Trabajo Final Análisis Inteligente de Datos

Yudy Carolina Daza Caro

17 de agosto de 2021

1.0. Descripción de la base de datos

El nombre de la base de datos a analizar es Forest Covertype data[1].

El dataset consiste en medidas cartográficas de áreas silvestres en el Bosque Nacional Roosevelt en el norte de Colorado en USA y el tipo de covertura vegetal presente. En la mayoría de los casos las bases de este tipo usadas para clasificación consisten en imágenes satelitales y son empleadas para segmentar zonas con distintos usos (bosques, agricultura, ciudades etc). En el caso de la presente base de datos, el objetivo es clasificar áreas sin o con muy poca intervención humana.

El dataset contiene un conjunto de mediciones realizadas sobre celdas de 30 X 30 metros que consisten en 12 medidas cartográficas, que incluyen 4 áreas silvestres, 40 tipos de suelo y 7 grandes tipos de covertura vegetal. La base de datos tiene 581012 observaciones sin *missings* en ninguna variable. En la tabla1.1 se detalla la descripción del dataset.

Nombre	Tipo de dato	Medida	Descripción
Elevation	cuantitativo	m	Elevación en metros
Aspect	cuantitativo	azimut	Orientación
Slope	cuantitativo	grados	Pendiente
Horizontal_Distance To_Hydrology	cuantitativo	m	Distancia hor. a la superficie de agua mas cercana
Vertical_Distance To_Hydrology	cuantitativo	m	Distancia ver. a la superficie de agua mas cercana
Horizontal_Distance To_Roadways	cuantitativo	m	Distancia hor. al camino mas cercano
Hillshade_9am	cuantitativo	0-255	Indice de sombreado 9am, verano
Hillshade_Noon	cuantitativo	0-255	Indice de sombreado 12pm, verano
Hillshade_3pm	cuantitativo	0-255	Indice de sombreado 3pm, verano
Horizontal_Distance To_Fire_Points	cuantitativo	m	Distancia hor. a incendio mas cercano
Wilderness_Área (4 columnas binarias)	categórica	0,1	Designación de area silvestre
Soil_Type (40 columnas binarias)	categórica	0,1	Designación de tipo de suelo
Cover_Type (7 tipos)	categórica	1 a 7	Designación de covertura forestal

Cuadro 1.1: Descripción de los datos.

Los 7 tipos de covertura forestal son:

- Spruce/Fir (picea -pino-/abeto)
- Lodgepole Pine (pino)
- Ponderosa Pine (pino)
- Cottonwood/Willow (sauces)
- Aspen (alamos)
- Douglas-fir (pino oregón)
- Krummholz (vegetación atrofiada y deformada, árbol "bandera")

[1]Remote Sensing and GIS Program Department of Forest Sciences College of Natural Resources Colorado State University

Variables Categóricas

Una de las variables consiste en medidas del tipo de suelo. En el dataset esta variable se mapeo a 40 one-hot variables. Cada variable corresponde a un tipo de suelo, catalogado con un código que incluye información de la zona climática y zona geológica. El suelo mas abundante tiene el doble de frecuencia que el siguiente mas abundante y su frecuencia esta 4 órdenes de magnitud por encima del suelo menos frecuente (Ver Fig.1.1). En vista de que son muchas variables individuales a ser consideradas y no pueden ser combinadas (no sin conocimiento experto en geología) para reducir su numero no serán consideradas en el análisis. Otra de las variables categóricas es la zona silvestre del parche, de la cual se ilustra su abundancia en la Fig.1.2. Las áreas silvestres 3 y 4 son alrededor de 8 veces más abundantes en el dataset que las 1 y 2. También se podria usar esta variable para clasificar pero se escoge la clase de cobertura forestal que es más explícita. No se usará la variable de áreas silvestres en el análisis.

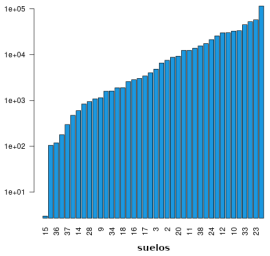

Figura 1.1: Distribución tipos de suelo.

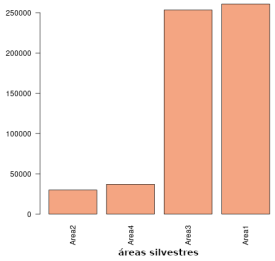

Figura 1.2: Distribución tipos de área silvestre.

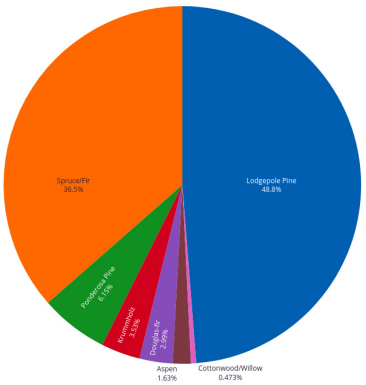

Figura 1.3: Abundancia de cada tipo de bosque.

En la Fig.1.3 se detalla la abundancia de cada clase de cobertura forestal. Se puede observar que los datos estan bastante desbalanceados en cuanto a clases de cobertura, el bosque de Spruce/Fir y el de Lodgepole Pine son los más representados.

Variables numéricas

En la Fig.1.4 se pueden observar las distribuciones de las variables numéricas para cada tipo de cobertura, todas son sesgadas. La variable Aspect es bimodal (da la Orientación de las caras del relieve). La variable Hillshade_3pm es la más centrada (probablemente las sombras que proyectan las montañas a esa hora debido al sol en verano son menores). En todas las variables excepto en Elevation las clases siguen la misma tendencia, en esta variable las distribuciones para cada grupo estan centradas distinto.

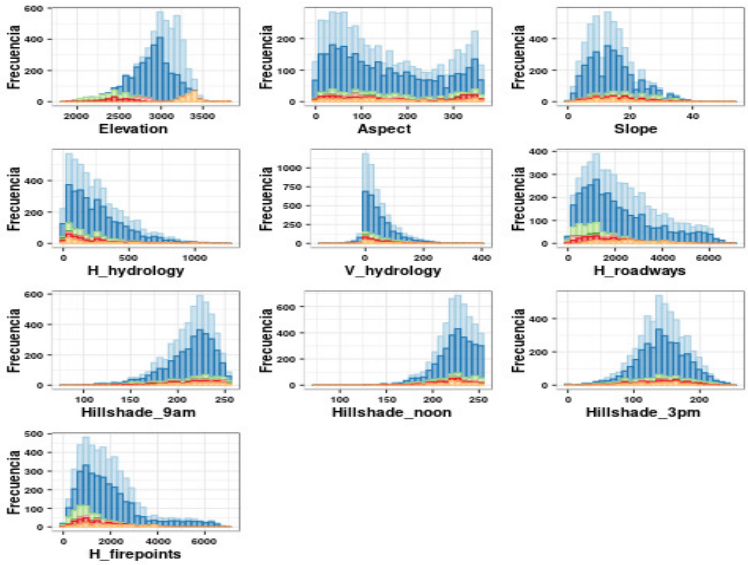

Figura 1.4: Histogramas de las variables numéricas.

Correlaciones

Calculando las correlaciones que tienen las variables entre si, obtenemos la Fig.**??**. Se observa que las variables Hillshade estan relacionadas con Aspect, Slope y entre ellas. Esto tiene que ver con como se calculan las cantidades hillshade que son los sombreados que se dibujan sobre los mapas cartográficos. Se supone una iluminación simulada que depende de la orientación a la fuente de luz, la cual esta basada en las variables Aspect y Slope. Las cantidades distancia vertical y horizontal a cursos de agua también estan relacionadas asi como la distancia horizontal a caminos y distancia a puntos de fuego.

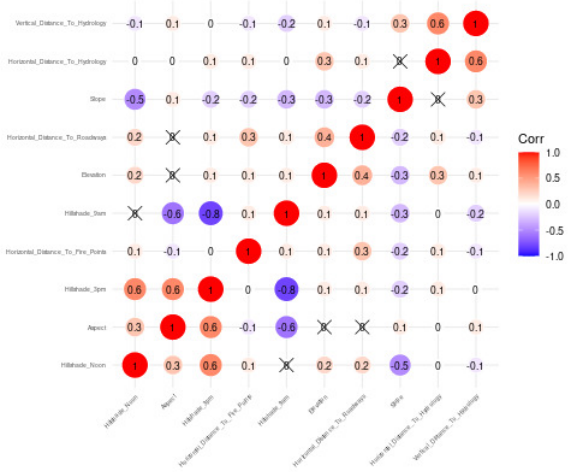

Figura 1.5: Correlaciones entre las variables numéricas. Las x corresponden a valores sin significancia.

Con base en esto se acotan las variables a usar a Elevation, Aspect, Slope, Horizontal_Distance_Roadways y Horizontal_Distance_to_Hydrology.

Análisis de componentes principales

Al realizar el PCA sobre las variables seleccionadas, se encuentra la Fig.1.6 cuyos ejes explican un 54,9 % de la varianza total, estas componentes según sus coeficientes, son de forma. La variable Slope está descorrelacionada de la de distancia a superficies de agua (Horizontal_Distance_To_Hydrology) y las variables Aspect (la orientación de la cara de la montaña) y Elevation tienen correlación baja para los datos considerados. Los observaciones de cada grupo estan muy superpuestas, pero se vislumbra que los puntos correspondientes a los grupos 3 a 6 tienden a estar concentrados a la derecha del gráfico, en valores bajos de elevación, distancia a caminos y a superficies de agua. Hay una división preliminar de los datos en dos grandes grupos.

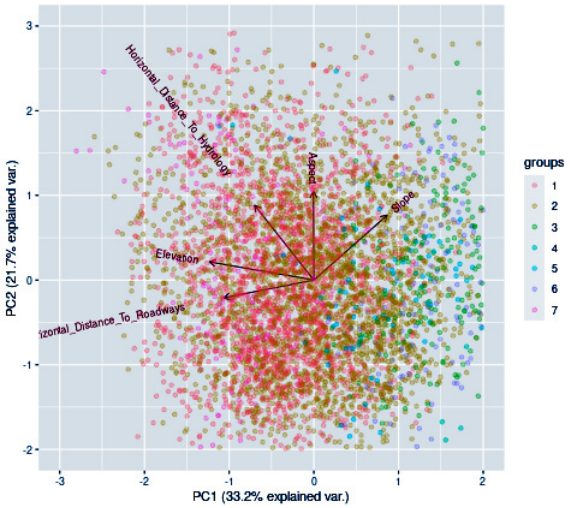

Figura 1.6: Análisis de componentes principales para 5 variables del dataset con las observaciones agrupadas por clase de cobertura forestal.

Hasta acá se han encontrado las variables mas relevantes y se ha encontrado una clasificación gruesa que contrapone dos grandes grupos basados en sus valores de elevación y su distancia a caminos y cursos de agua. En lo siguiente se realizará una clasificación más fina por varios métodos.

1.1. Clasificación Supervisada

Análisis de Discriminante lineal

El primer metodo a implementar es el LDA. Para que tenga validez el método deben cumplirse los supuestos de normalidad multivariada para cada nivel de cada variable, y de homocedasticidad.

Para evaluar normalidad multivariada usamos el test de Shapiro. Al evaluar con este test se obtuvo que se rechaza la normalidad multivariada para todas las variables en todos sus niveles. Al realizar de nuevo el test para las variables escaladas, rechaza de nuevo para todas, excepto para la clase "3".

Para considerar comparar grupos y encontrar diferencias, hay que determinar que las medias son distintas, para eso se realiza el test de Hotelling. Este test tiene como requisito la normalidad multivariada, que se rechazó al tomar el test de Shapiro. Aún así se lo evaluó obteniendo un p-valor de 0, rechazando por ende que las medias son iguales y dando vía a la comparación. En lo siguiente se evaluó el test de M de Box para determinar si se cumplen los requerimientos de igualdad de varianzas y covarianzas, el resultado fue un p-valor $< 2{,}2e-16$, rechazandose así la homocedasticidad. En tal caso deberá realizarse un Análisis Discriminante Lineal Robusto (QDA).

Sin embargo, al realizar el QDA se obtuvieron peores resultados que con el LDA, incluso escalando los datos. Con datos escalados se obtuvieron errores del 82,52 % con el QDA frente a 31,56 % con el LDA. Esos resultados no mejoraron incluso transformando las variables con un log o considerando un par de variables más en el análisis. En vista de que la cantidad de datos para este dataset no fué un problema, se seleccionaron un conjunto de 5000 observaciones para el conjunto de validación y 2500 para el conjunto de test. En la gráfica de la Fig.1.7 se detalla de a pares el resultado del análisis de discriminante para los datos escalados. Se observa que la variable que mejor explica la separación de los grupos parece ser la variable elevation.

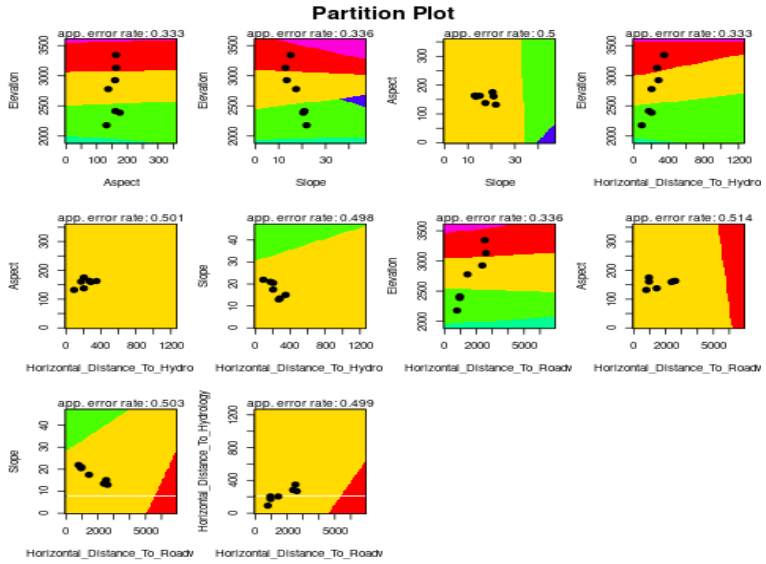

Figura 1.7: Análisis discriminante lineal para los datos

Máquina de soporte vectorial

Se implementó un método adicional de clasificación supervisada, el SVM. Se ensayó con varios kernel obteniéndose mejor resultado con el kernel gaussiano con un error medio en la predicción del 24,64 % en la clasificación ingenua frente a 28,89 % de error con el conjunto de test. No obstante al graficar, debido al gran numero de clases la

visualización no resulta muy clara. También se intentó implemental el método solo con puntos pertenencientes a las clases mayoritarias y pese a que el error mejora, la visualización no lo hace. En el notebook de R adjunto se pueden apreciar estas figuras.

1.2. Clasificación no Supervisada

Algoritmos jerárquicos

Como método de clasificación no supervisada se construyeron dendrográmas mediante single-linkage, average-linkage, complete-linkage, centroid-linkage y método de Ward. Los coeficientes de correlación cofrenética obtenidos se consignan en la Tabla1.2.

complete-linkage	0.7768551
average-linkage	0.7736942
single-linkage	0.1769797
método de Ward	0.6320018
centroid-linkage	0.7774966

Cuadro 1.2: Coeficientes de correlación cofrenética

Se observa que el método que divide con más naturalidad los datos es el del centroide, seguido por el complete-linkage. Para visualizar la división que este último método hizo de los datos, se grafican los grupos clasificados sobre los scatter-plot de a pares de las variables (ver Fig1.8) seleccionando las gráficas donde los grupos se ven más claros y se superponen menos.

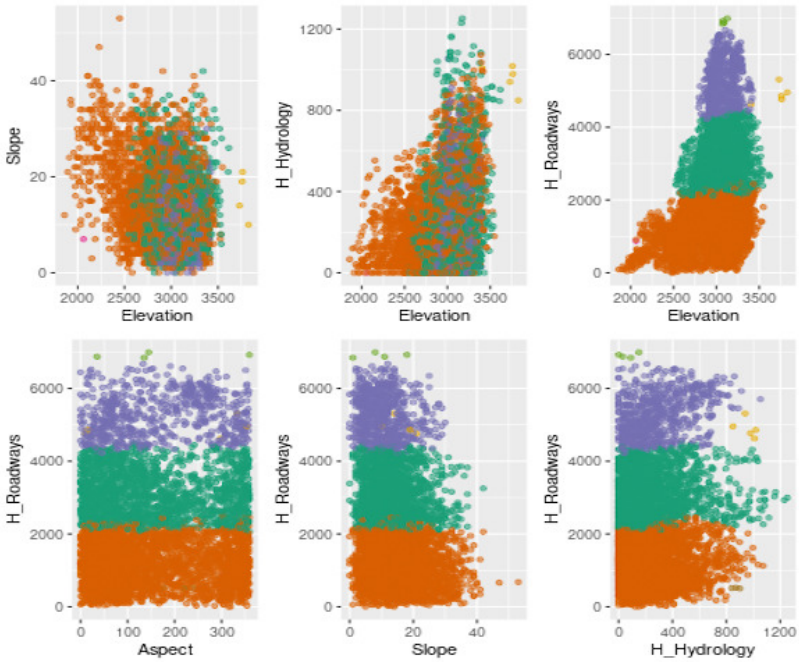

Figura 1.8: Partición basada en el algoritmo centroid-linkage

El número de clusters inicial que se dió fué de 7, pero se observan en general 3 clusters mayoritarios bien separados cuando se consideran combinaciones de las variables Elevation, Horizontal_Distance_To_Roadways, Horizontal_Distance_To_Hydrology y Slope. Por ende, bajo este método, estas serían las variables bajo las cuales se separan mejor las observaciones.

Algoritmos no jerárquicos

Como algoritmo de partición no jerárquico, se usó el k-means (ver Fig.1.9)

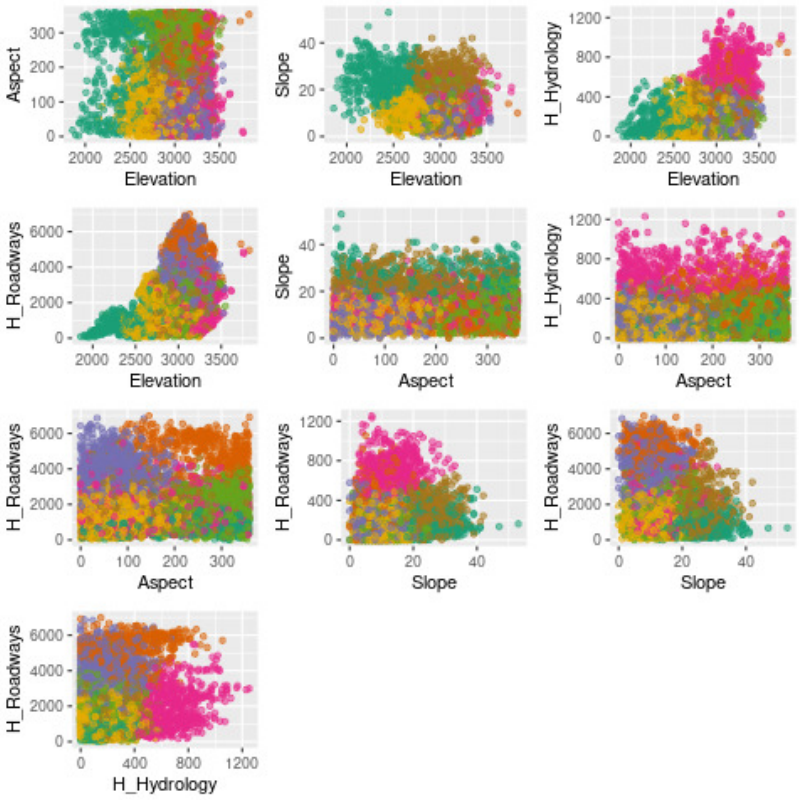

Figura 1.9: Partición basada en el algoritmo k-means

Este algoritmo detecta mas grupos en los datos, aunque se puede afirmar que las variables Elevation, Horizontal_Distance_To_Roadways e Horizontal_Distance_To_Hydrology siguen siendo variables interesantes para definir separaci'on más clara entre los datos.

1.3. Conclusiones

Se estudió el problema de clasificación de coberturas forestales con datos cartográficos. El interés principal fué tratar de detectar las variables relevantes para el problema, asi como las que pudiesen derivar más fácilmente en una catalogación rápida de las observaciones y una más fina, que diera cuenta de algunas relaciones relevantes entre los datos con miras a detectar grupos de vegetacion forestal. Se redujeron las variables del dataset con ayuda de métodos de ingeniería de datos y se encontró con un análisis de

discriminante lineal que la variable que da la altitud de los parches es muy importante para detectar el tipo de cobertura boscosa. Con métodos de clasificación no supervisada, se encontró que posiblemente la interacción de esta variable con las que dan cuenta de la distancia a cursos de agua y caminos y la pendiente sean también determinantes a la hora de clasificar el bosque. Esta conclusión está en congruencia con las derivadas del análisis de componentes principales realizado al principio.

1.4. Referencias

- La base de datos https://archive.ics.uci.edu/ml/datasets/Covertype
- Acerca de las variables cartográficas del dataset https://pro.arcgis.com/es/pro-app/latest/help/analysis/raster-functions/hillshade-function.htm
- Sobre el calculo de Hillshade https://www.e-education.psu.edu/geog480/node/490