**Entrega Final Proyecto IA “Forest Cover Type Prediction”**

Bryan Zuleta Vélez

Sol Yajhaira Linares Mateus

Daniel Alejandro Yepes Mesa

Facultad de Ingeniería

Universidad de Antioquia

Introducción a la inteligencia artificial para ciencias e ingeniería

Profesor: Raul Ramos Pollan

12 de Noviembre de 2022



**Forest Cover Type Prediction**

INTRODUCCIÓN

La cobertura forestal es crucial para la salud del suelo, el ciclo del agua, el clima y la calidad del aire, de ahí la importancia de analizar cuál es la cobertura forestal adecuada para cada bosque; en el caso de este proyecto, analizaremos esto para los bosques que se encuentran en la región 2 del servicio forestal de estados unidos, este análisis se realiza teniendo en cuenta principalmente la función de sus suelos, y otros factores.

Para este proyecto se va a generar una predicción de la cobertura forestal (basándonos en la categoría donde predomina la cubierta arbórea) de un bosque a partir de variables estrictamente cartográficas. Los datos reales se obtienen a partir de los datos del Sistema de Información de Recursos de la Región 2 del Servicio Forestal de EE.UU (USFS).

Los resultados posibles para esta clasificación son:

1. Spruce/Fir
2. Lodgepole Pine
3. Ponderosa Pine
4. Cottonwood/Willow
5. Aspen
6. Douglas-fir
7. Krummholz

EXPLORACIÓN DESCRIPTIVA DEL DATASET

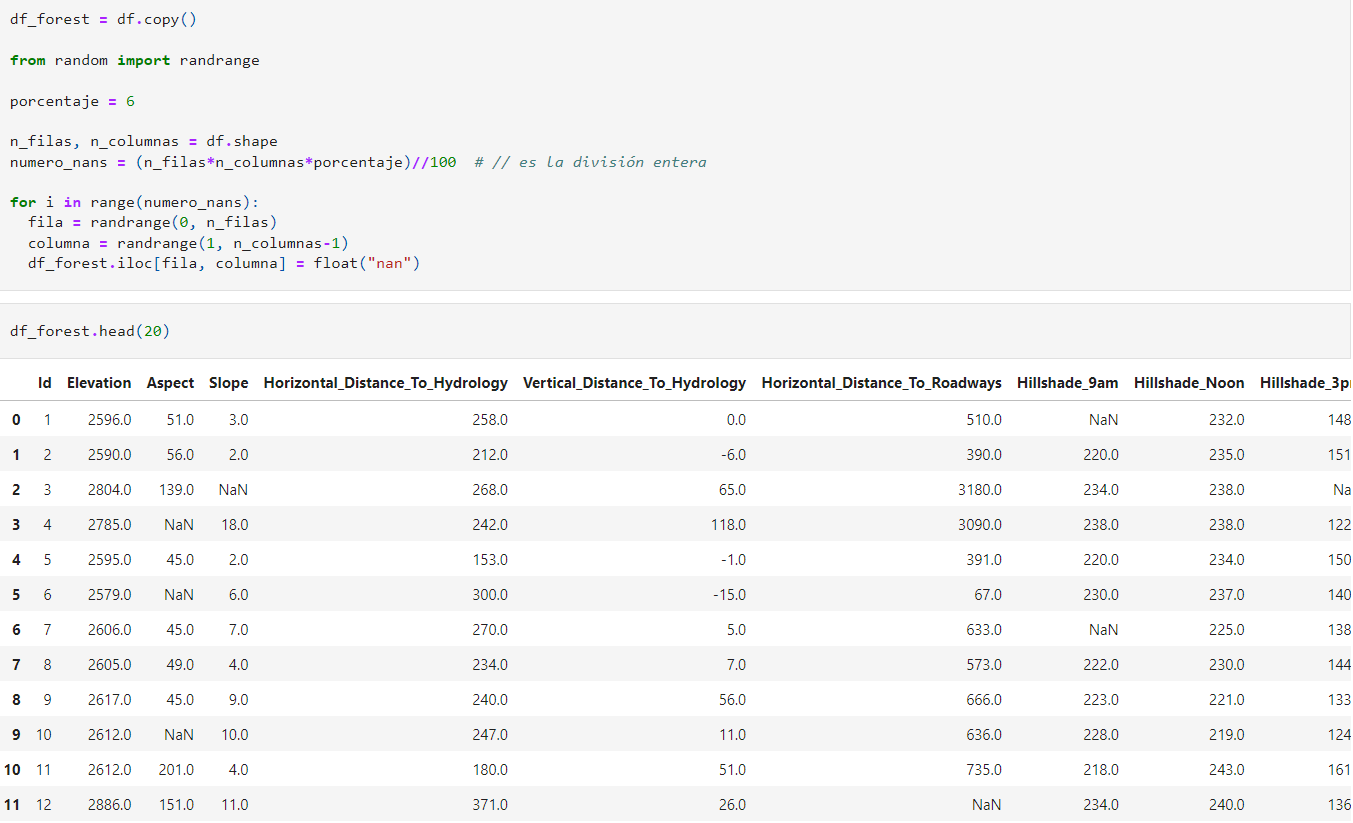
Para iniciar con el procesamiento de los datos vamos a usar tres archivos *.csv*; el primer archivo será ***Train.csv***, donde obtendremos información como:

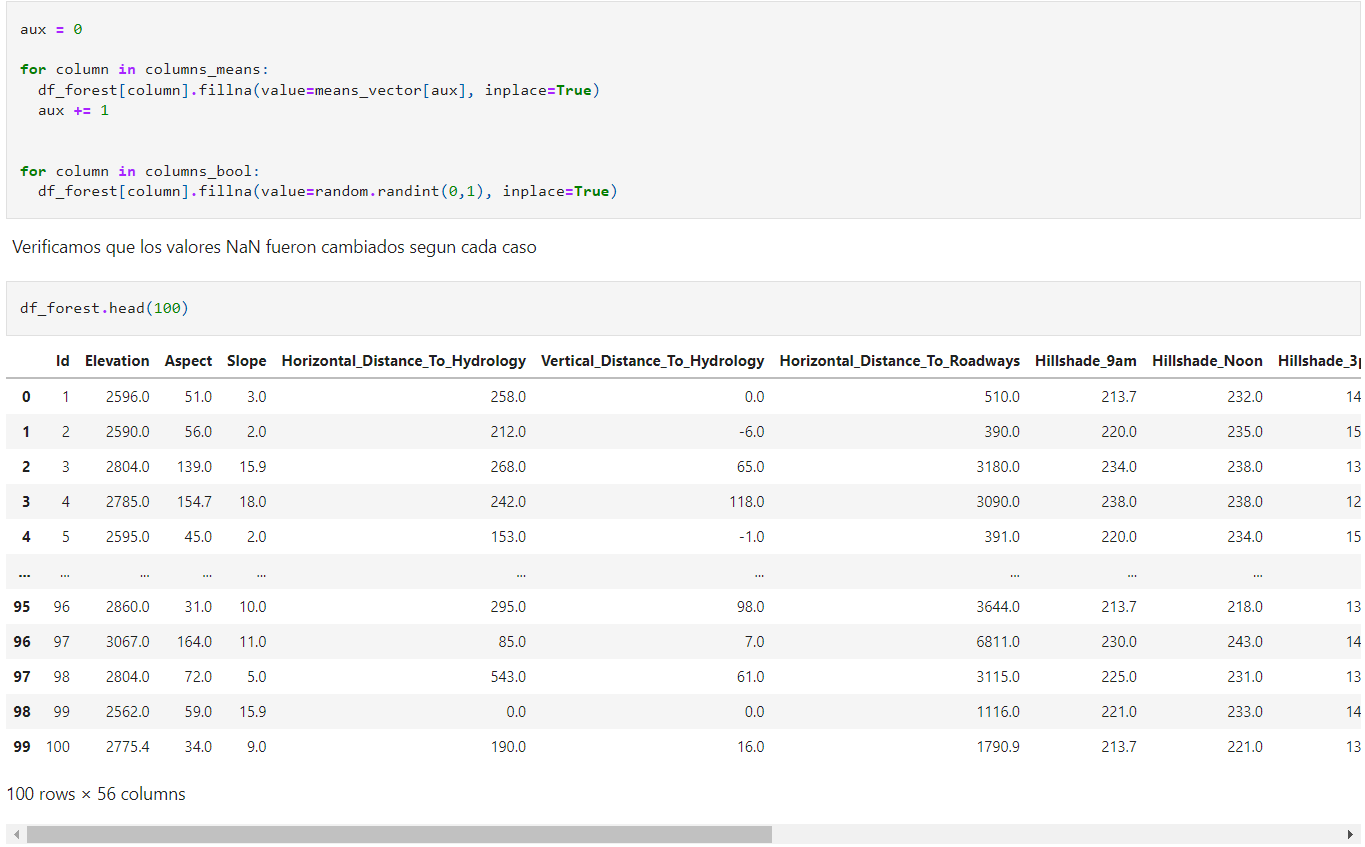
1. Elevation - Elevación en metros
2. Aspect - Aspecto en grados acimut
3. Slope – Pendiente en grados
4. Horizontal\_Distance\_To\_Hydrology - Horz distancia a las características de agua superficial más cercanas
5. Vertical\_Distance\_To\_Hydrology - Vert distancia a las características de agua superficial más cercanas
6. Horizontal\_Distance\_To\_Roadways - Horz Dist a la carretera más cercana
7. Hillshade\_9am (0 to 255 index) - Índice de sombreado a las 9 a.m., solsticio de verano
8. Hillshade\_Noon (0 to 255 index) - Índice de sombreado al mediodía, solsticio de verano
9. Hillshade\_3pm (0 to 255 index) - Índice de sombreado a las 15:00, solsticio de verano Horizontal\_Distance\_To\_Fire\_Points - Horz Dist a los puntos de ignición de incendios forestales más cercanos
10. Wilderness\_Area (4 binary columns, 0 = absence or 1 = presence) - Designación de área silvestre
11. Soil\_Type (40 binary columns, 0 = absence or 1 = presence) - Designación del tipo de suelo
12. Cover\_Type (7 types, integers 1 to 7) - Designación de tipo de cubierta forestal

El segundo archivo que es ***Test.csv****,* a parte de la información contenida en train.csv, nos dará los valores de la columna *Cover\_Type* y por último el tercer archivo es **sampleSubmission** que contiene las observaciones aplicadas al archivo *Test.csv*, con el cual se obtiene el resultado de acierto del modelo desarrollado.

Para analizar el comportamiento de los datos y realizar los respectivos modelos se procedió a ubicar en una carpeta de drive los archivos test y train, para luego crear un acceso directo al colab respectivo para iniciar con el procesamiento de este.

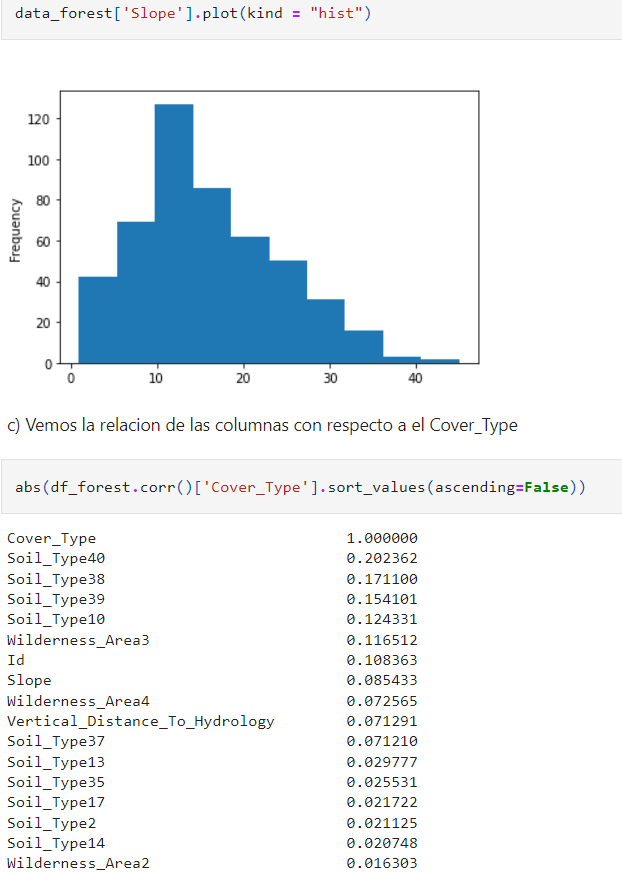
Luego de dicho procesado se analizó el comportamiento de cada variable de manera independiente, verificando la existencia de datos atípicos o fuera de lo establecido, en este proceso notamos que nuestro Dataset no contenía datos faltantes, así que usamos un método para generar un numero fijo de datos nulos.

Debido al tamaño de nuestro dataset, también a que solo tenemos una columna fija (ID) y a la combinación entre números enteros y booleanos en las columnas que teníamos que rellenar para el entrenamiento, se crearon dos ciclos *for*, uno es para aquellos datos que son números enteros que se llenó con el promedio de los valores y el otro para las columnas con valores booleanos que se llenó con valores de 1’s y 0’s.



A pesar del tamaño de la base de datos, no tuvimos problemas a la hora de procesar los datos, así que realizamos un entrenamiento para ver su funcionamiento y revisar si estaba adaptada a lo que se nos solicitaba.

Para analizar los datos se generó un histograma que mostraba en grupos de 5 la frecuencia de los datos, además usamos una función para relacionar los datos de la columna Cover\_Type, y el resultado fue:



en donde se puede notar que el tipo de suelo 38, 39, 40, 10 y 35 son los que más están relacionados con la cobertura que se está buscando, aunque tiene relación con todos, los mencionados son los que más tienen fuerza con referencia al Cover\_Type.

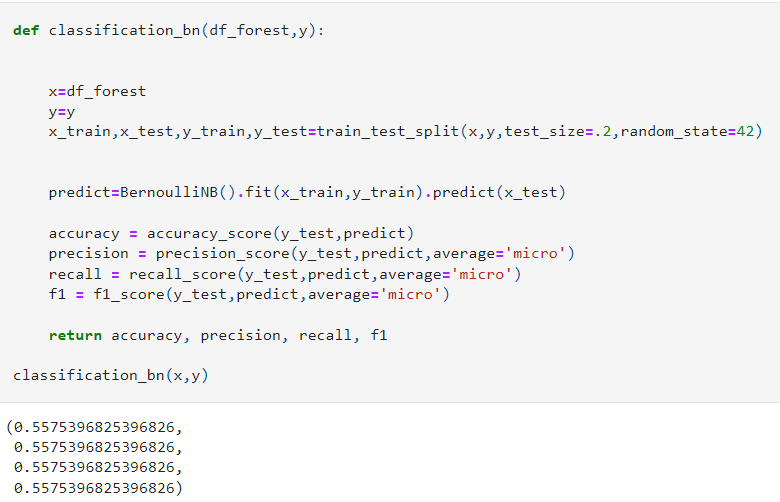
ITERACIONES DE DESARROLLO

Para realizar predicciones relacionadas con el comportamiento futuro de los datos, es decir conocer las diferentes coberturas de suelo y tipos de suelo, se realizaron múltiples modelos como:

Metodo GaussianoNB



Método Bernoulli



Entre otros, de los cuales podemos destacar los siguientes:

DecisionTree



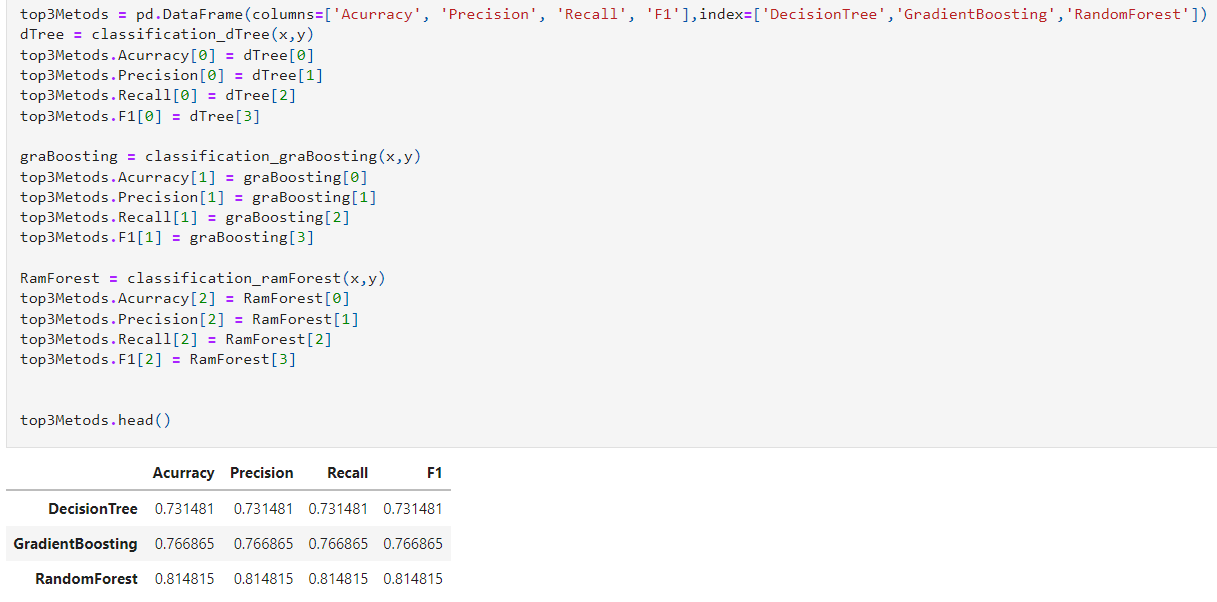
GradientBoosting 

y RandomForest



que fueron los que nos dieron mejores resultados.

Con el fin dar un mejor análisis creamos un *dataframe* con los datos que se obtienen de estos tres modelos y concluimos que el más preciso en el análisis era el modelo RandomForest.

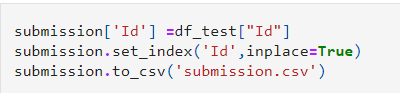


La métrica empleada será multi-class classification accuracy, definimos la exactitud (accuracy en inglés) como el ratio entre las predicciones correctas (suma de verdaderos positivos y verdaderos negativos) y las predicciones totales. Scikit-Learn implementa la métrica sklearn.metrics.accuracy\_score que puede utilizarse en clasificación binomial y multiclase y que devuelve el porcentaje de predicciones correctas. Lógicamente, el clasificador ideal tendría una exactitud de 1 (todas las muestras serían bien clasificadas) y el peor clasificador posible tendría una exactitud de 0 (ninguna muestra sería bien clasificada).

Teniendo esto claro, procedemos a analizar el modelo RandomForest, y realizar el entrenamiento, con los datos obtenidos del procesamiento del dataset.



Y por último, generamos un archivo .csv con los resultados.



CONCLUSIONES

Al comparar los 3 modelos del dataframe podemos ver que la exactitud está por encima del 0.6, y el más cercano a 1 es el modelo RandomForest, así que será el que nos dará resultados más confiables para nuestro análisis.

Con este resultado podemos notar que el Cover\_Type que se sugiere en la mayoría de bosques ubicados en la región 2 del USFS de EE.UU es tipo 2, que son plantas herbáceas, es decir una cobertura en su gran mayoría de pasto con árboles.

Para finalizar si vemos los resultados que obtuvimos en nuestro entrenamiento y en el análisis del modelo final, podemos notar que en el de entrenamiento los el cover\_type indicado era el 5, pero ya con los datos del modelos de Radom Forest, notamos que realmente, el tipo de cobertura indicada es la 2.