PUCE Virtual
EDUCACIÓN FLEXIBLE

Adquisición, gestión y gobernanza de datos

Proceso de Descubrimiento del Conocimiento

Clase 1

MAESTRÍA EN
SISTEMAS DE INFORMACIÓN
Mención Data Science

La excelencia no se improvisa

INTRODUCCIÓN

En esta sesión, nos adentraremos en la naturaleza de los datos y su importancia en el contexto de la adquisición, gestión y gobernanza de la información. Comenzaremos explorando los tipos de atributos que componen los datos, distinguiendo entre atributos cualitativos y cuantitativos, y cómo estos influyen en la interpretación y el análisis de la información. Además, clasificaremos los datos según su estructura en tres categorías: estructurados, semiestructurados y no estructurados. Esta clasificación no solo es crucial para la organización de los datos, sino que también impacta directamente en la forma en que se pueden utilizar y gestionar en diferentes sistemas.

Posteriormente, nos enfocaremos en el Proceso de Descubrimiento del Conocimiento (KDD), que implica una serie de etapas destinadas a transformar los datos en conocimiento útil. Analizaremos cada fase del proceso, desde la selección y el preprocesamiento de datos hasta la interpretación de los resultados. Este enfoque nos permitirá comprender cómo los diferentes tipos de datos y atributos se integran en el KDD, facilitando la toma de decisiones informadas y estratégicas en las organizaciones. Así, un entendimiento profundo de estos conceptos será fundamental para optimizar la gestión y gobernanza de los datos en cualquier entorno empresarial.

RDA 1

Utilizar cada una de las etapas del proceso de Descubrimiento del Conocimiento en el contexto de Big Data.

Proceso de Descubrimiento del Conocimiento

El proceso de descubrimiento del conocimiento es esencial en el ámbito académico y científico, ya que permite a los investigadores adquirir nuevas comprensiones y desarrollar teorías que explican fenómenos complejos. Este proceso involucra la recolección sistemática de datos, el análisis crítico de la información y la formulación de hipótesis que pueden ser probadas y validadas. A través de metodologías rigurosas, los investigadores transforman datos en conocimiento aplicable, contribuyendo así al avance de diversas disciplinas. Además, el descubrimiento del conocimiento no es un esfuerzo solitario; se nutre de la colaboración entre diferentes áreas y de la integración de múltiples perspectivas, lo que enriquece el proceso y aumenta su relevancia en la solución de problemas contemporáneos (Popper, 1959; Kuhn, 1962).

Datos

Los datos son la materia prima del análisis informático y se definen como hechos, cifras o elementos de información que, en su forma más básica, carecen de significado. Se recolectan de diversas fuentes, como encuestas, transacciones comerciales y dispositivos de medición. Su presentación puede variar, abarcando desde texto y números hasta imágenes y audio. La clave de su utilidad radica en la capacidad de transformar estos datos en información valiosa a través de un análisis adecuado, lo que permite a las organizaciones tomar decisiones informadas y estratégicas (Smith, 2022).

El contexto juega un papel fundamental en la interpretación de los datos. Por ejemplo, un conjunto de números aislados, como "50, 75, 100", no tiene un significado relevante hasta que se relaciona con un contexto específico. Si estos números representan las ventas de un producto en tres meses diferentes, su análisis puede revelar tendencias de compra y comportamientos del consumidor. Por lo tanto, la contextualización es esencial para extraer el valor significativo de los datos, lo que permite a las empresas anticiparse a las necesidades del mercado y adaptarse a ellas (Johnson, 2023).

Tipos de Atributos

Los atributos son características o propiedades que se utilizan para describir objetos, sujetos o fenómenos en la investigación. Es fundamental comprender los tipos de atributos para el análisis de datos, ya que influyen en las metodologías de recolección y en la interpretación de los resultados. Generalmente, los atributos se dividen en dos categorías principales: cualitativos y cuantitativos.

Atributos Cualitativos

Los atributos cualitativos son aquellos que describen características no numéricas y se centran en la calidad de los elementos observados. Estos atributos se pueden dividir en dos subcategorías: nominales y ordinales.

Atributos Nominales

Estos atributos representan categorías sin un orden específico. Son simplemente etiquetas que identifican diferentes grupos. Por ejemplo, el género de los participantes en un estudio (masculino, femenino, no binario) es un atributo nominal. Otro ejemplo es el tipo de fruta (manzana, naranja, plátano), donde cada categoría es igual en términos de jerarquía y no se puede decir que una fruta es "mejor" que otra solo por su nombre.

Atributos Ordinales

A diferencia de los atributos nominales, los ordinales tienen un orden inherente. Esto significa que las categorías pueden clasificarse de acuerdo con algún criterio. Un ejemplo común es la escala de satisfacción del cliente, donde las respuestas pueden incluir categorías como muy insatisfecho, insatisfecho, satisfecho y muy satisfecho. Aunque se puede establecer un orden, las diferencias entre los niveles de satisfacción no son necesariamente uniformes. Es importante señalar que, en este tipo de atributos, aunque el orden es significativo, no se pueden realizar operaciones matemáticas complejas debido a la falta de una distancia definida entre las categorías (Bryman, 2016).

Atributos Cuantitativos

Los atributos cuantitativos se centran en la medición numérica y son fundamentales en investigaciones que requieren análisis estadísticos. Estos se dividen en atributos discretos y continuos.

Atributos Discretos

Son aquellos que solo pueden tomar valores enteros y representan cantidades contables. Un ejemplo típico es el número de estudiantes en una clase. No se puede tener 20.5 estudiantes, por lo que este atributo solo puede tomar valores enteros como 20, 21 o 22. Otro ejemplo es el número de hijos en una familia. En este caso, la variable puede ser 0, 1, 2, etc., pero no se puede tener un número fraccionario de hijos (Field, 2018).

Atributos Continuos

Estos atributos pueden tomar cualquier valor dentro de un rango determinado y se utilizan para medir características que tienen una escala infinita de posibilidades. Un ejemplo es la altura de una persona, que puede medirse en centímetros e incluir decimales (por ejemplo, 175.5 cm). Otro caso es el tiempo que una persona tarda en completar una tarea, que también puede incluir fracciones de segundo. Los atributos continuos son importantes en investigaciones que requieren análisis más sofisticados, ya que permiten la aplicación de diversas técnicas estadísticas, como la regresión lineal (Trochim, 2020).

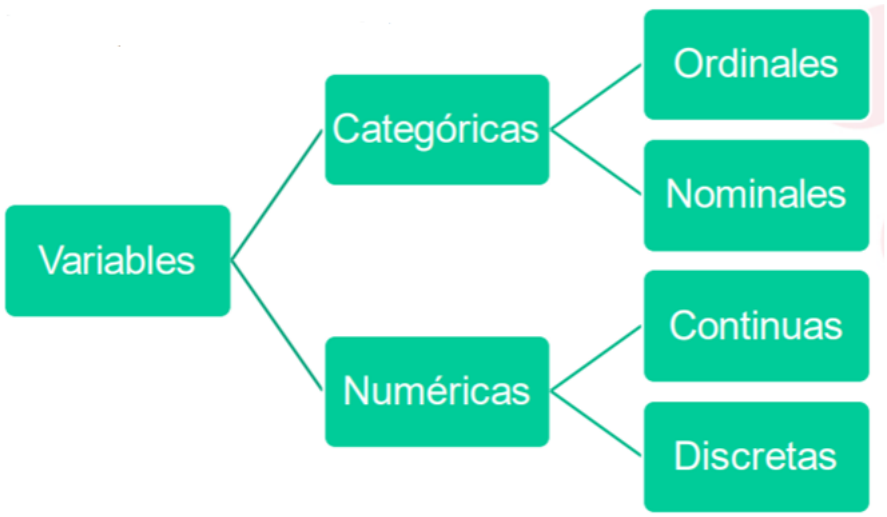

Figura 1. Tipos de Variables. Elaboración Propia

Importancia de los atributos en la investigación

La clasificación de los atributos es esencial para el diseño de estudios y la recolección de datos. Comprender si un atributo es cualitativo o cuantitativo ayuda a los investigadores a seleccionar las herramientas de análisis adecuadas. Por ejemplo, si los datos son cualitativos, es más probable que se utilicen métodos como el análisis de contenido o la codificación temática. En contraste, los datos cuantitativos pueden requerir análisis estadísticos como ANOVA o regresiones.

Además, el manejo adecuado de los tipos de atributos puede influir en la validez y confiabilidad de un estudio. Una elección incorrecta de los atributos puede llevar a interpretaciones erróneas y, por ende, a conclusiones equivocadas (Creswell, 2014). Por ello, es crucial que los investigadores dediquen tiempo a comprender la naturaleza de sus datos y a clasificar adecuadamente los atributos desde el inicio del proceso de investigación.

Tipos de Datos Según su Estructura

En el campo de la investigación y el análisis de datos, la comprensión de los diferentes tipos de datos según su estructura es fundamental para la recolección, el análisis y la interpretación de la información. Los datos pueden clasificarse generalmente en tres categorías: datos estructurados, datos no estructurados y datos semiestructurados. Cada uno de estos tipos tiene características únicas que determinan cómo se pueden manejar y analizar.

Datos Estructurados

Los datos estructurados son aquellos que se organizan en un formato definido y predecible. Esto incluye datos que se pueden almacenar en bases de datos relacionales, donde la información se presenta en filas y columnas. Cada columna representa un atributo y cada fila representa un registro. Este formato permite un fácil acceso y manipulación de los datos mediante consultas SQL (Structured Query Language).

Ejemplos de datos estructurados incluyen:

- **Bases de datos de clientes**: Información como nombres, direcciones, números de teléfono y correos electrónicos se organiza en tablas, facilitando la búsqueda y el análisis.
- **Hojas de cálculo**: Datos financieros, como ingresos y gastos, organizados en columnas y filas que permiten realizar cálculos y análisis sencillos.
- **Sistemas de gestión de inventarios**: Listas de productos con atributos como SKU, cantidad, precio y ubicación, que pueden ser fácilmente consultadas y actualizadas.

La estructura rígida de estos datos permite el uso de herramientas analíticas avanzadas, lo que facilita la generación de informes y el análisis de tendencias. Sin embargo, esta rigidez también significa que cualquier cambio en el formato de los datos puede requerir modificaciones significativas en las bases de datos y en las aplicaciones que las utilizan (Feldman & Sanger, 2007).

id	peso	talla	imc	presión sistólica
1	76	1,8	23,5	130
2	88	1,7	30,4	135
3	69	1,75	22,5	120
4	68	1,65	25,0	129
5	54	1,6	21,1	142
6	70	1,9	19,4	135
7	62	1,5	27,6	133
8	75	1,9	20,8	142
9	76	1,8	23,5	144
10	65	1,7	22,5	136

Tabla 1. Matriz de datos biométricos. Elaboración Propia

Características de los datos estructurados [https://aws.amazon.com/es/what-is/structured-data/]

Datos No Estructurados

A diferencia de los datos estructurados, los datos no estructurados no tienen un formato predefinido, lo que dificulta su análisis y almacenamiento en bases de datos tradicionales. Este tipo de datos incluye información que puede variar en longitud y formato, y a menudo requiere técnicas avanzadas de procesamiento para extraer información valiosa.

Figura 2. Ejemplos de datos no estructurados. Elaboración Propia

Ejemplos de datos no estructurados incluyen:

Los datos no estructurados son un tipo de información que no se organiza en un formato predefinido o rígido, lo que dificulta su almacenamiento y análisis en bases de datos tradicionales. Este tipo de datos representa una gran parte de la información generada en el mundo moderno, y su valor radica en la riqueza de insights que pueden ofrecer cuando se analizan adecuadamente. A continuación, se presentan ejemplos concretos de datos no estructurados en diversos contextos.

1. **Contenido de Redes Sociales**

Las redes sociales son una fuente abundante de datos no estructurados. Publicaciones, comentarios, fotos y videos son ejemplos de información que se genera constantemente y que carece de un formato estructurado. Por ejemplo, un comentario, en una publicación de Facebook sobre un producto, contiene texto libre que refleja las opiniones y emociones del usuario, pero no sigue un formato predefinido.

Un ejemplo: una empresa puede recopilar comentarios de clientes sobre un nuevo producto a través de redes sociales para realizar un análisis de sentimientos. Utilizando técnicas de procesamiento de lenguaje natural (NLP), pueden extraer insights sobre la percepción del producto en el mercado, identificando tendencias positivas o negativas (Manning et al., 2008).

2. **Documentos de Texto**

Los documentos de texto, como informes, artículos, correos electrónicos y archivos PDF, son otro ejemplo común de datos no estructurados. Estos documentos contienen información valiosa, pero su formato libre dificulta el análisis automatizado.

Un ejemplo: en una organización, los correos electrónicos pueden contener información crítica sobre decisiones empresariales, quejas de clientes o propuestas de proyectos. Utilizando técnicas de minería de texto, se pueden analizar los correos electrónicos para identificar patrones de comunicación y áreas de mejora en el servicio al cliente (Feldman & Sanger, 2007).

3. **Multimedia**

Los datos multimedia, que incluyen imágenes, videos y grabaciones de audio, son ejemplos claros de datos no estructurados. La cantidad de información contenida en un solo archivo de imagen o video puede ser vasta, pero no está organizada de manera que se pueda acceder fácilmente a su contenido.

Un ejemplo: en el ámbito de la salud, los hospitales pueden almacenar imágenes médicas, como radiografías y resonancias magnéticas. Aunque estas imágenes son fundamentales para el diagnóstico, su análisis requiere técnicas avanzadas, como el reconocimiento de patrones y el aprendizaje automático, para extraer información útil (Litjens et al., 2017).

4. **Registros de Sensores**

Los dispositivos de IoT (Internet de las Cosas) generan datos a partir de sensores, que a menudo son no estructurados. Estos datos pueden incluir lecturas de temperatura, humedad y otros parámetros ambientales que no se organizan en tablas o columnas.

Un ejemplo: un sistema de monitoreo ambiental puede recopilar datos de sensores distribuidos en una ciudad. Estos datos pueden analizarse para identificar patrones en la calidad del aire o las condiciones climáticas, lo que ayuda a las autoridades a tomar decisiones informadas sobre políticas ambientales (Kumar & Singh, 2018).

Datos Semiestructurados

Los datos semiestructurados representan un punto intermedio entre los datos estructurados y los no estructurados. Este tipo de datos contiene elementos que no tienen un formato

rígido, pero que incluyen etiquetas o marcas que proporcionan cierta organización. Esto permite que los datos sean más flexibles y, al mismo tiempo, más accesibles para el análisis.

Ejemplos de datos semiestructurados incluyen:

- **XML y JSON**: Formatos de intercambio de datos que utilizan etiquetas para organizar la información, permitiendo que los datos se interpreten de manera más flexible que en una base de datos estructurada.
- **Correos electrónicos**: Aunque el contenido puede ser texto libre (no estructurado), los encabezados (remitente, destinatario, fecha) proporcionan una estructura que permite cierta organización.
- **Datos de sensores**: Información recopilada de dispositivos IoT (Internet de las Cosas) que puede incluir una combinación de datos numéricos y texto descriptivo.

Los datos semiestructurados son cada vez más comunes en la era digital, donde la flexibilidad y la interoperabilidad son esenciales para el manejo de grandes volúmenes de información (Baker & Thomas, 2020). Esto permite a las organizaciones integrar datos de diversas fuentes y facilitar el análisis sin la necesidad de una estructura rígida. La siguiente imagen muestra un ejemplo de un archivo JSON.

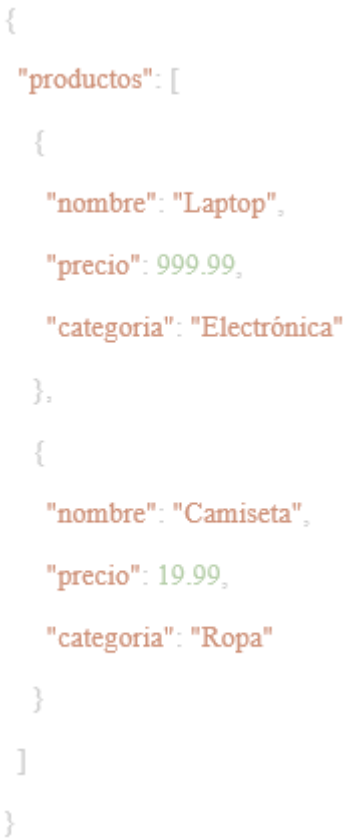

Figura 3. Ejemplo JSON. Elaboración Propia

Aquí, aunque no hay una estructura de tabla estricta, los datos son accesibles y comprensibles gracias a las etiquetas.

Proceso de Descubrimiento del Conocimiento (KDD)

El proceso de KDD (Knowledge Discovery in Databases), o descubrimiento de conocimiento en bases de datos, es una metodología esencial en la era de la información, donde la capacidad de transformar grandes volúmenes de datos en información útil se ha convertido en una ventaja competitiva clave. A medida que las organizaciones generan y recopilan datos en cantidades sin precedentes, el proceso de KDD permite identificar patrones, tendencias y conocimientos ocultos que pueden influir en la toma de decisiones estratégicas.

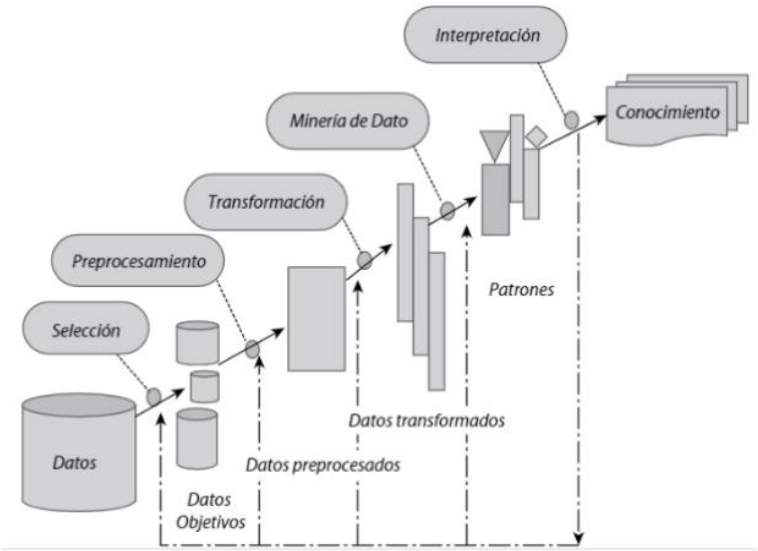

Figura 4. Proceso de Descubrimiento del Conocimiento.

Tomado de: https://saludelectronica.com/descubrimiento-de-conocimiento-en-bases-de-datos-kdd/

A continuación, se describen las etapas del proceso de KDD y se proporcionan ejemplos ilustrativos que destacan su relevancia en diversas industrias.

1. **Selección de Datos**

La primera etapa del proceso de KDD es la selección de datos, que consiste en identificar y recopilar los datos relevantes para el análisis. Esta fase es fundamental porque la calidad y pertinencia de los datos seleccionados impactarán directamente en el éxito del descubrimiento de conocimiento. Durante esta etapa, los analistas deben considerar qué fuentes de datos utilizarán, que pueden incluir bases de datos relacionales, archivos de

texto, registros de sensores y datos provenientes de redes sociales (Fayyad, Piatetsky-Shapiro y Smyth, 1996).

Un ejemplo: en un estudio sobre el comportamiento de compra de los consumidores, una cadena de supermercados podría seleccionar datos de ventas de sus sistemas de gestión, información demográfica de los clientes extraída de encuestas y comentarios de clientes de redes sociales. Al combinar estas fuentes, la empresa puede obtener una visión más completa de su base de clientes y sus preferencias.

2. **Preprocesamiento de Datos**

Una vez que se han seleccionado los datos, la siguiente etapa es el preprocesamiento. Este proceso incluye la limpieza, transformación y preparación de los datos para su análisis. Las actividades en esta fase son cruciales, ya que aseguran que los datos sean de alta calidad y estén listos para el análisis. El preprocesamiento puede involucrar la eliminación de duplicados, el tratamiento de valores perdidos, la normalización de datos y la conversión de variables.

Un ejemplo: en el caso de una cadena de supermercados, es probable que algunos registros de ventas tengan valores perdidos en el campo de la dirección del cliente. Durante el preprocesamiento, el equipo de análisis podría imputar estos valores basándose en datos similares de clientes que sí tienen información completa. Además, podrían normalizar los precios de los productos para que se presenten en una escala uniforme, facilitando el análisis posterior (Han, Kamber y Pei, 2011).

3. **Transformación de Datos**

La transformación de datos es la fase en la que los datos preprocesados se convierten en un formato adecuado para el análisis. Esto puede incluir la reducción de dimensionalidad, la creación de nuevas variables o la agregación de datos. Esta etapa busca resaltar las características más relevantes de los datos, facilitando la identificación de patrones y tendencias.

Un ejemplo: la cadena de supermercados podría calcular el gasto promedio por cliente a partir de las transacciones históricas. Esta nueva variable permitiría al equipo de marketing segmentar a los clientes en función de su nivel de gasto, lo que facilitaría el diseño de campañas de marketing más efectivas dirigidas a diferentes grupos de clientes. Por ejemplo, podrían identificar a los clientes de alto gasto y ofrecerles promociones exclusivas para fomentar su lealtad (Mannila y Rousi, 2005).

4. **Minería de Datos**

La minería de datos es la etapa central del proceso de KDD, donde se aplican técnicas analíticas para extraer patrones y conocimiento de los datos transformados. Existen

diversas técnicas que se pueden aplicar en esta fase, incluidas la clasificación, la agrupación, la regresión y la asociación. La elección de la técnica adecuada depende del tipo de problema que se está abordando y del tipo de datos disponibles.

Un ejemplo: supongamos que la cadena de supermercados desea comprender qué productos tienden a ser comprados juntos. Para ello, podrían aplicar algoritmos de asociación, como el algoritmo Apriori, para identificar patrones en las transacciones de los clientes. Este análisis podría revelar, por ejemplo, que los clientes que compran pan también tienden a comprar mantequilla. Con esta información, la empresa podría crear promociones o exhibiciones de productos que fomenten estas compras combinadas, aumentando así sus ventas (Agrawal, Imielinski y Swami, 1993).

5. **Evaluación del Conocimiento**

Una vez que se han extraído los patrones o el conocimiento, es necesario evaluarlos para determinar su utilidad y relevancia. Esta etapa implica revisar los resultados obtenidos para asegurarse de que sean significativos y aplicables a los objetivos del negocio. La evaluación puede incluir la comparación de los resultados con conjuntos de datos de validación o la aplicación de métricas específicas para medir la calidad de los patrones encontrados.

Un ejemplo: después de realizar el análisis de asociación, la cadena de supermercados evaluaría si los patrones identificados son comercialmente significativos. Por ejemplo, si la asociación entre pan y mantequilla es fuerte y estadísticamente significativa, la empresa podría implementar una estrategia de marketing basada en estos hallazgos. Sin embargo, si los patrones encontrados no son suficientemente relevantes, podría ser necesario revisar los datos o el enfoque analítico utilizado (Kuo, 2016).

6. **Presentación y Visualización de Resultados**

La última etapa del proceso de KDD es la presentación y visualización de los resultados. En esta fase, se comunican los hallazgos a las partes interesadas mediante herramientas de visualización de datos. La presentación de los resultados debe ser clara y concisa, facilitando la comprensión de la información y su relevancia para la toma de decisiones.

Un ejemplo: la cadena de supermercados podría desarrollar un dashboard interactivo que muestre las tendencias de compra de los clientes, las asociaciones de productos y los segmentos de clientes. Esto permitiría a los ejecutivos de la empresa tomar decisiones informadas basadas en datos y ajustar sus estrategias de marketing y ventas en consecuencia. La visualización efectiva de los resultados no solo ayuda en la toma de decisiones, sino que también facilita la comunicación de insights a diferentes niveles dentro de la organización (Few, 2009).

A continuación, se muestra una porción de código para cargar datos estructurados en Python. El dataset usado tiene una estructura similar a esta:

```
vhigh,vhigh,2,2,small,low,unacc
vhigh,vhigh,2,2,small,med,unacc
vhigh,vhigh,2,2,small,high,unacc
vhigh,vhigh,2,2,med,low,unacc
vhigh,vhigh,2,2,med,med,unacc
vhigh,vhigh,2,2,med,high,unacc
vhigh,vhigh,2,2,big,low,unacc
vhigh,vhigh,2,2,big,med,unacc
```

El código es el siguiente:

```python
# Carga de librerías
import pandas as pd
# Importamos archivo csv
df = pd.read_csv('cars.csv', header=None)
# Visualizamos los primeros 5 registros
print(df.head())
# Renombramos las columnas en el mismo dataset
df.rename(
    columns = {
        0:'buying',
        1:'maint',
        2:'doors',
        3:'persons',
        4:'lug_boot',
        5:'safety',
        6:'class'
    }, inplace=True
)
# Visualizamos los datos cargados
print(df.head())
```

La explicación de cada uno de los pasos del proceso de KDD se puede encontrar aquí: https://www.youtube.com/watch?v=sF-iPHx9fHY

Referencias

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record, 22*(2), 207-216. https://doi.org/10.1145/57004.57035

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publications.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54. https://doi.org/10.1609/aimag.v17i3.1230

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Sage Publications.

Few, S. (2009). *Now you see it: Simple visualization techniques for quantitative analysis*. Analytics Press.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Johnson, A. (2023). Contexto y análisis de datos en el mundo empresarial. *Revista de Análisis de Datos, 15*(2), 45-58.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kuo, R. J. (2016). The impact of data mining on customer relationship management. *Journal of Business Research, 69*(12), 5234-5241. https://doi.org/10.1016/j.jbusres.2016.04.008

Mannila, H., & Rousi, M. (2005). The role of KDD in data mining. *ACM SIGKDD Explorations Newsletter, 7*(2), 33-39.

Popper, K. R. (1959). *The logic of scientific discovery*. Basic Books.

Smith, J. (2022). *La importancia de los datos en la toma de decisiones*. Editorial XYZ.

Trochim, W. M. (2020). *The research methods knowledge base* (3rd ed.). Cengage Learning.

PUCE Virtual
EDUCACIÓN FLEXIBLE

La excelencia no se improvisa

síguenos

MDASC - 12 - 241