Lectura fundamental

Proceso de ciencia de datos: del planteamiento del problema a la automatización del análisis

Contenido

Palabras clave: recuperación de datos, limpieza de datos, análisis exploratorio, visualización, modelado, evaluación, automatización.

1. Introducción

La ciencia de datos no comienza con algoritmos ni termina con un gráfico atractivo. Comienza, en realidad, cuando una organización, una institución o un equipo necesita comprender una situación compleja, tomar mejores decisiones o anticipar comportamientos a partir de evidencia. Por eso, su proceso no consiste en memorizar herramientas, sino en comprender una secuencia de decisiones intelectuales y técnicas.

En la práctica, muchos errores en proyectos analíticos no provienen de modelos sofisticados mal programados, sino de preguntas mal formuladas, datos incompletos, supuestos no verificados o visualizaciones poco claras. Greve, B. (2020) advierte "Missing data is the most common data quality problem" (s. p.), es decir, los datos faltantes constituyen el problema de calidad más frecuente en el trabajo con datos. Esa observación tiene una consecuencia pedagógica importante: antes de querer predecir, clasificar o segmentar, es necesario aprender a inspeccionar, limpiar, interpretar y justificar cada transformación.

Desde esa perspectiva, el proceso de ciencia de datos puede entenderse como un ciclo articulado. Se formula el propósito del estudio, se obtienen datos desde fuentes diversas, se verifican su estructura y su calidad, se exploran patrones, se construyen modelos, se evalúan sus resultados y, finalmente, se comunican conclusiones útiles. Cuando ese recorrido madura, también se automatiza parcialmente para hacerlo reproducible y sostenible.

2. Establecimiento del objetivo del estudio

El punto de partida de cualquier proyecto analítico es definir con precisión qué se quiere entender, explicar, predecir o mejorar. Esta etapa suele parecer sencilla, pero en realidad determina la calidad de todas las decisiones posteriores. Un objetivo bien formulado delimita la población de interés, las variables relevantes, el horizonte temporal, la métrica de éxito y el tipo de respuesta esperada. Sin ese marco, la recuperación de datos se vuelve dispersa, la limpieza se hace arbitraria y el modelado termina respondiendo preguntas distintas de las que motivaron el estudio.

En términos formativos, esta fase es central porque obliga a traducir una necesidad amplia del contexto en una pregunta analítica operativa. Greve, B. (2020) señala que modelos de proceso como CRISP-DM comienzan con la comprensión del negocio o del problema, es decir, con la formulación de las preguntas correctas desde la perspectiva de uso de los datos. En la misma línea, Das, S. R. (2025) subraya que el buen trabajo en ciencia de datos requiere contextualizar correctamente el entorno del análisis y construir modelos efectivos con juicio, no solo con cómputo.

Una pregunta como **¿cómo mejorar las ventas?** es demasiado vaga para guiar un estudio. En cambio, un planteamiento como **¿qué variables explican la caída trimestral de recompra en clientes nuevos y qué segmentos presentan mayor riesgo de abandono?** permite identificar mejor las fuentes de datos, el nivel de observación, el horizonte temporal y el tipo de técnicas a emplear. La calidad metodológica crece cuando el objetivo especifica unidad de análisis, verbo analítico y criterio de decisión. Así, el proceso deja de ser una recolección mecánica de datos y se convierte en una investigación orientada por evidencia.

También es necesario distinguir entre objetivos descriptivos, diagnósticos, predictivos y prescriptivos: un estudio descriptivo busca caracterizar qué está ocurriendo; uno diagnóstico, por qué ocurre; uno predictivo, qué puede pasar; y uno prescriptivo, sugiere acciones. Cada uno exige diferentes diseños, variables y criterios de evaluación. Esta diferenciación evita exigir a un conjunto de datos respuestas que no puede ofrecer. De hecho, Das, S. R. (2025) advierte que en grandes volúmenes de datos pueden aparecer correlaciones estadísticamente significativas pero engañosas, lo que obliga a no sobredimensionar las conclusiones.

Tabla 1. Componentes para formular un objetivo analítico sólido.

Componente	Pregunta orientadora	Ejemplo
Problema	¿Qué situación requiere explicación o mejora?	Disminución de retención de usuarios
Unidad de análisis	¿Sobre qué entidad se estudiará?	Cliente, transacción, sesión, producto
Alcance temporal	¿De qué periodo provienen los datos?	Últimos 12 meses
Variable objetivo	¿Qué resultado interesa observar?	Cancelación, compra, tiempo de respuesta
Criterio de éxito	¿Cómo se evaluará la utilidad del análisis?	Reducción del error, claridad del informe, mejora operativa

Fuente: elaboración propia

Así, de acuerdo con la definición clara de los componentes, el objetivo del estudio presentado en el ejemplo podría ser: analizar los datos de **clientes y sus sesiones de uso en los últimos 12 meses para identificar factores asociados a la disminución de la retención de usuarios**, por medio de la observación de **cancelación del servicio**, **compra y tiempo de respuesta**, con el fin de **generar información útil para mejorar las estrategias de retención**.

3. Técnicas de recuperación de datos

Después de establecer el objetivo, el siguiente reto consiste en obtener datos pertinentes, suficientes y trazables. La recuperación de datos no es un acto neutro: cada fuente refleja una forma de producción, almacenamiento y filtrado de la realidad. Por eso, antes de descargar un archivo o consultar un *endpoint*, conviene preguntarse quién generó los datos, con qué propósito, bajo qué reglas y con qué sesgos potenciales. Esa lectura crítica evita asumir que la disponibilidad equivale automáticamente a calidad o pertinencia.

La **recopilación de datos** es la etapa inicial del proceso de análisis en ciencia de datos y consiste en obtener información relevante a partir de diversas fuentes para su posterior procesamiento y análisis. Esta fase implica identificar dónde se encuentran los datos, cómo pueden accederse y en qué formato están disponibles. Una adecuada recopilación de datos es fundamental, ya que la calidad, disponibilidad y organización de la información influyen directamente en los resultados del análisis y en la confiabilidad de las conclusiones obtenidas.

Recuperar datos también implica documentar el proceso. Qué consulta se hizo, en qué fecha, con qué parámetros, sobre qué versión de la fuente y con qué filtros. Esta documentación no es un formalismo secundario: afecta la reproducibilidad, la auditoría y la posibilidad de repetir o actualizar el análisis. Greve, B. (2020) insiste en que el trabajo analítico iterativo debe poder reconstruirse desde los datos originales hasta el resultado final, especialmente cuando hay implicaciones sensibles o decisiones relevantes asociadas al análisis.

3.1. Fuentes de recopilación de datos

En el proceso de recopilación de datos, una de las primeras tareas consiste en identificar de dónde proviene la información que se utilizará para el análisis. En la actualidad, los datos pueden generarse y almacenarse en una gran variedad de fuentes digitales, lo que amplía las posibilidades para estudiar fenómenos sociales, científicos o empresariales. Comprender estas fuentes es importante porque cada una produce datos con características distintas y requiere métodos específicos para acceder a ellos.

Una fuente común de información son los **sistemas transaccionales**, utilizados por organizaciones para registrar actividades cotidianas como compras, pagos, reservas o registros de usuarios. Cada una de estas acciones genera datos que suelen almacenarse en bases de datos estructuradas, permitiendo posteriormente analizarlos para identificar patrones de comportamiento o evaluar el desempeño de procesos.

Otra fuente importante son los sensores y dispositivos conectados, que forman parte de lo que se conoce como **Internet de las Cosas (IoT).** Estos dispositivos recopilan información de manera automática sobre variables físicas o ambientales, como temperatura, ubicación, velocidad o niveles de contaminación. Los datos generados por sensores suelen producirse de forma continua y en grandes volúmenes, lo que permite monitorear fenómenos en tiempo real.

También existen fuentes de datos accesibles a través de **interfaces de programación de aplicaciones (APIs)**. Las APIs permiten que diferentes sistemas informáticos compartan información de manera estructurada a través de *internet*. Muchas plataformas digitales, como redes sociales, servicios meteorológicos o sistemas de mapas, ofrecen APIs que permiten a investigadores y analistas recuperar datos de forma automatizada.

Finalmente, otra fuente frecuente de información son los **sitios *web* y plataformas digitales**, desde los cuales es posible recopilar datos mediante técnicas como el *web scraping*, que consiste en extraer información contenida en páginas *web* para su posterior análisis. Este tipo de recopilación suele aplicarse cuando los datos no están disponibles directamente mediante APIs o bases de datos.

En conjunto, estas fuentes —sistemas transaccionales, sensores, APIs y plataformas *web*— muestran la diversidad de orígenes desde los cuales pueden obtenerse datos en proyectos de ciencia de datos. Identificar adecuadamente la fuente de información constituye un paso clave para seleccionar las herramientas y técnicas más adecuadas para su recopilación y análisis.

Cómo mejorar...

Una referencia útil para ampliar este tema es el documento de la United Nations Statistics Division (2023), que describe las principales fuentes de generación y recopilación de datos en el entorno digital, como transacciones, sensores y plataformas digitales. Este recurso ofrece una visión clara sobre el origen de los datos y su importancia en procesos de análisis y toma de decisiones. Lo puede encontrar en el siguiente enlace: https://unstats.un.org/capacity-development/handbook/chapters/Ch8_Handbook_20230417.pdf?utm_source=chatgpt.com

3.2. Recuperación de datos estructurados

Las técnicas y fuentes de recopilación de datos pueden variar según el tipo de datos con el que se trabaje. En el caso de los datos estructurados, que suelen estar organizados en tablas con filas y columnas, es común obtenerlos desde bases de datos relacionales, sistemas transaccionales o archivos como hojas de cálculo y archivos CSV.

Las bases de datos relacionales, que constituyen uno de los sistemas más comunes para almacenar datos estructurados. En este tipo de bases de datos, la información se organiza en tablas, donde cada fila representa un registro y cada columna corresponde a un atributo o característica de la entidad que se está describiendo. Esta forma de organización facilita el almacenamiento, la consulta y el análisis de grandes volúmenes de información.

Por ejemplo, una empresa puede almacenar la información de sus clientes en una tabla que contenga atributos como el nombre, la ciudad o la edad. En este caso, cada fila representaría un cliente y cada columna describiría alguna característica de ese cliente.

Tabla 2. Ejemplo base de datos relacional

Estudiante	Programa	Semestre
Laura Gómez	Enfermería	3
Juan Pérez	Ingeniería	5
Camila Ruiz	Psicología	2
...	...	...

Fuente: elaboración propia

Para gestionar este tipo de información se utilizan programas conocidos como Sistemas de Gestión de Bases de Datos (SGBD). Estos sistemas permiten almacenar, organizar y administrar grandes conjuntos de datos, además de facilitar el acceso a la información cuando es necesario realizar consultas o análisis. Entre los sistemas de gestión de bases de datos más utilizados se encuentran MySQL, PostgreSQL, SQL Server y Oracle.

Los SGBD permiten realizar diferentes operaciones sobre los datos, como crear nuevos registros, consultar información existente, actualizar valores o eliminar datos. Estas operaciones suelen conocerse como operaciones CRUD (*Create*, *Read*, *Update* y *Delete*), que representan las acciones básicas que se pueden realizar sobre una base de datos.

Para interactuar con estas bases de datos se utiliza comúnmente SQL (Structured Query Language), un lenguaje diseñado para consultar y manipular información almacenada en tablas. Mediante SQL es posible seleccionar datos específicos, filtrar registros o agrupar información según determinados criterios.

Por ejemplo, la siguiente consulta permite obtener los nombres y programas de los estudiantes que se encuentran en tercer semestre:

```sql
SELECT Estudiante, Programa
FROM Estudiantes
WHERE Semestre = 3;
```

Este tipo de consultas constituye una de las formas más comunes de recopilar datos estructurados para su posterior análisis. En muchos proyectos de ciencia de datos, el proceso comienza extrayendo información desde bases de datos organizacionales mediante consultas SQL o mediante herramientas de análisis que permiten conectarse directamente a estos sistemas.

Posteriormente, los datos recuperados pueden ser procesados utilizando herramientas de análisis como Python, R, Excel o plataformas de visualización, lo que permite explorar la información, identificar patrones y generar conocimiento útil para la toma de decisiones.

3.3. Recuperación de datos semiestructurados

Los datos semiestructurados ocupan una posición intermedia entre los datos estructurados y los no estructurados. A diferencia de los datos organizados en tablas con filas y columnas, estos datos no siguen un esquema rígido; sin embargo, mantienen cierto nivel de organización que permite identificar relaciones entre sus elementos. Esta organización suele representarse mediante estructuras jerárquicas, lo que facilita su interpretación por parte de sistemas informáticos.

Entre los formatos más comunes de datos semiestructurados se encuentran JSON (JavaScript Object Notation) y XML (Extensible Markup Language). Estos formatos se utilizan ampliamente para intercambiar información entre aplicaciones, ya que permiten representar datos de forma flexible y legible tanto para humanos como para máquinas.

Una de las formas más comunes de recuperar este tipo de datos es mediante interfaces de programación de aplicaciones (APIs). Las APIs permiten que diferentes sistemas informáticos intercambien información a través de *internet* mediante solicitudes automáticas. Cuando una aplicación realiza una solicitud a una API, el servidor responde enviando datos, generalmente en formato JSON, que posteriormente pueden procesarse con herramientas de análisis. De acuerdo con Turrell (2024), las APIs constituyen una de las formas más eficientes de acceder a datos en línea, ya que permiten a los programas recuperar información directamente desde los servicios que la generan.

Por ejemplo, una API podría devolver información con una estructura como la siguiente:

```json
{
  "estudiante": "Laura Gómez",
  "edad": 21,
  "programa": "Enfermería",
  "cursos": [
    {"nombre": "Estadística", "creditos": 3},
    {"nombre": "Biología", "creditos": 4}
  ]
}
```

En este ejemplo se observa que los datos no están organizados en una tabla, sino en una **estructura jerárquica** donde algunos elementos contienen otros datos en su interior. Este tipo de organización es característico de los datos semiestructurados.

En Python, es posible recuperar y procesar este tipo de información utilizando librerías que permiten conectarse a servicios *web* y trabajar con datos en formato JSON. El siguiente ejemplo muestra de manera simplificada cómo obtener datos desde una API y convertirlos en un formato que pueda analizarse posteriormente.

```python
import requests
import pandas as pd

# URL de una API que devuelve datos en formato JSON
url = "https://api.example.com/datos"

# Realizar la solicitud a la API
respuesta = requests.get(url)

# Convertir la respuesta en formato JSON
datos = respuesta.json()

# Convertir los datos en una tabla para su análisis
df = pd.json_normalize(datos)

print(df.head())
```

En este ejemplo, el programa realiza una **solicitud a un servicio *web***, recibe una respuesta en **formato JSON** y posteriormente transforma esa información en una **estructura tabular que puede analizarse con herramientas de ciencia de datos**.

En la actualidad, muchos servicios digitales, plataformas en línea y repositorios de datos abiertos utilizan **APIs y formatos semiestructurados** para compartir información. Por esta razón, comprender cómo recuperar y procesar estos datos constituye una habilidad fundamental en proyectos de ciencia de datos.

3.4. Recuperación de datos no estructurados

Los datos no estructurados incluyen información que no posee una estructura definida, como textos, imágenes, audios o videos. Estos datos suelen recopilarse desde fuentes como redes sociales, páginas *web*, documentos digitales o repositorios multimedia. Para su obtención se utilizan técnicas como ***web scraping*, descarga de archivos o acceso a plataformas digitales** que permiten recuperar grandes volúmenes de información para su posterior procesamiento mediante herramientas especializadas.

Una de las técnicas más utilizadas para recuperar este tipo de información es el ***web scraping***, que consiste en **extraer automáticamente datos disponibles en páginas *web*** mediante programas que analizan su contenido. Este proceso permite recolectar información publicada en sitios *web* sin necesidad de copiarla manualmente, facilitando la creación de conjuntos de datos a partir de contenidos como noticias, reseñas, catálogos de productos o publicaciones en blogs (OpenStax, 2023).

El proceso de *web scraping* suele implicar varias etapas. En primer lugar, un programa accede a una página web y descarga su contenido, generalmente en formato HTML. Posteriormente, el sistema analiza la estructura del documento para identificar los elementos donde se encuentra la información relevante. Para ello se utilizan técnicas como análisis de HTML, expresiones regulares o navegación por etiquetas, que permiten localizar textos, enlaces o datos específicos dentro de la página (OpenStax, 2023).

Además del *scraping* directo de páginas *web*, otra fuente importante de datos no estructurados son las **redes sociales**. En estos entornos se generan grandes volúmenes de información en forma de comentarios, publicaciones, imágenes y videos que pueden ser analizados para estudiar tendencias, opiniones o comportamientos de los usuarios. La **recopilación de estos datos suele realizarse mediante APIs o herramientas de monitoreo**, que permiten acceder a contenidos publicados en las plataformas y analizarlos posteriormente.

Es importante considerar que la recuperación de datos no estructurados suele ser solo el primer paso del proceso. Una vez recopilada la información, **normalmente se requiere aplicar técnicas adicionales de procesamiento de lenguaje natural, visión por computador o análisis multimedia** para transformar estos datos en formatos que puedan ser analizados. Por ejemplo, los textos pueden convertirse en listas de palabras o temas relevantes, mientras que las imágenes pueden analizarse para identificar objetos o patrones visuales.

El siguiente ejemplo muestra cómo utilizar Python para recuperar el contenido de una página *web* y extraer el texto principal de un artículo.

```python
import requests
from bs4 import BeautifulSoup

# Solicita el contenido de una página web
url = "https://example.com"
respuesta = requests.get(url)

# Analiza el contenido HTML de la página
soup = BeautifulSoup(respuesta.text, "html.parser")

# Extrae todos los párrafos del documento
parrafos = soup.find_all("p")

# Imprime los primeros párrafos encontrados
for p in parrafos[:3]:
    print(p.text)
```

En este ejemplo, el programa descarga el contenido de una página *web* y utiliza la biblioteca **BeautifulSoup** para identificar los párrafos del documento HTML. Posteriormente, el texto extraído puede almacenarse y analizarse para identificar palabras frecuentes, temas o patrones dentro del contenido.

Para reflexionar...

Para profundizar en estos métodos de recuperación de datos, se recomienda a los estudiantes consultar el material de ***Das,S. R. (2025)***, donde se presentan distintas formas de acceso a datos como ***web scraping***, ***lectura de tablas en páginas web y consumo de APIs***. Este recurso permite comprender que, en muchos casos, los datos no se encuentran disponibles en archivos listos para el análisis, sino que deben recuperarse desde páginas *web* o servicios digitales.

Tips

El ***web* scraping** se utiliza cuando la información está visible en un sitio web, pero no existe una descarga estructurada, mientras que las **APIs** permiten acceder a los datos mediante mecanismos formales de consulta.

4. Estrategias de limpieza y preparación de datos

En los proyectos de ciencia de datos, la información recopilada rara vez se encuentra lista para ser analizada de forma inmediata. Con frecuencia los conjuntos de datos contienen valores faltantes, registros duplicados, formatos inconsistentes o errores de captura. Por esta razón, antes de realizar cualquier análisis es necesario aplicar procesos de **limpieza y preparación de datos**, cuyo objetivo es mejorar la calidad de la información y asegurar que los datos sean adecuados para su análisis posterior.

Diversos autores señalan que una parte significativa del trabajo en ciencia de datos se dedica a estas tareas. Por ejemplo, McKinney (2022) explica en *Python for Data Analysis* que la preparación de datos incluye procesos como la **corrección de errores, la reorganización de variables, la transformación de formatos y la integración de múltiples fuentes de información**. Estas actividades forman parte de lo que comúnmente se denomina ***data wrangling*** o preparación de datos.

De manera general, las estrategias de limpieza y preparación buscan garantizar que los datos sean **consistentes, completos y adecuados para el análisis**. Esto puede implicar eliminar registros duplicados, manejar valores faltantes, estandarizar formatos de fechas o categorías, corregir errores tipográficos o reorganizar las variables de un conjunto de datos. Una vez realizadas estas tareas, los datos pueden transformarse para facilitar su análisis mediante técnicas estadísticas, modelos de aprendizaje automático o procesos de visualización.

Aunque muchos de estos procedimientos son comunes a distintos tipos de información, las estrategias específicas pueden variar dependiendo del tipo de datos con el que se trabaje. En particular, la preparación de **datos estructurados** suele centrarse en la corrección de valores, la estandarización de variables y la reorganización de tablas, mientras que el tratamiento de **datos no estructurados** puede requerir técnicas adicionales para convertir textos, imágenes o audios en representaciones que puedan ser analizadas por herramientas computacionales.

Para reflexionar...

La preparación de datos no es una etapa menor ni una labor puramente mecánica. Greve, B. (2020) resume su importancia señalando que esta fase suele consumir gran parte del tiempo del trabajo analítico y que los problemas de calidad, si no se detectan a tiempo, terminan rompiendo pasos posteriores del flujo. Esa idea coincide con una experiencia común en proyectos reales: el modelo puede ser técnicamente correcto y aun así fracasar porque las entradas fueron mal interpretadas, incompletas o inconsistentes.

Tips

En ciencia de datos normalmente se habla de ***data cleaning + data preparation + data transformation***, dentro de lo que muchos libros llaman ***data wrangling.***

4.1. Limpieza de datos estructurados

Este tipo de información suele provenir de bases de datos relacionales, sistemas transaccionales o archivos como hojas de cálculo y archivos CSV. Aunque su estructura facilita el almacenamiento y la consulta, estos datos no están necesariamente libres de errores o inconsistencias, por lo que es necesario aplicar procesos de limpieza antes de realizar cualquier análisis.

La limpieza de datos consiste en **identificar y corregir problemas que pueden afectar la calidad de la información**. Según explica Wes McKinney en Python for Data Analysis, esta etapa incluye actividades como la detección de valores faltantes, la eliminación de registros duplicados, la corrección de formatos inconsistentes y la identificación de valores atípicos. Estas tareas permiten asegurar que los datos representen adecuadamente la realidad que se desea analizar.

Uno de los problemas más frecuentes en los datos estructurados es la presencia de **valores faltantes**. Estos pueden aparecer cuando un registro no fue completado correctamente o cuando la información no estuvo disponible en el momento de la captura. Dependiendo del contexto, estos valores pueden eliminarse, completarse mediante estimaciones o mantenerse explícitamente como datos faltantes para evitar introducir sesgos en el análisis.

Los valores faltantes exigen decisiones explícitas. No siempre deben eliminarse. En ocasiones conviene imputarlos, en otras marcarlos como categoría especial y en otras excluir registros, pero siempre justificando la elección. La clave es entender el mecanismo de ausencia y su impacto analítico. Si la falta de información no es aleatoria, un tratamiento ingenuo puede sesgar seriamente los resultados. Por eso, antes de actuar sobre los vacíos, se debe medir su magnitud, detectar patrones y relacionarlos con el contexto del problema.

Otro aspecto importante es la detección de **registros duplicados**, que ocurre cuando una misma observación se registra más de una vez dentro del conjunto de datos. La duplicación puede generar distorsiones en cálculos estadísticos, conteos o modelos predictivos, por lo que es común aplicar procedimientos que permitan identificar y eliminar estos registros.

También es frecuente encontrar inconsistencias en los **formatos de los datos**. Por ejemplo, una misma variable puede registrarse con diferentes formatos de fecha, unidades de medida o categorías escritas de distintas maneras. La limpieza de datos busca estandarizar estos formatos para asegurar que las variables puedan analizarse de manera consistente.

Finalmente, algunos conjuntos de datos pueden contener **valores atípicos o extremos**, conocidos como ***outliers***. Estos valores pueden deberse a errores de captura o a observaciones reales, pero poco frecuentes. Su identificación permite decidir si deben corregirse, eliminarse o mantenerse dentro del análisis dependiendo del contexto del estudio. Greve, B. (2020) propone definir rangos plausibles y revisar valores obviamente incorrectos; además, sugiere que técnicas como el método de recorte (*capping*) pueden reducir el efecto de extremos cuando el objetivo es modelar y no preservar con exactitud cada valor individual. Esa orientación es útil porque evita dos errores frecuentes: borrar indiscriminadamente valores altos o conservar datos absurdos solo por miedo a intervenirlos.

Además, para tratar valores atípicos Greve, B. (2020) recomienda hacer supuestos sobre los datos y verificarlos y, cuando aparezcan combinaciones extrañas, contrastarlas con personas conocedoras del proceso o del negocio. Esta sugerencia recuerda que la calidad de los datos no se resuelve solo con estadística, también requiere comprensión del dominio. Un valor raro puede ser un error o un caso legítimo, y distinguirlos depende de conocimiento contextual.

En conjunto, estas estrategias contribuyen a mejorar la calidad del conjunto de datos y constituyen un paso fundamental dentro del proceso de preparación de datos. Una vez completada esta etapa, los datos pueden transformarse y reorganizarse para facilitar su análisis estadístico o su utilización en modelos de aprendizaje automático.

Tabla 3. Problemas frecuentes de preparación de datos y respuesta sugerida

Problema	Ejemplo	Riesgo analítico	Respuesta sugerida
Valores faltantes	Campos vacíos en ingreso mensual	Sesgo y pérdida de información	Medir proporción, detectar patrón, imputar o excluir con justificación
Duplicados	Misma transacción repetida	Sobreestimación de volúmenes	Definir llave y depurar duplicados
Formatos inconsistentes	Fechas en varios formatos	Errores de unión o de series temporales	Estandarizar tipos y formatos
Valores inesperados	Edad = 250	Distorsión descriptiva y predictiva	Verificar reglas de negocio y plausibilidad
Valores atípicos (Outliers)	Monto extremadamente alto	Sensibilidad de métricas y modelos	Auditar, transformar o limitar según objetivo

Fuente: elaboración propia

El siguiente ejemplo muestra cómo identificar valores faltantes y eliminar registros duplicados utilizando la biblioteca **pandas**.

```python
import pandas as pd

# Cargar un archivo de datos
data = pd.read_csv("ventas.csv")

# Identificar valores faltantes en cada columna
print(data.isnull().sum())

# Eliminar registros duplicados
data = data.drop_duplicates()

# Mostrar las primeras filas del conjunto de datos limpio
print(data.head())
```

En este ejemplo se realiza una revisión básica del conjunto de datos para identificar valores faltantes y eliminar registros duplicados. Estas operaciones forman parte de las tareas iniciales de limpieza que suelen aplicarse antes de realizar análisis más avanzados.

Aunque los lenguajes de programación como Python o R son ampliamente utilizados para la limpieza y preparación de datos, también existen herramientas que permiten realizar estas tareas de manera visual. Por ejemplo, aplicaciones como **OpenRefine** facilitan la exploración y corrección de inconsistencias en conjuntos de datos, permitiendo identificar valores duplicados, estandarizar categorías o transformar formatos de manera interactiva. De igual forma, programas ampliamente utilizados como Microsoft Excel incorporan herramientas para la preparación de datos, entre ellas **Power Query**, que permite importar información desde múltiples fuentes y aplicar procesos de limpieza y transformación de manera automatizada.

Tabla 4. Herramientas utilizadas en la preparación de datos

Herramienta	Tipo de uso	Ventaja principal
Excel / Power Query	Visual	Fácil para principiantes
OpenRefine	Limpieza avanzada	Detectar inconsistencias
Python / R	Limpieza avanzada con programación	Automatización y escalabilidad

Fuente: elaboración propia

4.2. Transformación de datos

Una vez que los datos han sido limpiados y se han corregido posibles errores o inconsistencias, es necesario realizar procesos de transformación de datos. Esta etapa consiste en modificar, reorganizar o adaptar la información para que pueda ser utilizada de manera adecuada en análisis estadísticos, visualizaciones o modelos de aprendizaje automático.

La transformación de datos puede implicar diversas operaciones, como cambiar el formato de las variables, agrupar registros, crear nuevas variables a partir de las existentes o reorganizar la estructura del conjunto de datos. Estas transformaciones permiten que los datos se ajusten mejor a los objetivos del análisis y facilitan su procesamiento mediante herramientas computacionales.

Según explica Wes McKinney en Python for Data Analysis, muchas tareas de análisis requieren reorganizar los datos antes de aplicar métodos estadísticos o modelos predictivos. Por ejemplo, puede ser necesario convertir categorías en variables numéricas, combinar información de diferentes tablas o resumir grandes conjuntos de datos mediante agregaciones.

Entre las transformaciones más comunes en el análisis de datos se encuentran:

- Conversión de tipos de datos, por ejemplo, transformar texto en números o convertir fechas a un formato estándar.
- Creación de nuevas variables, derivadas de cálculos realizados sobre variables existentes.
- Agrupación o agregación de datos, que permite resumir información mediante promedios, conteos u otros indicadores.
- Reorganización de la estructura de los datos, por ejemplo, pasar de formatos de datos "largos" a "anchos" o viceversa.
- Codificación de variables categóricas, necesaria cuando se utilizan algoritmos que requieren variables numéricas.

Estas transformaciones forman parte del proceso general de preparación de datos, cuyo objetivo es construir un conjunto de datos adecuado para responder a la pregunta de análisis planteada al inicio del estudio.

El siguiente ejemplo muestra cómo crear una nueva variable y agrupar datos utilizando la biblioteca pandas.

```python
import pandas as pd

# Cargar datos
data = pd.read_csv("ventas.csv")

# Crear una nueva variable calculando el ingreso total
data["ingreso_total"] = data["precio"] * data["cantidad"]

# Agrupar por categoría de producto y calcular el total de ventas
ventas_categoria = data.groupby("categoria")["ingreso_total"].sum()

print(ventas_categoria)
```

En este ejemplo se crea una nueva variable que representa el ingreso total de cada registro y posteriormente se agrupan los datos por categoría para obtener un resumen del total de ventas.

4.3. Preparación de datos no estructurados

Los datos no estructurados, como textos, imágenes, audios o videos, requieren procesos de preparación diferentes a los utilizados con datos estructurados. En particular, cuando se trabaja con información textual, es necesario aplicar técnicas que permitan transformar el lenguaje natural en una representación que pueda ser procesada por herramientas computacionales.

La preparación de datos textuales forma parte de lo que se conoce como **minería de textos**, un conjunto de métodos orientados a analizar grandes colecciones de documentos para identificar patrones, relaciones o información relevante. En términos generales, la minería de textos busca transformar datos textuales en datos estructurados que puedan analizarse mediante métodos estadísticos o algoritmos de aprendizaje automático (IBM).

De acuerdo con Fernández Áviles, G. y Montero, J. (2024), la preparación adecuada de los textos es un paso fundamental antes de realizar cualquier análisis, ya que los resultados dependen en gran medida de cómo se segmenten y procesen los documentos.

Entre las tareas más comunes en la preparación de datos textuales se encuentran:

- **Limpieza del texto**, que consiste en eliminar caracteres especiales, signos de puntuación o elementos irrelevantes.
- **Tokenización**, proceso mediante el cual el texto se divide en unidades más pequeñas como palabras o frases.
- **Eliminación de palabras vacías (*stopwords*)**, es decir, palabras muy frecuentes que no aportan significado relevante al análisis, como “el”, “la” o “de”.
- **Conteo o frecuencia de palabras**, que permite identificar los términos más frecuentes dentro de un conjunto de documentos.
- **Transformación del texto en representaciones numéricas**, como matrices de términos o vectores, necesarias para aplicar algoritmos de análisis de datos.

Estas transformaciones permiten convertir información escrita en estructuras que pueden analizarse mediante métodos computacionales. Por ejemplo, al representar un conjunto de documentos como una matriz de palabras y frecuencias, es posible identificar temas recurrentes, analizar opiniones o detectar patrones de lenguaje dentro de grandes colecciones de textos.

El siguiente ejemplo muestra cómo limpiar un texto y contar la frecuencia de palabras utilizando Python.

```python
import re
from collections import Counter

texto = "La ciencia de datos permite analizar grandes volúmenes de datos y encontrar pat
rones en la información."

# Convertir el texto a minúsculas
texto = texto.lower()

# Eliminar signos de puntuación
texto = re.sub(r"[^\w\s]", "", texto)

# Dividir el texto en palabras
palabras = texto.split()

# Contar la frecuencia de cada palabra
frecuencia = Counter(palabras)

print(frecuencia)
```

En este ejemplo, el texto se transforma a minúsculas, se eliminan signos de puntuación y posteriormente se divide en palabras para contar cuántas veces aparece cada una. Este tipo de procesamiento constituye uno de los pasos iniciales en el análisis de datos textuales.

Para comprender mejor las principales etapas involucradas en la preparación de datos textuales, la tabla 5 resume algunas de las tareas más comunes que se realizan antes de aplicar técnicas de análisis o minería de textos. Estas etapas permiten transformar el lenguaje natural en una forma que pueda ser procesada por herramientas computacionales, facilitando la identificación de patrones, temas o tendencias dentro de grandes colecciones de documentos.

Tabla 5. Etapas de preparación de datos textuales

Etapa	Descripción
Limpieza	Eliminación de caracteres innecesarios
Tokenización	División del texto en palabras
Eliminación de stopwords	Eliminación de palabras poco informativas
Representación numérica	Conversión del texto en matrices o vectores

Fuente: elaboración propia

En conjunto, las estrategias de limpieza, preparación y transformación de datos constituyen una etapa fundamental dentro del proceso de análisis de datos. A través de estas actividades es posible mejorar la calidad de la información, corregir inconsistencias y adaptar los datos para su posterior análisis. Dependiendo del tipo de datos con el que se trabaje —estructurados o no estructurados—, las técnicas utilizadas pueden variar, pero todas comparten el objetivo de construir conjuntos de datos confiables y adecuados para responder a las preguntas de investigación o apoyar procesos de toma de decisiones. Estas etapas preparan el camino para aplicar métodos analíticos más avanzados, como modelos estadísticos, técnicas de minería de datos o algoritmos de aprendizaje automático.

5. Análisis exploratorio de datos y fundamentos de visualización

Una vez que los datos han sido recopilados, limpiados y preparados, comienza una de las etapas más reveladoras del proceso de análisis: explorar la información. El **análisis exploratorio de datos** (Exploratory Data Analysis, EDA) no busca todavía confirmar una hipótesis definitiva ni optimizar un modelo predictivo, sino f**amiliarizarse con la estructura del conjunto de datos**, identificar patrones, detectar posibles anomalías y comprender cómo se comportan las variables disponibles. En términos formativos, esta etapa permite que el analista **aprenda a observar los datos antes de sacar conclusiones**.

Durante la exploración inicial es común plantear preguntas básicas que ayudan a comprender la información disponible. Por ejemplo: ¿cuántos registros contiene el conjunto de datos?, ¿qué variables son numéricas y cuáles categóricas?, ¿cómo se distribuyen los valores?, ¿existen valores extremos o concentraciones particulares?, ¿qué categorías aparecen con mayor frecuencia?, o ¿qué relaciones preliminares pueden observarse entre diferentes variables?

Las respuestas a estas preguntas permiten construir una primera comprensión del conjunto de datos y orientan decisiones posteriores relacionadas con la transformación de variables, la selección de métodos de análisis o el desarrollo de modelos analíticos.

Para responder a estas preguntas, el análisis exploratorio suele combinar **resúmenes estadísticos y representaciones visuales de los datos**. El cálculo de indicadores como promedios, medianas o frecuencias permite obtener una visión general del comportamiento de las variables, mientras que las representaciones gráficas ayudan a identificar patrones que pueden ser difíciles de detectar únicamente en tablas de datos extensas. No obstante, el desarrollo detallado de herramientas estadísticas como la **estadística descriptiva, las distribuciones de probabilidad o las pruebas de hipótesis** se abordará con mayor profundidad más adelante, en otro material, donde se estudiarán los fundamentos estadísticos necesarios para interpretar los datos de manera rigurosa.

La visualización de datos cumple un papel importante dentro del análisis exploratorio porque permite **hacer visibles las relaciones y patrones presentes en la información**. El libro *Introduction to Data Science* de Rafael Irizarry resalta que una parte fundamental del análisis consiste en mostrar los datos de **manera clara**, ya que algunos resúmenes numéricos pueden ocultar estructuras importantes presentes en la información. Por ejemplo, representar los datos mediante gráficos puede revelar agrupaciones, tendencias o valores atípicos que no serían evidentes al observar únicamente promedios o totales.

Entre las visualizaciones más utilizadas en el análisis exploratorio se encuentran los **diagramas de dispersión**, que permiten examinar relaciones entre variables numéricas, así como gráficos que muestran la distribución de los datos o comparaciones entre grupos. Estas representaciones ayudan a cuestionar suposiciones iniciales y a comprender mejor la complejidad del fenómeno estudiado. Sin embargo, en esta sección la visualización se introduce únicamente como herramienta de apoyo para la exploración de datos.

En síntesis, el análisis exploratorio constituye una etapa clave dentro del proceso de análisis de datos, ya que permite **comprender el comportamiento del conjunto de datos** antes de aplicar métodos analíticos más avanzados. Una exploración cuidadosa ayuda a detectar posibles problemas en la información, identificar patrones iniciales y formular preguntas más precisas que orienten las decisiones analíticas en etapas posteriores.

Aunque el análisis exploratorio no sigue necesariamente un procedimiento rígido, suele desarrollarse a través de una serie de actividades que permiten examinar progresivamente los datos y comprender sus características principales. La tabla 6 resume algunas de las etapas más comunes de este proceso. Estas etapas ofrecen una guía general para organizar la exploración inicial de los datos antes de avanzar hacia análisis estadísticos más formales, los cuales se estudiarán con mayor profundidad más adelante.

Tabla 6. Etapas básicas del análisis exploratorio

Etapa	Propósito	Herramientas, medidas o técnicas utilizadas
Comprender el conjunto de datos	Identificar número de registros y tipos de variables	Conteo de registros, identificación de tipos de variables, revisión de metadatos
Examinar distribuciones	Observar cómo se comportan los valores de cada variable	Frecuencias, histogramas, medidas de tendencia central
Detectar anomalías	Identificar errores o valores atípicos	Diagramas de caja, análisis de valores extremos
Explorar relaciones	Analizar posibles relaciones entre variables	Diagramas de dispersión, tablas cruzadas, correlaciones
Formular hipótesis	Plantear preguntas para análisis posteriores	Comparaciones entre grupos, segmentación de datos, análisis preliminar de tendencias

Fuente: elaboración propia

6. Creación y evaluación de modelos

Una vez que los datos han sido recopilados, limpiados, preparados y explorados, es posible avanzar hacia una etapa central del análisis de datos: la **creación de modelos**. En términos generales, un modelo es una representación simplificada de la realidad que permite **describir relaciones entre variables, explicar fenómenos observados o realizar predicciones sobre nuevos datos** (James, Witten, Hastie, & Tibshirani, 2023).

En el contexto de la ciencia de datos, los modelos se construyen utilizando **métodos estadísticos o algoritmos de aprendizaje automático** que identifican patrones en los datos disponibles, sin embargo, modelar no significa necesariamente aplicar técnicas complejas de inteligencia artificial. En esencia, modelar consiste en construir una representación analítica que permita **describir, clasificar, estimar o predecir un fenómeno a partir de datos**. La elección del modelo depende del objetivo del análisis y del tipo de variable que se desea estudiar (Das, S. 2025).

De forma general, los problemas de modelado en ciencia de datos pueden agruparse en tres enfoques principales: **clasificación, predicción y segmentación** (James, Witten, Hastie, & Tibshirani, 2023).

6.1. Modelos de clasificación

Los modelos de clasificación se utilizan cuando el objetivo es asignar cada observación a una categoría previamente definida. En este tipo de problemas, **la variable de interés es categórica**, es decir, representa clases o grupos posibles.

Este enfoque se aplica en numerosos contextos. Por ejemplo, un sistema puede clasificar correos electrónicos como *spam o no spam*, identificar si una transacción es potencialmente fraudulenta o determinar si un equipo requiere mantenimiento preventivo. En el ámbito de la salud, los modelos de clasificación también pueden emplearse para apoyar el diagnóstico de determinadas enfermedades a partir de variables clínicas.

Para evaluar el desempeño de los modelos de clasificación se utilizan diferentes métricas. Una de las herramientas más utilizadas es la **matriz de confusión**, que compara las categorías predichas por el modelo con las categorías reales observadas en los datos (Das, S. R. 2025). En esta matriz, los aciertos suelen ubicarse en la diagonal principal, mientras que los demás valores representan errores de clasificación. Este análisis permite identificar en qué categorías se producen más errores y comprender mejor el comportamiento del modelo.

A partir de la matriz de confusión también pueden calcularse indicadores como la **exactitud**, que representa el porcentaje de casos correctamente clasificados. Estas métricas permiten evaluar qué tan bien funciona el modelo y si es adecuado para el problema que se desea analizar.

6.2. Modelos de predicción o regresión

Los modelos de predicción, también llamados modelos de regresión, se utilizan cuando la variable que se desea estimar es **numérica o continua**. Su objetivo es calcular un valor aproximado a partir de otras variables disponibles en los datos. Para ello, estos modelos suelen construirse utilizando **datos históricos**, es decir, registros de observaciones pasadas que permiten identificar relaciones entre variables y utilizarlas para estimar valores futuros o desconocidos.

Por ejemplo, un modelo puede estimar el consumo energético de una instalación, el tiempo de entrega de un servicio, la demanda futura de un producto o el costo estimado de una reparación. En todos estos casos, el modelo busca identificar relaciones entre variables que permitan generar estimaciones razonables para nuevos datos.

La evaluación de los modelos de predicción suele basarse en medidas de **error**, que permiten comparar los valores estimados por el modelo con los valores reales observados (James, Witten, Hastie, & Tibshirani, 2023). Entre los indicadores más comunes se encuentran el error medio o el error cuadrático medio, que cuantifican qué tan lejos se encuentran las predicciones del modelo respecto a los valores reales.

Estas métricas ayudan a determinar si el modelo produce estimaciones suficientemente precisas para el contexto en el que se desea aplicar.

6.3. Modelos de segmentación o agrupamiento

En **algunos casos no existe una variable objetivo definida previamente**. En estas situaciones se utilizan técnicas de segmentación o agrupamiento, cuyo propósito es identificar **grupos de observaciones con características similares dentro del conjunto de datos** (James, Witten, Hastie, & Tibshirani, 2023).

Este tipo de análisis permite descubrir **estructuras o patrones ocultos en los datos**. Por ejemplo, puede utilizarse para identificar perfiles de comportamiento de usuarios, segmentar productos según características similares o agrupar regiones geográficas con patrones de consumo semejantes.

A diferencia de los modelos de clasificación o predicción, en la segmentación no existe una respuesta correcta conocida previamente. Por esta razón, **la evaluación de estos modelos suele basarse en la coherencia y utilidad de los grupos identificados**. En algunos casos se utilizan medidas que evalúan la similitud entre los elementos de cada grupo o la separación entre grupos diferentes, aunque también es frecuente complementar la evaluación con interpretación experta del dominio analizado.

6.4. Evaluación general del modelo

Independientemente del tipo de modelo utilizado, una práctica común en ciencia de datos consiste en dividir el conjunto de datos en dos partes: un **conjunto de entrenamiento**, utilizado para construir el modelo, y un **conjunto de prueba**, que permite evaluar su comportamiento con datos que no fueron utilizados durante el proceso de entrenamiento (James, Witten, Hastie, & Tibshirani, 2023).

Este procedimiento ayuda a detectar problemas como el **sobreajuste**, que ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad para generalizar resultados a nuevas observaciones (Das, S. R. 2025).

Además de las métricas cuantitativas, la evaluación de un modelo también puede considerar aspectos como la **interpretabilidad**, la estabilidad frente a nuevos datos y su utilidad para apoyar procesos de toma de decisiones en contextos reales.

A continuación, se presenta una tabla que resume los principales enfoques de modelado utilizados en ciencia de datos, el tipo de variable que se analiza en cada caso y algunos ejemplos de aplicación. Este esquema permite visualizar de manera sintética cómo los modelos pueden orientarse a diferentes objetivos analíticos, como clasificar observaciones, estimar valores numéricos o identificar agrupaciones dentro de los datos. En conjunto, estos enfoques muestran que la creación de modelos consiste en seleccionar herramientas analíticas adecuadas para comprender un fenómeno o apoyar la toma de decisiones a partir de la información disponible.

Tabla 7. Tipos generales de modelos en ciencia de datos

Tipo de modelo	Tipo de variable objetivo	Ejemplo de aplicación
Clasificación	Categórica	Detectar correos spam
Predicción (regresión)	Numérica	Estimar consumo energético
Segmentación	No hay variable objetivo	Identificar perfiles de usuarios

Fuente: elaboración propia

7. Técnicas de presentación de resultados e informes

Una vez finalizadas las etapas de preparación, exploración y modelado de datos, es necesario comunicar los resultados obtenidos de manera clara y comprensible. En los proyectos de ciencia de datos, el análisis no concluye cuando se construye un modelo o se calculan indicadores, sino cuando los hallazgos pueden interpretarse y utilizarse para apoyar la toma de decisiones.

La presentación de resultados consiste en **organizar, sintetizar y explicar la información obtenida durante el análisis**, de forma que pueda ser comprendida por diferentes audiencias, como equipos técnicos, responsables de gestión o tomadores de decisiones. De acuerdo con los autores de R for Data Science, el proceso de análisis de datos incluye no solo la exploración y el modelado, sino también la **comunicación clara de los resultados**, de manera que otras personas puedan interpretar y aprovechar los hallazgos del estudio.

Existen diferentes formas de presentar los resultados de un análisis de datos. Algunas de las más utilizadas incluyen **informes escritos, tablas de resultados, gráficos y visualizaciones,** así como **tableros interactivos o *dashboards*** que permiten explorar la información de manera dinámica. Estas herramientas ayudan a resumir grandes volúmenes de información y facilitan la interpretación de los resultados obtenidos durante el análisis.

En muchos casos, los informes de análisis de datos incluyen elementos como una breve descripción del problema estudiado, las fuentes de datos utilizadas, las principales etapas del análisis, los resultados obtenidos y las conclusiones o recomendaciones derivadas del estudio. Presentar esta información de manera ordenada permite comprender no solo los resultados, sino también el proceso seguido para llegar a ellos.

En la elaboración de informes también es importante distinguir entre **hallazgos, evidencia y recomendaciones**. El hallazgo corresponde a lo que muestran los datos después del análisis; la evidencia está representada por las tablas, métricas o visualizaciones que respaldan ese hallazgo; y la recomendación consiste en la interpretación o acción sugerida a partir de dicha evidencia. Diferenciar estos niveles contribuye a estructurar los informes de forma más clara y evita que los datos, las interpretaciones y las conclusiones se confundan entre sí.

Asimismo, el lenguaje y el nivel de detalle del informe deben adaptarse a la audiencia a la que está dirigido. Mientras que un público técnico puede requerir mayor detalle metodológico, otros actores, como responsables de gestión o tomadores de decisiones, suelen necesitar explicaciones más enfocadas en las implicaciones de los resultados, los posibles riesgos identificados y las alternativas de acción derivadas del análisis.

En relación con la presentación visual de los resultados, las gráficas y tablas utilizadas en los informes deben seguir los mismos principios de claridad y precisión que se aplican durante el análisis exploratorio. Como señala Rafael Irizarry en *Introduction to Data Science*, una buena visualización debe facilitar la interpretación de los datos, evitar distorsiones y permitir comparaciones adecuadas. En la práctica, esto implica utilizar títulos informativos, ejes comprensibles, etiquetas suficientes y un orden significativo de las categorías, evitando elementos visuales innecesarios que puedan distraer del mensaje principal.

Finalmente, un informe de análisis de datos también debe reconocer las **limitaciones del estudio.** Aspectos como la presencia de datos faltantes, supuestos realizados durante la preparación de la información, variables no disponibles o un desempeño moderado del modelo pueden influir en la interpretación de los resultados. Explicitar estas limitaciones contribuye a fortalecer la credibilidad del análisis, ya que permite comprender con mayor claridad el alcance de las conclusiones obtenidas.

En síntesis, la presentación de resultados constituye una etapa clave dentro del proceso de análisis de datos, ya que permite **transformar los resultados técnicos en información comprensible y útil para la toma de decisiones**. En otras lecturas del curso se profundizará en las estrategias y herramientas para la comunicación efectiva de soluciones de ciencia de datos, incluyendo principios de visualización, diseño de informes y construcción de narrativas basadas en datos.

8. Automatización y optimización del flujo de trabajo de ciencia de datos

En proyectos de análisis de datos, es común que ciertos procesos deban repetirse de forma periódica, por ejemplo, cuando los datos se actualizan con frecuencia o cuando el análisis forma parte de una operación continua dentro de una organización. En estos casos, resulta conveniente **automatizar algunas etapas del flujo de trabajo**. Automatizar no significa eliminar el juicio humano en el análisis, sino reducir tareas manuales repetitivas, disminuir la probabilidad de errores y asegurar que los resultados se generen de manera consistente cada vez que se ejecuta el proceso.

La automatización suele ser una consecuencia natural de haber estructurado adecuadamente las etapas del análisis de datos. Cuando los procesos de **recuperación, limpieza, exploración, modelado y comunicación de resultados** están claramente definidos y documentados, es posible transformarlos en procedimientos que se ejecuten de manera sistemática mediante scripts o herramientas de análisis de datos.

Un aspecto fundamental para lograrlo es la **reproducibilidad del análisis**. Según Greve, B. (2020) es recomendable conservar intactos los datos originales, utilizar código en lugar de procedimientos manuales siempre que sea posible y documentar cada paso desde la obtención de los datos hasta la generación de resultados. Estas prácticas permiten mantener trazabilidad sobre el proceso analítico y constituyen la base para construir flujos de trabajo automatizados y confiables.

La optimización del flujo de trabajo también implica organizar el proceso en etapas claras y modulares. Por ejemplo, un flujo automatizado puede incluir la **descarga periódica de datos desde una fuente externa, la verificación de su estructura, la aplicación de rutinas de limpieza, el cálculo de indicadores, la actualización de visualizaciones y la generación de reportes.** Asimismo, pueden incorporarse validaciones automáticas que alerten cuando ocurran cambios en la estructura de los datos, aumenten los valores faltantes o se detecten variaciones importantes en los resultados obtenidos.

En proyectos más avanzados, estas prácticas conducen a la construcción de **pipelines de análisis reproducibles**, en los que cada etapa del proceso se ejecuta de forma ordenada y documentada. Aunque las herramientas utilizadas pueden variar, el objetivo es que cualquier miembro del equipo pueda comprender el flujo de trabajo, reproducir el análisis y obtener resultados comparables cuando se incorporen nuevos datos.

En el contexto formativo del curso, comprender la lógica de la automatización y la optimización del flujo de trabajo permite desarrollar habilidades relacionadas con la **organización del proceso analítico, la reproducibilidad y la responsabilidad en el manejo de datos**, aspectos fundamentales en la práctica profesional de la ciencia de datos.

9. Conclusiones

El proceso de ciencia de datos debe entenderse como una secuencia integrada de decisiones, en la que cada etapa se relaciona con las demás. La formulación clara del objetivo del estudio orienta la recuperación de datos, su preparación, el análisis exploratorio, la construcción de modelos y la forma en que se comunican los resultados.

A lo largo de la unidad se abordaron las principales etapas de este proceso: la obtención de datos desde diversas fuentes, su limpieza y preparación, la exploración y visualización inicial, así como la creación y evaluación de modelos. Estas fases permiten comprender la estructura de la información, identificar patrones relevantes y construir representaciones analíticas útiles para describir o predecir fenómenos.

Finalmente, se destacó la importancia de comunicar los resultados de manera clara y documentada, así como de automatizar y optimizar el flujo de trabajo, de modo que el análisis pueda reproducirse y actualizarse con nuevos datos. En conjunto, estas prácticas permiten desarrollar un enfoque metodológico riguroso para el análisis de datos y fortalecer la capacidad de aplicar la ciencia de datos en contextos reales.

Referencias

Das, S. R. (2025). *Data science: Theories, models, algorithms, and analytics*. GitHub Pages. https://srdas.github.io/MLBook/DataScience.html

Fernández Áviles, G. y Montero, J. (2024). *Fundamentos de ciencia de datos con R*. McGraw Hill. https://cdr-book.github.io/index.html

Greve, B. (2020). *A beginner's guide to clean data*. GitBook. https://b-greve.gitbook.io/clean-data/

IBM. (s.f.). *¿Qué es la minería de texto?* IBM Think. https://www.ibm.com/es-es/think/topics/text-mining?utm_source=chatgpt.com

Irizarry, R. (2025). *Introducción a la ciencia de datos*. GitHub Pages / Harvard Data Science Lab. https://rafalab.dfci.harvard.edu/dslibro/index.html

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023). *An Introduction tp Statistical Learning*. Springer. https://www.statlearning.com/

McKinney, W. (2022). *Python for Data Analysis*. Wes McKinney. https://wesmckinney.com/book/data-cleaning

OpenStax. (2023). *Principios de la ciencia de datos*. OpenStax. https://openstax.org/books/principles-data-science/pages/preface

Turrell, A. (s.f.). *Python for Data Science*. Arthur Turrell. https://aeturrell.github.io/python4DS/welcome.html

Información técnica

Autora: Ana María Palomino Marín

Asesor Pedagógico: Juan Sebastian Patarroyo Ocampo

Diseñadora Gráfica: Ana Viviana Cortés