# Modelos Y Simulación De Sistemas 1

## Proyecto De Semestre - Entrega 1


**Responsables**

Eliana Janneth Puerta Morales

Valentina Muñoz Rincón

Juan Fernando Lopera Muñoz

Medellín - Antioquia 2023

# Descripción del problema

En esta competición, se solicita predecir el tipo de cobertura forestal predominante (el tipo principal de vegetación arbórea) a partir de variables cartográficas estrictamente, en lugar de datos obtenidos por sensores remotos. El tipo de cobertura forestal real para una celda de 30 x 30 metros determinada se obtuvo a partir de datos del Sistema de Información de Recursos de la Región 2 del Servicio Forestal de EE. UU. (USFS). Las variables independientes se derivaron posteriormente a partir de datos obtenidos del Servicio Geológico de EE. UU. y del USFS. Los datos se encuentran en su forma original (sin escalar) y contienen columnas binarias de datos para variables independientes cualitativas, como áreas silvestres y tipos de suelo.

Esta área de estudio abarca cuatro áreas silvestres ubicadas en el Bosque Nacional Roosevelt del norte de Colorado. Estas áreas representan bosques con disturbios mínimamente causados por actividades humanas, de modo que los tipos de cobertura forestal existentes son más resultado de procesos ecológicos que de prácticas de gestión forestal.

# Dataset

El dataset que se va a utilizar es (https://www.kaggle.com/competitions/forest-cover-type-prediction/) el cual consta de 15.120 instancias y 56 columnas.

### Descripción de columnas

- **Elevation**: Se refiere a la altitud en metros.
- **Aspect**: Representa la orientación en grados azimutales.
- **Slope**: Indica la inclinación en grados del terreno.
- **Horizontal_Distance_To_Hydrology**: Es la distancia horizontal a las características de agua superficial más cercanas.
- **Vertical_Distance_To_Hydrology**: Es la distancia vertical a las características de agua superficial más cercanas.
- **Horizontal_Distance_To_Roadways**: Muestra la distancia horizontal a la carretera más cercana.
- **Hillshade_9am**: Es un índice de sombreado a las 9 a.m. durante el solsticio de verano.
- **Hillshade_Noon**: Es un índice de sombreado al mediodía durante el solsticio de verano.
- **Hillshade_3pm**: Es un índice de sombreado a las 15:00 durante el solsticio de verano.
- **Horizontal_Distance_To_Fire_Points**: Indica la distancia horizontal a los puntos de ignición de incendios forestales más cercanos.

- **Wilderness_Area (4 columnas, 0 = Ausencia o 1 = Presencia)**: Designa la presencia o ausencia en cuatro áreas silvestres distintas.
- **Soil_Type (40 columnas, 0 = Ausencia o 1 = Presencia)**: Designa la presencia o ausencia de cuarenta tipos diferentes de suelo.
- **Cover_Type (7 tipos, Enteros del 1 a 7)**: Representa siete tipos diferentes de cubierta forestal mediante números enteros del 1 al 7.

**Los datos categóricos:**

Área silvestre (Wilderness_Area), Tipo de suelo (Soil_Type) y Tipo de cubierta forestal (Cover_Type).

**Los datos simulados:**

El dataset no cumplía con el requisito de que al menos 3 columnas del dataset le faltaran el 5% de los datos, por lo tanto procedimos a realizar una simulación aleatoria en la cuál quitamos más del 5% de los datos de las siguientes columnas:

- Horizontal_Distance_To_Hydrology
- Vertical_Distance_To_Hydrology
- Horizontal_Distance_To_Roadways
- Horizontal_Distance_To_Fire_Points

# Métrica

La métrica que se va a utilizar en este proyecto es exactitud (accuracy en inglés). Va a calcular la precisión del modelo a la hora de predecir cual es el tipo de cobertura forestal en Colorado, EE.UU. Esta métrica se trata de la proporción de predicciones correctas en relación con todas las predicciones realizadas, en otras palabras, va a medir cuántas de las predicciones son correctas en comparación con el total de predicciones realizadas.

La exactitud está definida por la siguiente fórmula:

$$Accuracy = \frac{Número de predicciones correctas}{Número total de predicciones}$$

**Primer criterio:**

El modelo debe tener una exactitud del 70% para ser considerado óptimo, ya que permitirá observar la tendencia de las coberturas forestales en las zonas determinadas.