Primera entrega proyecto IA “Forest Cover Type Prediction”

Daniel Alejandro Yepes Mesa

Sol Yajhaira Linares Mateus

Juan Camilo Castañeda Ospina

Facultad de Ingeniería, Universidad de Antioquia

Introducción a la inteligencia artificial para las ccias e ingenierías

Profesor: Raúl Ramos Pollan

5 de Julio de 2022

1-Descripción del problema predictivo:

El problema selección consiste en generar una predicción de la categoría a la cual que pertenece la cobertura forestal (el tipo predominante de 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).

Para esta clasificación puede haber 7 resultados posibles los cuales son:

1 - Spruce/Fir

2 - Lodgepole Pine

3 - Ponderosa Pine

4 - Cottonwood/Willow

5 - Aspen

6 - Douglas-fir

7 - Krummholz

2-DataSet:

Para esto vamos a usar el dataset de kaggle esta competición https://www.kaggle.com/competitions/forest-cover-type-prediction/data , el cual contiene 3 archivos de tipo csv que se consideran útiles.

El archivo **Train.csv** el cual contiene 15.120 observaciones y contiene la siguiente información:

Elevation - Elevación en metros

Aspect - Aspecto en grados acimut

Slope – Pendiente en grados

Horizontal_Distance_To_Hydrology - Horz distancia a las características de agua superficial más cercanas

Vertical_Distance_To_Hydrology - Vert distancia a las características de agua superficial más cercanas

Horizontal_Distance_To_Roadways - Horz Dist a la carretera más cercana

Hillshade_9am (0 to 255 index) - Índice de sombreado a las 9 a.m., solsticio de verano

Hillshade_Noon (0 to 255 index) - Índice de sombreado al mediodía, solsticio de verano

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

Wilderness_Area (4 binary columns, 0 = absence or 1 = presence) - Designación de área silvestre

Soil_Type (40 binary columns, 0 = absence or 1 = presence) - Designación del tipo de suelo

Cover_Type (7 types, integers 1 to 7) - Designación de tipo de cubierta forestal

El archivo **Test.csv** contine la misma información del archivo train.csv a excepción de la columna Cover_Type la cual corresponde a la predicción a desarrollar.

El archivo **sampleSubmission** las observaciones correctas aplicadas al archivo test.csv con el cual se puede obtener el resultado del acierto del modelo desarrollado.

3-Metricas

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).

Formalmente, la exactitud está definida por la siguiente función:

$$accuracy\_score(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i)$$

Donde y representa los valores reales, ŷ representa las predicciones y n es el número de muestras.

4-primer criterio de desempeño

Se decide que el desempeño del modelo para que sea optimo tenga un acierto del 60% ya que permite ver una tendencia de los tipos de la cobertura forestal en una zona específica.

Referencias

-Forest Cover Type Prediction | Kaggle. (2014). Retrieved 04 July 2022, From https://www.kaggle.com/competitions/forest-cover-type-prediction/overview/description

-Exactitud | Interactivechaos. (2021). Retrieved 04 July 2022, From https://interactivechaos.com/es/manual/tutorial-de-machine-learning/exactitud