Trabajo Final

Eduardo Raúl Herrero Rivero

Instituto Superior Politécnico Córdoba

Tecnicatura en Ciencia de Datos e Inteligencia Artificial, cohorte 2020.

Materia: Minería de Datos

Docente: Omar Horta Fischer

Córdoba, 2023

Trabajo Final

Minería de Datos

# Presentación:

La reutilización de programas es un concepto importante en la programación de software que se refiere a la práctica de utilizar código existente en la creación de un nuevo programa o en la modificación de uno ya existente. En lugar de crear todo el código desde cero, los programadores pueden utilizar módulos, funciones, clases o bibliotecas previamente desarrolladas que ya han sido probadas y están disponibles para su uso.

*La minería de datos* es un proceso que implica descubrir patrones, relaciones y tendencias en grandes conjuntos de datos utilizando técnicas estadísticas y de aprendizaje automático. En términos simples, la minería de datos implica encontrar información valiosa y útil en grandes cantidades de datos.

Otras definiciones:

* Según la definición del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE), la minería de datos es "el proceso de descubrir patrones significativos en grandes conjuntos de datos".
* Para el software de minería de datos RapidMiner, la minería de datos es "el proceso de explorar y analizar grandes cantidades de datos para descubrir patrones y relaciones significativas que pueden ser útiles para tomar decisiones empresariales".
* En palabras del experto en minería de datos, Gregory Piatetsky-Shapiro, la minería de datos es "el proceso de aplicar métodos matemáticos y estadísticos a grandes conjuntos de datos para extraer patrones ocultos y relaciones significativas".
* Para el autor y científico de datos, David Hand, la minería de datos es "el proceso de extracción de información útil de grandes conjuntos de datos".

# Objetivos:

Desarrollar en forma práctica un caso de búsqueda de patrones y conceptualizar la minería de datos.

# Instrucciones:

1. Conceptualizar a partir de las siguientes preguntas:
2. ¿Qué significa "patrones significativos" en la definición del IEEE?
3. ¿Cómo se pueden aplicar los métodos matemáticos y estadísticos en la minería de datos?
4. ¿Cómo se define "información útil" en la definición de David Hand?
5. ¿Cómo la minería de datos difiere de la analítica de datos o el big data?
6. ¿Qué son los algoritmos de aprendizaje automático y cómo se relacionan con la minería de datos?

# Respuestas:

a. En la definición del IEEE, "patrones significativos" se refiere a relaciones, tendencias o características que se encuentran en grandes conjuntos de datos y que pueden ser importantes o útiles para la toma de decisiones. Estos patrones pueden ser información oculta o no obvia que puede proporcionar información valiosa para los usuarios.

b. Los métodos matemáticos y estadísticos se aplican en la minería de datos para analizar grandes conjuntos de datos y descubrir patrones o tendencias. Los métodos incluyen la regresión, la clasificación, el clustering y el análisis de asociación, entre otros. Estos métodos ayudan a identificar patrones y relaciones en los datos y a hacer predicciones sobre futuros resultados.

c. En la definición de David Hand, "información útil" se refiere a la información que se extrae de los grandes conjuntos de datos que puede ser aplicada para tomar decisiones en un contexto empresarial o de otro tipo. Por ejemplo, esta información puede incluir tendencias de ventas, preferencias del cliente o patrones de comportamiento en línea.

d. La minería de datos se enfoca en descubrir patrones significativos en grandes conjuntos de datos, mientras que la analítica de datos se enfoca en el análisis de datos para identificar tendencias y patrones que pueden ayudar a las empresas a tomar decisiones. Por otro lado, el big data se refiere a la gestión y análisis de grandes cantidades de datos que son demasiado grandes para ser manejados por herramientas de software tradicionales.

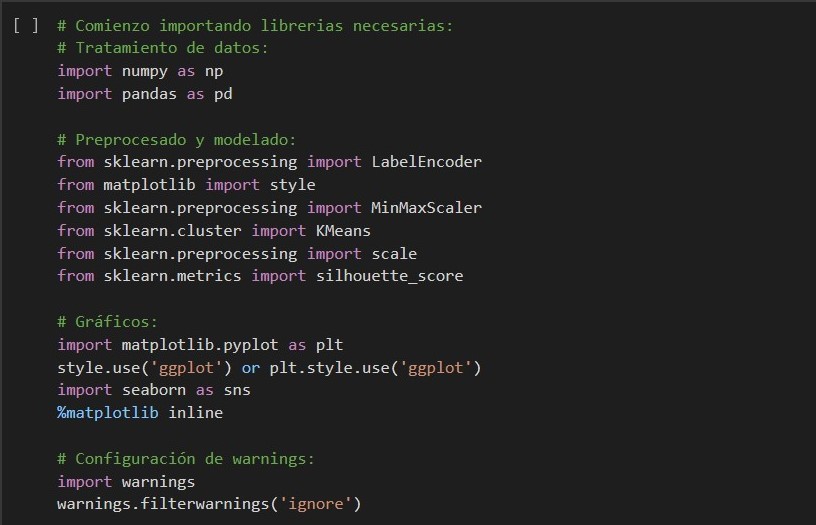
e. Los algoritmos de aprendizaje automático son un tipo de técnica estadística que se utiliza en la minería de datos para identificar patrones en los datos y hacer predicciones. Estos algoritmos se basan en datos de entrenamiento y aprenden a medida que se les da más información. La minería de datos y los algoritmos de aprendizaje automático están estrechamente relacionados, ya que los algoritmos se utilizan en la minería de datos para descubrir patrones y relaciones en los datos.

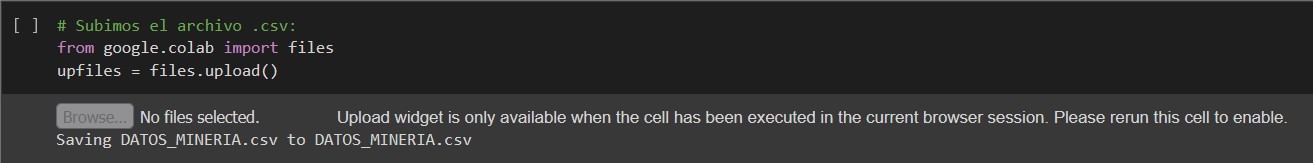
1. Se deberá reutilizar un algoritmo ya creado y un modelo que detecta determinados objetos dentro de una imagen obtenida de redes sociales y/o internet. Para realizar este trabajo siga el caso ejemplo y trate de generar un valor agregado, como gráficos.

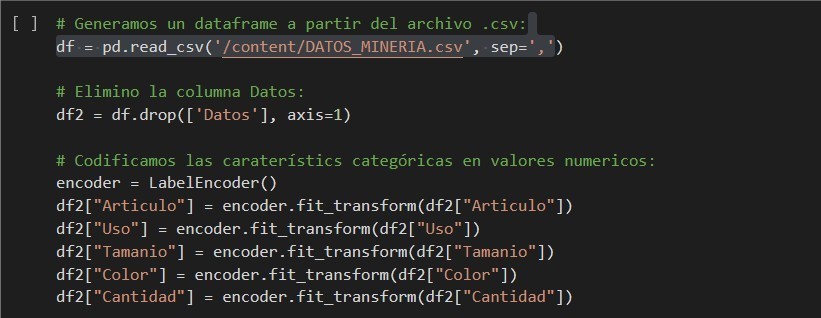
# Respuesta:

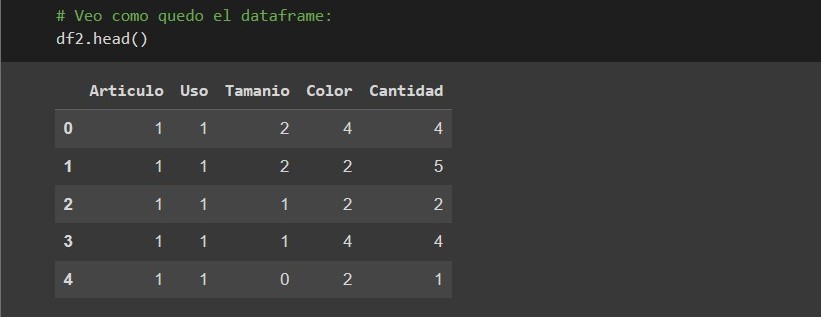
Link al repositorio con el ejercicio en Python:

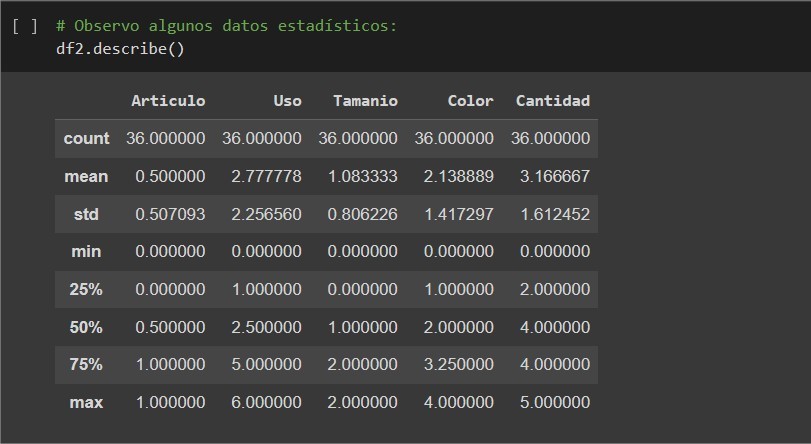
[**https://github.com/EdCsDsIA/Mineria\_de\_Datos.git**](https://github.com/EdCsDsIA/Mineria_de_Datos.git)

****

****

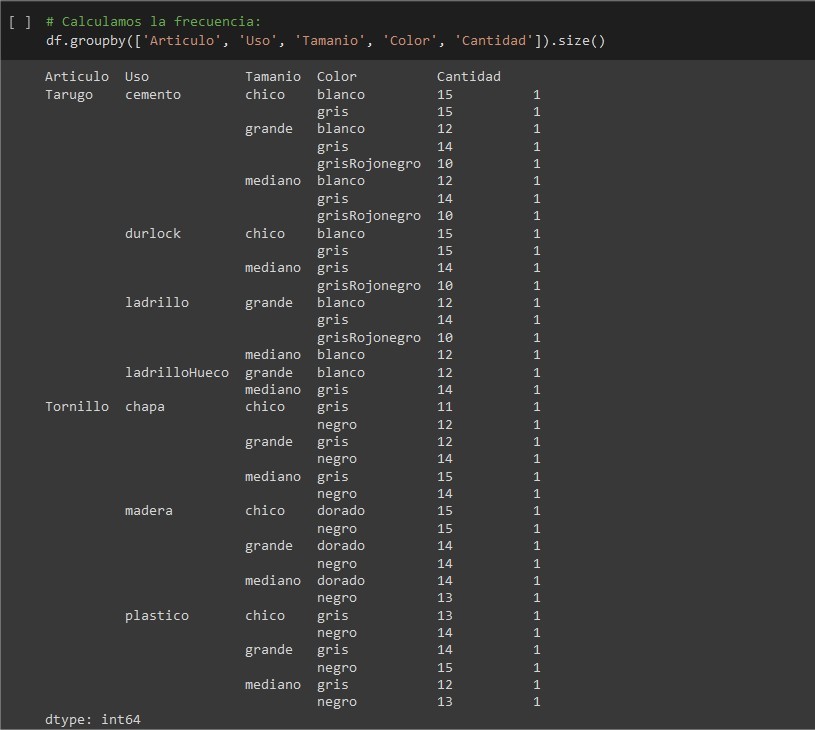
****

****

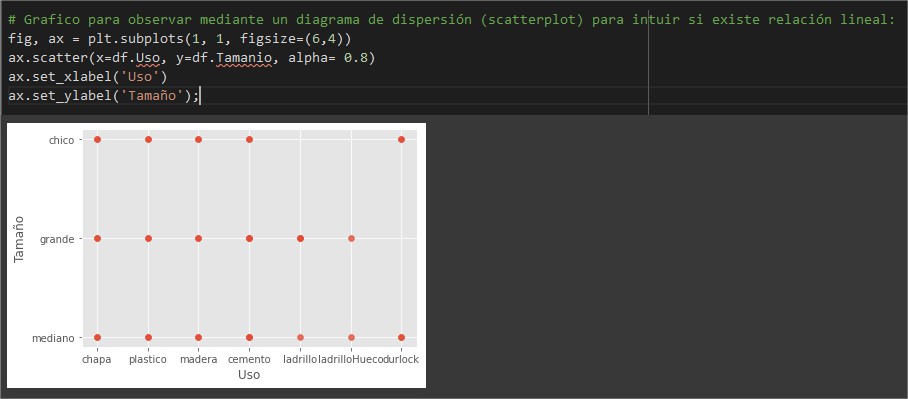
****

Se puede interpretar lo siguiente para cada columna:

* Articulo: La columna contiene 36 valores. No se puede inferir mucho sobre los datos, ya que parece ser una columna categórica binaria que solo puede tomar valores 0 y 1.
* Uso: La columna contiene 36 valores. La media es 2.78, lo que sugiere que el uso promedio de los artículos es de alrededor de 3 veces. La desviación estándar es de 2.26, lo que indica una gran variabilidad en los datos. El valor mínimo es 0, lo que sugiere que algunos artículos no se usaron en absoluto. El valor máximo es 6, lo que indica que algunos artículos se usaron mucho más que otros.
* Tamanio: La columna contiene 36 valores. La media es 1.08, lo que sugiere que el tamaño promedio de los artículos es de alrededor de 1. La desviación estándar es de 0.81, lo que indica una variabilidad moderada en los datos. El valor mínimo es 0, lo que sugiere que algunos artículos pueden ser muy pequeños. El valor máximo es 2, lo que sugiere que algunos artículos pueden ser relativamente grandes.
* Color: La columna contiene 36 valores. La media es 2.14, lo que sugiere que el color promedio de los artículos es de alrededor de 2. La desviación estándar es de 1.42, lo que indica una gran variabilidad en los datos. El valor mínimo es 0, lo que sugiere que algunos artículos pueden no tener un color definido. El valor máximo es 4, lo que sugiere que algunos artículos pueden tener un color muy distintivo.
* Cantidad: La columna contiene 36 valores. La media es 3.17, lo que sugiere que la cantidad promedio de los artículos es de alrededor de 3. La desviación estándar es de 1.61, lo que indica una variabilidad moderada en los datos. El valor mínimo es 0, lo que sugiere que algunos artículos pueden no estar disponibles en grandes cantidades. El valor máximo es 5, lo que sugiere que algunos artículos están disponibles en grandes cantidades.

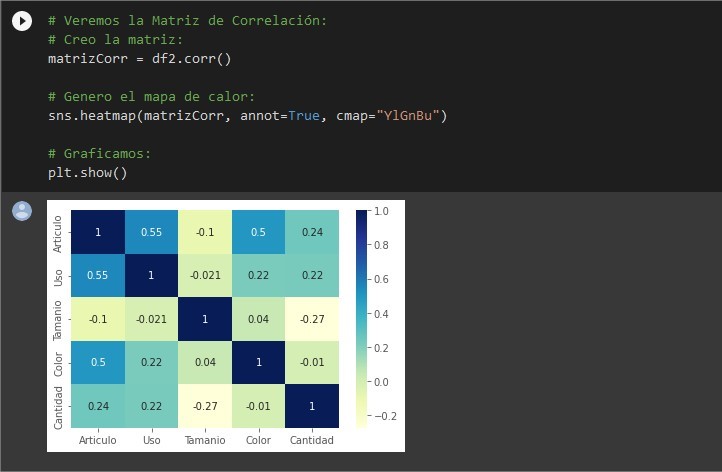
****

La frecuencia indica cuántas veces se repite cada combinación de valores únicos de las variables categóricas Articulo, Uso, Tamanio, Color y Cantidad. En este caso, se han agrupado los valores de estas variables en función de su frecuencia y se ha calculado cuántas veces aparece cada combinación única en el conjunto de datos. Por ejemplo, la combinación de valores Articulo="Tarugo", Uso="cemento", Tamanio="chico", Color="blanco" y Cantidad=15 aparece una sola vez en el conjunto de datos.



A simple vista, parece que hay una tendencia a que los materiales más grandes se utilizan para usos más intensivos. También podemos observar que el tamaño de los materiales de madera y chapa es principalmente pequeño o mediano, mientras que el tamaño de los ladrillos, cemento y durlock tiende a ser mediano o grande.

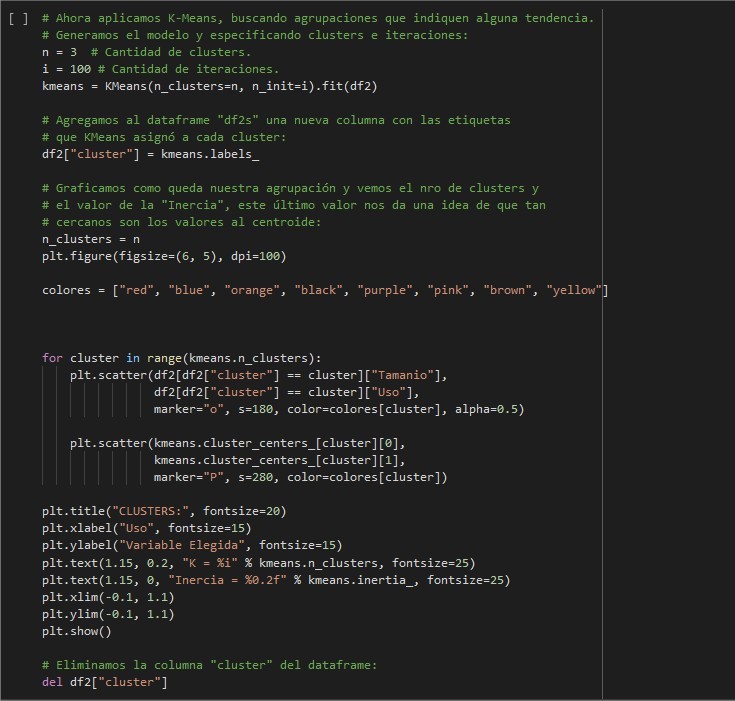
Sin embargo, hay que tener en cuenta que este gráfico solo muestra una relación aparente y no indica necesariamente una correlación lineal entre las variables. Para confirmar si hay una relación lineal, sería necesario calcular la correlación entre las variables y verificar si es significativa.

****

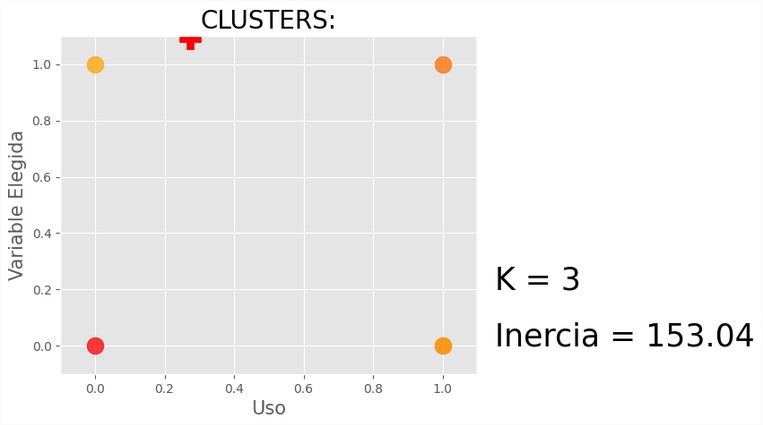
La matriz que se muestra es una matriz de correlación, que indica cómo están relacionadas linealmente las variables del DataFrame. Los valores de la matriz de correlación varían entre -1 y 1, donde los valores más cercanos a 1 indican una correlación positiva fuerte entre las variables, y los valores más cercanos a -1 indican una correlación negativa fuerte entre las variables. Un valor de 0 indica que no hay correlación lineal entre las variables.

Al analizar la matriz de correlación dada, se puede ver que la variable "Articulo" está positivamente correlacionada con "Uso" y "Color", lo que sugiere que es probable que haya una relación entre el tipo de artículo y su uso o color. Además, "Uso" y "Cantidad" también están positivamente correlacionados, lo que sugiere que los artículos que se usan con mayor frecuencia también se compran en mayores cantidades.

Sin embargo, la variable "Tamanio" está débilmente correlacionada con todas las otras variables, lo que sugiere que no hay una relación fuerte entre el tamaño del artículo y las otras variables en el DataFrame.

****

La visualización resultante muestra cómo se agrupan las observaciones del conjunto de datos según dos variables, "Tamanio" y "Uso", y se les asigna un color diferente a cada cluster. Los puntos marcados con una "P" indican los centroides de cada cluster.

La inercia, medida que se muestra en el gráfico, es una medida de qué tan cercanas están las observaciones en un cluster con su centroide. Cuanto menor sea la inercia, más compacto será el cluster y mejor será el agrupamiento. Creo que no se consigue una agrupación útil con este set de datos.****