Segunda entrega del proyecto

POR:

Santiago Ramírez Pérez

1017240851

Verónica Palacio Muñoz

1017247211

Felipe Sánchez Londoño

1026156025

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raúl Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2023

1. Descripción y avance del proyecto:

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). Usaremos el archivo train que contiene las siguientes columnas (datos cartográficos):

- 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.

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 (Cover_type) son:

1. Picea/Abeto
2. Pino lodgepole
3. Pino ponderosa
4. Álamo/Sauce
5. Álamo temblón
6. Abeto de Douglas
7. Krummholz

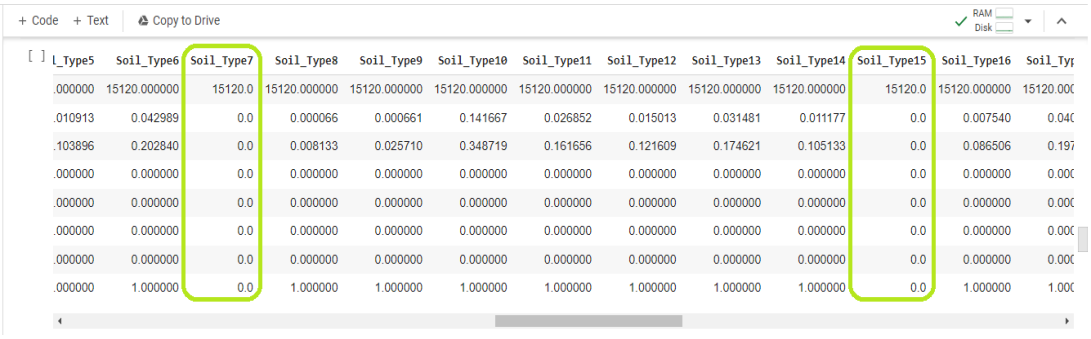
Luego de comprender la tarea que tenemos frente a nosotros, procedemos a revisar, depurar y analizar los datos contenidos en el archivo train.csv.

Comenzamos descargando los archivos en GoogleColab desde Kaggle, y con ayuda de la librería pandas leemos y estudiamos los datos del archivo train. Donde primeramente supimos que en cada columna de train hay guardados 15.120 datos, todos de tipo entero, ningún flotante, sin datos faltantes y sin datos nulos, lo que es una buena señal hasta ahora porque nos da cuenta del buen trabajo que se hizo al recolectar los datos.

Procedemos a revisar más a fondo el data frame y le hacemos una descripción estadística, donde de cada columna de train obtendremos los siguientes datos:

- Conteo de datos almacenados.
- Media aritmética.
- Desviación estándar.
- Valor mínimo.
- Percentil 25%.
- Percentil 50%.
- Percentil 75%.
- Valor máximo.

En este punto nos encontramos con algo curioso, en las columnas tipo de suelo #7 y tipo de suelo #15 nos devuelve valores de cero en todo el análisis estadístico…



l_Type5	Soil_Type6	Soil_Type7	Soil_Type8	Soil_Type9	Soil_Type10	Soil_Type11	Soil_Type12	Soil_Type13	Soil_Type14	Soil_Type15	Soil_Type16	Soil_Typ
.000000	15120.000000	15120.0	15120.000000	15120.000000	15120.000000	15120.000000	15120.000000	15120.000000	15120.000000	15120.0	15120.000000	15120.000
.010913	0.042989	0.0	0.000066	0.000661	0.141667	0.026852	0.015013	0.031481	0.011177	0.0	0.007540	0.040
.103896	0.202840	0.0	0.008133	0.025710	0.348719	0.161656	0.121609	0.174621	0.105133	0.0	0.086506	0.197
.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000
.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000
.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000
.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000
.000000	1.000000	0.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.0	1.000000	1.000

Figura 1. Tabla de descripción estadística de train.csv, señalando columnas con datos igual a cero.

Esto nos deja una dificultad y es si estos valores igual a 0 son porque no se hizo una correcta medición o simplemente pusieron estos valores para no dejar los campos vacíos y tener inconvenientes con datos nulos. Esta es la mayor dificultad que nos hemos encontrado hasta ahora en el proyecto. Además, aunque tengamos incertidumbre en los datos almacenados en train debemos trabajar con lo que tenemos hasta el momento, así que verificamos el sesgo en cada columna para una posterior corrección. Verificamos que en cada tipo de corteza designada en el archivo train hay 2160 instancias para cada una de los 7 tipos de cobertura forestal.

Ahora procedemos a obtener los datos de las distintas áreas silvestres (4) y recordemos que estos datos son del Bosque Nacional Roosevelt del norte

de Colorado (Estados Unidos). Obtuvimos un diagrama de tortas donde nos dice que el área silvestre con mayor cantidad de datos y la que menos datos tiene asignados son: Comanche_Peak_WA y Neota_WA respectivamente.

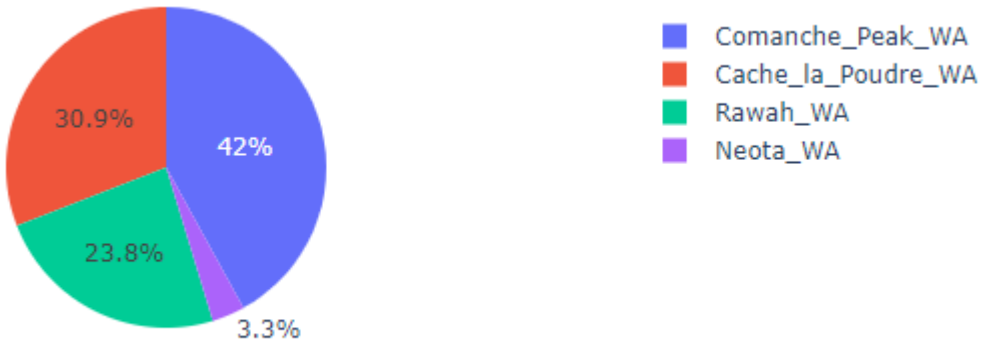


Figura 2. Diagrama de torta de la cantidad de datos en cada área silvestre.

Realizando un histograma corroboramos esta información dándonos cuenta que en cada área silvestre no se encuentran presentes los 7 tipos de cortezas, y verificando que el área Comanche Peak WA tiene más datos, existiendo en esta área 6 tipos de corteza de 7.

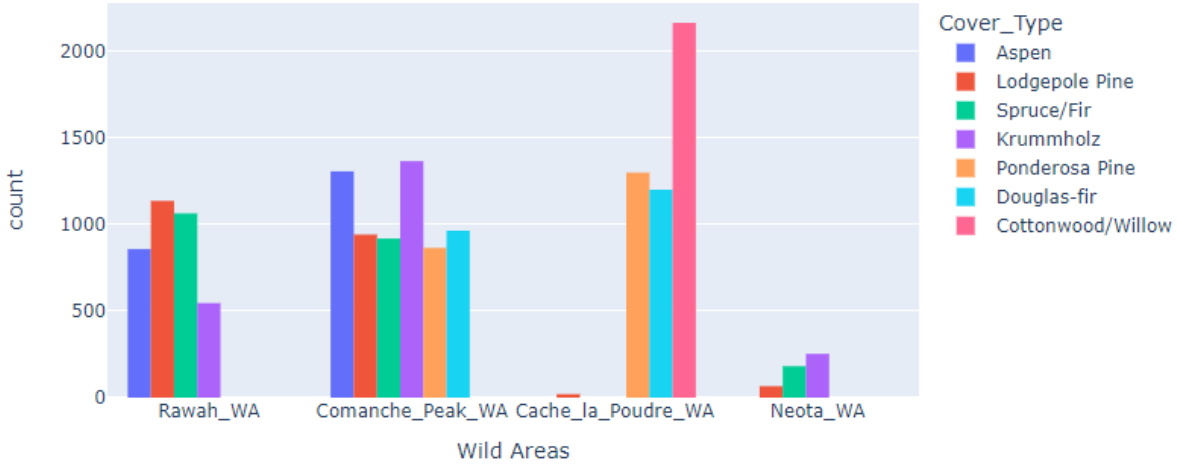


Figura 3. Histograma de cantidad datos de cada tipo de corteza en cada área silvestre.

Finalmente, con ayuda de la librería sklearn procedemos a generar modelos de clasificación. Habiendo 9 tipos de modelos predictivos y en la búsqueda del mejor optamos por conocer cuál método de clasificación se ajusta mejor a nuestro proyecto, en este paso obtuvimos la siguiente información luego de generar un bloque de código con un algoritmo de clasificación que probará cada método y nos mostrará su respectivo accuracy sobre el archivo train.

	AccuracyScore	PrecisionScore	RecallScore	f1_Score
RandomForest	0.866733	0.866733	0.866733	0.866733
KNeigbors	0.804894	0.804894	0.804894	0.804894
GradientBoosting	0.802249	0.802249	0.802249	0.802249
DecisionTree	0.793651	0.793651	0.793651	0.793651
SVC	0.625661	0.625661	0.625661	0.625661
Bernoulli	0.610780	0.610780	0.610780	0.610780
Gaussian	0.600860	0.600860	0.600860	0.600860
LogisticRegr	0.474537	0.474537	0.474537	0.474537
AdaBoost	0.346561	0.346561	0.346561	0.346561

Figura 4. Tabla de precisión de los métodos de clasificación sobre el archivo train.csv.

Con esto notamos que los primeros tres métodos son los que tienen mayor nivel de precisión, ahora el reto es generar predicciones con los primeros métodos (Los más acertados) y revisar, las veces que sean necesarias, con cual método llegamos más cerca a la predicción exacta.

2. 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