

**Universidad Peruana de Ciencias Aplicadas**

Ciencias de la Computación

Administración De La Información (CC50-2201-CC52)

2022-1

Profesor :

Reyes Silva, Patricia Daniela

**Trabajo Final**

Integrantes:

Lopez Gonzalez, Maria Isabel - U201724423

Trujillo Mori , Jeanpier Alexander - U201523565

Contreras Inostroza, Eduardo Junior - U201414103

Junio, 2022

# Objetivos del Proyecto

No cabe dudas, de que Youtube ha sido una de las redes sociales más grandes de la década, siendo el sitio web para video más grande del mundo y uno de los sitios webs en general más visitados globalmente. El impacto que ha tenido Youtube en miles de personas y hasta ha sido el medio de comunicación de eventos de gran relevancia global.

Es por ello que estaremos analizando las estadísticas de tendencias de este sitio web, pues es claro que la información acerca de qué videos tienen éxito en la plataforma puede indicar que intereses tiene el público y apoya la toma de decisiones.

Estaremos analizando varias preguntas en diferentes métricas de los videos. Siendo las métricas por categorías de videos, por canales, por tiempo transcurrido y por geografía del país.

# Caso de análisis

## Origen de los Datos

El dataset ha trabajar se llama “Trending YouTube Video Statistics” obtenido desde la web Kaggle, aunque tiene una variación de datos agregados por los \*\*\* profesores del cursos de administración de la información de la UPC\*\*. En dataset original ayuda a determinar los videos más populares del año, YouTube usa una combinación de factores que incluyen la medición de las interacciones de los usuarios (cantidad de vistas, acciones compartidas, comentarios y Me gusta) .

## Casos de Usos

*¿Para quién sería importante el análisis de estos datos?*

El dataset “Trending YouTube Video Statistics” es importante para los productores de contenido , de esta manera pueden hacer un análisis que los ayudará a tener mayor impacto en las redes sociales.

*¿Quién o quiénes se benefician?*

Los beneficiados de este dataset , son el público y creadores de contenido , ya que de esta manera pueden ofrecer un mejor catálogo para los consumidores de contenido.

*¿A quién va dirigido el desarrollo del proyecto?*

Para los creadores de contenido , ya sean autor independiente, un profesional especialista en alguna modalidad interesante que puede ser enseñada, un profesor o incluso un bloguero que habla sobre temas específicos en su blog , de esta manera podrán posicionarse en la categoría que más deseen.

# Conjunto de Datos

Estaremos analizando dos datasets el principal es el MXvideos.csv que cu

| **Columna** | **Descripción** | **Tipo** |
| --- | --- | --- |
| video\_id | Identificación única para cada video | Numérico |
| trending\_date | Fecha en la que el video estuvo en tendencia | Fecha |
| title | Titulo del video | Categórica |
| channel\_title | Nombre del canal que ha publicado el video | Categórica |
| category\_id | Identificación del la categoría del video puede ser un valor del 1 al 43 | Categórica |
| publish\_time | Fecha en la que se publicó el video | Fecha |
| tags | Etiquetas del video | Categórica |
| views | Número de vistas en el video | Numérico |
| likes | Número de me gustas en el video | Numérico |
| dislikes | Número de no me gustas en el video | Numérico |
| comment\_count | Número de comentarios en el video | Categórica |
| thumbnail\_link | URL de la imagen del thumbnail del video | Categórica |
| comments\_disabled | Indica si los comentarios están desactivados para este video valor puede ser VERDADERO O FALSO | Categórica |
| ratings\_disabled | Indica si los ratings están desactivados para este video. El valor puede ser VERDADERO O FALSO | Categórica |
| video\_error\_or\_removed | Indica si el video da error o ha sido removido. El valor puede ser VERDADERO O FALSO | Categórica |
| description | Descripción corta acerca del video | Categórica |
| state | Nombre del Estado perteneciente al país | Categórica |
| lat | Latitud geográfica de ubicación del Estado | Categórica |
| lon | Longitud geográfica de ubicación del Estado | Numérico |
| geometry | Registra las coordenadas de las geometrías donde se ubica el Estado dentro del planeta. | Numérico |

# Análisis Exploratorio de los Datos

## Cargar de datos

Primero cargamos el archivo principal GBvideos\_cc50.csv donde se encuentra la información principal del archivo. Usamos el parámetro quotechar='"' para asegurarnos que tanto en los título como en las descripciones la función read\_csv ignore las comas que se encuentren dentro de las comillas.

| df = pd.read\_csv('./GBvideos\_cc50.csv', quotechar='"') |
| --- |

## Inspeccionar los datos

### Visualizar los datos en el dataset

### Pre-procesar los datos

* Considerar crear una nueva columna para la descripción de la categoría a partir del archivo json respectivo (si faltase alguna categoría, obtenerla desde el archivo Json de US).
* Trabajar con las variables trending\_date y publish\_time como fechas y no como de tipo objeto. Ambas deben tener el mismo formato (yyyy-mm-dd o dd-mmyyyy). Adicionalmente, se puede crear una columna para la hora de publicación.
* Verificar datos faltantes. Si el dato faltante es el video\_id, entonces, se removerán dichas observaciones. En otro caso, analizar la aplicación de alguna técnica para el tratamiento de los datos faltantes.
* No olvidar explicar que técnica fue utilizada para eliminar o completar los datos faltantes.
* Identificación de los datos atípicos u outliers (si los hubiera)
* Explicación y aplicación de la(s) técnica(s) utilizada(s) para transformar los datos atípicos (si los hubiera)

### Identificación de datos faltantes Identificación de datos atípicos

## Visualizar los Datos

### Por Categorías de Videos

#### ¿Qué categorías de videos son las de mayor tendencia?

| fig, ax = plt.subplots()  plt.figure(figsize = (14,20))  \_ = sns.barplot(x="category\_name", y="No\_of\_videos", data=cdf, ax=ax)  \_ = ax.set\_xticklabels(ax.get\_xticklabels(), rotation=75)  \_ = ax.set(xlabel="", ylabel="No. de videos")  \_ = ax.set\_title("No. de videos por categoria") |
| --- |
|  |
| Descripción: Es claro que la categoría de Música es la que tiene la mayor cantidad de videos en tendencias. Casi 14.000 videos en tendencias. Le sigue la categoría de Entretenimiento con unos 9.000 vídeos. Se evidencia que estas dos categorías contienen más del 50% de todos los videos en tendencias. |

#### ¿Qué categorías de videos son los que más gustan? ¿Y las que menos gustan?

| # Crear la distribución  cld\_df = df[['category\_name']].copy()  cld\_df['dislikes\_dis'] = np.log(df['dislikes'] + 1)  cld\_df['likes\_dis'] = np.log(df['likes'] + 1)  # Mostrar (mismo código para like y dislikes)  fig, ax = plt.subplots()  plt.figure(figsize = (14,20))  g = sns.boxplot(x='category\_name', y='likes\_dis', data=cld\_df, palette="Set1", ax=ax)  \_ = ax.set(xlabel="", ylabel="Likes (dis)", title="Distribución de Me Gustas por categoría")  \_ = ax.set\_xticklabels(ax.get\_xticklabels(), rotation=75)  fig.set\_size\_inches(15, 8, forward=True) |
| --- |
|  |
| Descripción: Por un lado, se evidencia que la categoría de Música es la que tiene la mayor cantidad de me gustas y parece ser la que más gusta, del mismo modo, es la categoría con la mayor cantidad de no me gustas. Debido al análisis anterior es entendible que a esta ser la categoría con el mayor número de video en tendencias tenga el mayor número de interacciones y valores outliers.  Por otro lado, la categoría de entretenimiento tiene la mayor cantidad de valores de no me gusta fuera del rango promedio. Es decir, a pesar de que la mayoría de los videos se encuentran en un rango estable de no me gustas, los videos menos gustados en este dataset de tendencia se encuentran en esta categoría. |

#### ¿Qué categorías de videos tienen la mejor proporción (ratio) de “Me gusta” / “No me gusta”?

| Debido a la variabilidad de los likes y dislikes en un video en relación a otro, primero crearemos un nuevo df con dos nuevas columnas que recolectan la data del ratio por columna de la categoria, que será representada por los me gusta y los no me gusta |
| --- |
| rld\_df = df[['category\_name']].copy()  rld\_df['likes'] = ((df.isnull().sum() / len(df))\*100).sort\_values(ascending = False)  rld\_df['dislikes'] = ((df.isnull().sum() / len(df))\*100).sort\_values(ascending = False) |
| fig, ax = plt.subplots()  plt.figure(figsize = (14,20))  g = sns.boxplot(x='category\_name', y='dislikes', data=rld\_df, palette="Set1", ax=ax)  \_ = ax.set(xlabel="", ylabel="Dislikes , title="Ratio de No me gusta por categoría")  \_ = ax.set\_xticklabels(ax.get\_xticklabels(), rotation=75)  fig.set\_size\_inches(15, 8, forward=True) |
| fig, ax = plt.subplots()  plt.figure(figsize = (14,20))  g = sns.boxplot(x='category\_name', y='likes', data=rld\_df, palette="Set1", ax=ax)  \_ = ax.set(xlabel="", ylabel="likes , title="Ratio de me gusta por categoría")  \_ = ax.set\_xticklabels(ax.get\_xticklabels(), rotation=75)  fig.set\_size\_inches(15, 8, forward=True) |
| Como se puede apreciar en el gráfico , el que tiene mayor proporción por categoría de videos , son las vistas , ya que dependiendo del contenido de la categoría los usuarios comentan si fue o no su agrado de la categoría |

### 

#### ¿Qué categorías de videos tienen la mejor proporción (ratio) de “Vistas” / “Comentarios”?

| Debido a la variabilidad de los likes y dislikes en un video en relación a otro, primero crearemos un nuevo df con dos nuevas columnas que recolectan la data del ratio por columna de la categoría, que será representada por los me vistas y los comentarios |
| --- |
| rld\_df = df[['category\_name']].copy()  rld\_df['views'] = ((df.isnull().sum() / len(df))\*100).sort\_values(ascending = False)  rld\_df['comment\_count'] = ((df.isnull().sum() / len(df))\*100).sort\_values(ascending = False) |
| fig, ax = plt.subplots()  plt.figure(figsize = (14,20))  g = sns.boxplot(x='category\_name', y='views', data=rld\_df, palette="Set1", ax=ax)  \_ = ax.set(xlabel="", ylabel="views , title="Ratio de vistas por categoria")  \_ = ax.set\_xticklabels(ax.get\_xticklabels(), rotation=75)  fig.set\_size\_inches(15, 8, forward=True) |
| fig, ax = plt.subplots()  plt.figure(figsize = (14,20))  g = sns.boxplot(x='category\_name', y='comment\_count', data=rld\_df, palette="Set1", ax=ax)  \_ = ax.set(xlabel="", ylabel="comment\_count , title="Ratio de comentarios por categoria")  \_ = ax.set\_xticklabels(ax.get\_xticklabels(), rotation=75)  fig.set\_size\_inches(15, 8, forward=True) |
| Como se puede apreciar en el gráfico , el que tiene mayor proporción por categoría de videos , son las vistas , ya que dependiendo del contenido de la categoría los usuarios comentan si fue o no su agrado de la categoría |

### 

### Por tiempo transcurrido

#### ¿Cómo ha cambiado el volumen de los videos en tendencia a lo largo del tiempo?

| #by\_time = df.groupby("trending\_date")  df2 = df.groupby('trending\_date')['views'].sum().plot(legend=True,rot=30)  plt.legend(["Cantidad de vistas"]); |
| --- |
|  |
| Aquí se puede apreciar la cantidad de vistas en cantidades elevadas ala 9, como escala. Podemos ver como incrementa el volumen de vistas mientras el tiempo(fechas) avanzan, y al final tiende a hacer una curva oscílente. A continuación le haremos un Zoom para ver su comportamiento, para esto elegiremos las fechas de auge y lo ordenaremos de mayor a menor |
| df2 = df.groupby('trending\_date')['views'].sum().nlargest(10).plot(legend=True,rot=30) |
|  |
| Si ordenaramos a los que mayores vistas tiene de mayor a menor vemos la fecha de su auje y como fue disminuyendo descendientemente, (se estan mostrando los 10 primeros) |

### Por Canales de YouTube

#### ¿Qué Canales de YouTube son tendencia más frecuentemente? ¿Y cuáles con menos frecuencia?

| by\_chane = df['channel\_title'].value\_counts().to\_frame().reset\_index()  by\_chane.rename(columns={"index": "Canal", "channel\_title": "Cantidad de videos"}, inplace=True)  by\_chane.head() |
| --- |
|  |
| Como vemos Este es el DataFrame de los canales con más videos en tendencia, y como se puede apreciar el canal "The Tonight Show Starring Jimmy Fallon" es el que anda liderando con 208 videos en tendencia, para lograr esto se realizó un groupby en "canal" y un count del mismo (cantidad de videos en tendencia) |
| fig, ax = plt.subplots()  plt.figure(figsize = (14,20))  \_ = sns.barplot(x="Canal", y="Cantidad de videos", data=by\_chane[:20], ax=ax, palette="Set1")  \_ = ax.set\_xticklabels(ax.get\_xticklabels(), rotation=30)  \_ = ax.set(xlabel="", ylabel="No. de videos Trending")  \_ = ax.set\_title("Canales", fontsize=20)  fig.set\_size\_inches(15, 8, forward=True) |
| ´ |
| Aquí lo podemos observar en un gráfico de barras, con la misma información del DataFrame |

### 

### Por la geografía del país

#### ¿En qué Estados se presenta el mayor número de “Vistas”, “Me gusta” y “No me gusta”?

El código para los tres es similar así que solo se colocará el código una vez.

| fig = px.scatter\_mapbox(views\_map, lat="lat", lon="lon",  zoom=4, height=450, color="views",  hover\_name="state", size="views", size\_max=25,  color\_continuous\_scale=px.colors.sequential.Rainbow  )  print( views\_map.sort\_values(by=['views'], ascending=False).head(2) )  fig.update\_layout(margin={"r":0,"t":0,"l":0,"b":0})  fig.show() |
| --- |
| state lon lat views  39 Southampton -1.399976849 50.90003135 5570080346  0 Aberdeen -2.079987021 57.17039797 5482827742  41 Stockton On Tees -1.230013063 54.58037518 5458518084  38 South Yorkshire -1.499996583 53.36667666 5448546504  46 West Midlands -1.499996583 52.42040367 5368026416  state lon lat views  2 Belfast -5.960034425 54.60001223 3532467750  26 Merseyside -2.917997886 53.41600181 3646696350 |
| En el gráfico se puede visualizar que existen tres zonas que se acercan a los 5.5B de visitas. El estado con el mayor número de vistas es Southampton con 5.570.080.346 de vistas y el segundo con el mayor número es Aberdeen con 5.482.827.742. Los estados con menos vistas son Belfast y Merseyside respectivamente. |

#### 

| state lon lat likes  38 South Yorkshire -1.499996583 53.36667666 132686060  46 West Midlands -1.499996583 52.42040367 121579342  23 Leicester -1.133248943 52.62997744 121366689  12 Devon -3.529950197 50.70040529 120366933  33 Perthshire And Kinross -3.469979697 56.40034161 119876590  state lon lat likes  2 Belfast -5.960034425 54.60001223 84331627  31 Omagh -7.300004315 54.60001223 85009092 |
| --- |
| En este gráfico se evidencia aún más claramente la zona con la mayor cantidad de Me gustas es South Yorkshire con más de 130M de Me gustas le sigue después West Midlands con 121M. Las regiones que menos interactuaron con el Me gusta fueron Belfast y Omagh. |

| state lon lat dislikes  1 Bath And North East Somerset -2.350022218 51.3837486 12622353  39 Southampton -1.399976849 50.90003135 8617820  38 South Yorkshire -1.499996583 53.36667666 8451143  19 Inverclyde -4.750030763 55.93329002 8390226  17 Glasgow -4.250707236 55.87440472 8385133  state lon lat dislikes  7 Cardiff -3.22500757 51.49999473 3906929  26 Merseyside -2.917997886 53.41600181 3910656 |
| --- |
| Pocas regiones interactúan con el botón de No Me Gusta pero entre todas destaca `Bath And North East Somerset` con una gran cantidad de No Me Gustas en relación al resto de los estados. Exactamente la diferencia entre el primero y el segundo estado es de más de `4M` de No Me Gustas. |

Adicionalmente, al cliente le gustaría conocer si:

### Predicciones

#### ¿Es factible predecir el número de “Vistas” o “Me gusta” o “No me gusta”?

| Se asume que la cantidad de “Vistas” o “Me gusta” o “No me gusta” , depende de la popularidad del video |
| --- |
| dfs=df.sort\_values(by=['views'], ascending=False)  dfs[['views','likes','dislikes']] |
|  |

| No es factible predecir la cantidad de "vistas" , "me gusta", "no me gusta" , de un video porque el factor puede variar dependiendo del contenido del video , algunos pueden tener mayor vistas y tener mas "no me gusta" de lo que un video de menor vistas pueda tener mayor "me gustas" |
| --- |
| mask = (df['views'] >= df['likes'])  df[['likes','dislikes']]  dfa = df[df['likes'] > df['dislikes']]  dfb[['views','likes','dislikes']] |

#### ¿Los videos en tendencia son los que mayor cantidad de comentarios positivos reciben?

#### 

| Se asume que la cantidad de comentarios negativos se da cuando hay un alto número de dislikes (más alto que el de los likes), que los comentarios fueron deshabilitados o que se deshabilitaron la calificación por rating |
| --- |
| dfs=df.sort\_values(by=['views'], ascending=False)  dfs[['views','likes','dislikes','comments\_disabled','ratings\_disabled']] |
|  |

| Como se puede ver los videos con mayores views que son los que mas en tendencia están no tienen ese comportamiento |
| --- |
| mask = (df['likes'] >= df['dislikes'])  df[['likes','dislikes']]  dfa = df[df['likes'] < df['dislikes']]  dfb = dfa[df['comments\_disabled']==True]  '''  Notas del editor:  se usaria "DFC" pero como se ve en el resultado no hay valores con True en nuestro filtrado  '''  dfc = dfb[df['ratings\_disabled']==True]  dfb[['views','likes','dislikes','comments\_disabled','ratings\_disabled']] |
|  |
| Como se puede ver no necesariamente los videos con mayores views o que más tendencia marcan son los que tienen comentarios negativos |

# Modelización y Evaluación de datos

* Identificar qué variables en el conjunto de datos son susceptibles a ser modeladas.
* Describir el "conocimiento" que se intenta extraer a partir de la aplicación de un modelo de datos.
* Utilizar algún algoritmo innovador (según la pregunta a responder) para crear un modelo de datos (e. algoritmo de regresión lineal).
* Obtener métricas y realizar una evaluación del resultado obtenido en el modelo.

# Conclusiones del Proyecto

Algunos de los resultados extraídos del análisis son los siguientes:

* Las categorías más importantes para la plataforma de youtube y con más vistas son las de música y entretenimientos.
* La categoría de The Tonight Show Starring Jimmy Fallon es considerada la categoría con mayor tendencia en Youtube.
* Las regiones de Gran Bretaña con la mayor cantidad de vistas es

# Anexos

Repositorio de github: <https://github.com/thavs-college-repos/ea-2022-1-cc51>