# Corte de Proyecto de Programación Avanzada

## Tema: Control de versiones de datos

## Participantes: Raidel Antonio Valdés Roseñada

Yohan Sebastián González Veriel

## Descripción breve del proyecto

El principal objetivo del proyecto es el Control de versiones de datos, como su nombre lo indica, pero específicamente a un conjunto de datos de entrenamiento de un modelo de aprendizaje automático. El modelo deberá ser capaz de detectar los datos nuevos que se le introduzcan y no pertenezcan a ninguno de los clústeres que él tiene definido. Luego, dichos datos serán agrupados según sus características comunes en nuevos clústeres y almacenados en una base de datos.

## Requerimientos del proyecto

¿Qué deben lograr con el proyecto? Se pueden asociar con los objetivos. Se escriben por puntos y en infinitivo.

-Desarrollar un modelo de aprendizaje automático

-Agrupar los datos que no pertenezcan a los clústeres de entrenamiento del modelo en nuevos clústeres.

-Lograr que el modelo genere nuevas versiones del conjunto de datos de entrenamiento inicial, utilizando la herramienta DVC a partir de un conjunto de datos y los nuevos datos que ingresen se ya no pertenezcan a este, y así realizar una predicción más precisa con la creación de nuevos clústeres.

-Crear interfaz de manera que el usuario sea capaz de nombrar los nuevos clústeres.

## Concepción de la solución

Esquema con la arquitectura del software especificando cada componente del software. Breve descripción de cada uno de los componentes del software indicando qué función cumplen.

Selección

Dato

Modificación

Registro

Programa

Método de Machine Learning: Define el método que se implementará en el programa para la selección de la categoría a la que pertenecerá el dato ingresado.

Conjunto de datos de entrenamiento: conjunto de datos que dará soporte al método de Machine Learning para el proceso de selección.

DVC: programa mediante el cual se le dará tratamiento a los datos ingresados que no pertenezcan a las categorías actuales de selección en base al conjunto de entrenamiento actual y los datos que no pertenezcan a este (los cuales ira registrando), este proceso se realizará condicionalmente dada una cantidad relevante de datos no pertenecientes, de esta manera modificará nuestro conjunto de entrenamiento, creando así nuevos clústeres a partir de estos nuevos datos.