77.50 Introducción a los sistemas inteligentes

Trabajo Práctico Final

Integrantes:

Alumno	padron
Llauró, Manuel Luis	95736
Rial, Sebastián Andrés	90309
Blanco, Sebastian Ezequiel	98539

Fecha de Entrega: 24/06/2019

GitHub:

https://github.com/BlancoSebastianEzequiel/7750-TPFinal

Índice

1. Fase 1: Comprensión del negocio

1.1. Determinar los objetivos del negocio

1.1.1. Escenario actual

Al momento de escribir este informe, la empresa en cuestión tiene interés en desarrollar un canal de youtube, y requiere medir de alguna forma las probabilidades de éxito de llevar a cabo videos populares.

1.1.2. Objetivos del negocio

El objetivo sera el de lograr encontrar un algoritmo, y sus parámetos correspondientes, que sean capaces de identificar videos que vayan a ser exitosos en un futuro, y que atributos contribuyen a ello, a partir de un set de datos como el que utilizamos en este trabajo.

1.1.3. Criterios de éxito del negocio

El proyecto se considerará exitoso si se logra un algoritmo capaz de identificar si un video es exitoso o no, con una precision de al menos el 70 %, y que tenga una efectividad de al menos el 30 % en los videos que son exitosos. Ademas se deben detectar al menos 2 atributos que influyan en que los videos del canal de youtube sean exitosos o no, y en qué nivel influye cada una de ellos. Se considerará que un video es exitoso si los mismos obtienen una cantidad de vistas mayor a 2 millones por video.

1.2. Evaluación de la situación

1.2.1. Inventario de recursos

Para desarrollar el presente proyecto se cuenta con una amplia gama de recursos que asegura un desarrollo de calidad y confianza del mismo. Se cuenta con un set de datos extraído de la pagina de kaggle sobre videos de youtube que fueron subidos a la plataforma. Además se cuenta con herramientas de software de análisis y visualización de datos líderes en el mercado, como Pandas. Por último, se cuenta con personal altamente calificado para la correcta interpretación de los mismos.

1.2.2. Requisitos, supuestos y restricciones

1.2.2.1 Requisitos

Contar con datos suficientes y sobre todo representativos de la plataforma youtube

1.2.2.2 Supuestos

Los datos en estudio son lo suficientemente correctos como para poder sacar conclusiones confiables a partir de ellos.

1.2.2.3 Restricciones

Se cuenta solamente con datos de videos subidos a youtube Estados Unidos. No se asegura que los resultados sean válidos para otros países.

1.2.3. Riesgos y contingencias

Si bien se puede llevar a cabo un análisis lo más riguroso posible, siempre existe la posibilidad de que un video pueda no ser exitoso porque hay que tener en cuenta que el exito de cada video puede depender de muchos factores de incertidumbre, y si bien se cumplen patrones, no hay nada que asegure al 100 % el éxito de los mismos. En caso de detectar que un video no está teniendo el éxito esperado, habrá que recurrir a las diversas métricas que pueda ofrecer Youtube a desarrolladores y analizar la situación

1.2.4. Terminología

1.2.4.1 Glosario de términos del negocio

Youtube: es una plataforma que te permite subir videos a partir de la creación de un canal.

Canal de youtube: Es el equivalente a crearse una cuenta en cierta aplicacion, en la cual se almacenan los videos subidos

Categoría: define el nombre de un grupo de vides con cualidades comunes.

likes/dislikes: Es un atributo de un video que muestra la cantidad de usuarios que le dieron like al video en cuestion

Vistas: Es la cantidad de veces que el video fue visto

Video exitoso: Aquel cuya cantidad de vistas es mayor a 200000

1.2.4.2 Glosario de términos de la minería de datos

Atributo: dato sobre alguna característica de las observaciones.

Atributo relevante: atributo que juega un papel principal en la clasificación, por lo que la clase dependerá en alguna medida de qué valor tenga.

Registro: fila que representa una observación, está compuesto de atributos.

Dataset: conjunto de datos a ser utilizados para la ejecución de los algoritmos de Data Mining, está compuesto de registros.

Filtrado de atributos: específica al dataset formado considerando sólo los atributos relevantes.

Regla: es una implicación, que representa una acción mediante una condición. Sigue la estructura "Si..., entonces...".

Soporte: es la relación entre la cantidad total de registros del dataset que cumplen la regla y la cantidad de observaciones procesadas.

Confianza: es la relación entre la cantidad total de observaciones de la clase mayoritaria que cumplen la regla y la cantidad de observaciones que fueron afectadas por esa misma regla.

Captura: es la relación entre la cantidad de observaciones de la clase mayoritaria que cumplen la regla y la cantidad de observaciones procesadas pertenecientes a esa misma clase.

Coeficiente de correlación de Pearson: es la estadística de prueba que mide la relación estadística, o asociación, entre dos variables continuas. Es conocido como el mejor método para medir la asociación entre variables de interés porque se basa en el método de covarianza. Da información sobre la magnitud de la asociación, o correlación, así como la dirección de la relación

1.2.5. Costos y beneficios

El beneficio del proyecto es detectar las características que hacen a un video de youtube sea exitoso. De esta manera se pueden tener en cuenta ciertos parametros para poder desarrollar un video que exitoso basandose en la historia. Al ser un trabajo final educativo, no hay costos.

1.3. Determinar objetivos de Minería de Datos

1.3.1. Objetivo de minería de datos

El objetivo de minería de datos es el análisis de los datos obtenidos a partir de la información disponible, buscando obtener así información relevante que permita predecir las condiciones bajo las cuales una aplicación es exitosa.

1.3.2. Criterios de éxito de minería de datos

Selección de al menos 4 reglas con captura y soporte mayor o igual al 3 % y con una confianza mayor al 60 %

1.4. Realizar el Plan del Proyecto

1.4.1. Plan de proyecto

- Recolección de datos: 5 horas.
- Preparación de datos: 5 horas.
- Ejecución del algoritmo de Inducción: 4 horas.
- Análisis de resultados de algoritmo de Inducción: 4 horas.
- Combinación de resultados: 3 horas.
- Elaboración de reporte: 7 horas.

1.4.2. Validación inicial de las herramientas

Se utilizarán las siguientes herramientas:

- Python
- Pandas
- Weka
- Jupiter Notebook
- Numpy

2. Fase 2: Comprensión de los datos

2.1. Recolectar los datos Iniciales

Los datos utilizados durante el transcurso del proyecto fueron obtenidos gratuitamente en el sitio de Kaggle y el mismo se encuentra en formato csv.

2.1.1. Reporte de la recolección de datos iniciales

- El data set se encuentra en la carpeta `src/data/` del proyecto.
- Se uso pandas como metodo de recoleccion de datos.
- Para poder levantar el set de datos se usa como separado una coma para separar por atributos

2.2. Descubrir datos

2.2.1. Reporte de descripción de datos

Se cuenta con dos datasets: por un lado se tienen datos estadísticos y generales de los videos descritos en 16 columnas; por otro lado hay un archivo json con el nombre de cada categoria segun el numero asignado. A continuación, los nombres de los features con su correspondiente descripción y tipo de dato:

Mombre de variable	tipo de dato	Descripcion
video_id	alfanumerica	identifica a cada video
trending_date	date	Es una fecha en un dia especifico luego de la publicacion del video
title	alfanumerica	Es el titulo del video
channel_title	alfanumerica	Es el nombre del canal del usuario que publico el video
category_id	numeric	it is the number that represent a category name
publish_time	timestamp	momento exacto de la publicacion del video
tags	alfanumerico	Etiquetas que se pueden agregar al video
views	numeric	Cantidad de vistas del video en el momento de la fecha de trending_date
likes	numeric	Cantidad de me gusta que el video tiene dados por usuarios de la plataforma youtube
dislikes	numeric	Cantidad de no me gusta que el video tiene dados por usuarios de la plataforma youtube
comment_count	numeric	Cantidad de comentarios del video
thumbnail_link	url	enlace a
comments_disabled	booleano	Dice si tiene o no los comentarios habilitados
ratings_disabled	booleano	Dice si tiene o no los likes/dislikes habilitados
video_error_or_removed	booleano	Dice si el video sufrio un error o fue borrado en el momento de la fecha de trending_date
description	alfanumeric	Texto escrito por el creador del video que describe el mismo

2.3. Exploración de los datos

2.3.1. Reporte de exploración de datos

2.3.1.1 Analisis de categorias

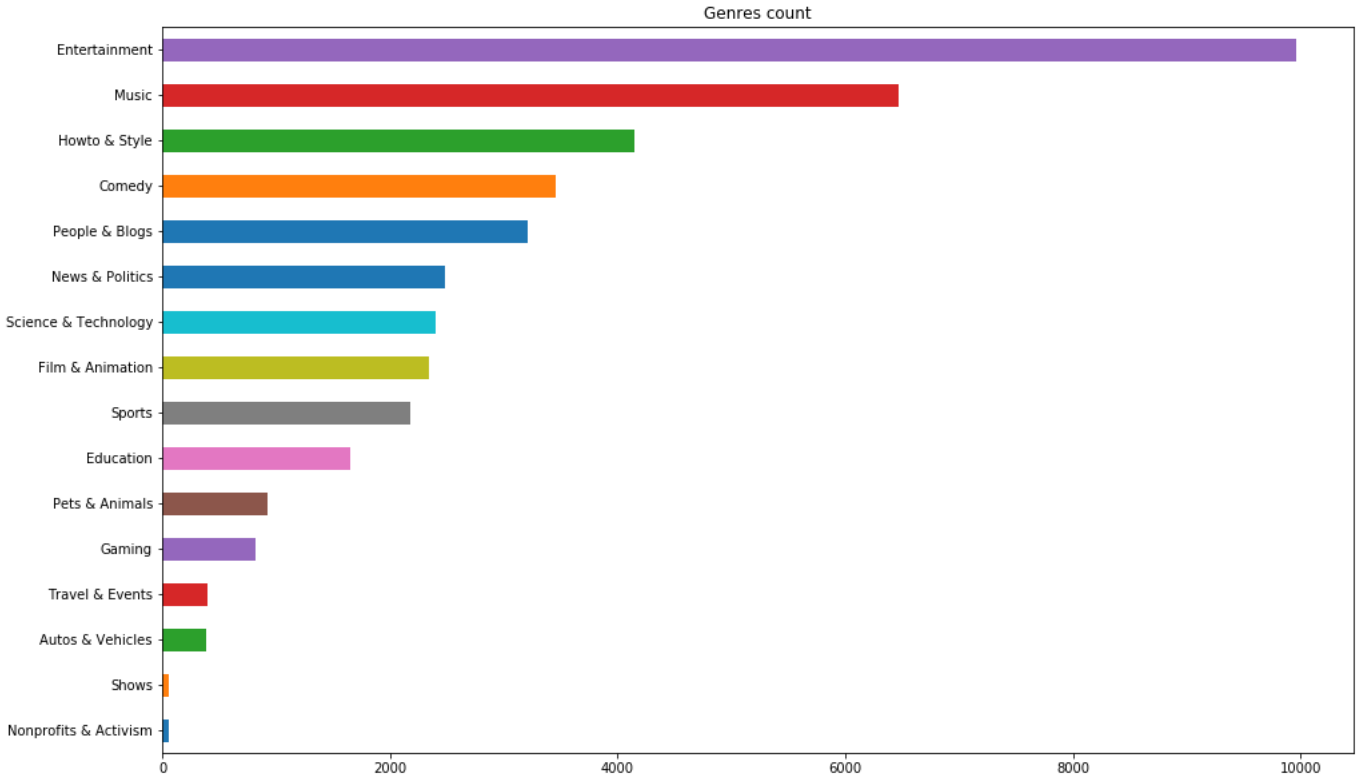

Figura 1: Cantidad de videos segun la categoria

Podemos observar que las categorias de entretenimiento y musica son las que mas predominan, sobre todo entretenimeinto lo cual tiene sentido ya que son de las categorias que mas se tienden a consumir

2.3.1.2 Analisis de vistas en funcion de likes

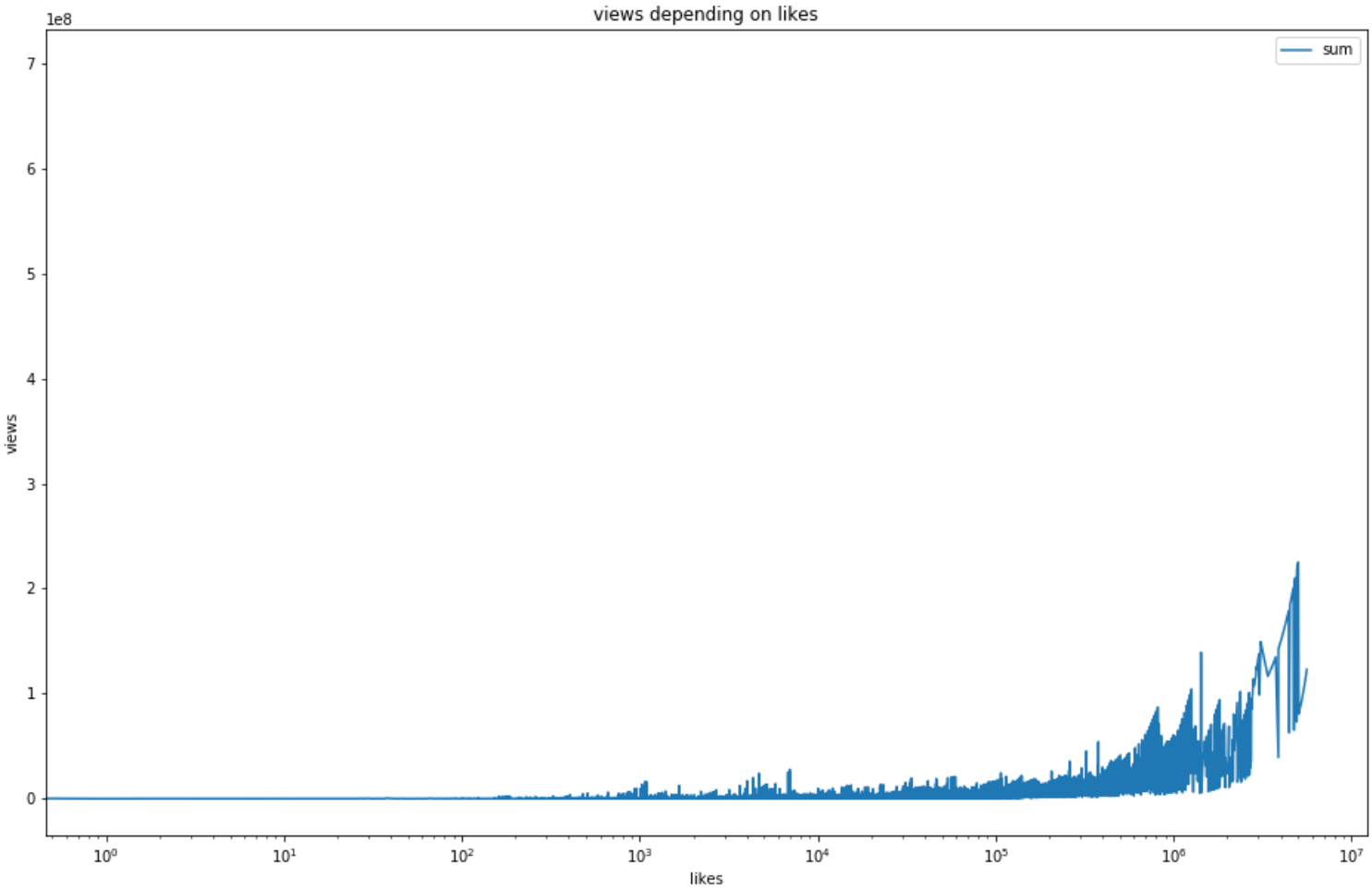

Figura 2: Cantidad de vistas segun la cantidad de likes

Podemos observar que hay un punto de quiebre en el cual a partir de cierta cantidad de likes los videos tienden a aumentar su cantidad de vistas

2.3.1.3 Analisis de vistas en funcion de la diferencia de dias entre la fecha de tendencia y la de publicacion

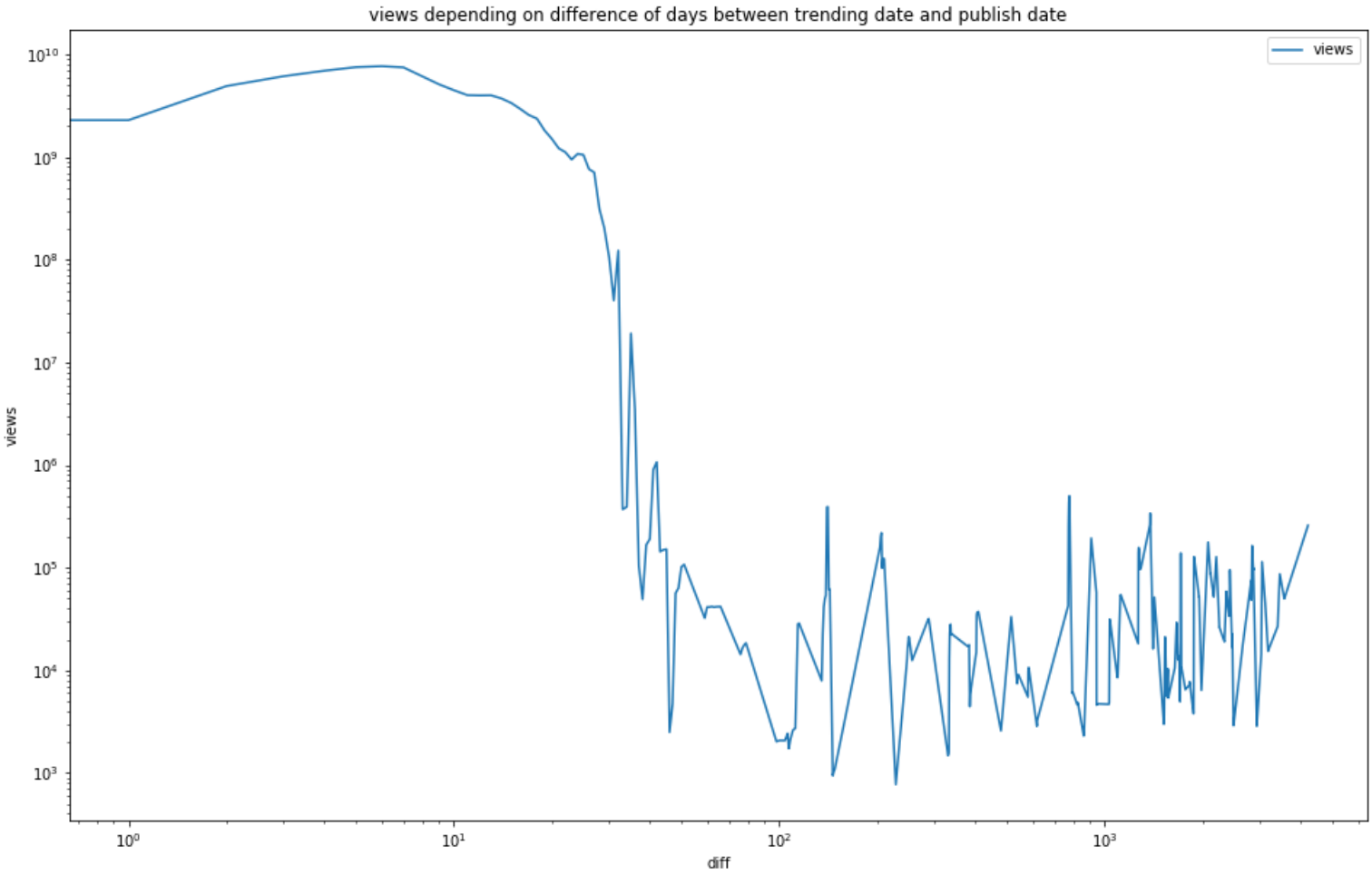

Figura 3: Cantidad de vistas a medida que pasan los dias de su publicacion

Podemos observar que el mayor incremento en la cantidad de vistas es a los pocos dias de la fecha de publicacion del video. Luego vemos que las vistan tienden a decaer

2.3.1.4 Analisis de vistas en funcion de la diferencia de likes y dislikes

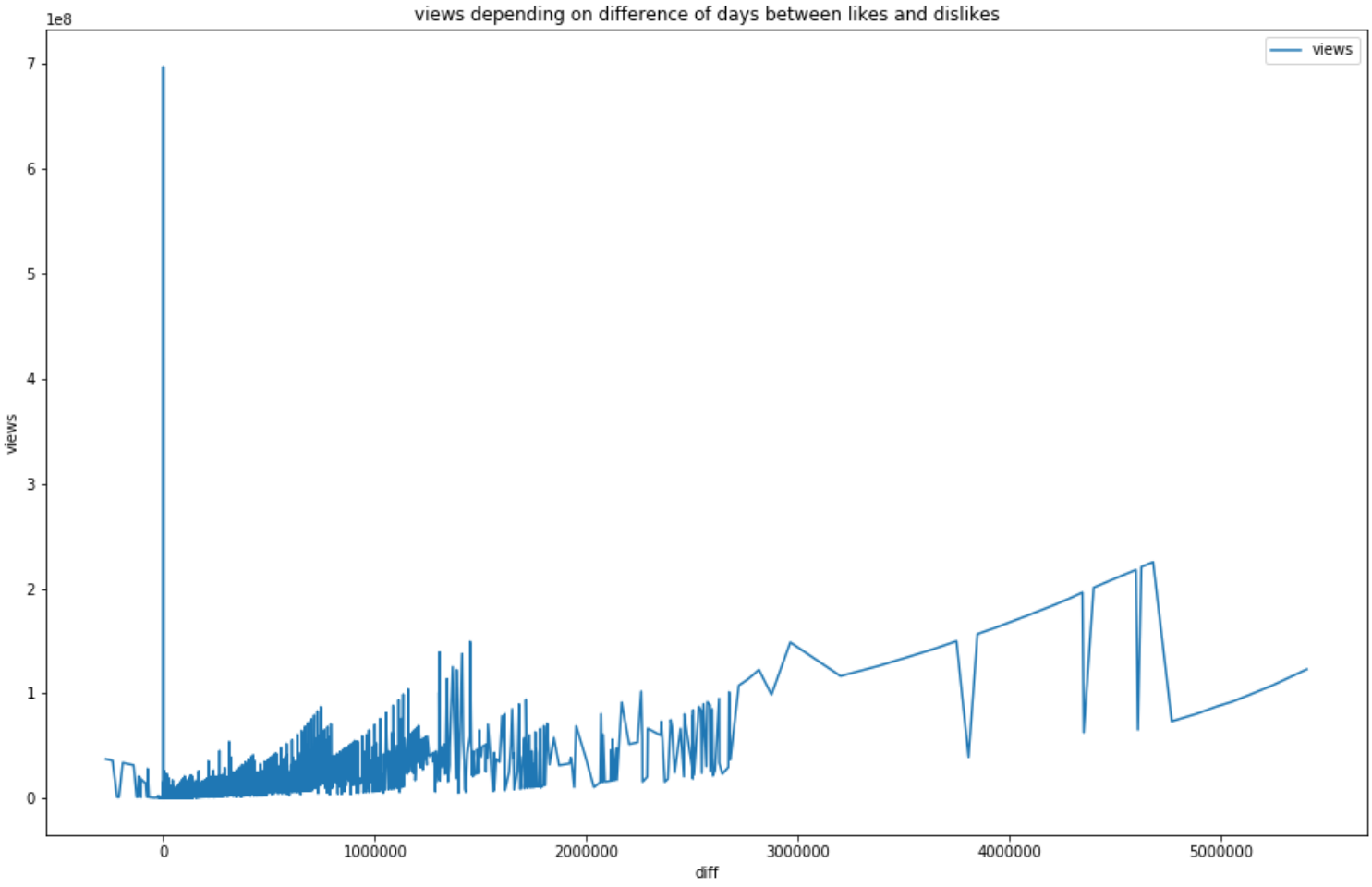

Figura 4: Cantidad de vistas segun la diferencia entre likes y dislikes

Podemos observar a medida que aumenta la distancia entre likes y dislikes, aumentan la cantidad de vistas. Esto es logico ya que en general un video tiene mas vistas cuanto mas gente descata que el mismo le gusto.

2.3.1.5 Analisis del progreso de views del video con mayor views del set de datos.

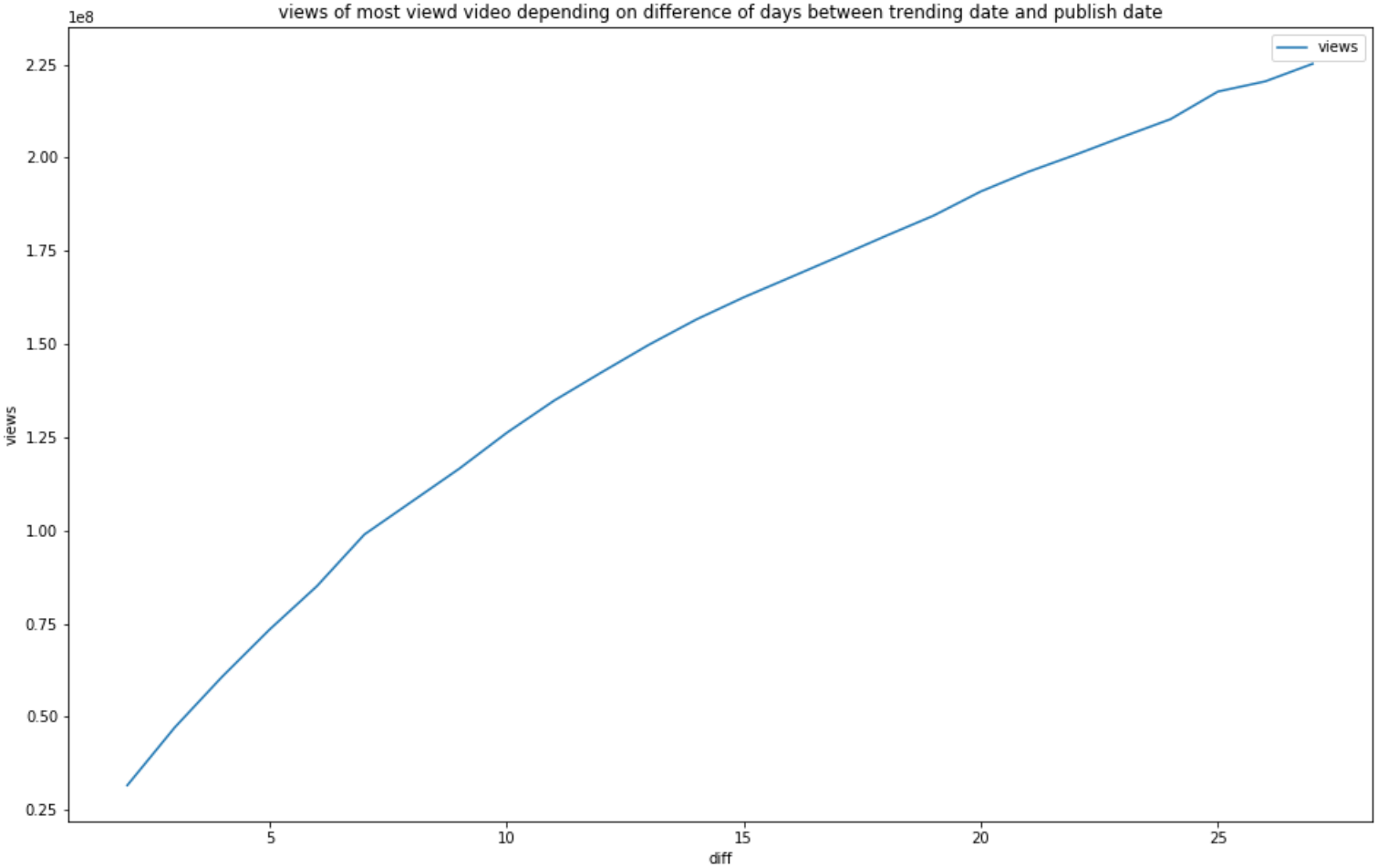

Figura 5: Progreso en dias de las vistas del video mas vistas

Podemos observar que a medida que pasan los dias respecto de la fecha de publicacion del video con mas vistas del set de datos, las vistas del mismo suben linealmente, lo cual tiene sentido ya que cuando un video es exitoso, sus vistan tienen a aumentar progresivamente.

2.3.1.6 Analisis del progreso de views de los videos con mayor, mediano y menor vistas del set de datos.

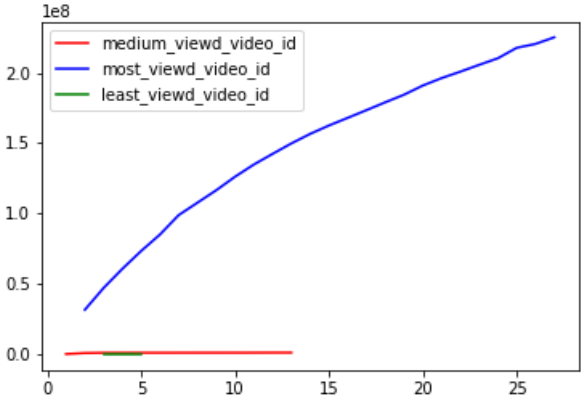

Figura 6: Progreso en dias de las vistas del video mas, mediano y menos vistas

Podemos observar que el video menos visto y el video cuya cantidad de vistas es la mediana del set de datos tienen un progreso de variacion de vistas en el tiempo parecido. Y el video mas visto muestra una amplia diferencia. Esto tiene sentido ya que para tener un progreso como el que tiene el video mas visto se tiene que tener ciertas caracteristicas parecidas. Esto nos dice que hay un grupo pequeño cercano al video mas visto que logra este progreso. y el resto se parece al menos visto

2.3.1.7 Analisis de la cantidad de views segun el largo del titulo del video.

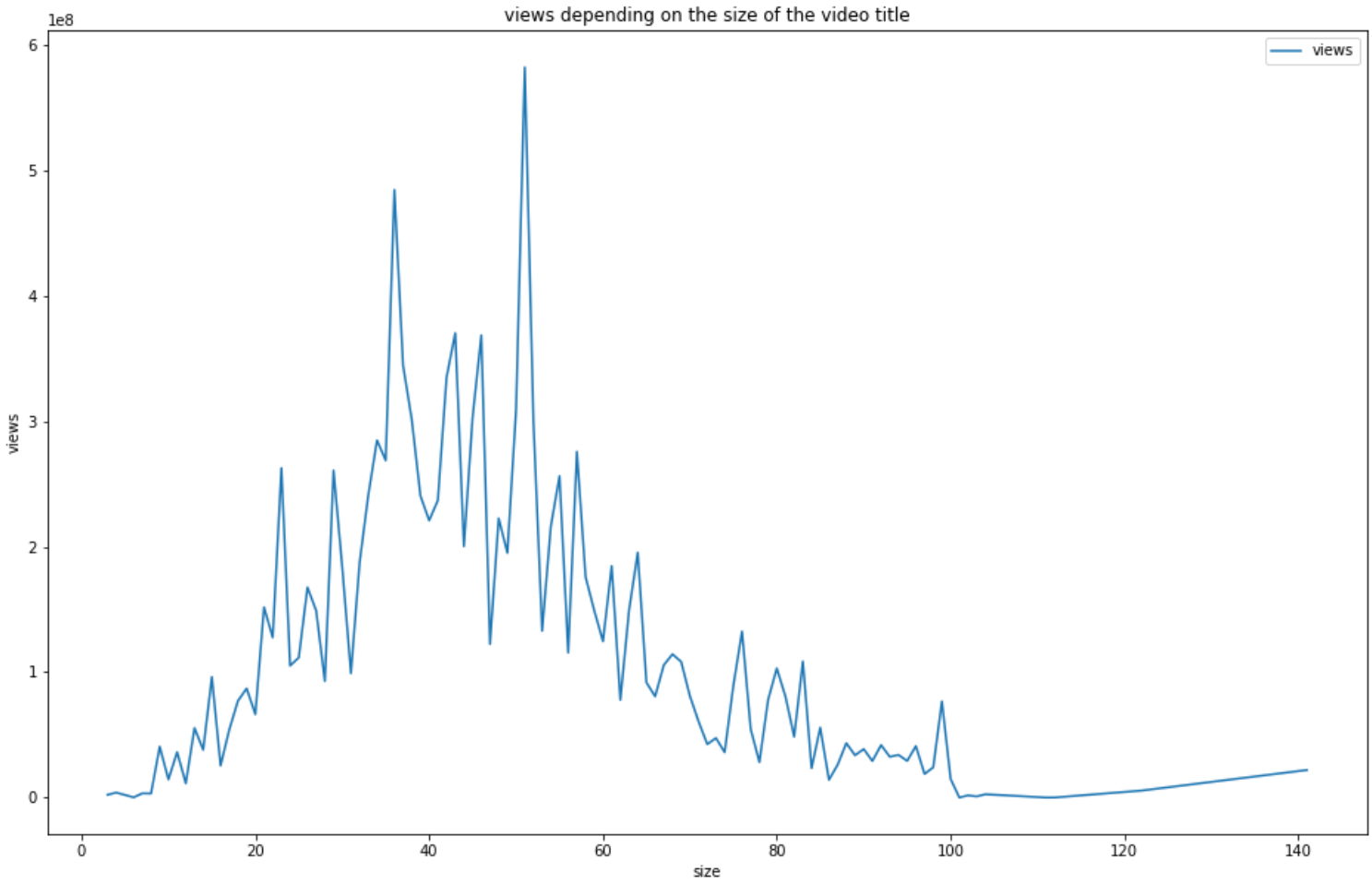

Figura 7: Cantidad de vistas segun el largo del video

Podemos observar que el largo del titulo tiene un rango de valores para su largo en el cual la cantidad de vistas del video logra un valor maximo. Esto tiene sentido ya que el ser humano tiende a leer cosas cortas que generen impacto y es por eso que este rango esta cercano al cero. Tambien si el titulo es muy corto es logico que las vistas no sean altas ya que quizas signifique que ese titulo corto no describe con impacto el video y por lo tanto no genera una atraccion para que la gente lo mire.

2.3.1.8 Analisis la cantidad de videos borrados por categoria.

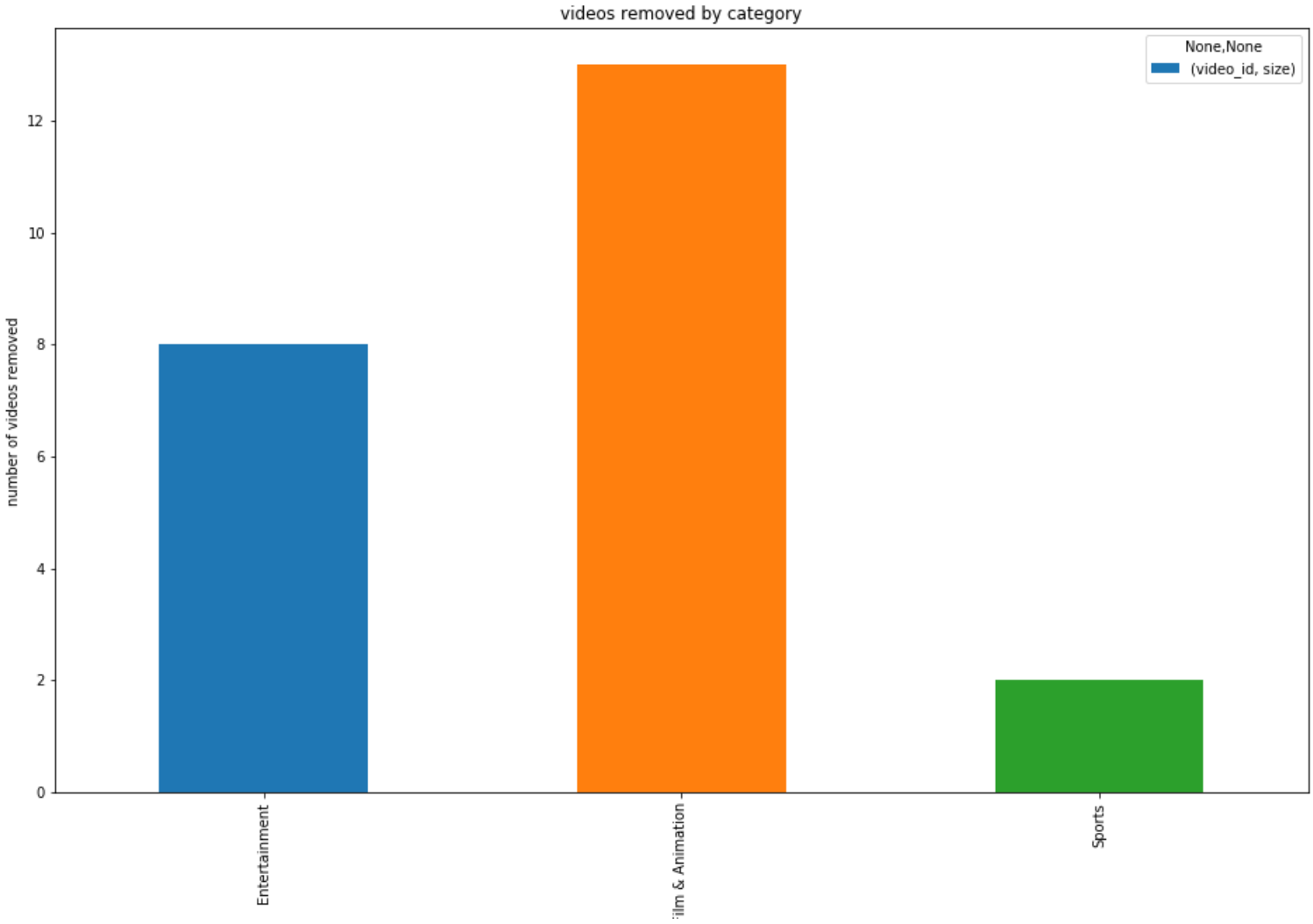

Figura 8: Cantidad de videos borrados por cateogria

Podemos observar que solo en las categorias entretenimeinto, deportes y animacion hay videos que fueron borrados. Esto es probable que se deba a problemas de copywright ya que youtube tiene ciertas resctricciones cuando se muestran imagenes que pertenecen a otras marcas, canciones o peliculas. Youtube ofrece un tiempo maximo para poner una cancion ajena hasta que te bloquean o desmonetizan el video por copywright. Lo mismo ocure con peliculas, o deportes.

2.3.1.9 Analisis de la cantidad de vistas de videos segun tenga o no descripcion.

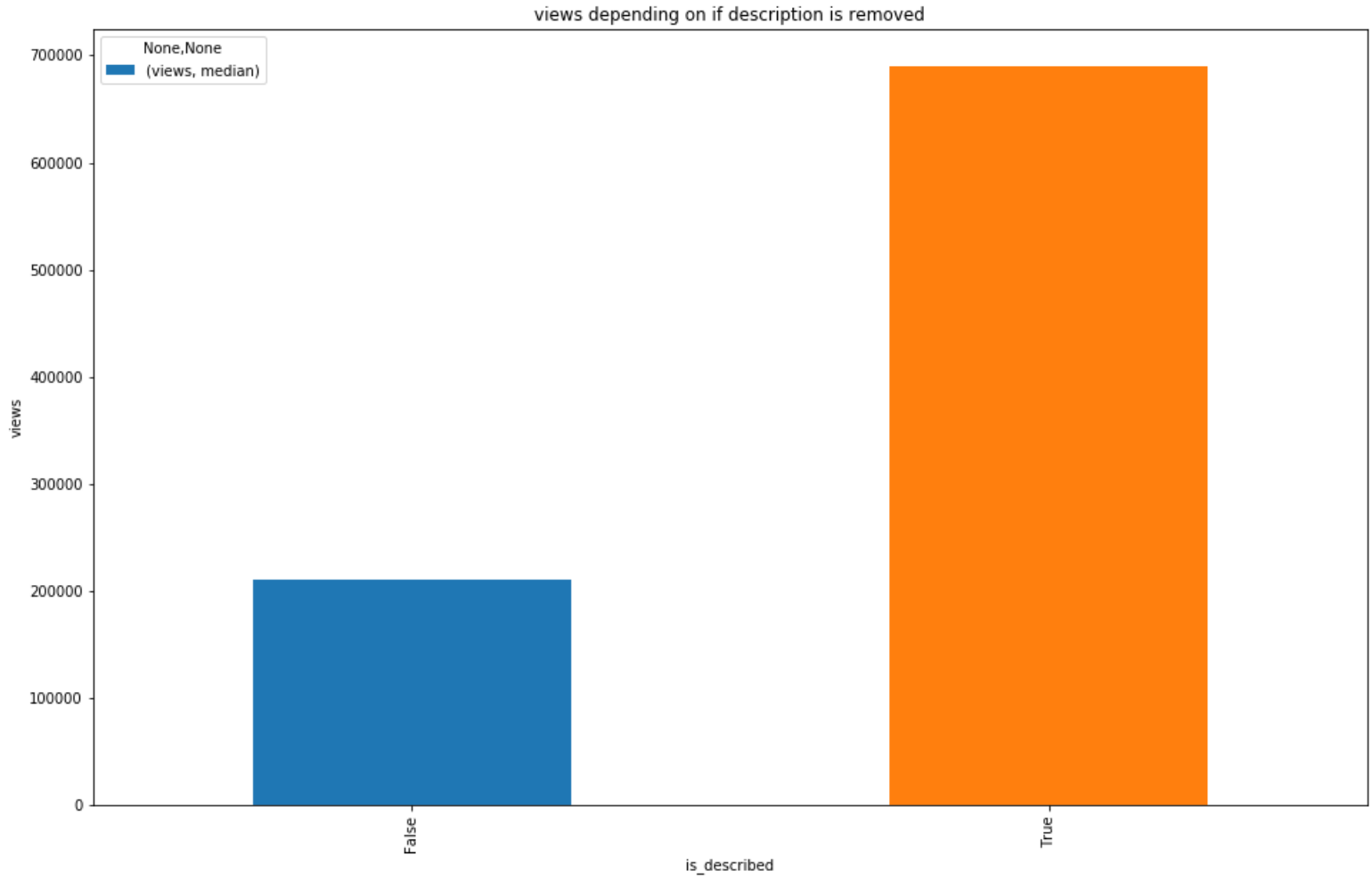

Figura 9: Cantidad de vistas segun su descripcion este o no habilitada

Podemos observar que hay una mayor cantidad de videos con vistas altas en la mediana lo cuales cuentan con una descrpicion. Y los videos que no tiene descripp-cion en general tienden a tener pocas vistas. Esto es probable que se deba a que cuando uno ve una lista de videos en yotube, se tiene una vista previa del mismo, se lee el titulo y un poco la descripcion. Esto es importante a la hora de decidir si uno le interesa ver ese video o no ya que te puede brindar una explicacion que te interese.

2.4. Verificación de calidad de datos

2.4.1. Reporte de calidad de datos

Los datos utilizados durante el transcurso del proyecto fueron obtenidos gratuitamente en el sitio de Kaggle y el mismo se encuentra en formato csv. Con millones de aplicaciones en la actualidad, el conjunto de datos se ha convertido en la clave para obtener las mejores los mejores videos que se publicaron en youtube. Este conjunto de datos contiene más de 40000 detalles de videos publicados en youtube.

Fecha de recolección de datos (de API): Abril 2019

3. Fase 3: Preparación de los datos

3.1. Reporte de calidad de los datos

A continuacion mostraremos un cuadro que muestra que atributos obtuvimos y de que tipo son:

Nombre de variable	tipo de dato	Descripcion
category_id	numeric	son cuatro numeros que representan un grupo de categorias
comments_disabled	booleano	Dice si tiene o no los comentarios habilitados
ratings_disabled	booleano	Dice si tiene o no los likes/dislikes habilitados
video_error_or_removed	booleano	Dice si el video sufrio un error o fue borrado en el momento de la fecha de trending_date
title_size	numeric	cantidad de caracteres que tiene el titulo del video
tags_quantity	numeric	cantidad de etiquetas que el video tiene
likes_ratio	numeric	porcentaje de likes respecto de dilikes
has_description	booleano	Dice si el video tiene descripcion o no
comments_per_view	numeric	Cantidad de cantidad de comentarios dividido cantidad de views
progess	numeric	Promedio de vistas segun pasan los dias desde la publicacion del video
is_successful	boolean	Dice si el video es exitoso o no

3.2. Seleccionar los datos

3.2.1. Inclusión/Exclusión de datos

Los datos que se terminaron utilizando fueron todos los de tipo alfanumerico, numérico y discreto.

3.3. Limpiar los datos

3.3.1. Reporte de limpieza de datos

Se limpiaron los siguientes campos:

- video_id
- trending_date
- title
- channel_title
- tags
- likes
- dislikes
- thumbnail_link
- description
- comment_count
- publish_time

Algunos de estos campos se usaron para crear nuevos campos .

3.4. Estructurar los datos

3.4.1. Derivación de atributos

- **title_size:** Longitud en cantidad de caracteres del atributo title
- **publish_time:** Se modifico el timestamp por un formato tal: YYYY-MM-DD
- **tags_quantity:** cantidad de tags
- **likes_ratio:** Ratio de likes y dislikes calculado como: likes/(likes+dislikes)

- **has_description:** Es verdadero si el video tiene descripcion.
- **comments_per_view:** Cantidad de comentarios dividido la cantidad de vistas del video
- **progess:** Promedio de vistas por dia que transcurre desde la publicacion del video

3.4.2. Generación de registros

Se genero el registro `is_sucessfull` que fue calculado en funcion de si el video tiene mas de 2 millones de vistas

3.5. Integrar los datos

3.5.1. Unificación de los datos

Se unieron las datos del dataset con el json de categorias para conocer la relacion entre el valor numerico y el nombre de la categoria pero como el data ser ya venia con el valor numerico no fue necesario integrar los datos ya que nosotros necesitamos el valor numerico. En cambio, esto fue necesario para el analisis ya nos sirve para saber de que categorias hablamos.
Luego hicimos una discretizacion de grupos de categorias debido a que teniamos demasiadas categorias posibles, los resultados obtenidos de esta manera no son satisfactorios ya que son demasiado + completos. Analizamos diferentes formas de agruparlos y llegamos a la conclusion de que esta forma que utilizamos es apropiada para el tratamiento que haremos, ya que las categorias tienen una alta cohesion, tienen muchas cosas similares entre si, y hay una alta probabilidad de que las categorias que pertenecen al mismo grupo tengan reglas similares.

Por lo tanto a continuacion mostramos el diccionario discretiado que nos dice a que grupo de categorias pertenece cada numero

3.5.1.1 Diccionario de categorias

Nombre	Category_id
Film and Animation	0
Gaming or Music	1
Videoblogging or People and Blogs orComedy or Entertainment	2
News and Politics or Education	3
others	4

3.6. Formato de los datos

3.6.1. Reporte de formato de los Datos

publish_time: Se modifico el timestamp por un formato tal: YYYY-MM-DD

4. Fase 4: Modelado

4.1. Seleccionar una técnica de modelado

Para este trabajo decidimos utilizar dos técnicas de modelado, J48 Decision Tree y Multilayer Perceptron. El motivo por el que elegimos estas técnicas es porque son algoritmos supervisados, lo cual es adecuado para el problema que intentamos resolver, y porque estos algoritmos, al funcionar sobre principios diferentes, reaccionan de distinta forma ante distintos tipos de datos de entrada, permitiéndonos de esta forma compararlos, y elegir el más apropiado. Se Aplicará cada una de estas técnicas, variando sus parámetros para obtener mejores resultados. La clase para estos algoritmos será el atributo `is_successfull`, el cual surgirá de aplicar una función de threshold a la cantidad de views de cada video. Para evitar que el algoritmo tenga información relacionada directamente con la clase, quitaremos el atributo `views` del set de datos para aplicar los algoritmos.

4.1.1. Técnica de modelado

J48 Decision Tree: Se utilizará el algoritmo supervisado J48, el cual consiste en un árbol de decisión utilizado para la clasificación del set de datos.
Multilayer Perceptron: Se utilizará el algoritmo supervisado Multilayer Perceptron, el cual consiste en una Red Neuronal del tipo backtracking, utilizado para la clasificación del set de datos.

4.1.2. Supuestos de modelado

Todos los datos usados para el modelado son numéricos, booleanos

4.2. Generar el diseño de las pruebas

4.2.1. Diseño de las pruebas

Para ambas técnicas de modelado se eligió utilizar un 70 % del dataset como set de entrenamiento y un 30 % del dataset comos set de prueba, elegido al azar (se mezcla al azar todo el set de pruebas y se elige el 70 % de los registros que se encuentran primero). Se escogió de esta forma para darle suficientes datos a la red neuronal. En total quedan 1906 registros para prueba y 4445 registros para entrenamiento.

4.3. Construir el modelo

4.3.1. Configuración de parámetros

4.3.1.1 Arbol J48

Para el caso del árbol J48 debemos ajustar los parámetros de nivel de confianza y cantidad mínima de elementos por hoja.

Nivel de Confianza
Un nivel de confianza demasiado alto provocará que el algoritmo no desestime ninguna regla lo cual llevará a un posible overfit.

Cantidad mínima de Elementos por hoja
Lo mismo puede ocurrir en caso de usar un valor demasiado bajo de cantidad minima de elementos por hoja.
Luego de Experimentar con distintos valores para cada caso, llegamos a los siguientes valores que nos dan los mejores resultados:

4.3.1.1 Parámetros para el arbol 1

Parametro	valor
Confidence Level	0.1
Minium elements per Leaf	3

4.3.1.1 Parámetros para el arbol 2

Parametro	valor
Confidence Level	0.3
Minium elements per Leaf	20

Luego de varios test, notamos que el valor de Minium elements per leaf que mejor resultado en términos de precisión era 60, pero el árbol quedaba demasiado grande como para comprenderlo fácilmente. Decidimos subirlo a 80 para obtener un árbol razonable, perdiendo una cantidad despreciable de precisión. Subirlo a mas de 80 ya podaba mucho el árbol y generaba pérdidas de precisión importantes.
De la misma forma elegimos un Confidence level de 0.01. Disminuirlo mas generaba una pérdida de precisión, aumentarlo nos resultaba en un arbol mucho mas grande.

4.3.1.2 Multilayer Perceptron

Para el caso del Multilayer Perceptron debemos ajustar la cantidad hidden layers, el learning rate, el momentum, el training time y el decay.

Hidden Layers
Si se colocan demasiadas hidden layers se incrementa mucho el tiempo de procesamiento, y ademas se corre el riesgo de overfiting. Si se colocan muy pocas puede que no sean suficientes para aproximar correctamente la función del problema.

Learning Rate
El learning rate y el momentum afectan la velocidad a la que intenta converger el algoritmo. Un learning rate o momentum demasiado elevador provocaran que el algoritmo no logre encontrar un mínimo local. Un learning rate o momentum demasiado pequeños provocan que el algoritmo no logre encontrar un mínimo global.

Training Time
Es la cantidad de pasadas de los datos a través de la red. Aumentarlo debería mejorar los resultados, sin embargo incrementa el tiempo de procesamiento y puede provocar overfiting.

Decay
Si está activado El algoritmo Cambia dinámicamente el learning rate según los resultados anteriores. Sirve para evitar que el algoritmo se pase por alto los mínimos, pero aumenta el tiempo de procesamiento.

Luego de Experimentar con distintos valores para cada caso, llegamos a los siguientes valores que nos dan los mejores resultados:

4.3.1.2 Parámetros

Parametro	**valor**
Hidden Layers	30
Momentum	0.2
Learing Rate	0.3
Training Time	500
Decay	Desactivado

4.3.2. Modelos

4.3.2.1 Resultados de la corrida 1 del arbol de decisión J48

4.3.2.1 Matriz de confusion

		is_successful false	is_successful true
0	era originalmente false	1706.0	51.0
1	era originalmente true	187.0	215.0

4.3.2.1 Resultados de la clasificación

		percentage	quantity
0	Correctly Classified Instances	88.9764%	1921
1	Incorrectly Classified Instances	11.0236%	238

4.3.2.1 Reglas

- **Regla 0**:

	Regla 0
0	Si progress <= 104370
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.814238475005	0.966074313409	0.682254763029

- **Regla 1**:

	Regla 1
0	Si progress > 104370
1	AND progress <= 268248
2	Entonces
3	is_successful = False

	captura	confianza	soporte
0	0.128963236724	0.659701492537	0.15824279641

- **Regla 2**:

	Regla 2
0	Si progress > 268248
1	AND progress <= 1912737
2	AND category_id = 0
3	Entonces
4	is_successful = True

	captura	confianza	soporte
0	0.0099202489788	0.910714285714	0.00881750905369

- **Regla 3**:

	Regla 3
0	Si category_id = 1
1	Entonces
2	is_successful = True

	captura	confianza	soporte
0	0.0420151721455	0.91914893617	0.0370020469217

- **Regla 4**:

	Regla 4
0	Si category_id = 2
1	AND tags_quantity <= 25
2	Entonces
3	is_successful = True

	captura	confianza	soporte
0	0.0248978797899	0.723163841808	0.0278696268304

- **Regla 5**:

	Regla 5
0	Si tags_quantity > 25
1	AND title_size <= 25
2	Entonces
3	is_successful = True

	captura	confianza	soporte
0	0.00272320560202	0.933333333333	0.00236183278224

- **Regla 6**:

	Regla 6
0	Si title_size > 25
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.0233417623031	0.535714285714	0.0352700362148

- **Regla 7**:

	Regla 7
0	Si category_id = 3
1	AND tags_quantity <= 21
2	AND comments_per_view <= 0.001495
3	Entonces
4	is_successful = True

	captura	confianza	soporte
0	0.0	0.0	0.00078727759408

- **Regla 8**:

	Regla 8
0	Si comments_per_view > 0.001495
1	AND progress <= 482712
2	Entonces
3	is_successful = False

	captura	confianza	soporte
0	0.00233417623031	0.8	0.00236183278224

- **Regla 9**:

	Regla 9
0	Si progress > 482712
1	Entonces
2	is_successful = True

	captura	confianza	soporte
0	0.000778058743435	0.8	0.00078727759408

- **Regla 10**:

	Regla 10
0	Si tags_quantity > 21
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.0	0.0	0.0029916548575

- **Regla 11**:

	Regla 11
0	Si category_id = 4
1	AND title_size <= 81
2	AND likes_ratio <= 0.853659
3	Entonces
4	is_successful = True

	captura	confianza	soporte
0	0.0	0.0	0.00188946622579

- **Regla 12**:

	Regla 12
0	Si likes_ratio > 0.853659
1	AND title_size <= 41
2	Entonces
3	is_successful = True

	captura	confianza	soporte
0	0.00739155806263	0.745098039216	0.00803023145961

- **Regla 13**:

	Regla 13
0	Si title_size > 41
1	AND likes_ratio <= 0.984742
2	AND title_size <= 53
3	Entonces
4	is_successful = False

	captura	confianza	soporte
0	0.0033067496596	0.85	0.00314911037632

- **Regla 14**:

	Regla 14
0	Si title_size > 53
1	Entonces
2	is_successful = True

	captura	confianza	soporte
0	0.00408480840303	0.6	0.00551094315856

- **Regla 15**:

	Regla 15
0	Si likes_ratio > 0.984742
1	Entonces
2	is_successful = True

	captura	confianza	soporte
0	0.0	0.0	0.000944733112896

- **Regla 16**:

	Regla 16
0	Si title_size > 81
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.00427932308889	0.846153846154	0.00409384348921

- **Regla 17**:

	Regla 17
0	Si progress > 1912737
1	Entonces
2	is_successful = True

	captura	confianza	soporte
0	0.0	0.0	0.0176350181074

4.3.2.1 Arbol

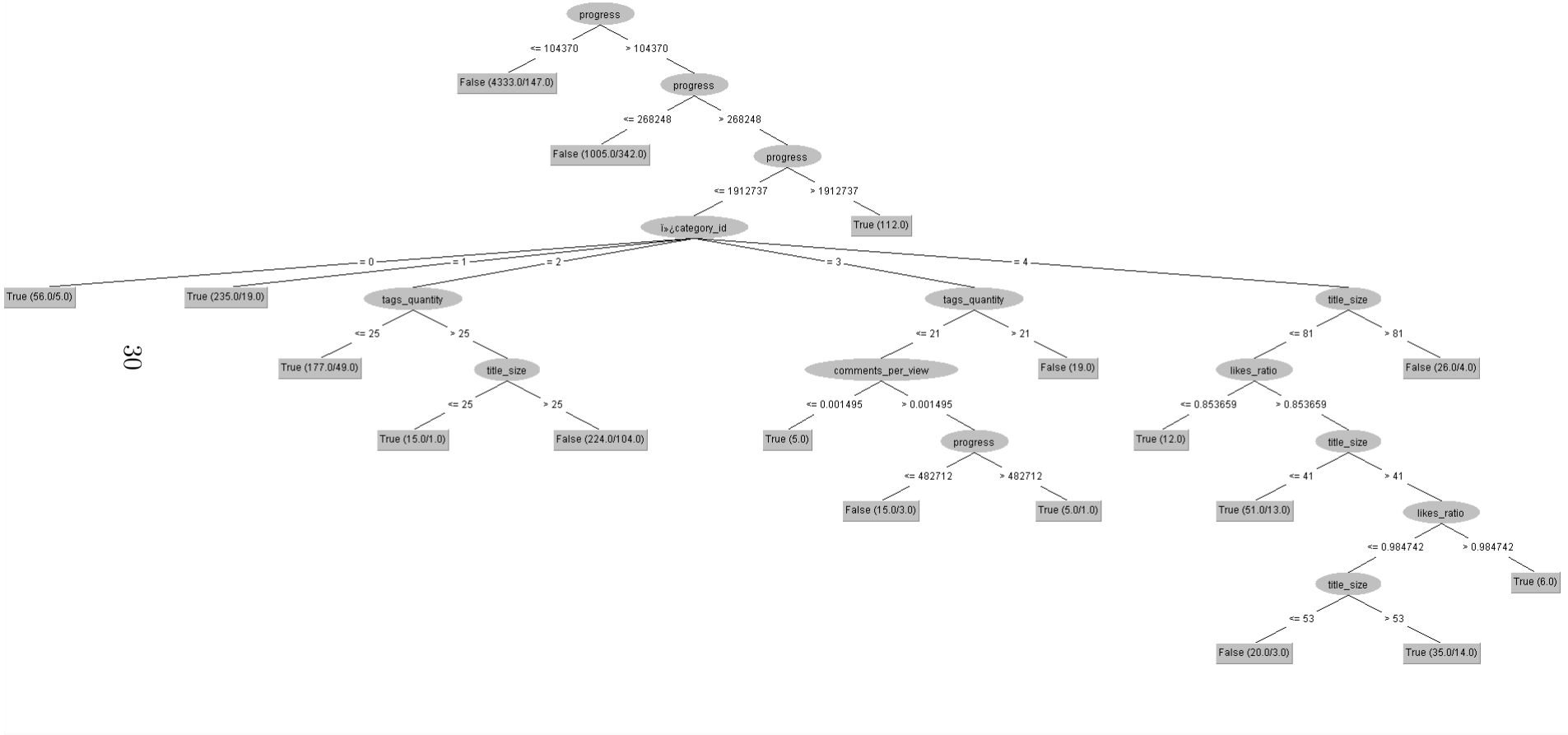

Figura 10: Arbol de desicion

4.3.2.2 Resultados de la corrida 2 del arbol de decisión J48

4.3.2.2 Matriz de confusion

		is_successful false	is_successful true
0	era originalmente false	1689.0	68.0
1	era originalmente true	175.0	227.0

4.3.2.2 Resultados de la clasificación

		percentage	quantity
0	Correctly Classified Instances	88.7448%	1916
1	Incorrectly Classified Instances	11.2552%	243

4.3.2.2 Reglas

- **Regla 0**:

	Regla 0
0	Si progress <= 104370
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.814238475005	0.966074313409	0.682254763029

- **Regla 1**:

	Regla 1
0	Si progress > 104370
1	AND progress <= 268248
2	AND progress <= 192094
3	Entonces
4	is_successful = False

	captura	confianza	soporte
0	0.0990079751021	0.713884992987	0.112265784916

- **Regla 2**:

	Regla 2
0	Si progress > 192094
1	AND category_id = 0
2	Entonces
3	is_successful = True

	captura	confianza	soporte
0	0.00272320560202	0.736842105263	0.0029916548575

- **Regla 3**:

	Regla 3
0	Si category_id = 1
1	AND likes_ratio <= 0.9753
2	Entonces
3	is_successful = True

	captura	confianza	soporte
0	0.00272320560202	0.636363636364	0.00346402141395

- **Regla 4**:

	Regla 4
0	Si likes_ratio > 0.9753
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.00311223497374	0.666666666667	0.00377893245158

- **Regla 5**:

	Regla 5
0	Si category_id = 2
1	AND title_size <= 39
2	Entonces
3	is_successful = True

	captura	confianza	soporte
0	0.00622446994748	0.64	0.0078727759408

- **Regla 6**:

	Regla 6
0	Si title_size > 39
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.011087337094	0.6	0.0149582742875

- **Regla 7**:

	Regla 7
0	Si category_id = 3
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.00175063217273	0.75	0.00188946622579

- **Regla 8**:

	Regla 8
0	Si category_id = 4
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.00797510212021	0.585714285714	0.0110218863171

- **Regla 9**:

	Regla 9
0	Si progress > 268248
1	AND category_id = 0
2	Entonces
3	is_successful = True

	captura	confianza	soporte
0	0.0108928224081	0.918032786885	0.00960478664777

- **Regla 10**:

	Regla 10
0	Si category_id = 1
1	Entonces
2	is_successful = True

	captura	confianza	soporte
0	0.0531025092394	0.934931506849	0.0459770114943

- **Regla 11**:

	Regla 11
0	Si category_id = 2
1	AND progress <= 1912737
2	AND tags_quantity <= 25
3	Entonces
4	is_successful = True

	captura	confianza	soporte
0	0.0248978797899	0.723163841808	0.0278696268304

- **Regla 12**:

	Regla 12
0	Si tags_quantity > 25
1	AND title_size <= 38
2	Entonces
3	is_successful = True

	captura	confianza	soporte
0	0.00719704337677	0.698113207547	0.00834514249724

- **Regla 13**:

	Regla 13
0	Si title_size > 38
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.0204240420152	0.564516129032	0.0292867264998

- **Regla 14**:

	Regla 14
0	Si progress > 1912737
1	Entonces
2	is_successful = True

	captura	confianza	soporte
0	0.0	0.0	0.00614076523382

- **Regla 15**:

	Regla 15
0	Si category_id = 3
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.00622446994748	0.727272727273	0.0069280428279

- **Regla 16**:

	Regla 16
0	Si category_id = 4
1	AND title_size <= 81
2	AND likes_ratio <= 0.865798
3	Entonces
4	is_successful = True

	captura	confianza	soporte
0	0.00369577903132	0.95	0.00314911037632

- **Regla 17**:

	Regla 17
0	Si likes_ratio > 0.865798
1	AND title_size <= 41
2	Entonces
3	is_successful = True

	captura	confianza	soporte
0	0.00816961680607	0.763636363636	0.00866005353488

- **Regla 18**:

	Regla 18
0	Si title_size > 41
1	AND likes_ratio <= 0.971997
2	Entonces
3	is_successful = False

	captura	confianza	soporte
0	0.00486286714647	0.625	0.00629822075264

- **Regla 19**:

	Regla 19
0	Si likes_ratio > 0.971997
1	Entonces
2	is_successful = True

	captura	confianza	soporte
0	0.00291772028788	0.75	0.00314911037632

- **Regla 20**:

	Regla 20
0	Si title_size > 81
1	Entonces
2	is_successful = False

	captura	confianza	soporte
0	0.00427932308889	0.846153846154	0.00409384348921

4.3.2.2 Arbol

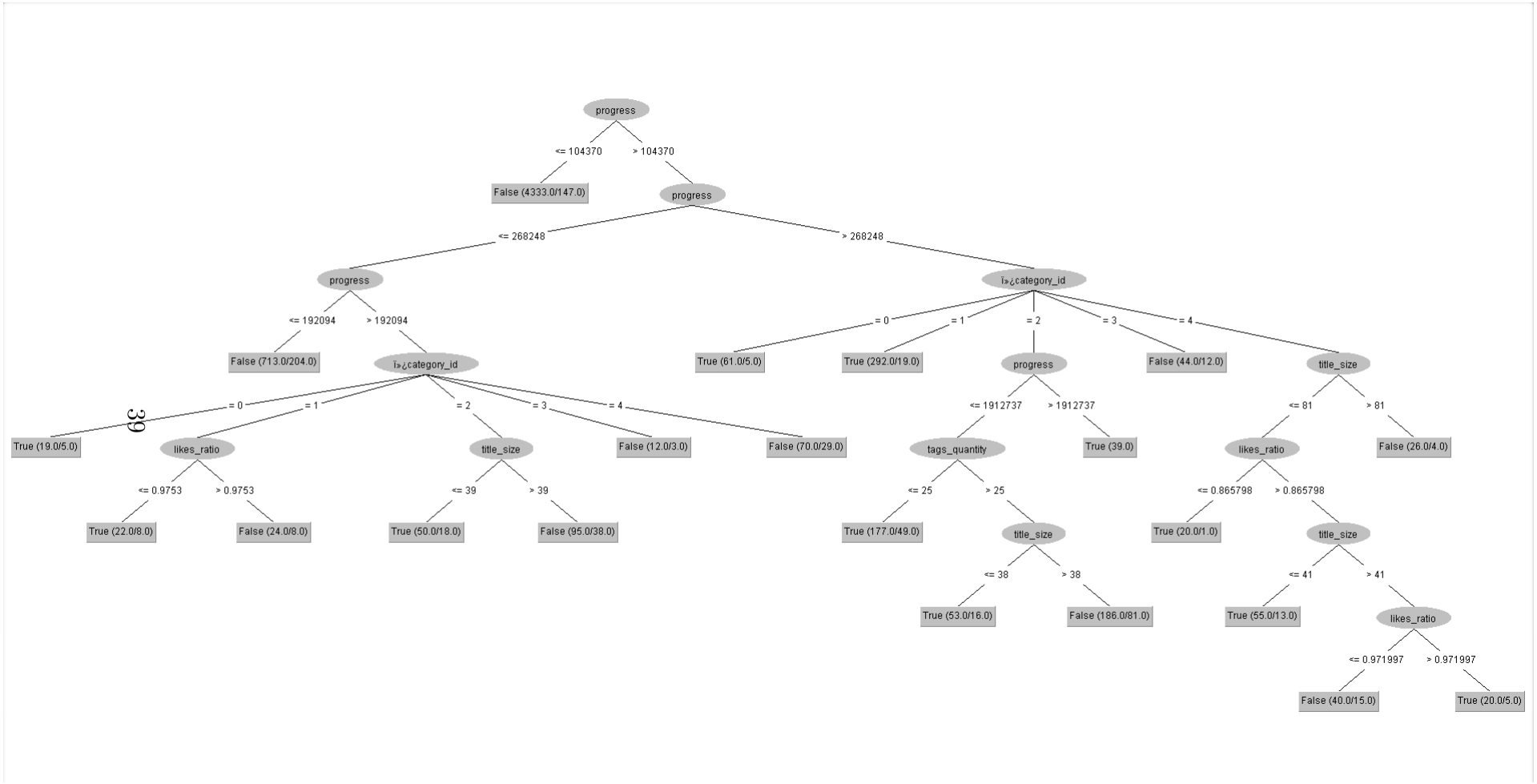

Figura 11: Arbol de desicion

4.3.2.3 Resultados de la corrida de Perceptron

4.3.2.3 Matriz de confusion

		is_successful false	is_successful true
0	era originalmente false	1660.0	97.0
1	era originalmente true	151.0	251.0

4.3.2.3 Resultados de la clasificación

		percentage	quantity
0	Correctly Classified Instances	88.5132%	1911
1	Incorrectly Classified Instances	11.4868%	248

4.3.3. Análisis de predicciones de clasificación

A continuación presentamos algunos gráficos que muestran la tendencia del algoritmo a clasificar correcta o incorrectamente los eventos, según sus atributos

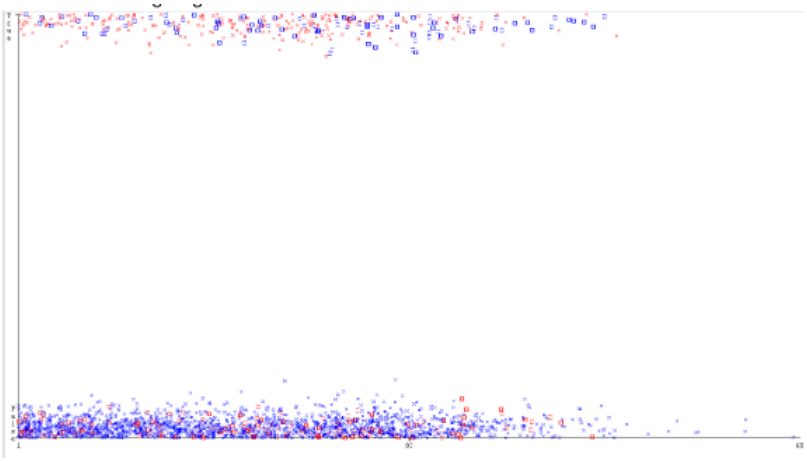

Figura 12: Clasif. Errors vs tag length

Podemos observar en este gráfico que los videos con muchos tags tienen menos chance de ser exitosos, algo que ya habíamos visto en la etapa de conocimiento

del negocio, notando además que para los videos que tienen muchos tags, aquellos clasificados como exitosos son en su mayoría falsos positivos, esta información puede ser útil para futuros análisis.

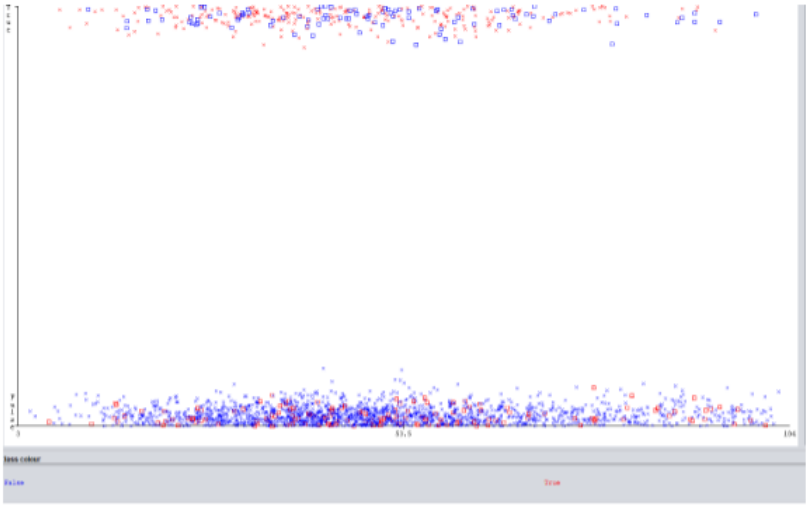

Figura 13: Clasif. errors vs title length

Vemos que los videos que tienen títulos demasiado largos tienen menos chance de ser exitosos, tal como nos muestran las reglas, y además vemos que el algoritmo identifica correctamente estos casos, teniendo un alto porcentaje de acierto tanto en positivos como en negativos.

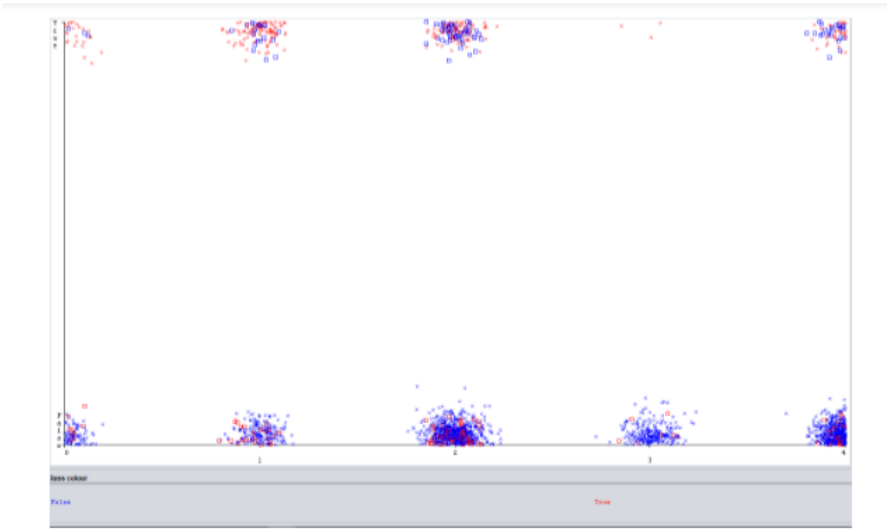

Figura 14: Clasif. errors vs category

En este gráfico vemos por un lado, que los grupos de categorías 3 y 4 tienen poca cantidad de videos exitosos, a pesar que tienen mucha cantidad de videos, y por otro

lado notamos que el algoritmo es muy efectivo clasificando videos pertenecientes a los grupos 0, 1 y 3, y no tan efectivo al clasificar los videos de grupos 2 y 4.

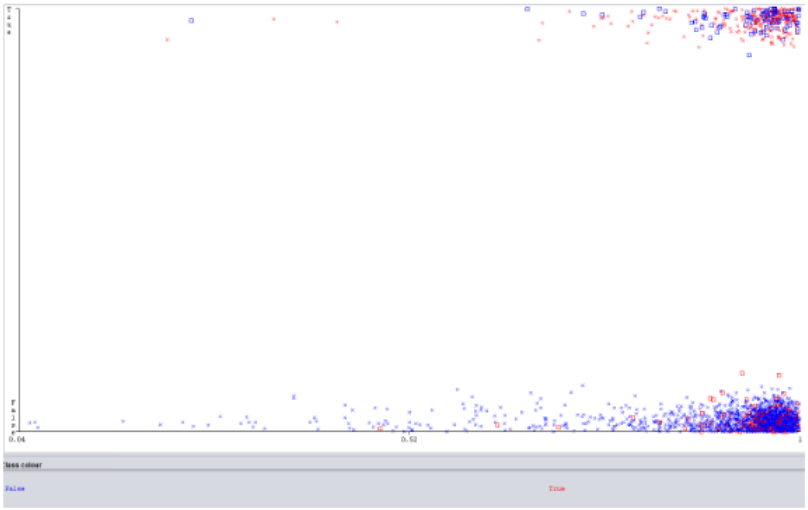

Figura 15: Clasif. errors vs like ratio

Este gráfico es interesante, vemos muy claramente que los videos que tienen poco like ratio tienen mucha menos chance de ser exitosos, y además el algoritmo los clasifica casi perfectamente, mientras que los videos con un like ratio mayor tienen más chance de ser exitosos, pero el algoritmo es menos efectivo al clasificarlos correctamente.

4.3.4. Descripción del modelo

4.4. Evaluar el modelo

4.4.1. Evaluación del modelo

4.4.1.1 Analisis de reglas

- **Arbol 1:**
  - **Regla 2**: Podemos apreciar que el progreso del video es un factor decisivo ya que el promedio de vistas por dia desde su publicacion determina el exito del video
  - **Regla 3**: Podemos apreciar que el hecho de que la categoria es importante porque nos dice que el exito del video se basa en que la gente prefiere cierto contenido relacionado a peliculas, animacion, comedia, etc
  - **Regla 4**: Podemos ver que la cantidad de tag juega un rol importante ya que juega como promotor del mismo, es decir, los tags facilitan la busqueda del video ya que los mismos se pueden buscar por tags, entonces estos te permiten llegar a la gente.
  - **Regla 5**: Podemos ver que el tamaño del titulo juega un rol principal ya que tener un maximo del largo del mismo nos dice que la curiosidad de la gente se despierta frente a titulos de cierto rango de tamaño.
- **Arbol 2:**
  - **Reglas 0 y 1:** Esta regla nos muestra que cuando un video no comienza con un crecimiento elevado, entonces es dificil que pueda llegar a ser un video exitoso, por lo que parece ser que los videos con bajo crecimiento tienen poca probabilidad de llegar a ser exitosos.
  - **Regla 2:** En esta regla podemos ver que los videos que comienzan con buen crecimiento y son de la categoria 0 tienen alta probabilidad de ser exitosos. Esto es razonable ya que la categoria 0 corresponde a los videos de peliculas y animacion, cuyo exito o no suele definirse en cuanto son subidos.
  - **Regla 3:** Esta regla muestra que los videos de cierta categoria tienen mas chances de ser exitoso. Ademas nos presenta como requisito que tenga una cierta cantidad de likes ratio, lo que nos muestra que la cantidad de likes influye en la popularidad y el exito de los videos.
  - **Regla 5:** Esta regla puede deberse a que la gente navega muy rapido buscando videos, y los titulos mas cortos se pueden leer mas rapidamente. A su vez, en la categoria que nos muestra de blogs y entretenimiento suelen ubicarse los videos de canales exitosos, los cuales por lo general

tienen titulos cortos y, al tener muchos seguidores el canal en si, tienen mas cantidad de views.

- **Regla 10:** Esta tiene una confianza muy alta y una captura razonable, pero simplemente nos muestra que los videos de categoria 1 suelen ser exitosos lo cual coincide con nuestro conocimiento previo ya que los videos de musica son los mas populares de youtube.
- **Regla 11:** Esta regla tiene una confianza y captura razonables, y nos muestra que los videos de entretenimiento y blogging que tienen menos de 25 tags tienen mas chances de ser exitosos. Una explicacion de esto podria ser por la inversa, es decir que los videos que tienen demasiados tags son menos exitosos, ya que probablemente son videos largos que abarcan muchos temas y por lo general menos views.
- **Regla 16:** Regla tiene poca captura y poco soporte, pero una confianza muy alta, lo que significa que identifica un nicho, y tiene sentido que son videos de una categoria que no es de las mas comunes, que tienen titulos muy largos, y una cantidad bajas de likes, por lo que parecen ser videos polemicos, que terminan siendo exitosos debido justamente a la polemica generado por estos mismos.
- **Regla 20:** Esta regla muestra algo que habiamos visto en el entendimiento del negocio y es que los videos con titulos extremadamente largos tienen menos chances de exito.
- **Regla 4, 6, 8, 13 y 18:** Estas reglas tienen poca confianza, por lo que no resulta muy apropiado tener en cuenta sus resultados ya que pueden llevar a errores.
- **Regla 7, 9, 12, 15, 17 y 19:** Estas reglas tienen poca captura y soporte, por lo que no son representativas de la poblacion, y tampoco parecieran señalar algun nicho ya que sus confianzas no son tan altas.

4.4.2. Revisión de la configuración de parámetros

Para lograr una buena construcción del árbol, se limitó la cantidad de nodos hoja a valores peequeños como 3 y 20 ya que nos da una cantidad de reglas razonable y a menos cantidad nos daban reglas mas chicas y en mas cantidad.
Se probó variar la cantidad mínima de muestras requerida para hacer un split, la cantidad mínima de muestras que debe haber en un nodo hoja, la cantidad máxima de niveles del árbol.

5. Fase 5: Evaluación

5.1. Evaluar Resultado

5.1.1. Valoración de los resultados de minería de datos

Luego de haber aplicado las técnicas introducidas previamente, se han llegado a inferir algunas reglas que dejan al descubierto algunas caracterśticas que aumentan la probabilidad de éxito de un video. Si bien queda lugar a más análisis debido a que con las reglas no se llega a cubrir un panorama completo de las cuestiones a tener en cuenta en el desarrollo de un video de youtube, se puede confirmar que se ha obtenido información que de otra forma sería difícil obtener. Ademas que los valores que ponemos a continuacion son solo aproximaciones y que simplemente deben ser usados no en forma estricta, sino como guía para mejorar las chances de que el video sea exitoso.
En general, revisando las reglas encontramos algunos patrones en común que categorizarían a un video como exitosa:

- El video tiene que tener un progreso de entre 268248 y 1912737 vistas por dia
- Tiene que tener categorias como cine, animacion, juegos, musica, comedia, entretenimiento
- La cantidad de tags tienen que ser alrededor de
- El largo del titulo tiene que ser de menor a 21
- Se tiene un entre un 80 y 96 porciento de likes respecto de dislikes

Por todo esto, se puede considerar que el proyecto de investigación fue exitoso.

5.1.2. Modelo aprobado

El árbol nos dió información útil. Cabe notar que visualizando el mismo se obtuvo un mejor panorama de los criterios buscados.

5.2. Proceso de revisión

5.2.1. Revisión del proceso

A partir del análisis exploratorio de datos, seguido de la generación de reglas, se ha llegado a cumplir con el objetivo planteado, por lo que en principio no quedan actividades pendientes.

5.3. Determinar Próximos pasos

5.3.1. Listado de posibles acciones

Como próximos pasos se podría dejar algún análisis más profundo en bases de datos de otros paises y no restringirse solo a la de Estados Unidos, ya que se supone que las reglas de éxito puede variar de país en país. Esto podría dar reglas muchísimo más claras.

6. Conclusiones

En base al set de datos, los resultados obtenidos, y los razonamientos a los que llegamos a partir de ellos, podemos concluir que los algoritmos utilizados son apropiados para la identificacion de videos potencialmente exitosos, y que tal algoritmo de clasificacion puede ponerse en produccion para futuros videos.
A su vez, se logro identificar ciertos patrones que contribuyen al exito o no de los videos y los mismos pueden utilizarse como referencia para la creacion de futuros videos con mayor potencial para ser exitosos.