Matemáticas Computacionales
Práctica 2: Estudio de una Base Datos

Profesor: Ángel Isabel Moreno Saucedo
Semestre Febrero - Junio 2021

1. Introducción

En esta práctica se estudiara un base de datos con la estadística descriptiva básica. Se analizara los atributos, tipos de datos, indicadores básicos de una y de dos variables. Se muestran algunos de los primeros pasos para luego implementar un preprocesamiento de datos para luego utilizarla en alguna experimentación o trabajo a tratar.

2. Base de datos: Identificación de vidrio

La base de datos de identificación del tipo de vidrio (Glass) es utilizada en la literatura en algoritmos de clasificación de aprendizaje máquina (*machine learning* en ingles) y el estudio de la clasificación de tipos de vidrio fue motivado por investigación criminológica. Las referencias [2] y [4] son algunos trabajos que utilizan esta base de datos.

2.1. Descripción del conjunto de datos

El dataset Glass cuenta con 214 observaciones que contiene ejemplos de 7 componentes químicos de vidrio. Esta base de datos. La base datos Glass tiene los siguientes 10 atributos:

1. Índice de refracción.
2. Na: sodio.
3. Mg: magnesio.
4. Al: aluminio.
5. Si: silicio
6. K: potasio.
7. Ca: calcio
8. Ba: bario.
9. Fe: hierro.
10. Tipo de vidrio.

El índice de refracción es un dato numérico, las unidades de medida de los componentes químicos del 2 al 9 es el porcentaje en peso en el óxido. El atributo tipo de vidrio es un número entero del 1 al 7 que clasifica la observación como:

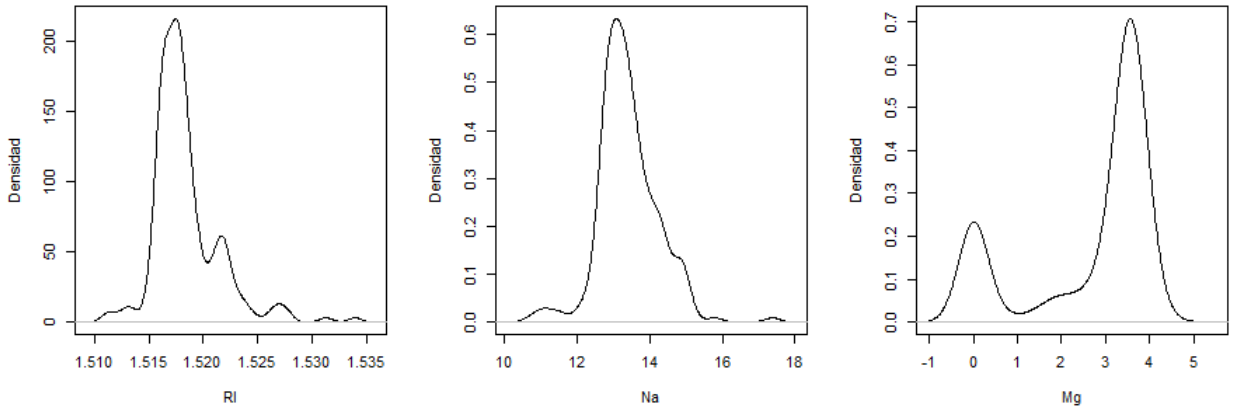

Figura 1: Gráficas de densidad del índice de refracción y los componentes químicos sodio y magnesio.

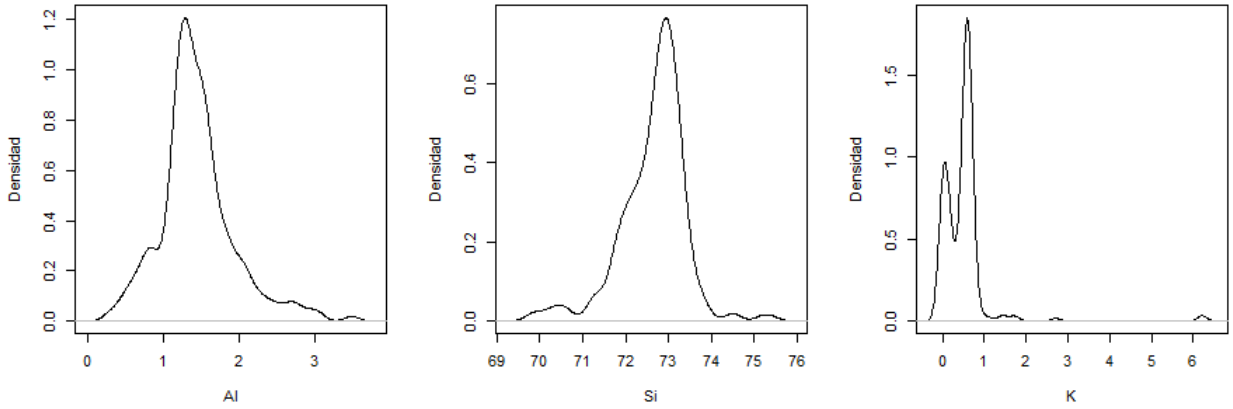

Figura 2: Gráficas de densidad de los componentes químicos aluminio, silicio y potasio.

1. building windows float processed
2. building windows non float processed
3. vehicle windows float processed
4. vehicle windows non float processed
5. containers
6. tableware
7. headlamps.

2.2. Estadística descriptiva de una variable

Los datos de índice de refracción tiene un mínimo de 1.511 y un máximo 1.534, con media y mediana idénticas con valor de 1.518. El componente químico sodio tiene un porcentaje mínimo de 10.73 y un máximo de 17.38, media de 13.41 y mediana de 13.30 y para el magnesio tiene porcentaje mínimo de 0 y una máximo de 4.49, media de 2.685 y media de 3.480. La Figura (1) muestra gráficas de densidad del indice de vidrio y los componentes sodio y magnesio.

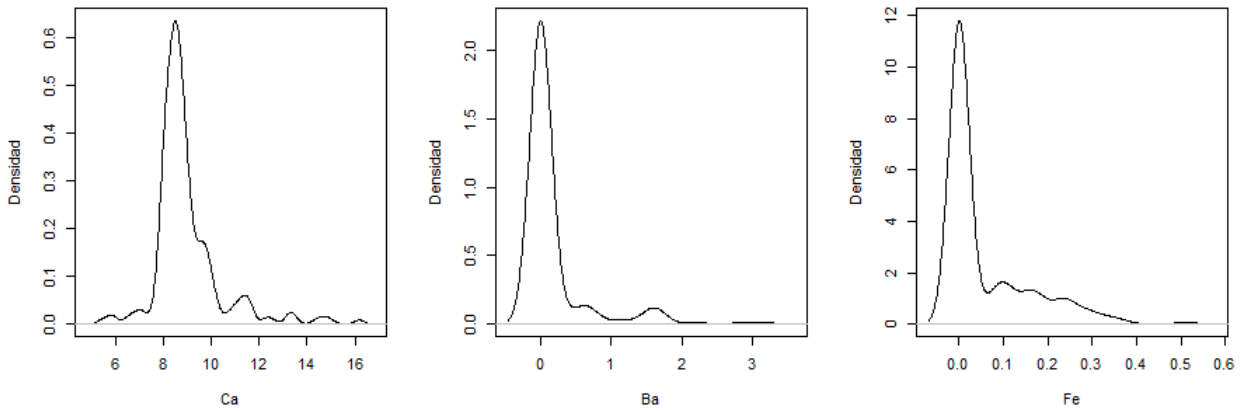

Figura 3: Gráficas de densidad de los componentes químicos calcio, bario y hierro.

Los datos del componente aluminio tiene sus datos entre [0.29, 3.5] con media de 1.445 y mediana de 1.36. El silicio se encuentra entre [69.81, 75.41] con media de 72.65 y mediana de 72.79. El potasio tiene datos entre [0, 6.21] con media de 0.4971 y mediana de 0.555. La Figura (2) muestra las gráficas de densidad de los componentes aluminio, silicio y potasio.

El químico calcio tiene un dato mínimo de 5.43 y un máximo de 16.19, con media de 8.957 y mediana 8.6. El bario tiene media de 0.175, mediana en el 0 y valores entre [0, 3.15]. El hierro tiene mínimo y máximo en 0 y 0.51 respectivamente, con media de 0.057 y mediana en 0. La Figura (3) muestra gráficas de densidad del químico calcio, bario y hierro.

El tipo de vidrio tipo 1 (building windows float processed) tiene 70 observaciones que corresponde el 32.71 %, el tipo 2 (building windows non float processed) tiene 76 observaciones que es el 35.51 %, el tipo 3 (vehicle windows float processed) tiene 17 observaciones que es el 7.94, el tipo 4 (vehicle windows non float processed) no tiene observaciones, el tipo 5 (containers) tiene 13 observaciones correspondiente al 6.07 %, el tipo 6 (tableware) tiene 9 observaciones que es el 4.20 % y el tipo 7 (headlamps) tiene 29 observaciones con porcentaje de 13.55 %. La Figura (4) muestra estos resultados.

2.3. Estadística descriptiva de dos variables

La Figura (5) muestra la correlación entre cada par de atributos utilizando el coeficiente de correlación de Pearson. Según Benesty et al. [1] este coeficiente mide correlación lineal y se encuentra entre el rango de [-1, 1], entre mas cerca este de los extremos se dice que las dos muestras están fuertemente correlacionadas, con la diferencia de que si éste es fuertemente correlacionado hacia el lado izquierdo (negativamente) indica relación inversa y viceversa, indica relación directa cuando es fuertemente correlacionado a la derecha (positivamente).

3. Tarea

Elija una base de datos de los dataset que se encuentran en `R`, ya sea del dataset que tiene `R` ya instalado o de la librería `mlbench`.

Realice un estudio de la base de datos elegida escribiendo un reporte que incluya introducción, base de datos y conclusiones. Dentro de la sección base de datos escriba para que se usa la base de datos y agrega 2 referencia de trabajos que la utilicen, añade subsecciones de descripción del conjunto de datos, estadística de una variable y de dos variables.

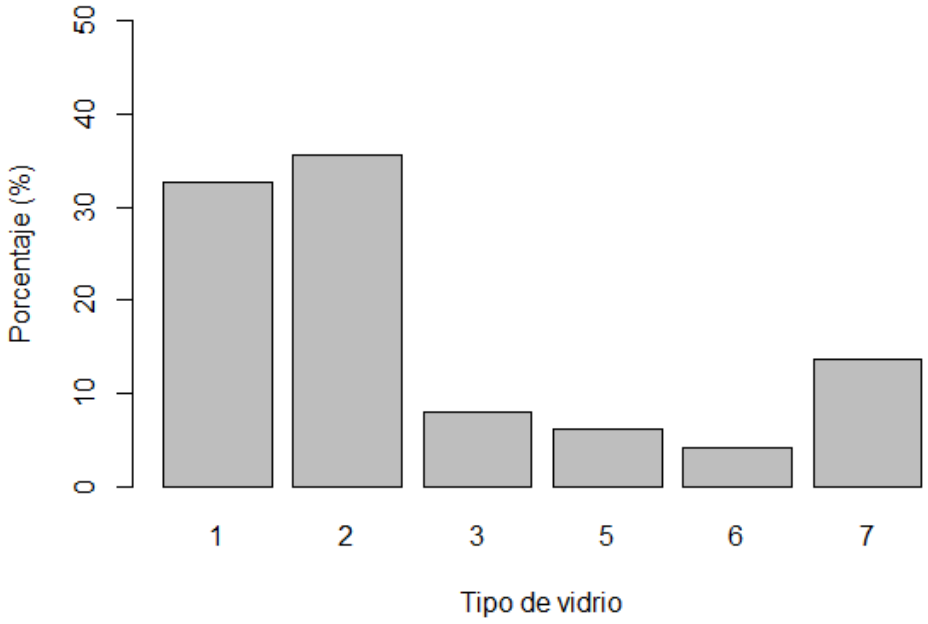

Figura 4

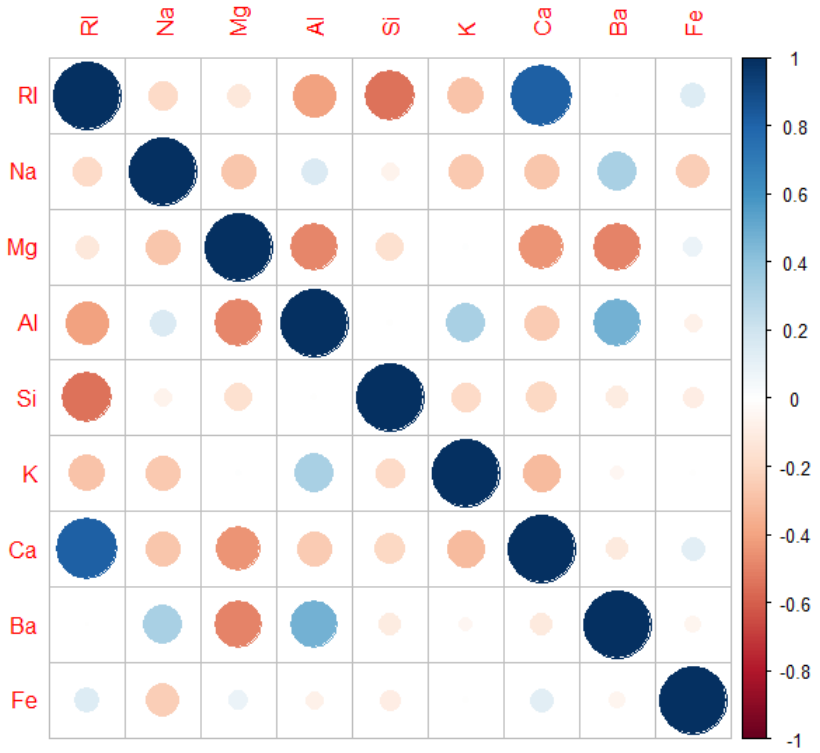

Figura 5

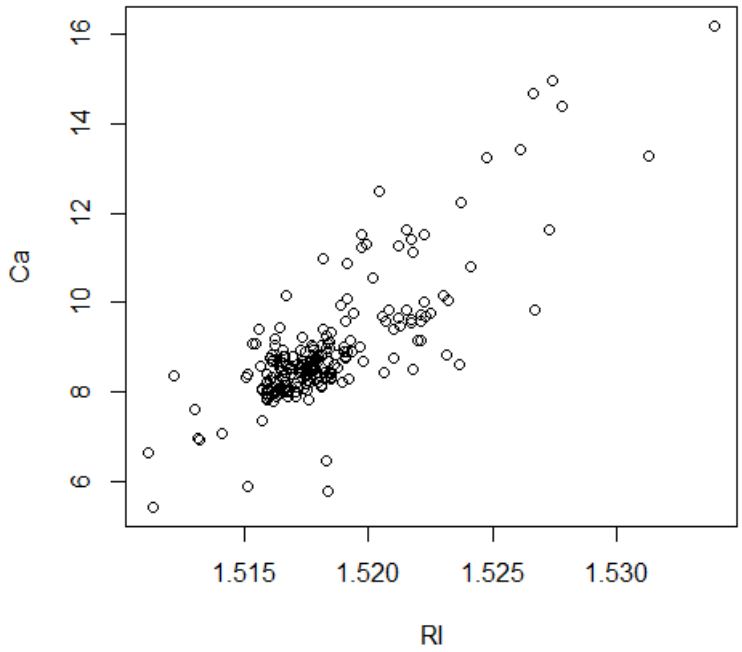

Figura 6

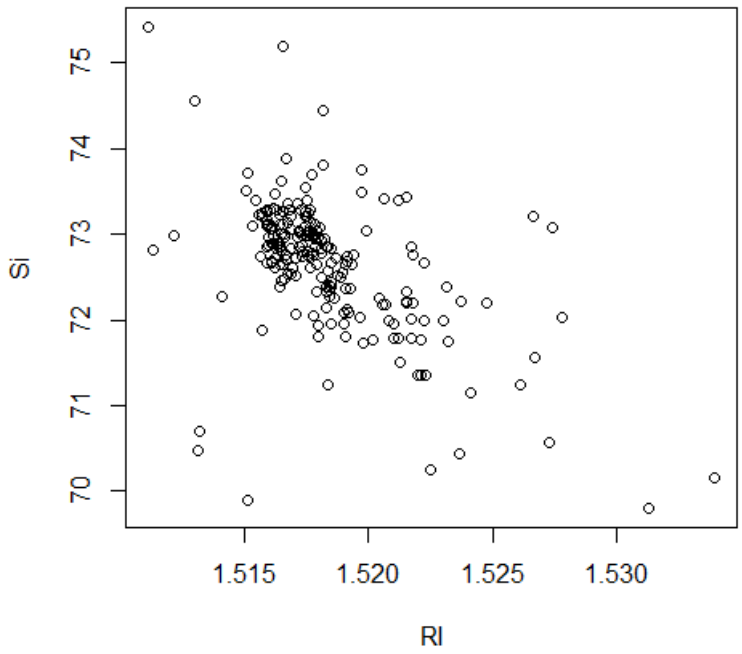

Figura 7

En la subsección estadística de una variable elabora una descripción de los atributos (columnas) apoyándote con una de las siguientes gráficas: histograma, gráfica de densidad o gráfica de caja bigotes (boxplot), para variables de tipo clase apóyate con gráfica de barras (barplot) ya sea de porcentaje o de cantidad de observaciones.

En la subsección estadística de dos variables crea la visualización de las correlaciones e identifica los atributos fuertemente correlacionados, luego añade, sí es posible, por lo menos 2 gráficos de dispersión (scatter plot) de los atributos fuertemente correlacionados indicando su relación.

Guarde el archivos .R de las visualizaciones y subirlo de su repositorio de Github.[3]

Referencias

[1] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[2] Krzysztof Krawiec. Genetic programming-based construction of features for machine learning and knowledge discovery tasks. *Genetic Programming and Evolvable Machines*, 3(4): 329–343, 2002.

[3] Ángel Moreno. Repositorio de Github. `https://github.com/angelmorenos`, 2021.

[4] Ping Zhong and Masao Fukushima. Regularized nonsmooth newton method for multi-class support vector machines. *Optimisation Methods and Software*, 22(1):225–236, 2007.