Segunda 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

2022

Introducción

En este texto se abordan las etapas de preprocesamiento de datos del problema de clasificación múltiple tratado, en el cual se tiene un dataset con diferentes tipos de características cartográficas a partir de las cuales se busca predecir la cubierta arbórea en un área. A continuación se describen las adecuaciones realizadas en dicho preprocesamiento (que consistió en la clasificación de los tipos de datos, la inducción del 5 % de datos nulos y la determinación de las variables más importantes del dataset).

1. Preprocesamiento de datos

1.1. Inducción de datos nulos y categóricos

1.1.1. Datos nulos

Ya que no existen datos nulos en el dataset, es necesario inducirlos. Al menos el 5 % de los datos han de ser nulos. Por lo tanto, la aproximación es la siguiente:

- Una columna o variable será seleccionada de acuerdo con un número entero aleatorio (e.g: si el número es 1, entonces la columna seleccionada será la primera; la columna 'id' no se tendrá en cuenta, pues es solo un identificador, solo se aplicará la aproximación para las columnas correspondientes a variables cartográficas).
- Con la columna seleccionada, el dato que será reemplazado por NaN será escogido con un nuevo número entero aleatorio, indicando su índice en el dataset.
- Los pasos anteriores serán repetidos hasta alcanzar la cantidad mínima de nulos del 5 %.

Elevation	121
Aspect	134
Slope	114
Horizontal_Distance_To_Hydrology	139
Vertical_Distance_To_Hydrology	147
Horizontal_Distance_To_Roadways	145
Hillshade_9am	150
Hillshade_Noon	160
Hillshade_3pm	137
Horizontal_Distance_To_Fire_Points	139
Wilderness_Area1	127
Wilderness_Area2	133
Wilderness_Area3	135
Wilderness_Area4	135
Soil_Types	5644
$\sum$	7560

Cuadro 1: Cantidad de datos nulos por columnas

1.1.2. Datos categóricos

Dado que se tiene el requerimiento de contar con al menos un 10 % de columnas categóricas, se realiza la verificación de esta condición, en la cual se constata que las columnas que indican el tipo de suelo presentan la siguiente convención:

- Se evidencia un 1 cuando se indica la presencia de este tipo de variable o característica topográfica en el individuo.
- Se evidencia un 0 cuando esta variable o característica topográfica no se encuentra en el individuo.

Esto quiere decir que las columnas con este tipo de formato constituyen columnas categóricas, ya que clasifican los datos por medio de 2 categorías fijas en función de si esta característica tipográfica se encuentra en el individuo o no. En particular, **las variables categóricas son las que llevan la etiqueta** 'Wilderness_Area','Soil_Type'.

1.1.3. Imputación de datos nulos

Según McDermeit et al.[1], cuando se encuentra que existe menos de un 5 % de datos faltantes, entonces estos se pueden sustituir por:

- **La Media:** Cuando hay distribución normal en los datos.
- **La Mediana:** Cuando los datos son sesgados.
- **La Moda:** Cuando la variable es categórica.

Es teniendo en cuenta estos criterios que determinamos escoger como método para la imputación de datos el reemplazo por media, mediana y moda (según los datos sean categóricos o no). El chequeo visual de la distribución de las columnas numéricas generó el resultado mostrado en (1). Ahora, se determinó que para $R^2 > 0{,}9$, es posible asumir que la columna sigue una distribución **normal** y, por lo tanto, aquellas que cumplan serán rellenadas con la media. Para las restantes, el proceso se realizará con la **mediana**.

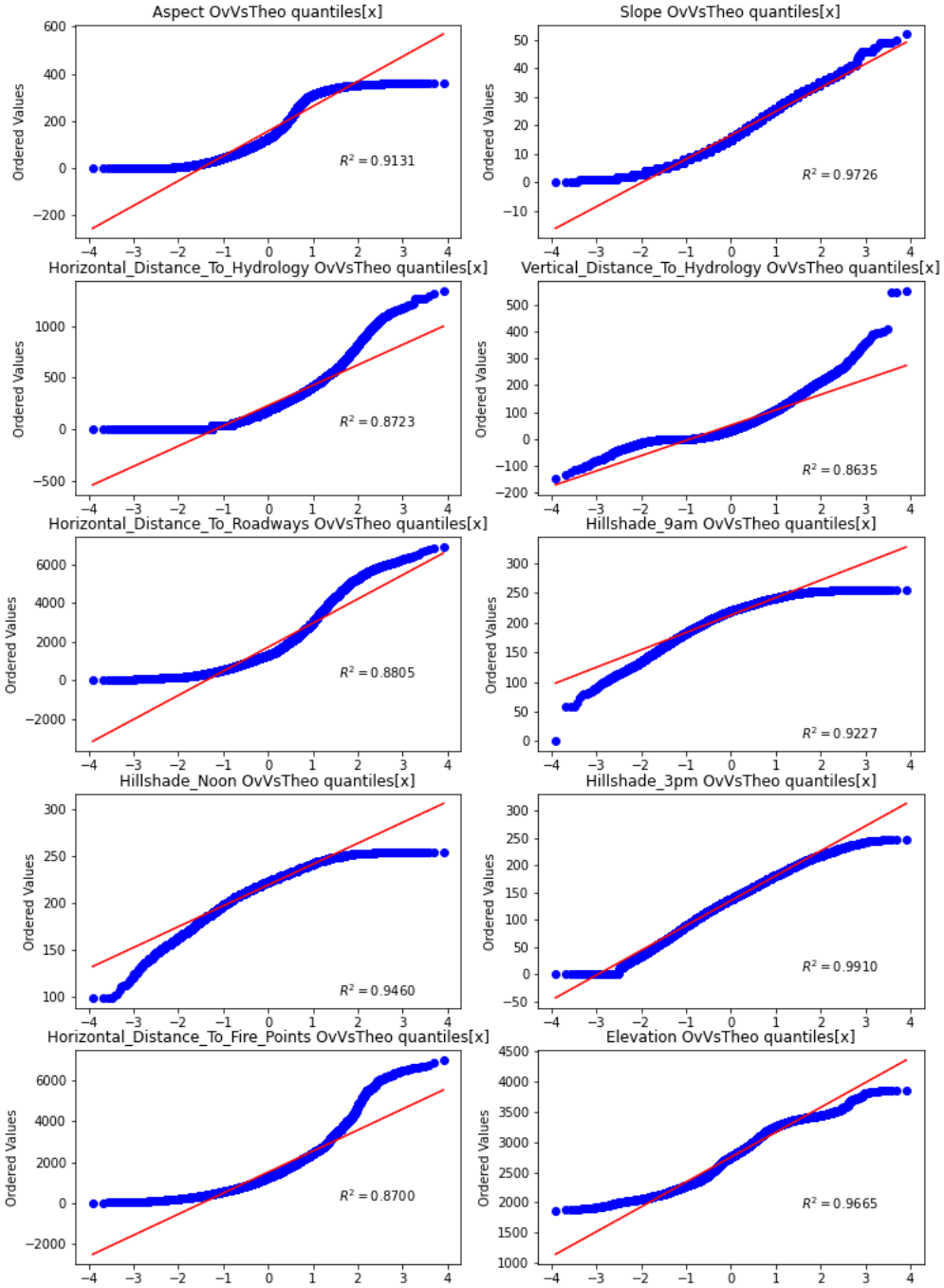

Figura 1: Gráficas Q-Q para test de normalidad con variables numéricas

1.2. Descripción estadística de variables

Variable	count	mean	std	min	25 %	50 %	75 %	max
Elevation	14995	2748.747	417.623	1863	2375	2751	3103	3849
Aspect	14975	156.662	110.109	0	65	125	261	360
Slope	14985	16.501	8.455	0	10	15	22	52
Horizontal_Distance_To_Hydrology	14974	227.217	210.144	0	67	180	324	1343
Vertical_Distance_To_Hydrology	14981	51.028	61.207	-146	5	32	79	554
Horizontal_Distance_To_Roadways	14980	1712.468	1324.166	0	760	1316	2266	6890
Hillshade_9am	14988	212.705	30.565	0	196	220	235	254
Hillshade_Noon	14980	218.959	22.817	99	207	223	235	254
Hillshade_3pm	14999	135.020	45.875	0	106	138	167	248
Horizontal_Distance_To_Fire_Points	14988	1510.200	1098.852	0	730	1256	1986	6993

Cuadro 2: Descripción estadística de las variables numéricas

Estas variables numéricas son de tipo *entero*. Se obviaron las variables categóricas, ya que al estar distribuidas entre solo 2 valores, no es necesario hacer una descripción tan detallada como en (2).

2. Inspección de variable objetivo

Siendo 'Cover_Type' la variable que se busca tener en cuenta tanto para entrenar al modelo como para determinar su exactitud, por medio de la siguiente figura se puede notar su naturaleza

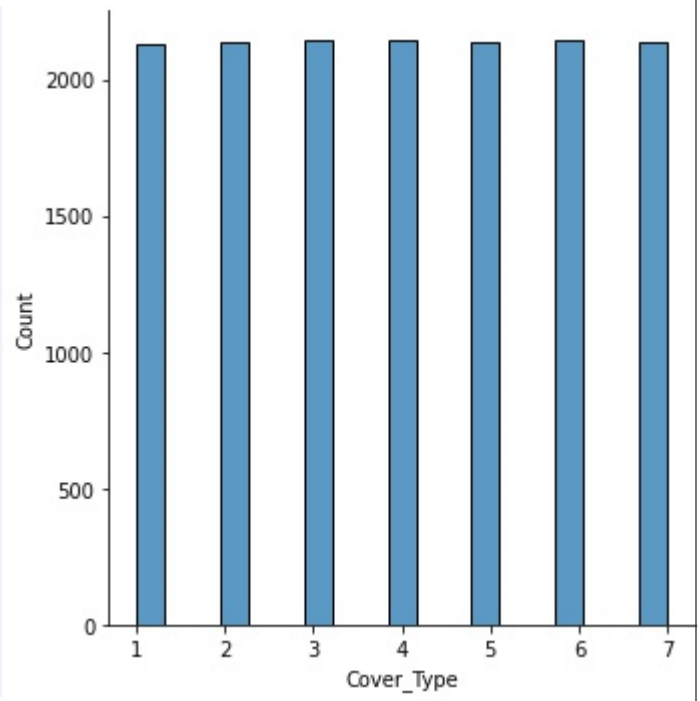

Figura 2: Distribución de los valores de variable 'Cover_Type'

Se aprecia que el tipo de predicción esperada debe ser **categórica**.

2.1. Primera selección de características: coeficiente de correlación

2.1.1. Preprocesamiento

Para el preprocesamiento de los datos, con el fin de determinar una primera relación de las variables cartográficas con 'Cover_Type', se utilizó el método de **Coeficientes de correlación**. Mediante este método se obtuvieron los siguientes resultados

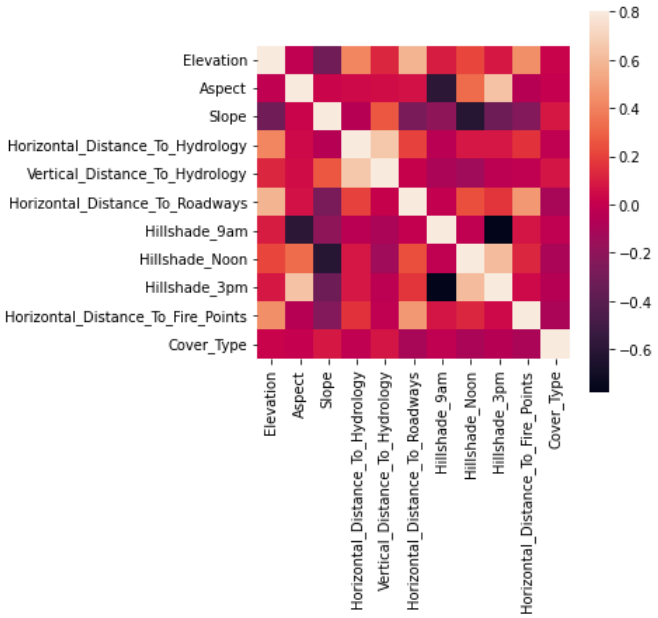

Figura 3: Correlaciones entre variables numéricas y 'Cover_Type'

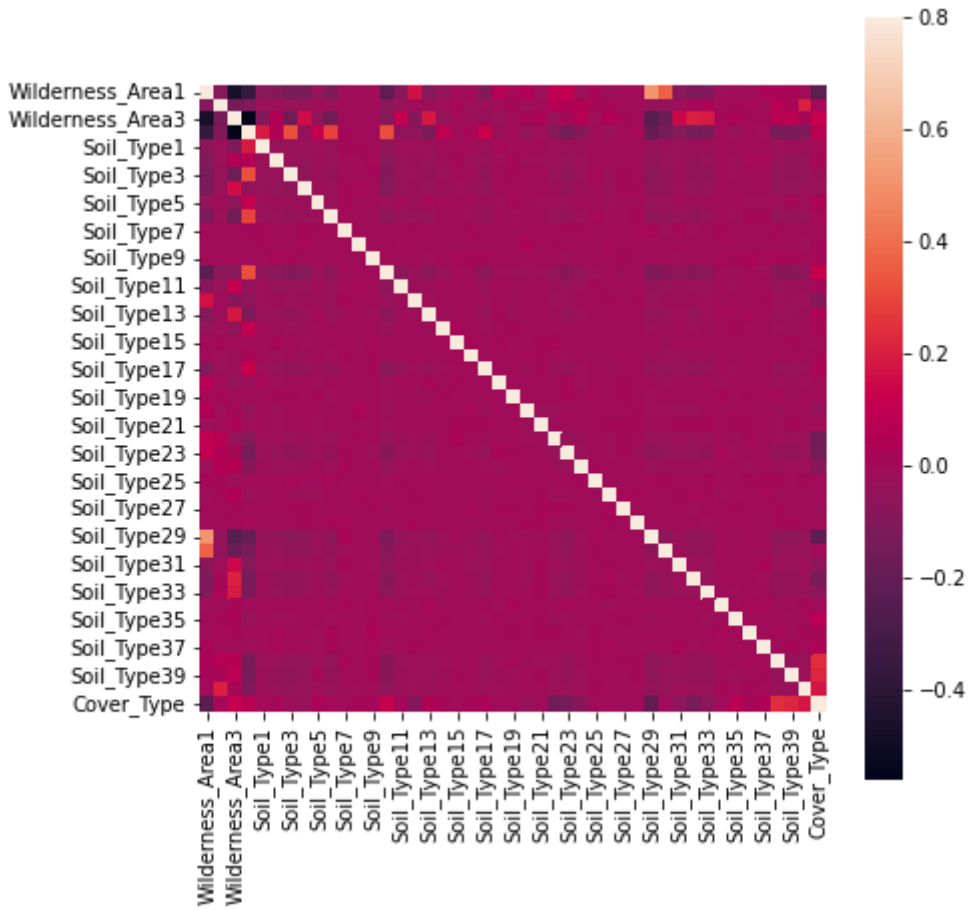

Figura 4: Correlaciones entre variables categóricas y 'Cover_Type'

Las variables cuya correlación con 'Cover_Type' es mayor al 10 % son

	Cover_Type
Wilderness_Area3	0.119821
Soil_Type10	0.123609
Soil_Type38	0.240434
Soil_Type39	0.219018
Soil_Type40	0.176921
Cover_Type	1.000000

Cuadro 3: Variables con correlación mayor a 10 %

Por lo tanto, se concluyó que es **necesario utilizar otros métodos a la hora de encontrar las relaciones entre las variables**, pues valores tan *bajos de correlación nos indican que la relaciones no son lineales.* Además, ya que existen coeficientes de correlación negativos, se puede asumir que es necesario descartar algunas variables por medio de los demás métodos de selección de carácteristicas de datos o *Feature selection*

Referencias

[1] M. McDermeit, R. Funk, and M. Dennis, "Data cleaning and replacement of missing values," *Chestnut Health Systems, Research and Training Lighthouse Institute. Retrieved March*, vol. 5, p. 2005, 1999.

[2] T. A. Factor, "Seven ways to make up data: Common methods to imputing missing data," 2017. [Internet; consultado 12-agosto-2022].