SEGUNDA ENTREGA DEL PROYECTO

POR

Natalia Marcela Henao

43.463.025

Andrés Felipe Duque Daza

1.152.706.282

Julián Camilo Duque Ospina

1.036.656.349

Introducción A la inteligencia Artificial

Raul Ramos Pollan

Universidad De Antioquia

Facultad De Ingeniería

Medellín, Octubre de 2023

1. Descripción

En este proyecto se deberá predecir el tipo de cubierta forestal (tipo de cubierta arbórea dominante) esto será tratado a partir de variables cartográficas precisas (a diferencia de los datos de teledetección). Utilizaremos el archivo de entrenamiento (train) el cual contiene las siguientes variables:

- 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 de la cual se recolectaron los datos son cuatro áreas silvestres ubicadas en el Bosque Nacional Roosevelt del norte de Colorado. Cada observación es un área de 30m x 30m. Se le pide que prediga una clasificación entera para el tipo de cubierta forestal de cada una de las áreas que está siendo evaluada. 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

2. Avance del proyecto

Luego de revisar y comprender las actividades a desarrollar que tenemos, procedemos a realizar una revisión exhaustiva, analizamos y depuramos los datos contenidos en el archivo de entrenamiento (train.csv)

Empezamos nuestro análisis descargando los archivos necesarios desde Kaggle directamente en Google Colab. Utilizando la biblioteca panda, procedimos a cargar y examinar

detenidamente los datos del archivo de entrenamiento (train.csv). Nuestra primera observación reveló que cada columna del conjunto de entrenamiento contiene exactamente 15,120 datos, todos ellos de tipo entero. Esta consistencia en el tipo de datos nos da confianza en la calidad de la recopilación de datos, ya que no encontramos ningún valor de tipo flotante, ni datos faltantes o nulos en ninguna de las columnas.

Esta coherencia y plenitud en los datos son señales muy positivas en esta etapa inicial del análisis. La ausencia de valores faltantes o nulos es especialmente alentadora, ya que indica que no hay lagunas en nuestros datos que puedan complicar nuestro análisis posterior.

Continuamos haciendo una revisión a fondo del data frame y realizamos una descripción estadística que nos permita obtener de cada una de las columnas de entrenamiento (train.csv) 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 unos datos inquietantes en las columnas de tipo de suelo 7 y 15 (Soil_Type7, Soil_Type15), en las cuales todos los datos en el análisis estadístico son 0.

Soil_Type5	Soil_Type6	Soil_Type7	Soil_Type8	Soil_Type9	Soil_Type10	Soil_Type11	Soil_Type12	Soil_Type13	Soil_Type14	Soil_Type15	Soil_Type16
15120.000000	15120.000000	15120.0	1 120.000000	15120.000000	15120.000000	15120.000000	15120.000000	15120.000000	15120.00000	15120.0	15120.000000
0.010913	0.042989	0.0	0.000066	0.000661	0.141667	0.026852	0.015013	0.031481	0.01117	0.0	0.007540
0.103896	0.202840	0.0	0.008133	0.025710	0.348719	0.161656	0.121609	0.174621	0.10513	0.0	0.086506
0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.0	0.000000
0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.0	0.000000
0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.0	0.000000
0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.0	0.000000
1.000000	1.000000	0.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00000	0.0	1.000000

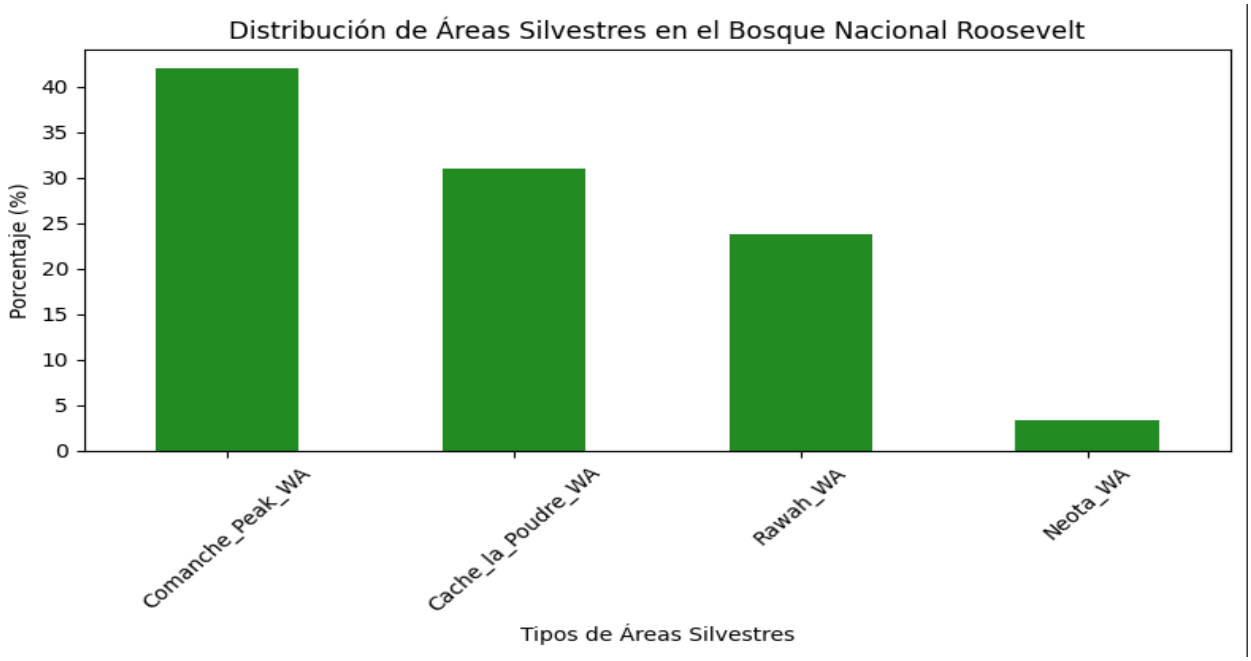
Nos enfrentamos a un desafío intrigante: la incertidumbre respecto a los valores iguales a 0 en nuestros datos. Es crucial determinar si estos ceros indican una medición incorrecta o si se han introducido intencionadamente para evitar campos vacíos y problemas asociados con los datos nulos. Esta ambigüedad representa una de las dificultades clave en nuestro proyecto en este punto. A pesar de esta incertidumbre, no tenemos más opción que trabajar con los datos tal como están hasta el momento.

Para abordar esta situación, verificamos el mensaje presente en cada columna para evaluar la necesidad de correcciones posteriores. Este análisis nos permitirá entender mejor la naturaleza de los valores iguales a 0 y determinar si tienen un impacto significativo en nuestro conjunto de datos.

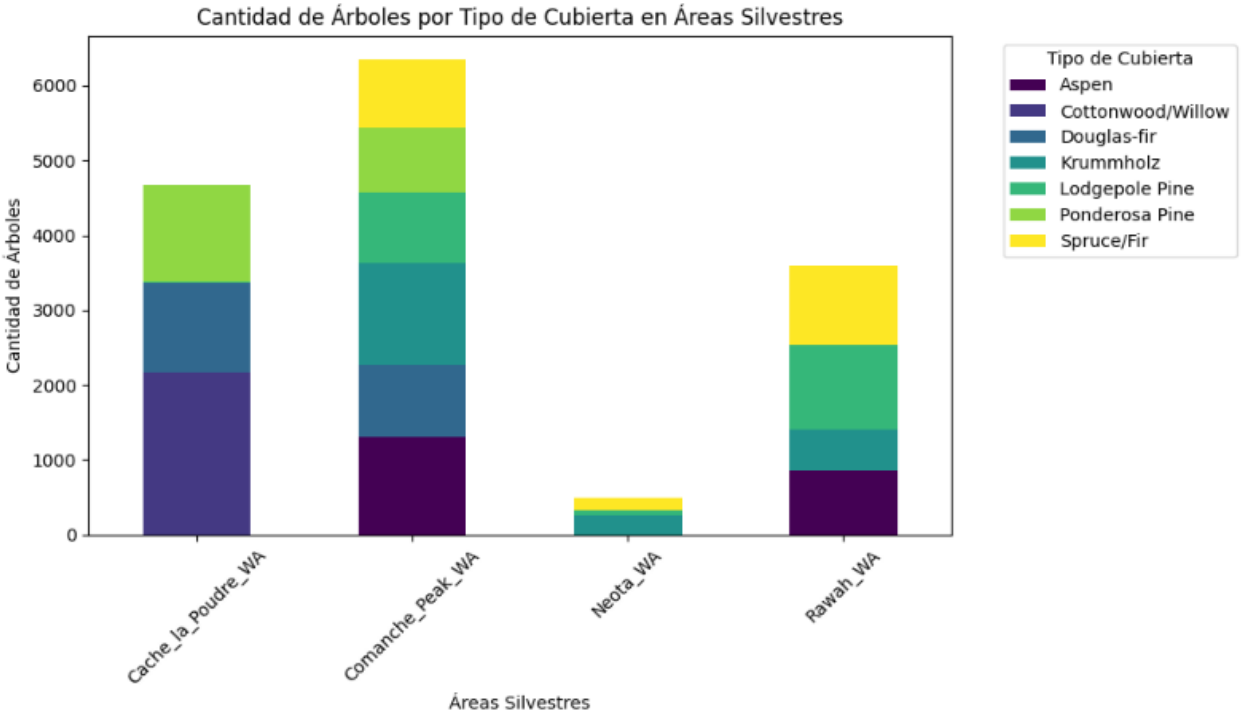
Además, aunque nos encontramos con esta incertidumbre en los datos almacenados en el conjunto de entrenamiento, seguimos avanzando con nuestra exploración. Realizamos un análisis detallado y descubrimos un patrón interesante: cada tipo de corteza forestal designado en el archivo de entrenamiento cuenta con exactamente 2,160 instancias. Esta uniformidad en la distribución de las clases es valiosa, ya que nos proporciona un equilibrio inicial en nuestro conjunto de datos, lo que puede ser crucial para el entrenamiento efectivo de nuestro modelo en etapas posteriores del proyecto.

A pesar de las incertidumbres iniciales, esta consistencia en la distribución de las clases es un hallazgo valioso. Nos brinda una base sólida sobre la cual construir nuestras investigaciones y análisis futuros.

Continuamos con la obtención de los datos de las distintas áreas silvestres (4). Obtenemos un gráfico de barras que nos muestra desde el área silvestre con mayor cantidad de datos hasta la con menor cantidad de datos:



Realizamos un histograma compuesto en el cual evidenciamos que los 7 tipos de corteza que se encuentran en el parque están distribuidos por las cuatro zonas y que la zona con mayor variedad de cortezas está en Comanche Peak WA teniendo 6 de las sietes variedades y Neota Wa solo 3.



Finalmente, utilizando la biblioteca Scikit-Learn, hemos avanzado a la fase de generación de modelos de clasificación. Dado que existen nueve tipos de modelos predictivos, nuestro objetivo es determinar cuál de estos métodos se adapta mejor a las necesidades de nuestro proyecto. Para lograr esto, hemos implementado un bloque de código que emplea un algoritmo de clasificación para probar cada método. Esta estrategia nos proporciona una evaluación precisa de la precisión de cada modelo en relación con los datos del archivo de entrenamiento.

Este proceso nos permite comparar y contrastar la eficacia de diversas técnicas de clasificación, lo que es crucial para seleccionar el modelo más adecuado para nuestro escenario específico. Esta fase de evaluación es fundamental, ya que nos brinda perspectivas valiosas para la selección del modelo final que emplearemos en nuestro proyecto de predicción de cobertura forestal. Estamos entusiasmados por los resultados que obtendremos y por la posibilidad de aplicar un enfoque sólido y bien fundamentado en nuestro análisis predictivo.

	AccuracyScore	PrecisionScore	RecallScore	f1_Score
RandomForest	0.869378	0.869378	0.869378	0.869378
KNeigbors	0.804894	0.804894	0.804894	0.804894
GradientBoosting	0.802249	0.802249	0.802249	0.802249
DecisionTree	0.787368	0.787368	0.787368	0.787368
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

3. Bibliografía

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