Primera entrega de proyecto

Introducción a la Inteligencia Artificial para las Ciencias e Ingenierías

Presentado por

José David Bustamante Sierra, Andrea Fernanda Muegues Pedraza

Presentado a

Prof. Raúl Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

Medellín

6 de julio de 2022

Introducción

A continuación se describen las características de un problema de clasificación multiple, para la resolución del cual se empleará un dataset obtenido en la plataforma Kaggle. Se describe resumidamente el dataset, se presenta la métrica de desempeño principal que habrá de emplearse y se hace una breve reflexión con respecto a los criterios de desempeño deseables en producción.

1. Descripción del problema

A partir de características cartográficas se busca predecir el tipo de cubierta arbórea predominante en un área. Para esto, se cuenta con información de los espacios naturales y tipo de suelo de cada región. La zona de estudio incluye cuatro áreas silvestres situadas en el Bosque Nacional Roosevelt del norte de Colorado. Las áreas silvestres son:

1. Área silvestre de Rawah.
2. Área silvestre de Neota.
3. Área silvestre de Comanche Peak.
4. Área silvestre de Cache la Poudre.

Cada observación consta de un área de 30m x 30m, para la cual se predice el tipo de cubierta forestal. Los siete tipos son:

1. Picea/Abeto.
2. Pino Lodgepole.
3. Pino Ponderosa.
4. Algodonero/ Sauce.
5. Aspen.
6. Abeto Douglas.
7. Krummholz.

2. Descripción del dataset

El dataset que se va a emplear es el de la competición de Kaggle Forest Cover Type Prediction. Este dataset cuenta con 565892 observaciones, y 56 columnas que contienen las características cartográficas de cada región observada, las cuales comprenden:

- Elevación en metros.
- Ascensión en grados acimutales.
- Pendiente en grados.
- Distancia horizontal a las fuentes de agua superficial más cercanas.
- Distancia vertical a las fuentes aguas superficiales más cercanas.
- Distancia horizontal a la calzada más cercana.
- Índice de sombra a las 9 am, solsticio de verano.
- Índice de sombra a mediodía, solsticio de verano.
- Índice de sombra a las 3 pm, solsticio de verano.
- Distancia horizontal a los puntos de ignición de incendios forestales más cercanos.
- Designación de área silvestre.
- Designación del tipo de suelo.
- Designación del tipo de cubierta forestal.

En cuanto a la variable que designa el tipo de suelo, esta se encuentra dividida en 40 columnas ('Soil_type#', con $\#= 1, 2, 3..,40$, empezando por Cathedral family), indicando la presencia o ausencia, 1 o 0 respectivamente, de los siguientes tipos de suelo:

1. Cathedral family - Rock outcrop complex, extremely stony.

2. Vanet - Ratake families complex, very stony.
3. Haploborolis - Rock outcrop complex, rubbly.
4. Ratake family - Rock outcrop complex, rubbly.
5. Vanet family - Rock outcrop complex complex, rubbly.
6. Vanet - Wetmore families - Rock outcrop complex, stony.
7. Gothic family.
8. Supervisor - Limber families complex.
9. Troutville family, very stony.
10. Bullwark - Catamount families - Rock outcrop complex, rubbly.
11. Bullwark - Catamount families - Rock land complex, rubbly.
12. Legault family - Rock land complex, stony.
13. Catamount family - Rock land - Bullwark family complex, rubbly.
14. Pachic Argiborolis - Aquolis complex.
15. unspecified in the USFS Soil and ELU Survey.
16. Cryaquolis - Cryoborolis complex.
17. Gateview family - Cryaquolis complex.
18. Rogert family, very stony.
19. Typic Cryaquolis - Borohemists complex.
20. Typic Cryaquepts - Typic Cryaquolls complex.
21. Typic Cryaquolls - Leighcan family, till substratum complex.
22. Leighcan family, till substratum, extremely bouldery.
23. Leighcan family, till substratum - Typic Cryaquolls complex.
24. Leighcan family, extremely stony.
25. Leighcan family, warm, extremely stony.
26. Granile - Catamount families complex, very stony.
27. Leighcan family, warm - Rock outcrop complex, extremely stony.
28. Leighcan family - Rock outcrop complex, extremely stony.
29. Como - Legault families complex, extremely stony.
30. Como family - Rock land - Legault family complex, extremely stony.
31. Leighcan - Catamount families complex, extremely stony.
32. Catamount family - Rock outcrop - Leighcan family complex, extremely stony.
33. Leighcan - Catamount families - Rock outcrop complex, extremely stony.
34. Cryorthents - Rock land complex, extremely stony.
35. Cryumbrepts - Rock outcrop - Cryaquepts complex.
36. Bross family - Rock land - Cryumbrepts complex, extremely stony.
37. Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.
38. Leighcan - Moran families - Cryaquolls complex, extremely stony.
39. Moran family - Cryorthents - Leighcan family complex, extremely stony.
40. Moran family - Cryorthents - Rock land complex, extremely stony.

El dataset se encuentra divido en 2 partes:

2.1. train.cvs

Contiene las características junto con el tipo de cobertura, y cuenta con 15120 observaciones y, junto con las variables cartográficas, 56 columnas. Será usado para entrenar al modelo.

2.2. test.csv

Consta de 550772 observaciones, pero no contiene la columna 'Cover_type', dando un total de 55 columnas.

Finalmente, en los datos no hay presencia de valores **"NaN"** o nulos, por lo que serán simulados.

3. Métricas de desempeño

3.1. Machine Learning

3.1.1. Exactitud multiclase

Se evaluará la exactitud de la clasificación multiclase de la predicción que se realice sobre cada una de las áreas de observación, con la cual se determinará el acierto de las predicciones realizadas con el modelo. Para esto, se utilizará en un inicio, la definición de exactitud multiclase:

$$\mathrm{ACC}_m = \frac{1}{N} \sum_{k=1}^{|G|} \sum_{x:g(x)=k} I(g(x) = \hat{g}(x))$$

3.1.2. Exactitud ponderada multiclase

Sin embargo, si durante la construcción del modelo se identifica que determinado tipo de cobertura tiene una mayor influencia en la métrica, se recurrirá a la exactitud ponderada:

$$\mathrm{ACC}_{mw} = \sum_{k=1}^{|G|} w_k \sum_{x:g(x)=k} I(g(x) = \hat{g}(x))$$

Tal que $\sum_{k=1}^{|G|} w_k = 1$ según la influencia de determinada clasificación en la exactitud ponderada.

3.2. De negocio

Como métrica de negocio puede usarse el ahorro en que se incurre con la correcta clasificación de los terrenos en contraposición con las técnicas de clasificación basadas en exploración física por medio de drones, equipos investigativos y observación directa.

4. Criterio de desempeño deseable en producción

Se espera que un buen desempeño del modelo repercuta en una menor cantidad de recursos destinados a mapear los bosques, reducción de tiempo verificando más fácilmente las predicciones del modelo mediante estudios de muestras representativas y un control más efectivo de las zonas, que suelen ser bastante extensas y, por lo tanto, difíciles de abarcar por el personal a cargo.