PROYECTO _ ENTREGA 1

- **DESCRIPCIÓN DEL PROBLEMA PREDICTIVO**

El desafío consiste en la clasificación de 72 clases de cobertura del suelo (por ejemplo, Bare Land ,Grassland without tree/shrub cover, Coniferous woodland, Potatoes, Vineyards) en cinco años diferentes (2006, 2009, 2012, 2015, 2018) utilizando un conjunto de datos que incluye más de 42,237 muestras in situ y 416 características derivadas de datos de teledetección e imágenes satelitales. Los participantes deberán predecir la clase de cobertura del suelo en un conjunto de pruebas que contiene alrededor de 42,271 muestras que no están presentes en el conjunto de datos de entrenamiento.

Con esto se tiene la oportunidad de ampliar los límites del conocimiento de vanguardia y los enfoques de ML aplicados a la clasificación de la cubierta terrestre utilizando un conjunto de datos de alta resolución temática.

Evaluación

Las presentaciones se evalúan utilizando la puntuación F1 ponderada entre las clases de cobertura del suelo previstas y esperadas para el conjunto de prueba:

F1 = 2 * (precisión * recall) / (precision + recall)

Para cada uno en el conjunto de prueba, deben predecir la clase de cobertura del suelo.

- **DATASET**

vamos a usar el dataset de kaggle esta competición (https://www.kaggle.com/competitions/oemc-hackathon-eu-land-cover-classification/data)

Para abordar este desafío de mapeo de la cubierta terrestre en Europa, se utilizará un conjunto de datos de alta resolución temática. A continuación, se proporciona una descripción detallada de este dataset:

Clases de Cobertura del Suelo: El conjunto de datos contiene un total de 72 clases de cobertura del suelo, que incluyen categorías como Bare Land, Grassland without tree/shrub cover, Coniferous woodland, Potatoes, Vineyards, entre otras. Estas clases representan diferentes tipos de paisajes y cobertura vegetal.

Años de Datos: El conjunto de datos abarca un período de cinco años: 2006, 2009, 2012, 2015 y 2018. Esto significa que se dispone de datos de clasificación de la cubierta terrestre para cada uno de estos años.

Muestras In Situ: Se cuenta con un conjunto de más de 42,237 muestras in situ. Estas muestras representan observaciones directas de la cubierta terrestre tomadas en el campo.

Características y Covariables: Para cada muestra, se han derivado un total de 416 características o covariables utilizando datos de teledetección e imágenes satelitales. Estas características proporcionan información adicional que se utilizará para la clasificación de la cubierta terrestre.

Conjunto de Prueba: Además del conjunto de entrenamiento, se proporciona un conjunto de prueba que contiene alrededor de 42,271 muestras. Estas muestras no están presentes en el conjunto de entrenamiento y se utilizarán para evaluar el rendimiento de los modelos de aprendizaje automático desarrollados.

Este dataset proporciona una base sólida para abordar el desafío de clasificar la cubierta terrestre en Europa a lo largo del tiempo. La combinación de múltiples clases, años de datos y características derivadas de teledetección ofrece un entorno de trabajo desafiante pero prometedor para el desarrollo de modelos de aprendizaje automático.

Conjunto de entrenamiento con 42,237 filas y 420 columnas, incluida la identificación de la muestra. Las columnas están formadas por seis campos de metadatos separados por _:

Campos de metadatos:

- F1 - Nombre de la variable: rojo
- F2 - Procedimiento variable incluido el nombre del producto: landsat.glad.ard
- F3 - Posición en la distribución de probabilidad: p50
- F4 - Resolución espacial: 30m
- F5 - Fecha de inicio: 25 de junio
- F6 - Fecha de finalización: 12 de septiembre

DESCRIPCIÓN DE LA COLUMNA:

Todas las columnas se pueden agregar en seis grupos temáticos según F1 y F2:

Satellite Images (Imágenes Satelitales - Reflectancia Espectral e Índices de Vegetación):

- blue_landsat.glad.ard_{..}
- blue_mod13q1_{..}
- evi_mod13q1.stl.trend.ols.alpha_{..}
- evi_mod13q1.stl.trend.ols.beta_{..}
- evi_mod13q1.stl.trend_{..}
- evi_mod13q1_{..}

- green_landsat.glad.ard_{..}
- mir_mod13q1_{..}
- ndvi_mod13q1_{..}
- nir_landsat.glad.ard_{..}
- nir_mod13q1_{..}
- red_landsat.glad.ard_{..}
- red_mod13q1_{..}
- swir1_landsat.glad.ard_{..}
- swir2_landsat.glad.ard_{..}

Temperature Images (Imágenes de Temperatura):

- lst_mod11a2.daytime_{..}
- lst_mod11a2.daytime.{month}_{..}
- lst_mod11a2.daytime.trend_{..}
- lst_mod11a2.daytime.trend.ols.alpha_{..}
- lst_mod11a2.daytime.trend.ols.beta_{..}
- lst_mod11a2.nighttime_{..}
- lst_mod11a2.nighttime.{month}_{..}
- lst_mod11a2.nighttime.trend_{..}
- lst_mod11a2.nighttime.trend.ols.alpha_{..}
- lst_mod11a2.nighttime.trend.ols.beta_{..}
- thermal_landsat.glad.ard_{..}

Climate Layers (Capas Climáticas):

- accum.precipitation_chelsa.annual_{..}
- accum.precipitation_chelsa.annual.3years.dif_{..}
- accum.precipitation_chelsa.annual.log.csum_{..}
- accum.precipitation_chelsa.montlhy_{..}
- bioclim.var_chelsa.{variable_code}_{..}

Accessibility & Distance Maps (Mapas de Accesibilidad y Distancia):

- accessibility.to.ports_map.ox.{variable_code}_{..}
- burned.area.distance_global.fire.atlas_{..}
- cost.distance.to.coast_gedi.grass.gis_{..}
- road.distance_osm.highways.high.density_{..}
- road.distance_osm.highways.low.density_{..}
- water.distance_glad.interanual.dynamic.classes_{..}

Digital Terrain Model (Modelo Digital del Terreno):

- elev.lowestmode_gedi.eml_{..}
- slope.percent_gedi.eml_{..}

Other Existing Maps (Otros Mapas Existentes):

- pop.count_ghs.jrc_{..}
- snow.duration_global.snowpack_{..}

- **MÉTRICAS DE DESEMPEÑO REQUERIDAS**

En este desafío de mapeo de la cubierta terrestre en Europa, se utilizarán métricas de desempeño específicas para evaluar la precisión y eficacia de los modelos de aprendizaje automático. Las métricas requeridas incluyen:

Puntuación F1 Ponderada (Weighted F1 Score): La métrica principal de desempeño será la puntuación F1 ponderada. Esta métrica tiene en cuenta tanto la precisión (precisión) cómo la exhaustividad (recall) del modelo en la clasificación de las clases de cobertura del suelo. La fórmula de la puntuación F1 ponderada es la siguiente:

$$F1=2*(precision*recall)/(precision+recall)$$

Esta métrica se calculó para cada clase de cobertura del suelo y se promedió de acuerdo con el peso de cada clase en el conjunto de datos de prueba.
El uso de la puntuación F1 ponderada es apropiado para este desafío, ya que proporciona una medida equilibrada de la precisión y la exhaustividad, lo que es esencial cuando se trata de clasificar múltiples clases de cobertura del suelo.

Es importante que los participantes se esfuercen por lograr una alta puntuación F1 ponderada en la clasificación de todas las clases de cobertura del suelo, ya que esto indicará un desempeño sólido en la tarea de mapeo de la cubierta terrestre.

Esperamos que esta métrica de desempeño ayude a los participantes a evaluar y comparar sus modelos de manera efectiva durante el desarrollo del proyecto.

- **DESEMPEÑO DESEABLE EN PRODUCCIÓN.**

El criterio de desempeño deseable en producción para este proyecto de mapeo de la cubierta terrestre en Europa es alcanzar una alta precisión y eficacia en la clasificación de las clases de cobertura del suelo en un entorno operativo en tiempo real. Esto significa que los modelos de aprendizaje automático desarrollados deben ser capaces de proporcionar resultados precisos y confiables al enfrentarse a datos de observación de la Tierra en situaciones del mundo real.

Para establecer un criterio específico de desempeño deseable en producción, se recomienda lo siguiente:

Precisión Global Alta: Se espera que el modelo alcance una alta precisión global en la clasificación de la cubierta del suelo, lo que significa que la mayoría de las predicciones deben ser correctas en general. Un objetivo razonable podría ser una precisión global superior al 90%.

Desempeño Consistente: El modelo debe mantener un buen desempeño en la clasificación de todas las clases de cobertura del suelo. No debe haber clases particularmente problemáticas con un rendimiento significativamente inferior. Se podría establecer un criterio de que la puntuación F1 ponderada promedio para todas las clases sea superior al 80%.

Eficiencia en Tiempo Real: El modelo debe ser capaz de realizar predicciones en tiempo real o con una latencia mínima, ya que se espera que se utilice en aplicaciones de observación de la Tierra en tiempo real.

Robustez ante Cambios Temporales: Dado que el conjunto de datos abarca cinco años diferentes, el modelo debe ser capaz de adaptarse y proporcionar resultados precisos en diferentes momentos temporales. Esto implica que el modelo no debe perder precisión a medida que se aplica a datos de años más recientes.

Escalabilidad: Si se espera que el modelo se utilice en una escala más amplia, debe ser escalable y capaz de manejar grandes volúmenes de datos de manera eficiente.