Proyecto

Anomalías en la Contratación Pública

Materia: Despliegue de soluciones analíticas

Integrantes:

Gustavo Cordero

Nelson Eduardo Melo Salamanca

Andrés Arias García

Noviembre 2023

# Planteamiento del problema

La corrupción es el principal problema que impacta a Colombia, esto ha llevado a que la ciudadanía pase de la indignación a la acción realizando movilizaciones en contra de esta y que en los noticieros los temas de corrupción sean el foco de todos los días. Cabe resaltar que el tipo de corrupción más frecuente es la administrativa, que se da por las fallas en los sistemas, procesos y procedimientos de la administración pública en Colombia y el cual es el centro del proyecto a ejecutar.

# Contexto

La corrupción es un problema generalizado en Colombia que cada año les cuesta a los colombianos [50 billones de pesos,](https://cnnespanol.cnn.com/2018/06/06/senado-de-colombia-da-luz-verde-a-consulta-popular-que-endurecera-los-castigos-para-los-corruptos-que-sigue-ahora/) unos 18.400 millones de dólares, según cifras entregadas por el contralor general Eduardo Maya Villazón a principios de este año.1 Le quita al país el 5% del Producto Interno Bruto, el 15% del presupuesto nacional, y que afecta por sobornos y coimas el 18% de la contratación pública de la región.

Además, la práctica del soborno está muy arraigada en Colombia. Según la Cuarta Encuesta Nacional Sobre Prácticas Contra el Soborno en Empresas Colombianas, citada [en un informe sobre corrupción de la Universidad Externado de Colombia,](http://dernegocios.uexternado.edu.co/negociacion/10-cosas-que-no-sabias-sobre-la-corrupcion-en-colombia/) lanzado la semana pasada en Bogotá, el 91% de los empresarios participantes consideró “que secretamente se ofrecen dádivas para obtener contratos” y esos montos pueden alcanzar en promedio el 17,3% del valor del contrato, según la misma encuesta.1 En Colombia, los tipos de corrupción más frecuentes, [según la Universidad Externado,](http://dernegocios.uexternado.edu.co/negociacion/10-cosas-que-no-sabias-sobre-la-corrupcion-en-colombia/) son el soborno, la apropiación de bienes públicos, la extorsión y el nepotismo.

El tipo de corrupción más frecuente es la administrativa, que se da por las fallas en los sistemas, procesos y procedimientos de la administración pública. Por ejemplo, cuando los funcionarios públicos aceptan sobornos que los ciudadanos pagan “para acelerar los procesos burocráticos o acceder a los servicios públicos a los que tienen derecho”.

# Pregunta de Negocio

¿Cómo identificar y segmentar las anomalías en los tiempos y procesos de ejecución de los contratos públicos del Estado Colombiano para mejorar la transparencia y eficiencia?

# Alcance del proyecto

Para el alcance del proyecto se tienen definidas una serie de actividades las cuales se realizarán durante la ejecución de cada semana de la materia.

1. Definición del problema y de la fuente de información.
2. Preparación y análisis de los datos.
3. Análisis descriptivo de los datos.
4. Definición del modelo analítico a utilizar.
5. Preparación de los procesos de despliegue y configuración de las herramientas en Git.
6. Ejecución del modelo.
7. Capa de visualización.
8. Manuales de usuario capa de visualización.

Durante la ejecución del alcance se usarán herramientas habilitadas y descritas en el curso, adicional la

fuente de datos es de datos abiertos y que es habilitada por el estado de Colombia.

# Conjunto de datos

El conjunto de datos que utilizaremos durante la ejecución del proyecto es una fuente de datos abiertos, dicha fuente es SECOP II Información de los contratos públicos de Colombia. La fuente de datos tiene las siguientes características:

Fecha de creación: 30 de septiembre de 2019

Fecha actualización de los datos: 27 de octubre de 2023

Nombre entidad: AGENCIA NACIONAL DE CONTRATACION PUBLICA -COLOMBIA COMPRA EFICIENTE-|08033

Categoría: Gastos Gubernamentales

Frecuencia de Actualización: Diario

Cobertura Geográfica: Nacional

Columnas: 71

Filas: 300.000

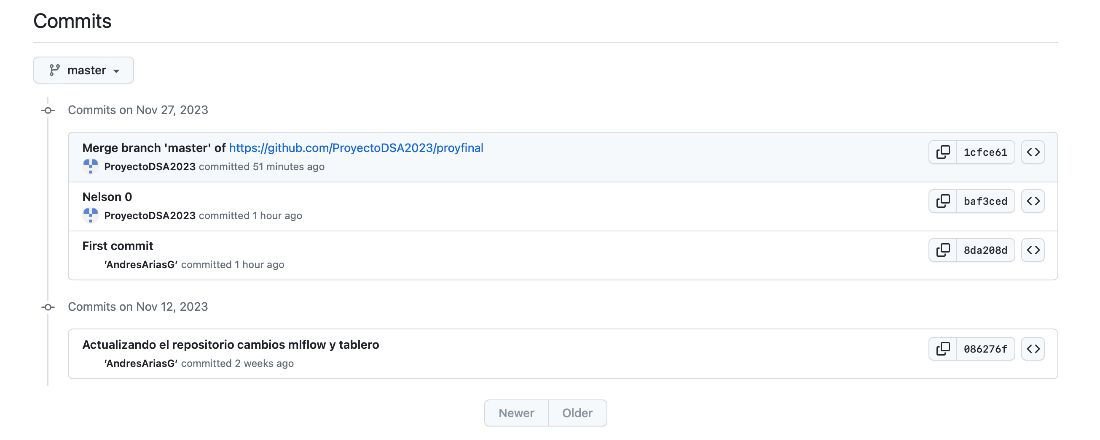
Peso: 380 MB.

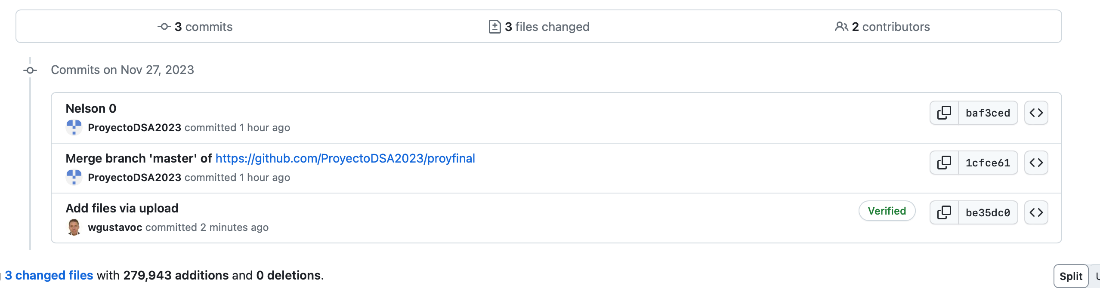
URL: <https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Contratos-Electr-nicos/jbjy-vk9h>

Nota: Debido al tamaño de los datos, y a las limitaciones del ejercicio se realizará un filtro de los datos que permita cumplir con las actividades del ejercicio del proyecto final.

# Posibles cambios realizados

1. Se realizo mejoras en los tableros, y en sus funcionalidades.
2. Se realizo ajustes en los despliegues de los tableros.
3. Se realizo proceso de trabajo remoto desde maquinas linux a git hub.

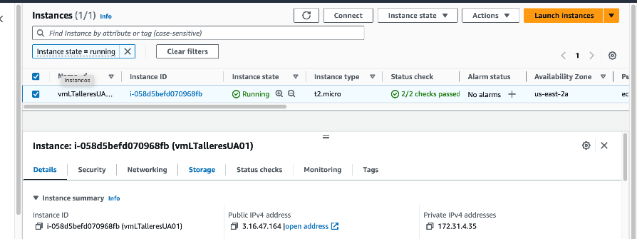




1. Se realizo la documentación de los tableros.
2. Se realiza ajustes en la publicación y despliegue en Git Hub: <https://github.com/ProyectoDSA2023/proyfinal.git>

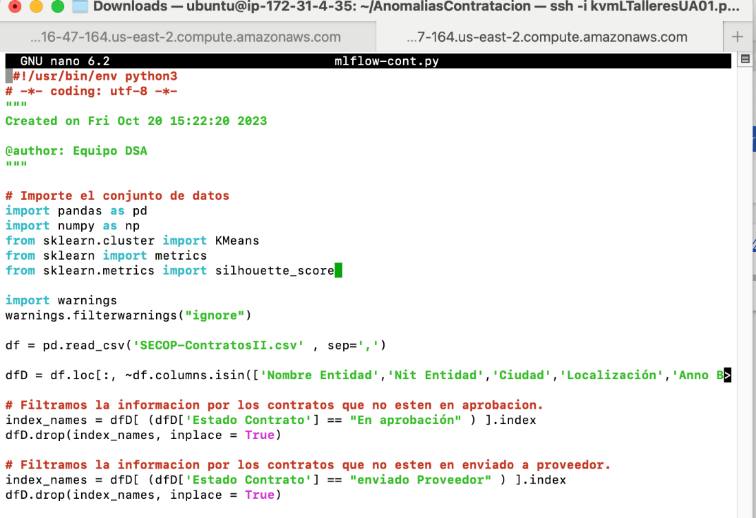
# Modelos desarrollados y evaluación

Se configuro una maquina en AWS para el desarrollo de todo el proyecto. Durante el desarrollo de las imágenes del modelo se idéntica la dirección IP que es la misma.



**MLFLOW**

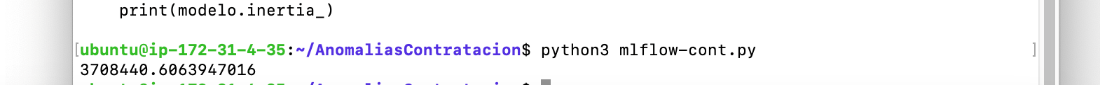
Para poder realizar este ejercicio, tuvimos que realizar una serie de configuraciones que nos permitieran correr la herramienta sobre Ubuntu, adicional el código del modelo de clusterización lo realizamos sobre un archivo de Python con extensión .py.

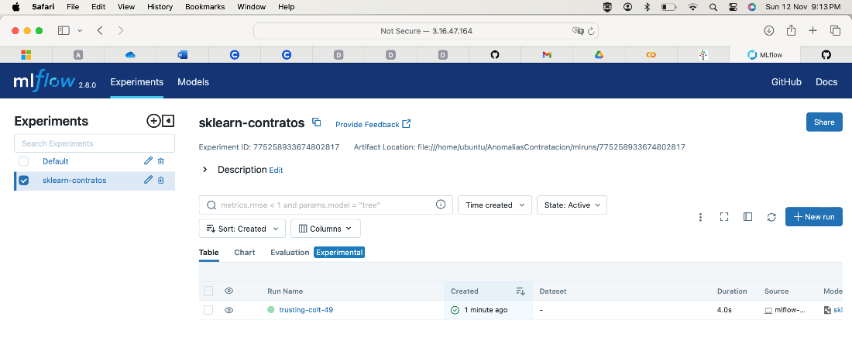


De acuerdo con el ejercicio propuesto realizamos el ejercicio del proyecto de segmentación de contratos, con la herramienta MLFLOW, para esto realizamos 3 ejercicios cambiando los parámetros del modelo que nos permitieron entender mejor su comportamiento:

Ejercicio 1:

Parámetros: n\_clusters=3, max\_iter=100



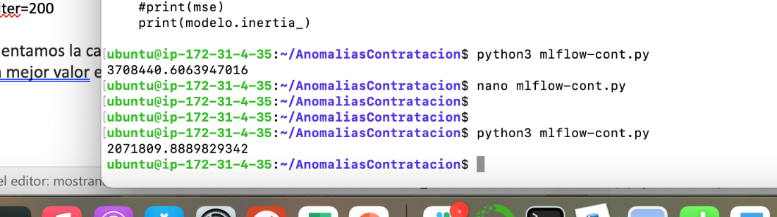


Como evaluación del modelo usamos la métrica de inertia\_, la cual nos permite identificar la capacidad de kmeans para agrupar un conjunto de datos. Esta medida va de la mano con la cantidad de clúster que se habiliten en los parámetros.

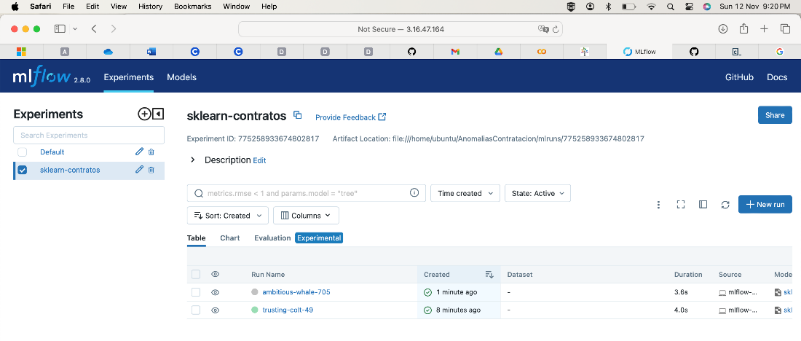
Ejercicio 2:

Parámetros: n\_clusters=4, max\_iter=200

Para esta nueva ejecución aumentamos la cantidad de clúster, y la cantidad de iteraciones. Como resultado nos arrojó un mejor valor en la capacidad de agrupación de los datos, con respecto al ejercicio anterior.



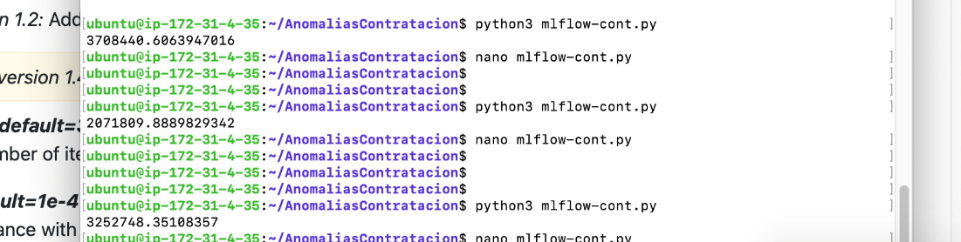
Al validar en MLFLOW, podemos ver la nueva ejecución.



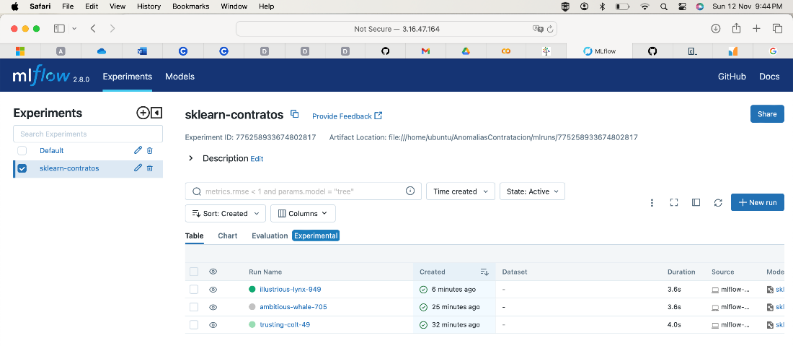
Ejercicio 3

Parámetros: init=**"random",** n\_clusters=4, max\_iter=100, n\_init= 20

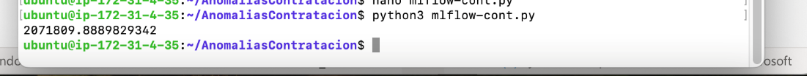
Para este último ejercicio realizamos un cambio de nuevo en los parámetros y adicionamos dos nuevos parámetros, para validar el comportamiento del modelo. Bajamos la cantidad de iteraciones y el resultado no fue tan positivo como la ejecución anterior.



Validamos la ejecución en la capa visual de MLFLOW para garantizar su ejecución.



En este último ejercicio esperábamos mejorar un poco más la capacidad de agrupación del cluster adicionando nuevos parámetros y cambiando los existentes, encontramos que no fue suficiente para mejorar el resultado del ejercicio. Realizamos un último ejercicio para entender el resultado y ampliamos la cantidad de interacciones y obtuvimos el mismo resultado del punto 2.



# Observaciones y conclusiones modelado

1. En MLFLOW encontramos restricciones para la generación de graficas de las métricas de KMEANS. Buscamos en foros y no encontramos una solución acertada que permitiera graficar la métrica propuesta para el ejercicio.
2. Como equipo decidimos limitar la cantidad de datos y el tamaño para la ejecución del ejercicio, al tener tanta volumetría encontramos limitación en los recursos de procesamiento.
3. Las herramientas seleccionadas nos dan la posibilidad de crear una tubería de datos (Pipeline) entre las salidas de los modelos y la visualización de los datos en un tablero de control (Dashboard), además, generan un desarrollo totalmente industrializado con entregas continuas.
4. Debido al volumen de información en SECOP genera una demora de unos segundos al actualizar o filtrar información en el tablero. Sin embargo, se logra visualizar el histograma con toda la información requerida.

# Descripción del tablero desarrollado

El tablero desarrollado toma como información de entrada un archivo .csv con el resultado de la simulación. Este archivo contiene la información más relevante de los contratos en el SECOP, y una columna final (generada por el modelo de simulación) con el resultado del clúster (o clasificación) para cada uno de los contratos.

El script que permite visualizar el tablero con esta información de SECOP se está ejecutando en la instancia de AWS con la siguiente información:

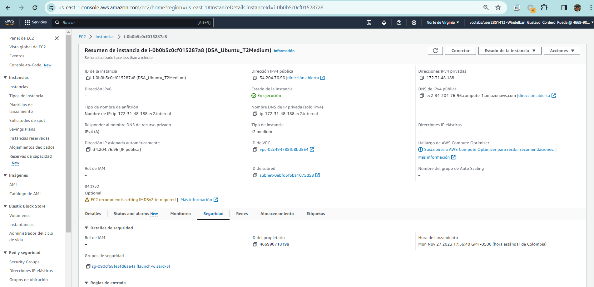


Fig. Definición de la instancia en AWS

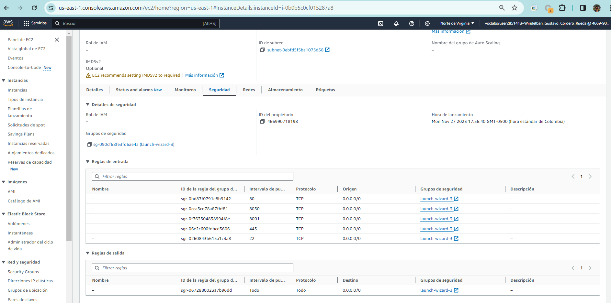


Fig. Instancia en AWS, puertos en uso

El tablero desarrollado durante esta fase del proyecto permite visualizar histogramas con los diferentes clústeres (o clasificación) generados por el modelo, y un listado con información para cada uno de los contratos en el SECOP tales como departamento, rama, clúster, etc.

Este tablero fue desarrollado usando Python y la librería dash, y a continuación podemos ver su ejecución desde la línea de comandos en la instancia de AWS

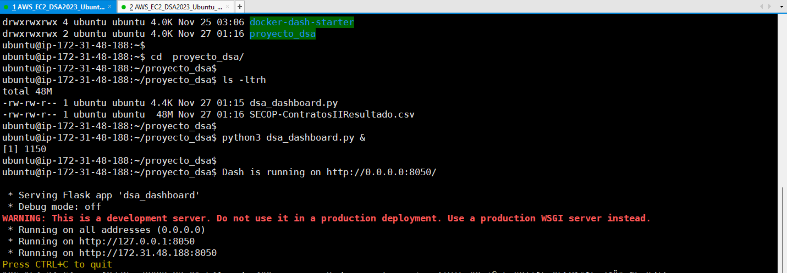
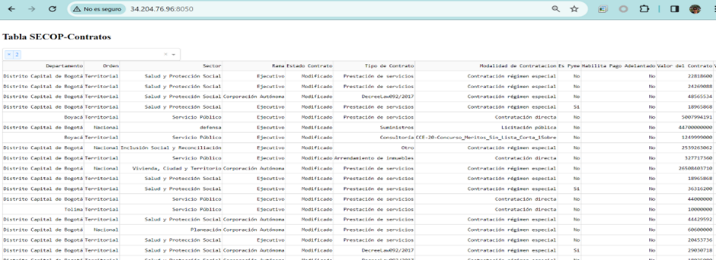


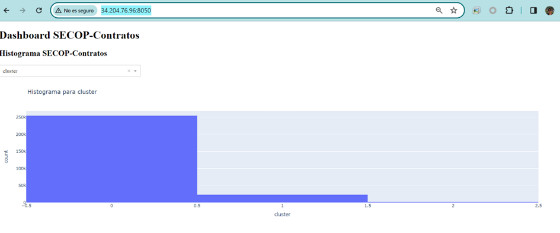
Fig. Ejecución de script en Python con tablero

Una vez ejecutado el script, podemos visualizar el tablero en cualquier navegador a través del siguiente enlace: <http://34.204.76.96:8050>

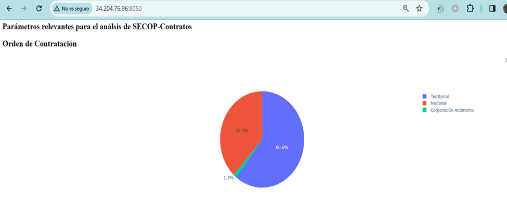
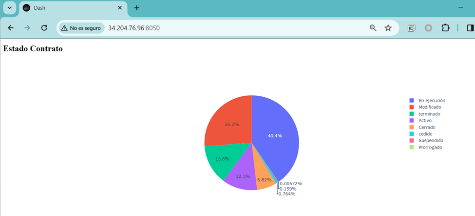
Allí se despliega la información de los contratos en el SECOP y los resultados del modelo con el clúster correspondiente para cada contrato:



La información en el histograma se puede filtrar por campos como clúster, sector, rama, etc. Por otra parte, la información en la tabla se puede filtrar por el clúster correspondiente para cada contrato lo cual permite visualizar similitudes como estado de contrato, y modalidad de contratación, que tienen una incidencia importante en cada clúster.



Se presenta presentan gráficos con indicadores y agregaciones de los datos, son relevantes para entender los análisis cobre la contratación.

La presentación gráfica de los resultados con gráficos de KPI permite a los usuarios finales realizar una rápida lectura e interpretación de las salidas de los modelos identificando el estado de los contratos seleccionados con los filtros.

# Observación y conclusión visualización

Se logró presentar de forma fácil e intuitiva los resultados de la analítica de datos agregando una capa de visualización. El tablero se convierte en una herramienta de uso diario para realizar un monitoreo efectivo sobre la contratación en el país.

# Trabajo en equipo

|  |  |  |  |
| --- | --- | --- | --- |
| **Actividad** | **Responsable** | **Fecha entrega** | **Estado** |
| Problema y contexto | Andrés Arias | Miércoles 08 de noviembre | Entregado |
| Pregunta de negocio y alcance | Andrés Arias | Miércoles 08 de noviembre | Entregado |
| Desarrollo mlflow y modelos | Andrés Arias | Sábado 11 de noviembre | Entregado |
| Desarrollo tablero | Nelson Melo y Gustavo Cordero | Sábado 11 de noviembre. | Entregado |
| Construcción de reporte | Andrés Arias, Nelson Melo, Gustavo Cordero | Sábado 11 de noviembre – domingo 12 de noviembre | Entregado |
| Ajustes y mejoras a los tableros | Nelson Melo – Gustavo Cordero | Lunes 20 de noviembre – Domingo 26 de noviembre | Entregado |
| Manuales de uso de los tableros | Gustavo Cordero | Domingo 26 de noviembre a lunes 27 de noviembre. | Entregado |
| Video resumen del proyecto | Andrés Arias, Nelson Melo, Gustavo Cordero | Lunes 27 de noviembre | Entregado |

Durante la ejecución de la materia, como equipo realizamos un trabajo por actividades y responsables buscando potenciar las fortalezas de todo el equipo buscando asegurar el aprendizaje y la correcta finalización de cada uno de los entregables finales. Cada integrante del equipo realizo un valioso aporte en cada uno de los entregables del proyecto.

Como equipo tuvimos algunos retos los cuales nos apoyamos para sacarlos adelante algunos de ellos fueron: Publicación git de forma remota, Compartir los tableros para su visualización entre otros.

# Anexos

1. ***Mflow-cont.py***: Archivo que permite la ejecución del modelo de clusterización realizado en kmeans.
2. ***script\_dash.py***: Archivo con la codificación del tablero de control (Dashboard). Incluye la creación de los objetos gráficos, su formato y conexión a los datos.
3. ***SECOP-ContratosIIResultado.csv:*** Fuente de datos resultado de la ejecución de los modelos. Se utiliza dentro del tablero de control (Dashboard) para modelar los datos y construir las visualizaciones.
4. Manuales de los tableros.
5. Vídeo de trabajo realizado y conclusiones: https://youtu.be/-C46qj2T1-4