# UNIVERSIDAD DE ANTIOQUIA


# Inteligencia Artificial para las Ciencias e Ingenierías

## Proyecto de semestre - Entrega 1

INTEGRANTES:

Silvio Otero Guzmán

Daniela Gonzáles Estrada

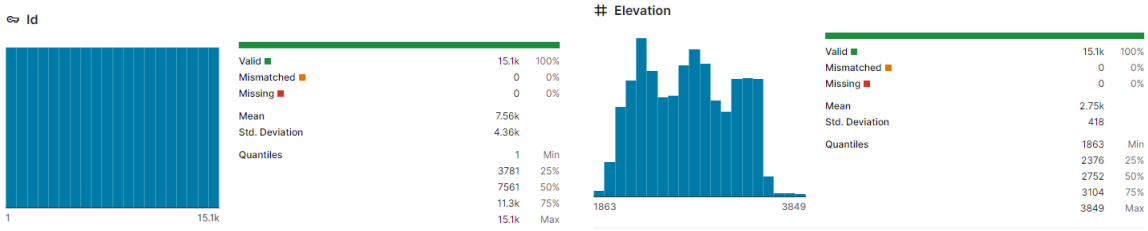
Medellín - Antioquia 2023

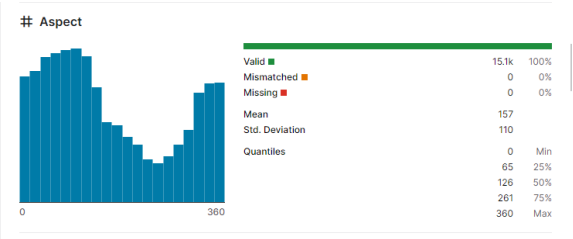
# Descripción del problema

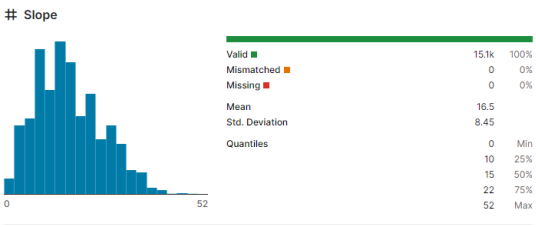
Los datos que se trabajarán fueron obtenidos a partir de cuatro áreas silvestres localizadas en el bosque nacional Roosevelt en carolina del norte. Áreas representadas en bosques con mínima interferencia humana, el follaje actual es más un resultado de procesos ecológicos que de prácticas de manejo de bosques. Con la información suministrada se puede determinar cuáles arboles predominan en el área de cobertura, los datos fueron obtenidos del US Forest Service y del US Geological Survey. Mediante esta información se desea generar un modelo predictivo para determinar que tipo de arboles tienden a la extinción por ser los de menor cantidad en las áreas analizadas, teniendo en cuenta que en los datos ofrecen información de distintos tipos de arboles y las condiciones en las que se encuentran.

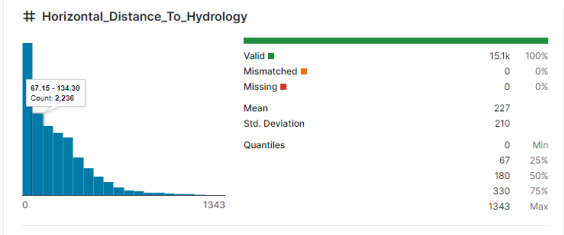
# Dataset

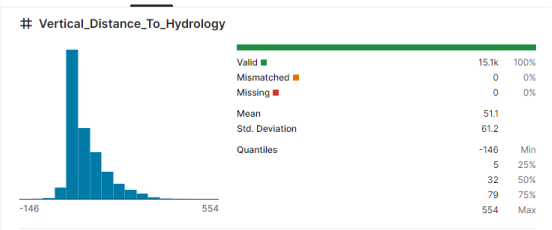
El dataset a utilizar será Forest Cover Type Prediction https://www.kaggle.com/competitions/forest-cover-type-prediction/data?select=train.csv que consta de 118 columnas y más de 15 mil instancias, cumpliendo los requisitos solicitados por el proyecto. Algunas de las distribuciones de los datos están dadas así:






# Métricas de desempeño requeridas

## Métricas de desempeño

Tomando en cuenta que el problema se ha planteado como una clasificación multi clase las métricas de desempeño a usar para entender de mejor manera el modelo pueden ser las mismas que las usadas con una clasificación binaria. La métrica se calcula para cada clase al procesarla como un problema de clasificación binaria después de agrupar todas las otras clases como pertenecientes a la segunda clase. A continuación, se calcula el promedio de la métrica entre todas las clases para obtener una métrica de promedio macro o de media ponderada. Las que se usarán son las siguientes:

- Accuracy
- Matriz de confusión
- Precisión
- Medición F1

## Métricas del negocio

La extinción de ciertas especies de arboles es en algunos casos un fenómeno inevitable, al conocer cuáles de ellos tienden a no ser capaces de adaptarse al cambio en el mundo y tener más posibilidad de extinción plantea una reevaluación en la escogencia de arboles con el fin de reforestar áreas con condiciones que podrían llegar a asemejarse a las mostradas en estos datos.

# Primer criterio sobre cuál sería el desempeño deseable en producción

Al estar hablando de datos útiles a la hora de toma de decisiones con respecto a la reforestación de áreas es muy importante unos resultados precisos ya que al ser proyectos a largo plazo los resultados negativos no serán visibles en poco tiempo y el impacto negativo sería aún mayor. Tomando esto en cuenta se hace necesaria una tasa de acierto mayor al 80%. Cuando de valor de sensibilidad se habla, se hace necesario que sea mayor al 75% ya que si el modelo predice que un árbol tenderá a la extinción este posiblemente sea descartado de futuros proyectos y uno que en realidad si tienda a la extinción sea agregado provocará que posiblemente este ultimo no sea capaz de llegar a etapas adultas y no se cumpla el propósito de reforestación.