Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Aprendizaje no supervisado.

* Melany Pérez Vergara
* Ignacio Gamboa Alvarado
* Gonzalo Vásquez Carvajal

Minera de Datos.

Alex González Fuentes

Índice.

[Introducción. 2](#_Toc170590738)

[Set de Datos. 3](#_Toc170590739)

[Diamantes. 3](#_Toc170590740)

[Implementación. 4](#_Toc170590741)

[K-medias. 4](#_Toc170590742)

[Bibliografía 5](#_Toc170590743)

# Introducción.

El concepto de aprendizaje no supervisado es imitar la capacidad humana de aprender sin supervisión explicita. El aprendizaje no supervisado es utilizado en el análisis de conjunto de datos sin etiquetas o categorías previas. Esta herramienta para el análisis de datos, en las machine learning agrupa los datos en conjuntos, lo que permite descubrir patrones.

Una forma de agrupar los datos es a través de clústeres, un clúster se puede definir como una agrupación de cosas similares posicionadas o que ocurren de manera cercana entre sí, existen variedad de algoritmos que permite clasificar en clústeres, esos son: exclusivos, superpuestos, jerárquicos y probabilístico.

El algoritmo de agrupación K-medias (K-means) es un tipo de algoritmo de clúster exclusivo, esto debido a que asume que los puntos de datos pueden existir solo en un clúster, por otro lado, los datos quedan asignados en K grupos, donde K son la cantidad de clúster con relación a la distancia al centro del centroide. Los clústeres del tipo jerárquico separan los puntos de datos en grupos separados, y estos a su vez se fusionan de manera iterativa hasta lograr un solo clúster.

Estos dos tipos de algoritmos de clasificación en clústeres serán utilizados en el posterior análisis del set de datos.

# Set de Datos.

## Diamantes.

El set de datos ‘diamantes’ nos presenta 53490 observaciones, y a su vez cuenta con 10 variables que son ‘*price*’ (precio), ‘*carat*’ (quilate), ‘*cut’* (corte), ‘*color’* (color), ‘*clarity’* (claridad), ‘*depth’* (profundidad), *‘table’* (tabla), x, y, z.

Debido a la naturaleza de algunas variables, podemos identificar que de ellas 7 son variables del tipo cuantitativas, y 3 del tipo cualitativa.

El set de datos nos presenta como se catalogaron los diamantes a partir de cada una de sus características físicas y se les asigno un precio a raíz de ello. A partir de la utilización del aprendizaje no supervisado, se busca establecer conjunto de datos que permitan predecir esta asignación de precios. Pero previo a ello, requiere un tratamiento de los datos, eliminando datos atípicos o que no sean correctos, de esta manera quedando el nuevo set de datos solo con 53420 datos.

De las variables cualitativas podemos obtener la siguiente tabla, que nos da datos estadísticos de cada variable:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Carat | Depth | Table | Price | X | Y | Z |
| Min. | 0.2000 | 43.00 | 43.00 | 326 | 3.730 | 3.680 | 1.07 |
| 1° cuartil | 0.4000 | 61.00 | 56.00 | 949 | 4.710 | 4.720 | 2.91 |
| Media | 0.7000 | 61.80 | 57.00 | 2401 | 5.700 | 5.710 | 3.53 |
| Promedio | 0.7977 | 61.75 | 57.46 | 3931 | 5.732 | 5.735 | 3.54 |
| 3° cuartil | 1.0400 | 62.50 | 59.00 | 5323 | 6.540 | 6.540 | 4.04 |
| Max. | 5.0100 | 79.00 | 95.00 | 18823 | 10.740 | 58.900 | 31.80 |

Tabla 1: datos estadísticos del set de datos diamantes.

# Implementación.

## K-medias.

Para la utilización de K-medias en el set de datos, y poder encontrar un clúster adecuado que nos permita encontrar un patrón de comportamiento de las variables estas se deben tratar de la tal manera de que pueda trabajar adecuadamente con el algoritmo.

library(factoextra)

library(ks)

diamantes\_aux<-diamantes

names(diamantes)

diamantes<-diamantes[,-c(3,4,7)]

summary(diamantes)

names(diamantes)

table(diamantes$cut)

diamantes$cut[diamantes$cut%in%"Fair"]<-"1"

diamantes$cut[diamantes$cut%in%"Good"]<-"2"

diamantes$cut[diamantes$cut%in%"Very Good"]<-"3"

diamantes$cut[diamantes$cut%in%"Premium"]<-"4"

diamantes$cut[diamantes$cut%in%"Ideal"]<-"5"

diamantes$cut<-as.numeric(diamantes$cut)

str(diamantes)

set.seed(123)

selectrows <- sample(1:nrow(diamantes),round(0.75\*nrow(diamantes)))

muestra <- diamantes[-selectrows,]

elResto <- diamantes[selectrows,]

Para la realización del análisis de los datos haciendo uso de k-medias se utilizan las librerías ‘*factoextra*’ la cual permite extraer y visualizar los resultados de análisis de datos multivariable, incluyendo varios tipos de análisis como el análisis de componentes principales (PCA), análisis de clústeres y análisis de correspondencia múltiple (R Basics, s.f.). También es necesaria la utilización de la librería ‘*KS*’ (*Kernel Smoothing*) utilizado para datos univariado y multivariados, incluyendo funciones de densidad, densidad derivada, distribución acumulada, clústeres modales, análisis discriminativo, regiones modales significativas y pruebas de hipótesis de dos ejemplos (RDocumentation, s.f.).

Para una realización representativa del análisis de los datos, se trabajará con el 25% del total de datos, se eliminarán las columnas de datos relacionados con el precio de los diamantes, al ser lo que buscamos predecir con los patrones encontrados, y se eliminan las columnas relacionadas con el corte y el color, ya que generan interferencias con el análisis.

# Bibliografía

Bhatia, A., Yu-Wei, & Chiu, D. (2017). *Machine learning with R cookbook:* (Segunda ed.). Reino Unido: Packt Publishing Ltd.

Boehmke, B., & Greenwell, B. M. (202). *Hands-On Machine Learning with R (Chapman & Hall/CRC The R Series).* Tylor & Francis Group.

Fernández Jauregui, A. (s.f.). *How to code kNN algorithm in R from scratch*. Obtenido de Ander Fernández Jauregui: https://anderfernandez.com/en/blog/code-knn-in-r/

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2024). *Introducción a la minería de datos.* Madrid, España: Pearson Educación S. A.

IBM. (s.f.). *¿Qué es el aprendizaje no supervisado?* Obtenido de IBM: https://www.ibm.com/es-es/topics/unsupervised-learning

Lantz, B. (2019). *Machine Learning with R. Expert techniques for* (Tercera ed.). Packt Publishing Ltd.

Lewis, N. (2017). *Machine Learning made easy with R: An Intuitive Step by Step Blueprint for Beginners.* CreateSpace Independent Publishing Platform.

R Basics. (s.f.). *The Ultimate Guide to the factoextra Package in R*. Obtenido de R Basics: https://rbasics.org/packages/factoextra-package-in-r/

RDocumentation. (s.f.). *ks (version 1.10.7)*. Obtenido de RDocumentation: https://www.rdocumentation.org/packages/ks/versions/1.10.7