Primera entrega de proyecto

POR:

Santiago Ramírez Pérez

Felipe Sánchez Londoño

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raúl Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2021

1. Planteamiento del problema

La naturaleza nos presenta su diversidad de diversas maneras, muchas veces no tenemos ideas de cuantas maneras estamos hablando, y esa pluralidad de formas, texturas, tamaños, etc… genera curiosidad por conocer sus características y de qué forma se podría representar esa diversidad. En este proyecto se debe predecir el tipo de cubierta forestal (el tipo predominante de cubierta arbórea) a partir de variables estrictamente cartográficas (a diferencia de los datos de detección remota). El dataset fue realizado y facilitado por el Sistema de Información de Recursos de la Región 2 del Servicio Forestal de EE. UU. (USFS) y Servicio Geológico de EE. UU. Los datos están en formato bruto (no escalados) y contienen columnas binarias de datos para variables independientes cualitativas, como áreas silvestres y tipo de suelo.

El área de estudio son cuatro áreas silvestres ubicadas en el Bosque Nacional Roosevelt del norte de Colorado. Cada observación es un parche de 30m x 30m. Se le pide que prediga una clasificación entera para el tipo de cubierta forestal. Los siete tipos son:

1 - Picea/abeto
2 - Pino Lodgepole
3 - Pino Ponderosa
4 - Álamo/Sauce
5 - Álamo temblón
6 - Abeto de Douglas
7 - Krummholz

El conjunto de entrenamiento (15120 observaciones) contiene tanto entidades como Cover_Type (el tipo de cubierta forestal). El conjunto de prueba contiene solo las funciones. Debe predecir Cover_Type para cada fila en el conjunto de prueba (565892 observaciones).

2. Dataset

El dataset que usaremos es de una competencia de kaggle en la cual se proporcionan datos cartográficos para cada celda de 30m x 30m de corteza forestal, estos son:

- Elevation - Elevación en metros.
- Aspect - Aspecto en grados de acimut.
- Slope - Pendiente en grados.
- Horizontal_Distance_To_Hydrology - Horz Dist a las características de agua superficial más cercanas.
- Vertical_Distance_To_Hydrology - Vert Dist a las características de agua superficial más cercanas.

- Horizontal_Distance_To_Roadways - Horz Dist a la carretera más cercana.
- Hillshade_9am (índice 0 a 255) - Hillshade índice a las 9 a. m., solsticio de verano.
- Hillshade_Noon (índice de 0 a 255) - Índice de sombreado al mediodía, solsticio de verano.
- Hillshade_3pm (índice de 0 a 255) - Índice de sombreado a las 3 p. m., solsticio de verano.
- Horizontal_Distance_To_Fire_Points- Dist Horz a los puntos de ignición de incendios forestales más cercanos.
- Wilderness_Area (4 columnas binarias, 0 = ausencia o 1 = presencia) - Designación de área silvestre.
- Soil_Type (40 columnas binarias, 0 = ausencia o 1 = presencia) - Designación de tipo de suelo.
- Cover_Type (7 tipos, números enteros 1 a 7) - Designación del tipo de cubierta forestal.

Para conocer cuáles son las 4 áreas silvestres y los 40 tipos de suelo los invito a miren la [página oficial de kaggle de este proyecto]

- https://www.kaggle.com/competitions/forest-cover-type-prediction/overview/description

- **train.csv:**

Es un archivo con los datos de entrenamiento(con 15120 instancias), descritos anteriormente.

- **test.csv**

Son los datos de prueba(con más de 500.000 instancias), que tiene la misma naturaleza que los datos de entrenamiento, en este caso hay que predecir la columna Cover_Type, con el tipo de corteza a la que pertenezca del 1 al 7.

- **sample_submission.csv**

Un archivo de envío de muestra en el formato correcto.

3. Métricas

La métrica de evaluación principal para el modelo será el porcentaje de precisión multiclases.

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Donde:

P: Condición positiva. El número de casos positivos reales en los datos.

N: Condición negativa. El número de casos negativos reales en los datos.

TP: True positive. Un resultado de prueba que indica correctamente la presencia de una condición o característica.
TN: True negative. Un resultado de prueba que indica correctamente la ausencia de una condición o característica.

FP: False positive. Un resultado de prueba que indica erróneamente que una condición o atributo en particular está presente.

FN: False negative. Un resultado de prueba que indica erróneamente que una condición o atributo en particular está ausente.

4. Desempeño

De este modelo esperamos que de poder integrarse en la investigación profesional pueda detectar eficientemente cada tipo de corteza dentro de las características en las que se hizo este proyecto, y poder seguir avanzando con otras características a medida que la inversión lo permita.

5. Bibliografía

- FOREST COVER TYPE PREDICTION – Use cartographic variables to classify forest categories | Kaggle. (2022). Retrieved 4 July 2022, from https://www.kaggle.com/competitions/forest-cover-type-prediction/overview/description