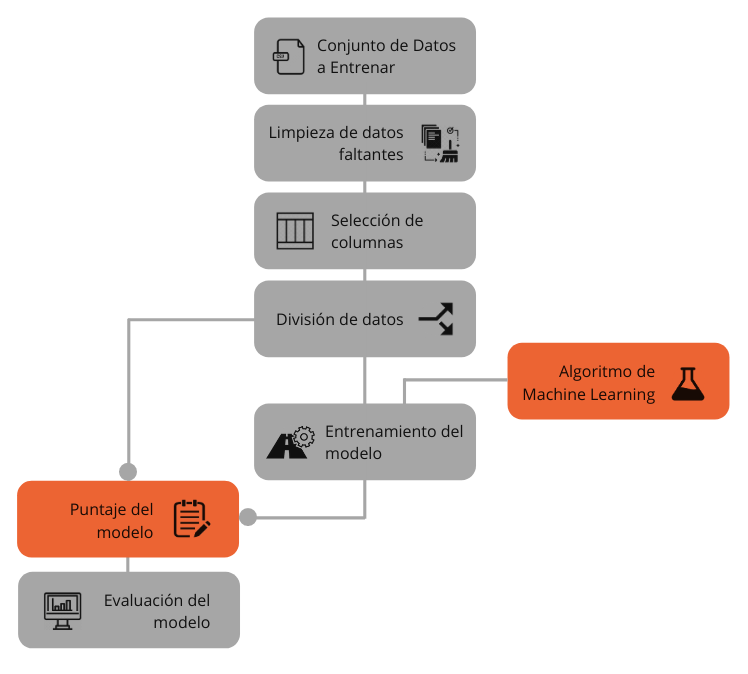
****

Figura 23: Proceso general de modelado predictivo

A continuación se detallan los modelos resultantes de los diferentes algoritmos en función a las variables a predecir; en todos los casos se realiza el mismo procedimiento, variando únicamente la configuración propia de cada algoritmo.

1. Para iniciar flujo del modelado, se selecciona el conjunto de datos o sea el CDE (Conjunto de datos para Entrenar).
2. Se aplica una limpieza de datos faltantes o nulos en caso de que corresponda.
3. Se seleccionan las columnas que serán tomadas en cuenta, en este caso son todas.
4. Se aplica una “separación de datos” que se utiliza para dividir el conjunto de datos en dos partes, la primera será para entrenar el modelo y la segunda para puntuarlo. La fracción de filas en el primer subconjunto de datos de salida será de 0.7 quedando para el segundo subconjunto 0.3.
5. Para entrenar el modelo se debe seleccionar una variable a predecir y asociarla a un algoritmo predictivo. En este caso se necesitan dos módulos para cada variable (Ingresos/Puntaje).
6. Se procede a configurar el respectivo algoritmo:

* **Linear Regression**
* Método de solución: **Mínimos cuadrados**, calcula el error como la suma del cuadrado de la distancia entre el valor real y la línea predicha, y ajusta el modelo minimizando el error cuadrático.
* Regularización de peso L2: **0.001**, se recomienda usar un valor distinto de cero para evitar el sobreajuste.
* **Boosted Decision Tree Regression**
* Crear en modo entrenador: **Único parámetro**.
* Número máximo de hojas por árbol: **20,** en función a este valor puede crecer el árbol y aumentar la precisión.
* Número mínimo de muestras por nodo hoja: **10,** al aumentar este valor aumenta el umbral para crear reglas nuevas.
* Velocidad de aprendizaje: **0.2,** el valor debe estarentre 0-1 y si el tamaño es demasiado grande, puede pasar por alto la solución más óptima.
* Número de árboles construidos: **100,** al crear más árboles de decisión, se puede obtener una mejor cobertura, pero el tiempo de entrenamiento aumenta.
* **Neural Network Regression**
* Crear en modo entrenador: **Único parámetro**.
* Especificación de capa oculta: **Caso completamente conectado**, crea un modelo mediante la arquitectura predeterminada de red neuronal.
* Número de nodos ocultos: **100,** el valor predeterminado es una capa oculta con 100 nodos.
* Velocidad de aprendizaje: **0.005,** define el paso llevado a cabo en cada iteración, antes de la corrección.
* Número de iteraciones de aprendizaje: **100,** número máximo de veces que el algoritmo procesa los casos de entrenamiento.
* El diámetro inicial del peso de aprendizaje: **0.1** (configuración por defecto)**.**
* El momentum: **0** (configuración por defecto)**.**

1. Se agregan módulos para obtener la puntuación del modelo. Para esto, se necesita el segundo subconjunto de datos de salida del CDE y el resultado del entrenamiento del algoritmo.
2. Finalmente se agregan módulos para evaluar el modelo resultante, donde se obtiene el rendimiento de cada algoritmo en función a la variable.

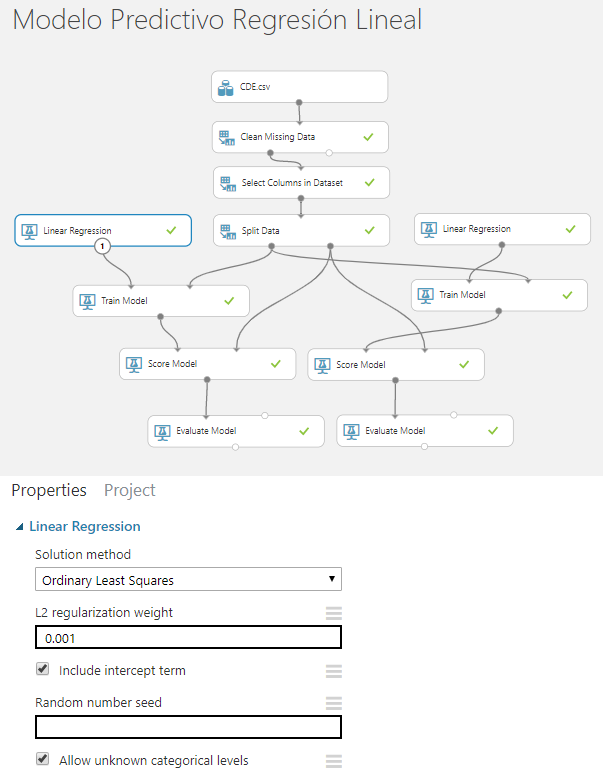


Figura 24: Modelo predictivo - Regresión Lineal

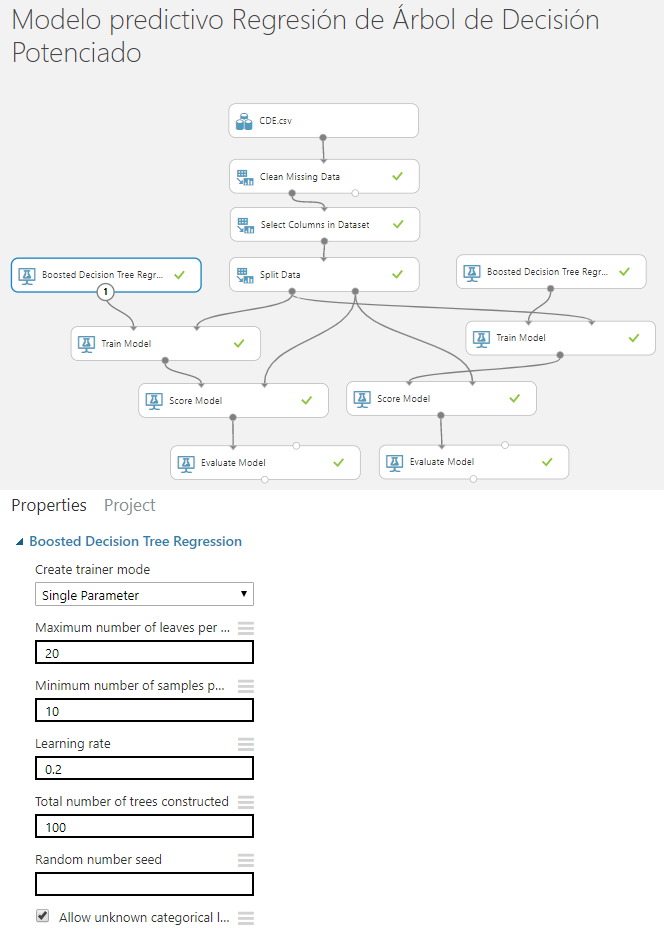


Figura 25: Modelo predictivo - Regresión de Árbol de Decisión Potenciado



Figura 26: Modelo predictivo - Regresión de Redes Neuronales

1. **Estimación de rendimiento**

Para evaluar el rendimiento de los modelos predictivos se tomarán en cuenta los siguientes valores resultantes:

* **Coeficiente de Determinación:** es también una manera estándar de medir cuánto se adapta el modelo a los datos. Se puede interpretar como la proporción de la variación que explica el modelo. Una mayor proporción es mejor en este caso, donde 1 indica un ajuste perfecto.
* **Error Relativo Cuadrado:** el término "error" representa la diferencia entre el valor predicho y el valor verdadero. Normalmente, se calcula el valor cuadrado de esta diferencia para capturar la magnitud total de errores en todas las instancias, dado que la diferencia entre el valor real y el predicho puede ser negativa en algunos casos. Las métricas de error miden el rendimiento de predicción de un modelo de regresión en cuanto a la desviación media de sus predicciones a partir de los valores reales. Los valores de error más bajos implican que el modelo es más preciso a la hora de realizar predicciones. [[51](#Mic)]

En la tabla 5 se detallan los resultados del Coeficiente de Determinación y Error Relativo Cuadrado, donde se puede concluir que la variable con mejor resultado para predecir es el **Puntaje**. Respecto al algoritmo, se puede observar que **Boosted Decision Tree Regression** obtiene el mejor resultado de rendimiento.

|  |  |  |
| --- | --- | --- |
| **Algoritmos *(Variable)*** | **Coeficiente de Determinación** | **Error Relativo Cuadrado** |
| Linear Regression *(Puntaje)* | 0.512922 | 0.487078 |
| Boosted Decision Tree Regression *(Puntaje)* | **0.812418** | **0.187582** |
| Neural Network Regression *(Puntaje)* | 0.765191 | 0.234809 |
| Linear Regression *(Ingresos)* | 0.037553 | 0.962447 |
| Boosted Decision Tree Regression *(Ingresos)* | 0.128192 | 0.871808 |
| Neural Network Regression *(Ingresos)* | -0.000215 | 1.000215 |

Tabla 5: Comparativa de estimación de rendimiento de los distintos modelos predictivos

# EXPERIMENTOS

Una vez seleccionado y entrenado el modelo, está listo para convertirse en un experimento predictivo para puntuar nuevos datos, es decir, el CDV (Conjunto de Datos para Validar). Al efectuar la conversión de un experimento de entrenamiento a un experimento predictivo, se prepara el modelo entrenado para implementarlo como servicio web. En la figura 27, se puede observar que el experimento predictivo tiene un diseño de diagrama de módulos donde se agrega como datos de entrada el CDV y del mismo se seleccionan todas las columnas, excepto la columna escogida como variable a predecir: **el puntaje**. Cabe recalcar que el CDV, tiene valores supuestos en las columnas Ingresos y Puntaje, que en función a los resultados del modelo entrenado se descarta una variable para que sea tomada como el valor a predecir.

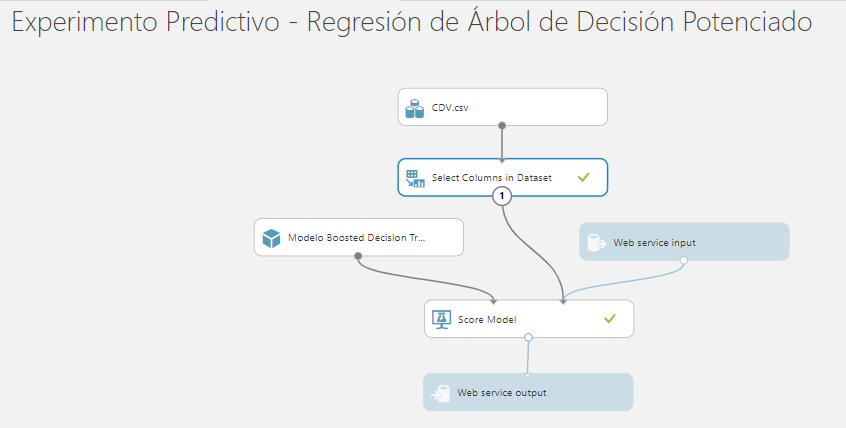


Figura 27: Experimento predictivo – Regresión de Árbol de Decisión Potenciado con CDV

El objetivo del experimento es obtener la predicción del puntaje del conjunto de datos y en función a ellos, analizar la relación de las variables seleccionadas con el valor predicho. A continuación se detalla en las figura 28, un fragmento de las predicciones generadas a nivel general y en la figura 29 filtradas por el país Argentina.



Figura 28: Predicción de Puntaje en el CDV



Figura 29: Predicción de Puntaje en el CDV filtrado por país

# RESULTADOS

Los resultados del experimento realizado con el CDV proveído por INCAA que se pueden observar en función a las variables seleccionadas y la variable predicha se enumeran a continuación:

1. **Relación Nombre/Puntaje:** Cómo se puede observar en la figura 30, la variable *Nombre* es analizada en función a la cantidad de coincidencias de los títulos de las películas y los puntajes obtenidos de las mismas. El gráfico obtiene una muestra de 15 títulos y con esto se interpreta que para este experimento el nombre de una película no es una variable determinante para la predicción del *Puntaje*, ya que no cuenta con los patrones necesarios para impactar en el modelo predictivo.

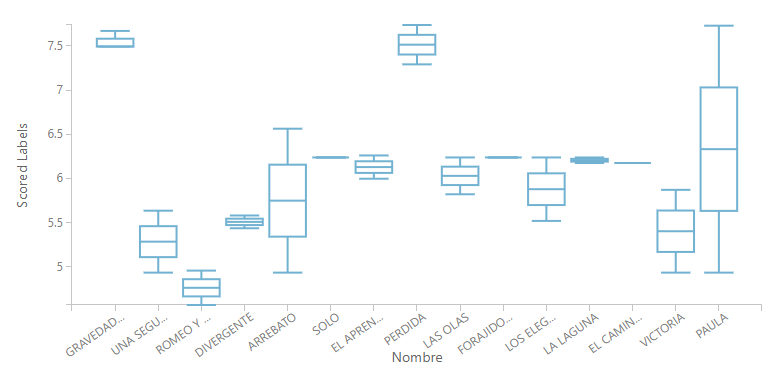


Figura 30: Multibox Plot - Relación entre variables Nombre y Puntaje

1. **Relación Categoría/Puntaje:** Los datos utilizados para este experimento pertenecen a una única categoría: *Cine*. La figura 31 muestra el valor mínimo, máximo, los cuartiles y los valores atípicos que existen dentro del conjunto de datos. Con estos resultados se interpreta que la variable *Categoría* si puede ser determinante para predecir el *Puntaje* debido a que al experimentar con datos pertenecientes a diferentes categorías de las ya definidas, se puede buscar resultados que se aproximen al valor deseado.

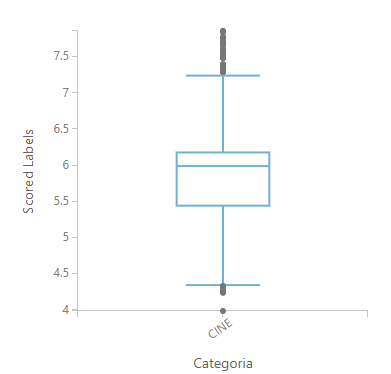


Figura 31: Multibox Plot - Relación entre variables Categoría y Puntaje

1. **Relación Fecha Lanzamiento/Puntaje:**Para identificar e interpretar la relación de la variable para predecir el *Puntaje* se debe agrupar el valor de la *Fecha Lanzamiento* por meses y/o estaciones*.* La figura 32, presenta una muestra de 15 fechas que exponen ciertos patrones agrupados, donde los meses de abril, octubre, noviembre y diciembre, o bien en las estaciones otoño y primavera presentan los puntajes más altos.

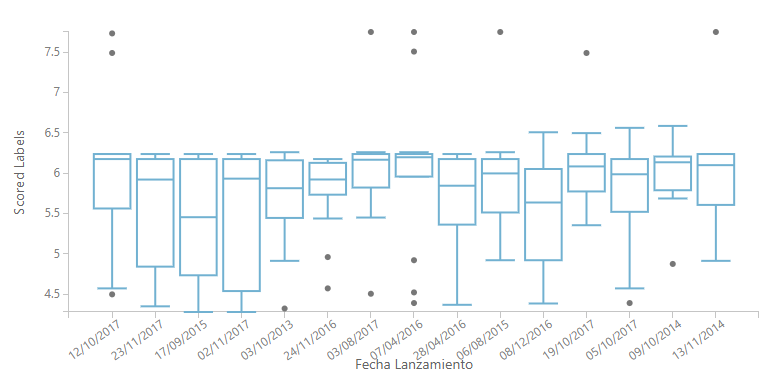


Figura 32: Multibox Plot - Relación entre variables Fecha Lanzamiento y Puntaje

1. **Relación Género/Puntaje:** La variable *Género* es determinante para la predicción del *Puntaje*. Como se puede observar en la muestra de la figura 33, el valor máximo, mínimo, los cuartiles y valores atípicos exponen que los géneros con puntajes más altos son: Drama, Comedia y Crimen*.* El género *No Definido* sería una excepción dentro del experimento.

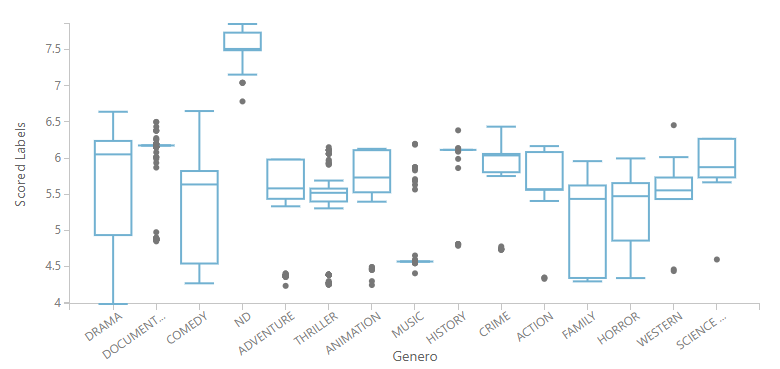


Figura 33: Multibox Plot - Relación entre variables Género y Puntaje

1. **Relación País/Puntaje:** Los datos utilizados para este experimento sólo tienen dos posibles valores como país: *Argentina* y *No Definido*. Con la figura 34, no se puede interpretar realmente que países obtienen el puntaje más alto, sin embargo no se descarta que la variable *País* sea relevante para la predicción del *Puntaje* debido a que el modelo fue entrenado con datos de diferentes países.

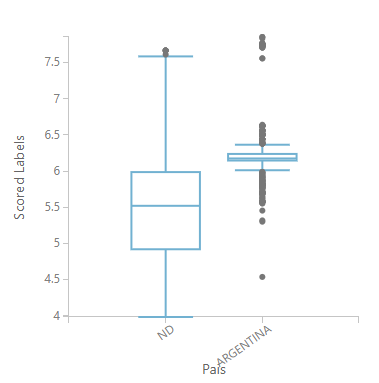


Figura 34: Multibox Plot - Relación entre variables País y Puntaje

1. **Relación Ingresos/Puntaje:** La figura 35 muestra un diagrama de dispersión de datos entre las variables *Ingresos* y *Puntaje*, donde se interpreta que no necesariamente la película que haya recaudado mayores ingresos es la mejor puntuada. La relación observada es que los puntajes oscilan entre 5.5 a 6.5 en función a los ingresos más altos.

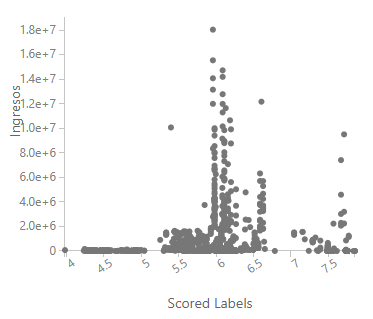


Figura 35: Gráfico de dispersión - Relación entre variables Ingresos y Puntaje

1. **Relación Canal/Puntaje:** Los datos utilizados para este experimento pertenecen a un único canal de distribución: *Sala de cines*. La figura 36 muestra el valor mínimo, máximo, los cuartiles y los valores atípicos que existen dentro del conjunto de datos. Con estos resultados se interpreta que la variable *Canal* se puede utilizar para experimentar con los distintos canales de distribución ya definidos hasta obtener resultados que se aproximen al valor deseado.

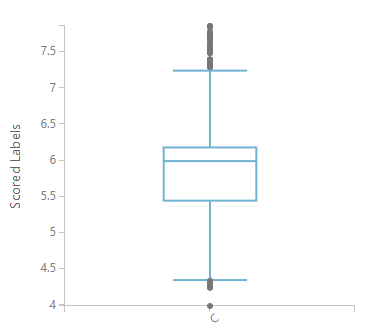


Figura 36: Multibox Plot - Relación entre variables Canal y Puntaje

El modelo predictivo entrenado resultante de soporte a la toma de decisiones en la industria del entretenimiento se detalla en la figura 37:

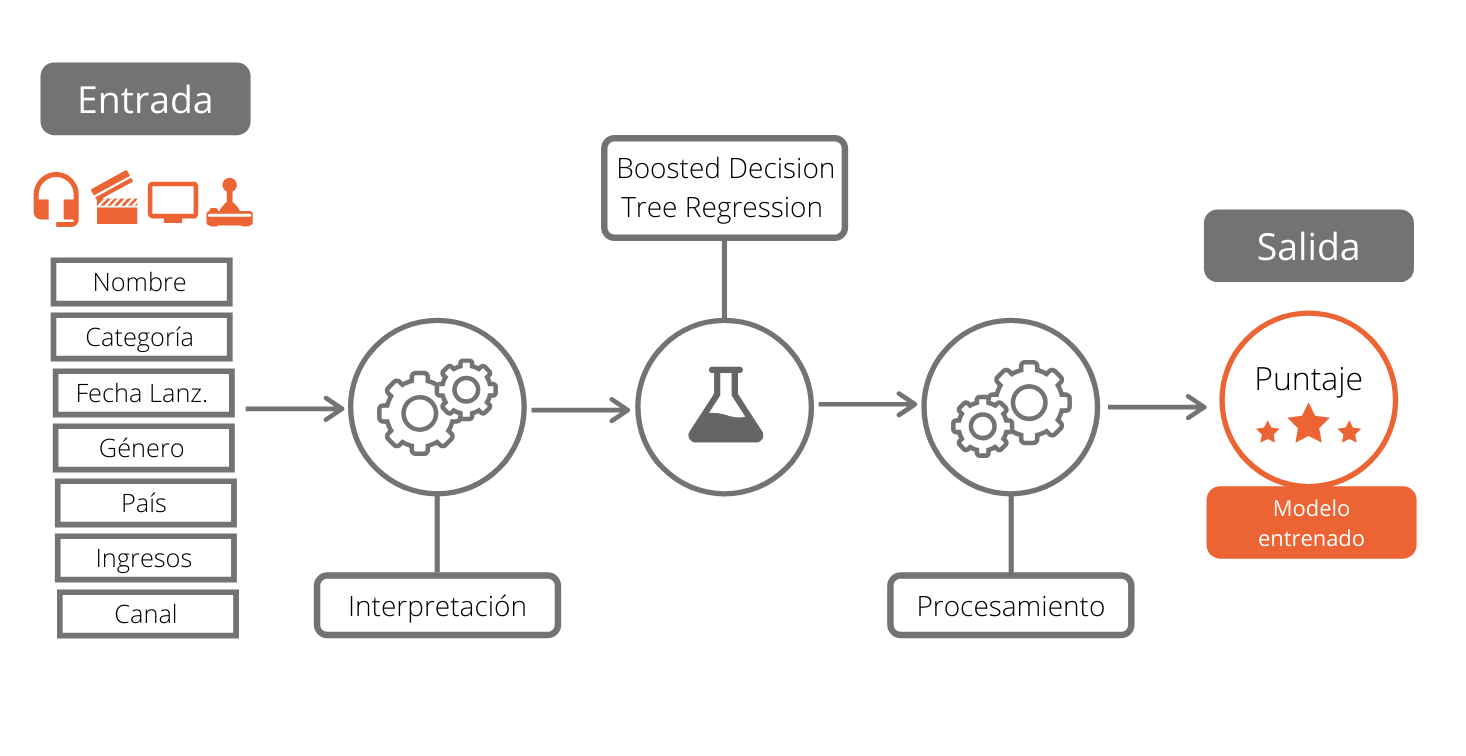


Figura 37: Modelo predictivo de soporte a la toma de decisiones en la industria del entretenimiento

Azure Machine Learning Studio compila una solución de análisis predictivo mediante un servicio web con interfaz que permite realizar pruebas con datos unitarios. A continuación se adjuntan 2 escenarios de predicción con datos ficticios:



Figura 38: Prueba de web service - Escenario de predicción N°1

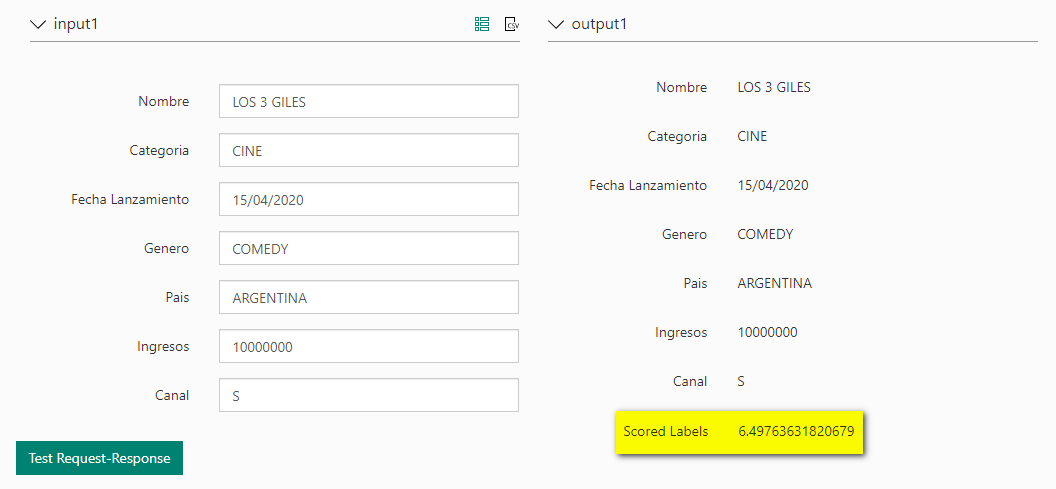


Figura 39: Prueba de web service - Escenario de predicción N°2

# PERSPECTIVAS DE CRECIMIENTO

Como futuras líneas de investigación y experimentación para aplicar a este modelo predictivo se proponen los siguientes puntos:

1. **Recolección de datos sobre ventas en la categoría *Música*:** El modelo predictivo actual fue entrenado con fuentes de datos de distintas categorías dentro de la industria del entretenimiento. Los datos sobre *cine*, *televisión* y *videojuegos* cuentan con información de ventas de las distintas producciones, sin embargo los datos utilizados para entrenar la categoría *música* cuentan únicamente con datos de popularidad de los artistas según la valoración de usuarios. Se propone recolectar datos referentes a las ventas semanales de los álbum de los distintos artistas, y con ello poder el predecir el puntaje en función de las ventas. Además se podría experimentar otro enfoque individualizado para la categoría *música,*que consistiría en predecir cuánto tiempo permanecerá un álbum en la lista de ventas, dados los primeros datos semanales, u observar si la posición de un nuevo álbum se puede predecir en una determinada semana en el futuro. [[52](#Son06)]
2. **Selección de variables específicas por categoría:** El modelo predictivo propuesto selecciona variables comunes dentro de las categorías definidas. Se propone expandir la posibilidad de variables específicas por categoría, para conseguir posibles nuevos factores que puedan afectar en la predicción. Por ejemplo el reparto de una película o serie de televisión puede influir directamente según la popularidad de los actores y actrices.

# CONCLUSIONES

La industria del entretenimiento no es un entretenimiento, es un negocio. El éxito en este entorno empresarial requiere más que grandes artistas y contenido deseable. Requiere de ciencia: la ciencia del análisis predictivo. La toma de decisiones de una nueva producción cinematográfica, televisiva, musical o de videojuegos, desde la idea inicial hasta la publicidad y canal de distribución, tiene que estar respaldada por un análisis previo basado en datos históricos que muestren tendencias, patrones y comportamientos que fundamenten dichas decisiones.

En el presente documento, se diseñó un modelo de análisis predictivo que pueda ajustarse a cuatro categorías identificadas dentro de la industria del entretenimiento: cine, televisión, música y videojuegos. El modelo tiene la finalidad de predecir de manera general el éxito de la producción mediante la variable *Puntaje* que se representa como un campo numérico de rango 0 a 10. Dicha variable es el resultado de procesar y entrenar el modelo predictivo con el algoritmo *Boosted Decision Tree Regression* de aprendizaje automático con la herramienta Azure Machine Learning Studio de Microsoft. El modelo propuesto recibe cómo parámetros de entrada un conjunto de variables pertenecientes a la nueva producción y permite experimentar con ellas hasta alcanzar la puntuación esperada.

En resumen, el análisis predictivo brinda apoyo para las decisiones humanas. La forma en que los modelos predictivos producen valor en una organización es simple en concepto; hacen posible tomar decisiones correctas, eficientes y con menos gasto. Para la industria del entretenimiento ofrece como ventaja principal el soporte a la toma acertada de decisiones en cuanto a una nueva producción inclusive antes de producirse; por otro lado, ofrece la posibilidad de explorar nuevos campos en el negocio e invertir en oportunidades con las expectativas y resultados esperados.

# BIBLIOGRAFÍA

x

|  |  |
| --- | --- |
| [1] | M., & Mcclarren, R. Gold. (2013) THE LESSONS OSCAR TAUGHT US: 1(2), 105–109. [Online]. <https://doi.org/10.1089/big.2013.0009> |
| [2] | N., & Silakari, S. Mishra. (2012) Predictive Analytics: A Survey, Trends, Applications. [Online]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.442.2759&rep=rep1&type=pdf> |
| [3] | LOGICALIS. Predictive Analytics - El impacto de la predicción para una organización. [Online]. <https://cdn2.hubspot.net/hubfs/273205/01-LOG/ANALYTICS/Ebooks%20y%20CTAs/Predictive%20Analytics/74_17_LOG_Predictive%20Analytics%20El%20impacto%20de%20la%20predicci%C3%B3n%20para%20una%20organizaci%C3%B3n.pdf> |
| [4] | J. Dysart. (2014) Predictive Analytics. National Underwriter Life & Health, 39–40. [Online]. <http://doi.org/10.1007/978-3-658-14539-2_5> |
| [5] | Deloitte. Social Analytics in Media & Entertainment - The three-minute guide. [Online]. <https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Deloitte-Analytics/dttl-analytics-us-da-3minMediaandEntertainment.pdf> |
| [6] | A. Aloysius S. Banumathi. (2017) PREDICTIVE ANALYTICS CONCEPTS IN BIG DATA- A SURVEY. [Online]. <http://dx.doi.org/10.26483/ijarcs.v8i8.4628> |
| [7] | Ricardo TIMARÁN PEREIRA and Andrés & JIMÉNEZ TOLEDO, Javier CALDERÓN ROMERO. (2013) Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil. [Online]. <http://revistasum.umanizales.edu.co/ojs/index.php/ventanainformatica/article/download/181/228/> |
| [8] | Gregory Piatetsky-Shapiro, Padhraic Smyth Usama Fayyad. (1996) Knowledge Discovery and Data Mining: Towards a Unifying Framework. [Online]. <https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf> |
| [9] | M. L. Garg Vaibhav Kumar. Predictive Analytics: A Review of Trends and Techniques. [Online]. <https://www.researchgate.net/publication/326435728_Predictive_Analytics_A_Review_of_Trends_and_Techniques> |
| [10] | Asit Kumar Das, Mausumi, Sashikala Mishra Debahuti Mishra. Predictive Data Mining: Promising Future and Applications. [Online]. <https://pdfs.semanticscholar.org/95ba/af3d56adb8ec8c2d7be0ec7b320f264cb27a.pdf> |
| [11] | Violeta Valcárcel Asencios. (2004) Data Mining y el Descubrimiento del Conocimiento. [Online]. <https://www.researchgate.net/publication/307181857_DATA_MINING_Y_EL_DESCUBRIMIENTO_DEL_CONOCIMIENTO/link/57c432b908aee5141be5bc8f/download> |
| [12] | Pradnya P. Sondwale. (2015) Overview of Predictive and Descriptive Data Mining Techniques. [Online]. <http://ijarcsse.com/Before_August_2017/docs/papers/Special_Issue/ITSD2015/58.pdf> |
| [13] | Musgrave Analytics Ltd. Predictive Analytics. [Online]. <https://www.musgraveanalytics.com/predictive-analytics> |
| [14] | Bernardo A. Huberman Sitaram Asur. (2010) Predicting the Future With Social Media. [Online]. <https://arxiv.org/pdf/1003.5699.pdf> |
| [15] | Jonas Krauss, Stefan Nann, Daniel Simon, and Kai Fischbach. (2008) PREDICTING MOVIE SUCCESS AND ACADEMY AWARDS THROUGH SENTIMENT AND SOCIAL NETWORK ANALYSIS. [Online]. <http://www.ickn.org/documents/Oscar_ECIS_Final_v1.3.pdf> |
| [16] | Thomas Goods. (2016) Social media is changing the entertainment industry. [Online]. <https://www.wesrch.com/gp/social-media-is-changing-the-entertainment-industry-1213> |
| [17] | SAS. Predictive Analytics: What it is and why it matters. [Online]. <https://www.sas.com/en_us/insights/analytics/predictive-analytics.html> |
| [18] | Wayne W. Eckerson. (2007) Predictive Analytic - Extending the Value of Your Data Warehousing Investment. [Online]. <http://hosteddocs.ittoolbox.com/sas-predictive-analytics-062508.pdf> |
| [19] | Dan Harris. (2017) 6 Ways Advanced Analytics Will Change Media & Entertainment. [Online]. <https://blog.dataiku.com/6-ways-advanced-analytics-will-change-media-entertainment> |
| [20] | DATAIKU. (2020) AI in Media& Entertainment. [Online]. <https://pages.dataiku.com/ai-media-entertainment> |
| [21] | Lucía Caballero. (2016) Yorokobu: Cómo predicen Spotify y Shazam quién será el próximo Justin Bieber. [Online]. <https://www.yorokobu.es/spotify-y-shazam/> |
| [22] | Laura Siri. (2016) El rol de Netflix en el ecosistema de medios y telecomunicaciones. [Online]. <http://revistahipertextos.org/wp-content/uploads/2016/11/El-rol-de-Netflix-en-el-ecosistema-de-medios-y-telecomunicaciones.-Siri.pdf> |
| [23] | Keyrus Spain. (2017) El éxito de Netflix, una cuestión de Big Data. [Online]. <https://keyrusspainblog.com/2017/06/29/el-exito-de-netflix-una-cuestion-de-big-data/> |
| [24] | Mansoor Iqbal. (2020) Netflix Revenue and Usage Statistics. [Online]. <https://www.businessofapps.com/data/netflix-statistics> |
| [25] | Statista. (2020) Netflix's revenue Q1 2011-Q1 2020. [Online]. <https://www.statista.com/statistics/273883/netflixs-quarterly-revenue/> |
| [26] | Analytics India Magazine. (2016) Analytics in Video Game Industry – Definitely a Game Changer. [Online]. <https://analyticsindiamag.com/analytics-in-video-game-industry-defintely-a-game-changer/> |
| [27] | Ed Burns. (2015) How The New York Times uses predictive analytics algorithms. [Online]. <https://searchbusinessanalytics.techtarget.com/feature/How-The-New-York-Times-uses-predictive-analytics-algorithms> |
| [28] | Evan Koblentz. (2020) New York Times Data Scientist Tells Students to Prioritize Client Communications. [Online]. <https://news.njit.edu/new-york-times-data-scientist-tells-students-prioritize-client-communications> |
| [29] | Oleksii Kharkovyna. (2019) Lights, Camera, Data Science: Predictive Analytics for Film Industry. [Online]. <https://towardsdatascience.com/lights-camera-data-science-predictive-analytics-for-film-industry-b7fe885780da> |
| [30] | Alexandru Niculescu-Mizil Rich Caruana. (2006) An Empirical Comparison of Supervised Learning Algorithms. [Online]. <https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf> |
| [31] | RIKARD KÖNIG. Predictive Techniques and Methods for Decision Support in Situations with Poor Data Quality. [Online]. <http://hb.diva-portal.org/smash/get/diva2:876907/FULLTEXT01.pdf> |
| [32] | Ying LU Yan-yan SONG. (2015) Decision tree methods: applications for classication and prediction. [Online]. <https://www.researchgate.net/publication/279457799_Decision_tree_methods_applications_for_classification_and_prediction> |
| [33] | Allen Y. Chang, Chih-Chung Ho Han-Chen Huang. Using Artificial Neural Networks to Establish a Customer-cancellation Prediction Model. [Online]. <https://www.researchgate.net/publication/290226043_Using_Artificial_Neural_Networks_to_Establish_a_Customer-cancellation_Prediction_Model> |
| [34] | Kaggle. [Online]. <https://www.kaggle.com/> |
| [35] | Kaggle from Wikipedia, the free encyclopedia. [Online]. <https://en.wikipedia.org/wiki/Kaggle> |
| [36] | Instituto Nacional de Cine y Artes Audiovisuales. [Online]. <http://www.incaa.gov.ar/> |
| [37] | The Movie Database. (TMDb) API - The Movie Database API. [Online]. <https://www.themoviedb.org/documentation/api> |
| [38] | Kaggle. (2017) TMDB 5000 Movie Dataset. [Online]. <https://www.kaggle.com/tmdb/tmdb-movie-metadata> |
| [39] | Kaggle. (2017) 1000 Netflix Shows. [Online]. <https://www.kaggle.com/chasewillden/netflix-shows> |
| [40] | Kaggle. (2019) Movie Ratings Dataset. [Online]. <https://www.kaggle.com/raidevesh05/movie-ratings-dataset> |
| [41] | Kaggle. (2018) Feature Film/TV Series. [Online]. <https://www.kaggle.com/nidaguler/feature-filmtv-series-sorted-by-popularity> |
| [42] | last.fm. [Online]. <https://www.last.fm/es/> |
| [43] | Kaggle. (2019) Music artists popularity. [Online]. <https://www.kaggle.com/pieca111/music-artists-popularity> |
| [44] | YouTube. YouTube para desarrolladores - Recursos de API. [Online]. <https://www.youtube.com/intl/es/yt/dev/api-resources/> |
| [45] | Kaggle. (2019) Trending YouTube Video Statistics. [Online]. <https://www.kaggle.com/datasnaek/youtube-new> |
| [46] | Kaggle. (2019) Video Games Sales Dataset. [Online]. <https://www.kaggle.com/sidtwr/videogames-sales-dataset> |
| [47] | TABLEAU SOFTWARE. Tableau Prep. [Online]. <https://www.tableau.com/es-es/products/prep> |
| [48] | INCAA. Fiscalización - Instituto Nacional de Cine y Artes Audiovisuales. [Online]. <http://fiscalizacion.incaa.gov.ar/index_entrada_promedio.php> |
| [49] | Cotizacion-Dolar.com.ar. Dólar Histórico - Cotización histórica del dólar en Argentina. [Online]. <https://www.cotizacion-dolar.com.ar/cotizaciones-dolar-historico.php> |
| [50] | Azure Machine Learning. [Online]. <https://azure.microsoft.com/es-es/services/machine-learning/> |
| [51] | Microsoft. Evaluación del rendimiento de un modelo en Azure Machine Learning Studio (clásico). [Online]. <https://docs.microsoft.com/es-es/azure/machine-learning/studio/evaluate-model-performance> |
| [52] | Malcolm Slaney, Jonathan Berger Song Hui Chon. (2006) Predicting Success from Music Sales Data — A statistical and adaptive approach. [Online]. <https://ccrma.stanford.edu/~shchon/pubs/AMCMM2006.pdf> |

x