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](#_Toc170654075)

[Set de Datos. 3](#_Toc170654076)

[Diamantes. 3](#_Toc170654077)

[Implementación. 4](#_Toc170654078)

[Análisis de componentes principales (PCA). 4](#_Toc170654079)

[K-Medias. 5](#_Toc170654080)

[Comparativa. 7](#_Toc170654081)

[Clústeres K=2 8](#_Toc170654082)

[Clústeres K=3 9](#_Toc170654083)

[Conclusiones 10](#_Toc170654084)

[Bibliografía 11](#_Toc170654085)

# 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.

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.

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.).

## Análisis de componentes principales (PCA).

En primera instancia se genera un análisis de componentes principales de cada una de las variables, incluyendo datos cualitativos, pasando estos a datos del tipo *dummy* para que puedan ser considerados en el análisis de los datos, obteniendo la siguiente grilla que nos muestra la significancia de cada una de las variables y nos permite reducir la cantidad de estas para un mejor análisis de los datos.

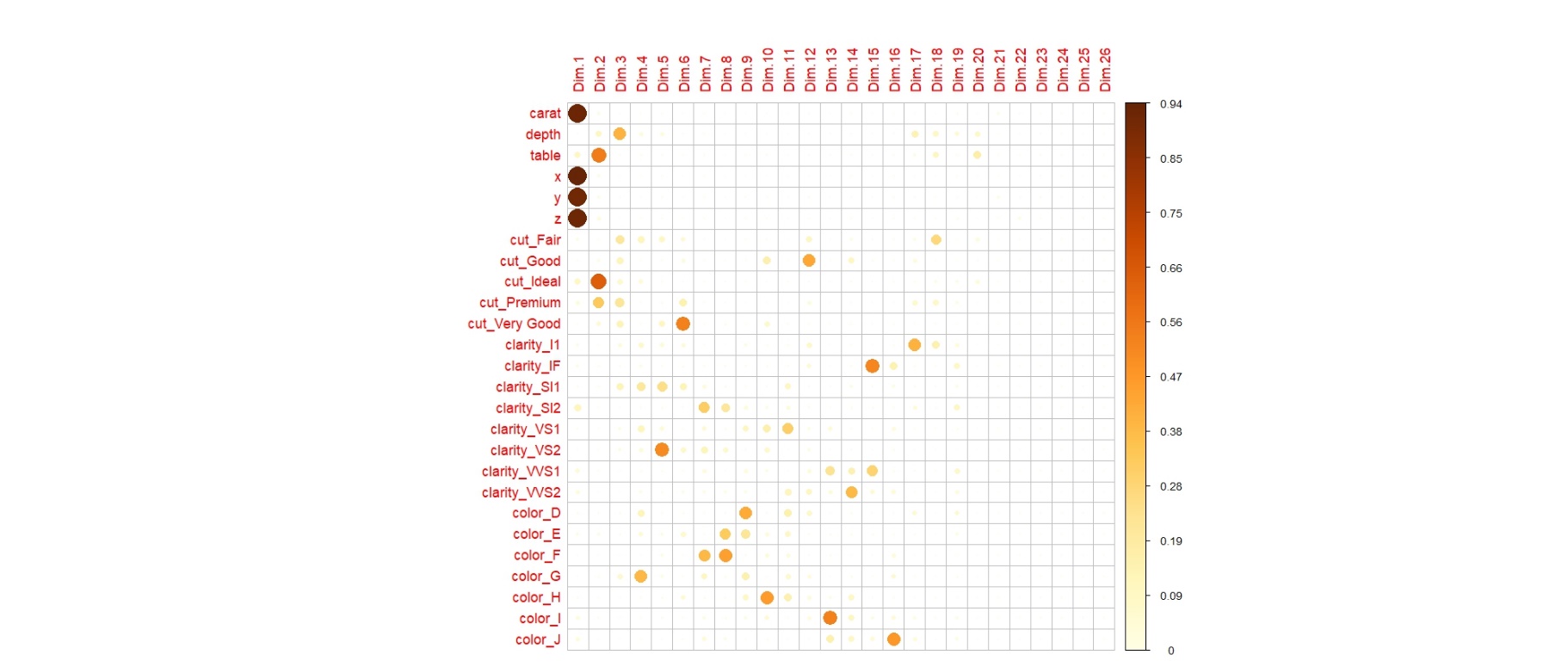


Ilustración 1: Representación gráfica de Análisis de componentes principales.

Al considerar que las variables *clarity* y *color* son variables del tipo nominal, podemos descartarlas como significativas en el análisis de datos, mientras que las variables *cut* al ser del tipo ordinal y genera un orden en la clasificación de los diamantes, se considera para continuar en el análisis de datos.

Para una realización representativa del análisis de los datos, se trabajará con el 25% del total de datos, esto debido a la cantidad de recursos necesarios para realizar un análisis total de los datos. Se eliminan las variables cualitativas nominales al ser considerados no significativos en el análisis de los datos para determinar el precio de los diamantes, al realizar un nuevo análisis de componentes principales, se obtiene la siguiente gráfica:

Gráfico, Gráfico de dispersión, Gráfico de burbujas

Descripción generada automáticamente

Ilustración 2: Nuevo PCA, sin variables cualitativas nominales.

## K-Medias.

Desde la librería *factoextra* se hace uso de la función *fviz\_nbclust()* dicha función nos permite determinar y visualizar en número óptimo de clústeres, mediante la utilización de diversos métodos. En la utilización del método *Elbow*, este nos indica que la cantidad optima de clústeres es de K=3, mientras que en la utilización del método *Silhoutte* indica que la cantidad optima de clústeres es de K=2, por otro lado, el método GAP indica de igual manera que la cantidad optima de clústeres es de K=2.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración 3: Cantidad de clústeres según método Elbow

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración 4: Cantidad de clústeres según método Silhoutte

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración 5: Cantidad de clústeres según método Gap

## Comparativa.

El set de datos de muestras utilizados para realizar el análisis de k-medias es necesario poder ahora asignarle como atributo a que agrupación pertenecen, esto con ayuda de gráficos de caja podemos ver como se comporta las agrupaciones asignadas principalmente con el precio de los diamantes, y posteriormente con cada una de las variables del set de datos.

### Clústeres K=2

Al realizar un análisis grafico de los clústeres, observamos que estos se encuentran realizando las comparativa de patrones entre las dimensiones Dim1 y Dim2.Forma

Descripción generada automáticamente

Ilustración 6: Cantidad de datos vinculados a cada clúster

Gráfico

Descripción generada automáticamente

Ilustración 7: Visualización de la agrupación de los datos

Gráfico

Descripción generada automáticamente

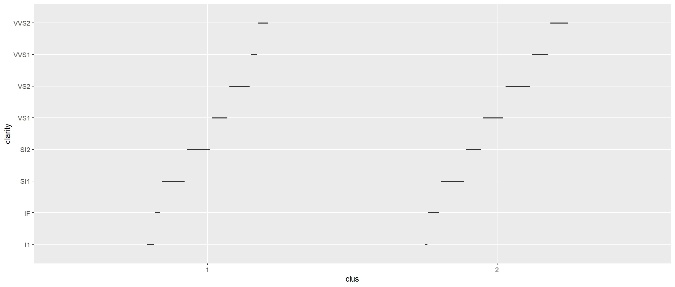


Ilustración 9: Clus VS Clarity

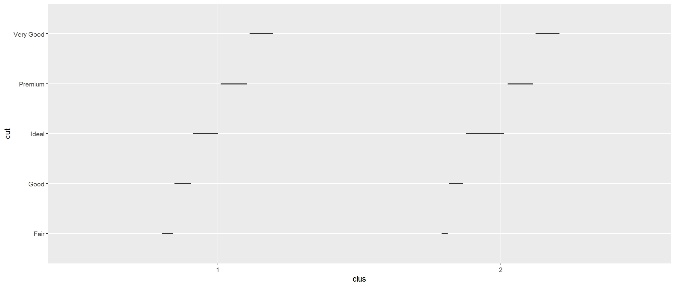


Ilustración 11: Clus VS Cut

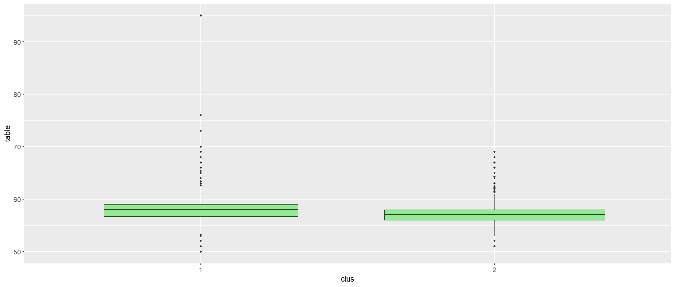


Ilustración 13: Clus VS Table

Ilustración 8: Clus VR Carat

Interfaz de usuario gráfica, Aplicación, Teams

Descripción generada automáticamente

Ilustración 10: Clus VS Color

Gráfico

Descripción generada automáticamente

Ilustración 12: Clus VS Depth

Gráfico

Descripción generada automáticamente

Ilustración 14: Clus VS Price

Gráfico, Gráfico de barras

Descripción generada automáticamente

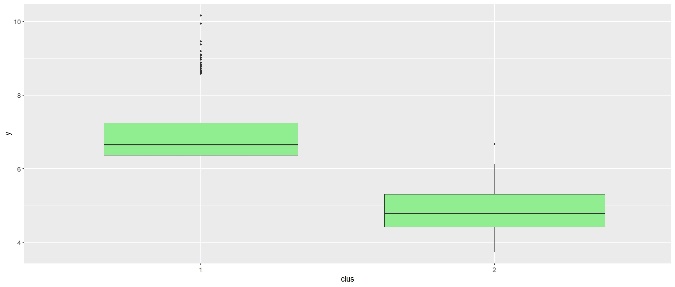


Ilustración 16: Clus VS Y

Ilustración 15: Clus VS X

Gráfico

Descripción generada automáticamente

Ilustración 17: Clus VZ Z

### Clústeres K=3

Realizando un procedimiento similar al análisis de los datos con clústeres k=2, en un clúster k=3 se comienzan a ver diferencias en el comportamiento de los datos:

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración 18: Distribución de los datos respecto a los clústeres.

Un dibujo de un animal

Descripción generada automáticamente con confianza baja

Ilustración 19: Distribución de las agrupaciones respecto a las dimensiones de los componentes principales

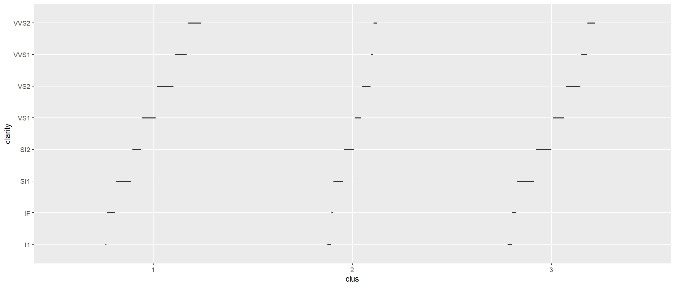


Ilustración 21: Clus VS Clarity

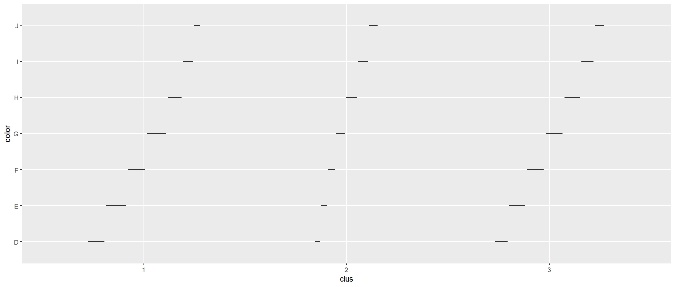


Ilustración 23: Clus VS Color

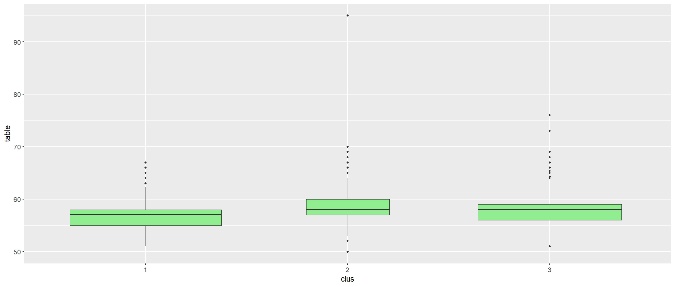


Ilustración 25: Clus VS Table

Gráfico, Gráfico en cascada

Descripción generada automáticamente

Ilustración 20: Clus VS Carat

Interfaz de usuario gráfica, Aplicación, Teams

Descripción generada automáticamente

Ilustración 22: Clus VS Cut

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración 24: Clus VS Depth

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico en cascada

Descripción generada automáticamenteIlustración 26: Clus VS Price

Ilustración 28: Clus VS Y

Gráfico, Gráfico en cascada

Descripción generada automáticamente

Ilustración 27: Clus VS X

Gráfico

Descripción generada automáticamente con confianza media

Ilustración 30: Clus VS Z

Observamos que la distribución de cada uno de los puntos de las dimensiones no cambia respecto a la anterior, más solamente la aparición de una tercera agrupación hace que los datos tengan una distribución ‘más equitativa’ entre clústeres.

Al realizar un análisis de los gráficos obtenidos, podemos observar que los clústeres seleccionados pueden explicar de manera eficaz las clasificaciones cualitativas de cada diamante (*Ilustración 9*, *Ilustración 10*, *Ilustración 11, Ilustración 21, Ilustración 22, Ilustración 23*). También se observa en el grafico Clus VS Price (*Ilustración 14, Ilustración 26*) relaciona de buena forma los precios y tiene una fuerte relación (o similar con la distribución de datos) respecto a la gráfica Clus VS Carat (*Ilustración 8, Ilustración 20*). La aparición de datos atípicos se debe a la naturaleza de los diamantes, debido a que pueden existir excepciones entre sus valores atípicos atener un atributo poco común.

# Conclusiones

En el proceso de tratamiento de se debe considerar que un exceso de variables a la hora de analizar los datos puede generar interferencias o ruidos en los algoritmos. De igual manera la cantidad de los datos impactas directamente en el tiempo de procesamiento y en algunos casos siendo inviable el análisis debido a la capacidad del equipo utilizado en el análisis, las capacidades tecnológicas son esenciales a la hora de aplicar este tipo de análisis.

Se considero la reducción de datos, debido a que en comportamiento de los datos en el análisis de componentes principales (*Ilustración 1*) tiene un comportamiento al extrapolar con el PCA realizado con los datos reducidos (*Ilustración 2*), indicando que es posible extrapolar el procedimiento y/o los resultados con el set de datos original y obtener resultados similares a los vistos en este informe.

Los clúster o agrupaciones de las diversas dimensiones de componentes principales que describen el comportamiento de los datos es una herramienta útil a la hora de querer establecer patrones de comportamiento que establezcan o relaciones estas agrupaciones a una o muchas variables. Se trabajo el análisis de los datos con clústeres con K=2 y K=3, aunque la distribución no cambia respecto a las dimensiones más representativas en componentes principales, si observamos que para un clúster con K=3 existe una mejor descripción del precio de los diamantes en las agrupaciones, encontrando datos atípicos en los grupos 1 y 3, mientras que el grupo 2 está mejor descrito y acotado sin encontrar datos atípicos en dicho grupo (*Ilustración 23*).

Por otro lado, se ve una fuerte relación entre el comportamiento del precio de los diamantes y la variable *Carat,* esto puede derivarse del análisis de componentes principales debido a que es una de las variables con más correlación con la Dim1 y se puede intuir que es un factor a importante a la hora de tazar un diamante.

# 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