AXIAL PRESS

Material de muestra para profesorado

Análisis de Datos con Python — Un enfoque funcional

Francisco Javier García Ros

Este documento contiene una muestra del libro "Análisis de Datos con Python", pensada para que el profesorado de Formación Profesional pueda evaluar su adopción en el aula.

Incluye:

- Capítulo 1 completo: "El Detective de Datos" (UT1)
- Prácticas resueltas de la UT1 con notas para el docente

El libro completo cubre las 10 Unidades de Trabajo del módulo 5101 (RD 566/2024) en 544 páginas, con ratio 40% teoría / 60% práctica, datasets reales y notebooks descargables para Google Colab.

info@axialpress.com · https://axialpress.com

Prácticas resueltas — UT1

A continuación se presenta la práctica resuelta de la UT1, con anotaciones para el docente. Esta práctica introduce al alumnado en la configuración de un entorno profesional de análisis de datos con Google Colab.

Práctica UT1.1: Configuración del Entorno Profesional de Análisis

Modalidad: Individual · Duración estimada: 1.5 horas · Recursos: Google Colab + Google Drive

Objetivos: estructurar un proyecto de análisis siguiendo buenas prácticas; gestionar datasets en Google Drive; preparar un cuaderno Colab reproducible; verificar el acceso a datos mediante código.

Celda 1: Documentación del proyecto (Markdown)

Contenido de la celda Markdown:

```
# Proyecto de Análisis: Titanic
* **Autor:** [Nombre del alumno]
* **Fecha:** [Fecha]
* **Objetivo:** Configurar el entorno de trabajo
  y verificar el acceso al dataset del Titanic
  para su posterior análisis.
```

Nota para el docente: verificar que el alumno ha personalizado sus datos y entiende la importancia de esta celda como "portada" del análisis.

Celda 2: Montaje de Google Drive (Código)

Código:

```
# Montar Google Drive para acceso persistente
from google.colab import drive
```

```
drive.mount('/content/drive')
```

Salida esperada:

```
Mounted at /content/drive
```

Nota para el docente: punto más común de error si el alumno no autoriza correctamente la conexión en la ventana emergente de Google.

Celda 3: Verificación de acceso al dataset

Código:

```
# Definir ruta y verificar existencia
path_titanic = '/content/drive/MyDrive/
    analisis_datos_python/proyectos/
    titanic_analysis/datasets/titanic.csv'

# Verificar acceso al fichero
!ls -lh {path_titanic}
```

Salida esperada:

```
-rw------- 1 root root 60K Sep 21 15:30 .../titanic.csv
```

Nota para el docente: la salida confirma que la ruta es correcta y el fichero existe (60KB). Insistir en la importancia de la ruta /content/drive/MyDrive/ como ruta base en Colab.

Celda 4: Reto opcional — Lectura sin Pandas

Este reto conecta con los conocimientos previos de Python del alumnado, demostrando que un CSV no es más que un fichero de texto con estructura.

Código:

```
import csv

print("Verificando el contenido del CSV...")
```

```
try:
    with open(path_titanic, 'r',
              encoding='utf-8') as f:
        reader = csv.reader(f)
        header = next(reader)
        row1 = next(reader)
        row2 = next(reader)
        print("\nCabecera:", header)
        print("Fila 1:", row1)
        print("Fila 2:", row2)
except FileNotFoundError:
    print("ERROR: fichero no encontrado.")
```

Salida esperada:

```
Verificando el contenido del CSV...
Cabecera: ['PassengerId', 'Survived', 'Pclass',
  'Name', 'Sex', 'Age', 'SibSp', ...]
Fila 1: ['1', '0', '3', 'Braund, Mr. Owen Harris',
  'male', '22', '1', '0', 'A/5 21171', '7.25', ...]
```

Nota para el docente: excelente para conectar con conocimientos previos. Refuerza el manejo de ficheros y librerías estándar de Python. Demuestra que un CSV es texto con estructura.

¿Te interesa adoptar este libro en tu módulo?

El libro completo "Análisis de Datos con Python" cubre las 10 Unidades de Trabajo del módulo 5101 (RD 566/2024) con:

- 544 páginas con ratio 40% teoría / 60% práctica
- Datasets reales: Titanic, Ames Housing, California Housing
- Notebooks Jupyter descargables para Google Colab
- Guías didácticas para cada unidad
- Ejercicios con criterios de éxito cuantitativos
- Proyecto final integrador evaluable

Disponible en Amazon (paperback y Kindle) y con repositorio GitHub complementario con notebooks y datasets.

Contacto para adopción institucional

info@axialpress.com

https://axialpress.com

1. El Detective de Datos

1.1. ¿Por qué la ciencia de datos cambia el mundo?

Vivimos en la era de los datos. Cada clic, cada compra, cada mensaje, cada sensor genera información. En los últimos dos años se han creado más datos que en toda la historia de la humanidad. Pero los datos en bruto son solo ruido. Lo que transforma ese ruido en decisiones millonarias, descubrimientos médicos o recomendaciones personalizadas es la **ciencia de datos**.

> **Información**
>
> **El impacto de los datos:**
>
> - **Crecimiento masivo:** En 2025 se superaron los **180 zettabytes** de datos generados globalmente, casi tres veces los 64 ZB de 2020.
> - **Demanda imparable:** El Foro Económico Mundial sitúa al analista de datos como el **puesto de trabajo con mayor crecimiento proyectado** hasta 2027.
> - **Ventaja competitiva:** Las empresas que usan análisis de datos son, de media, un **19% más rentables** que sus competidores.
> - **El "oro" del siglo XXI:** El mercado global de Big Data superó los **270.000 millones de dólares en 2024**.

Este módulo te convertirá en un **detective de datos**: alguien capaz de observar un problema, reunir pistas (los datos), interrogarlas con método científico y presentar conclusiones que cambien el rumbo de una empresa, un equipo deportivo o incluso una relación.

> **Contexto Profesional**
>
> El analista de datos es uno de los perfiles más demandados del mercado. Según LinkedIn, la demanda de profesionales con habilidades en análisis de datos ha crecido más del 650% en la última década. Los salarios de entrada en España oscilan entre 25.000 y 35.000 euros, llegando a superar los 60.000 para perfiles senior.

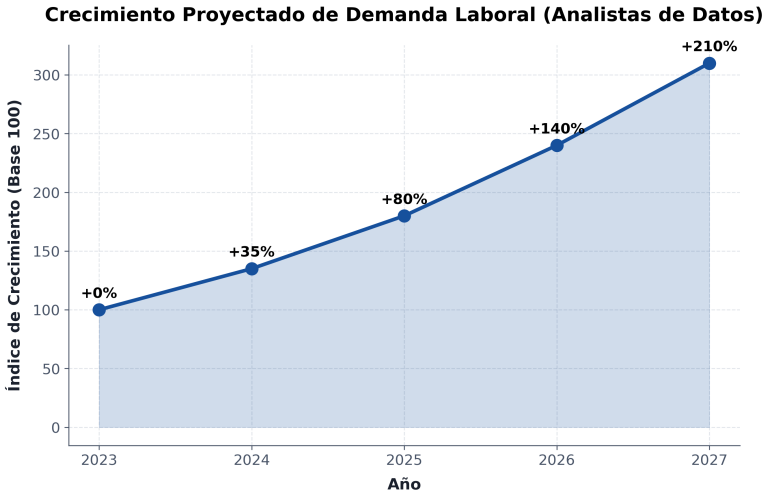

Figura 1.1.: Impacto laboral y crecimiento de la demanda

1.2. La Analogía del Detective

Imagina que trabajas en el departamento de investigación de una gran cadena de supermercados. El director te llama a su despacho con cara de preocupación: "Las ventas de nuestras tiendas en el norte han caído un 15% en los últimos tres meses. Nadie sabe por qué. Necesito respuestas."

Este es tu **caso**. Como detective de datos, tu trabajo no es adivinar ni dar opiniones. Tu método es riguroso:

1. **Recibes el caso:** Un problema de negocio real que necesita respuesta.
2. **Reúnes las pistas:** Datos de ventas, demográficos, meteorológicos, de la competencia...
3. **Interrogas las pistas:** Limpias los datos, buscas patrones, calculas correlaciones.
4. **Formulas hipótesis:** "La caída coincide con la apertura de un competidor discount."
5. **Verificas con evidencia:** Los datos confirman o refutan tu hipótesis.
6. **Presentas conclusiones:** Un informe claro que permite tomar decisiones.

La diferencia entre un detective de datos y alguien que "opina sobre números" es el **método**. Y ese método es exactamente lo que vas a aprender en este módulo.

Pro-Tip

> En el mundo profesional, las decisiones basadas en datos (*data-driven decisions*) tienen mucho más peso que las basadas en intuición. Aprender a comunicar tus hallazgos con evidencia te dará credibilidad en cualquier organización.

1.3. Roles Profesionales: Analista vs. Científico de Datos

Antes de continuar, es importante distinguir dos roles que a menudo se confunden, tal como se muestra en la Tabla 1.1:

Tabla 1.1.: Comparativa entre Analista y Científico de Datos

Aspecto	Analista de Datos	Científico de Datos
Pregunta principal	¿Qué pasó? ¿Por qué?	¿Qué pasará?
Enfoque	Descriptivo y diagnóstico	Predictivo y prescriptivo
Herramientas típicas	SQL, Excel, Python/Pandas, visualización	Python, R, Machine Learning (Aprendizaje Automático), estadística avanzada
Entregables	Dashboards, informes, KPIs	Modelos predictivos, algoritmos
Formación típica	Grado + especialización	Máster/Doctorado

El siguiente gráfico resume visualmente la diferencia entre ambos perfiles en términos de enfoque y herramientas.

En este módulo construiremos unos **cimientos sólidos de Analista de Datos**. Dominarás Python, Pandas, visualización y el flujo de trabajo profesional. Esto te abrirá las puertas tanto a posiciones de analista como a continuar hacia la ciencia de datos avanzada si así lo deseas.

1.4. Qué es la Ciencia de Datos

1.4.1. Una Fusión de Disciplinas

La **Ciencia de Datos** es una disciplina académica y profesional joven que combina múltiples campos preexistentes. Aunque el término *Data Science*

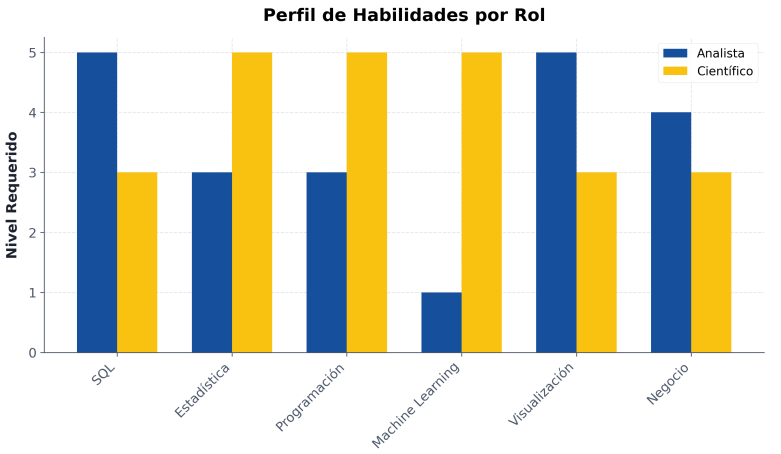

Figura 1.2.: Comparativa de perfiles: Analista vs Científico

aparece ya en 1974 (Peter Naur), fue William S. Cleveland quien en **2001** lo popularizó como disciplina independiente en su influyente artículo. Su adopción masiva explotó a partir de 2010, impulsada por la necesidad de analizar el **Big Data** que empresas y gobiernos estaban acumulando.

> **Información**
>
> **Definición:** La Ciencia de Datos es el campo interdisciplinar que utiliza métodos científicos, procesos, algoritmos y sistemas para extraer conocimiento e insights de datos estructurados y no estructurados.

1.4.2. El Origen del Big Data

A finales de los años 90, el **coste de los discos duros se desplomó**, permitiendo a las organizaciones almacenar enormes cantidades de información. Nació el concepto de **Big Data**: conjuntos de datos tan grandes y complejos que las herramientas tradicionales no podían manejarlos eficientemente.

El flujo histórico fue:

Discos baratos → Big Data → Cloud Computing → Herramientas de Análisis Masivo → Científicos de Datos → Equipos Multidisciplinares → Nuevos Insights

Empresas como Google, Yahoo! y Amazon inventaron una nueva arquitectura: el **Cloud Computing**. Dentro de la nube, una de las invenciones más importantes fue **MapReduce** (codificado en el software **Hadoop**). El paradigma cambió radicalmente:

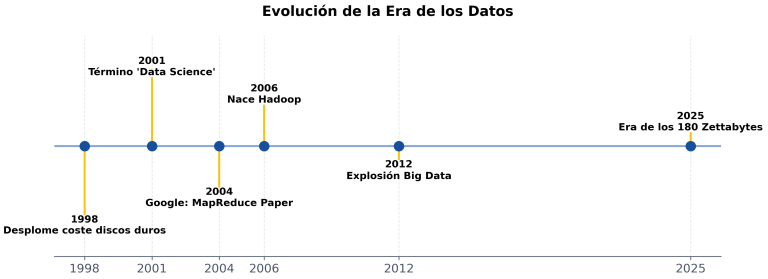

Figura 1.3.: Evolución tecnológica de la era de los datos

- **Antes:** Movíamos los datos hacia el algoritmo (abrir un CSV en un programa).
- **Ahora:** Movemos el algoritmo hacia los datos. Se envían copias del algoritmo para que se ejecuten donde residen los datos masivos.

> **Aviso**
>
> Trabajar con petabytes de información no es igual que abrir un archivo Excel. Requiere herramientas de computación distribuida. En este módulo, trabajaremos con datasets de tamaño medio que caben en la memoria de un ordenador normal, pero los principios de análisis que aprenderás son la base para escalar a Big Data.
> Hadoop y MapReduce son tecnologías que probablemente no usarás directamente en tu trabajo como analista junior. Sin embargo, entender este contexto histórico te ayudará a comprender por qué existen herramientas como Spark, Databricks o los servicios de datos en la nube.

1.5. Las 8 Disciplinas de la Ciencia de Datos

La ciencia de datos es una fusión de al menos **8 disciplinas**. Es muy raro que una sola persona sea experta en todas; de ahí la importancia del trabajo en equipo.

Figura 1.4.: Las 8 disciplinas que componen la ciencia de datos

> **Contexto Profesional**
>
> **El profesional en "T":** En el mercado laboral se valora el perfil "T-Shaped". Esto significa tener un conocimiento amplio y general de las 8 disciplinas (la barra horizontal de la T) pero ser un experto profundo en una o dos de ellas (la barra vertical). Como analista, tu "vertical" suele ser la programación y la estadística.

1.5.1. La tabla maestra de disciplinas

La tabla siguiente describe cada disciplina, qué proceso lleva a cabo y con qué herramientas y ejemplos reales se aplica.

Tabla 1.2.: Las 8 disciplinas de la ciencia de datos

Disciplina	Qué hace (proceso)	Herramientas clave	Ejemplo real
1. Ingeniería de datos	Construye la "fontanería": extrae y limpia datos.	SQL, Airflow, Spark	Unificar datos de 50 tiendas diferentes cada noche.
2. Método científico	Diseña el experimento: hipótesis y validación.	Lógica, A/B Testing	Probar si un descuento del 10% atrae más que "2x1".

Disciplina	Qué hace (proceso)	Herramientas clave	Ejemplo real
3. Matemáticas	La base teórica: álgebra lineal y cálculo.	NumPy, SciPy	Calcular la distancia entre perfiles de usuarios.
4. Estadística	El rigor: ¿son los resultados reales o azar?	R, Statsmodels	¿Es la subida de ventas significativa o casualidad?
5. Programación	El motor: automatiza y escala procesos.	Python, Pandas	Procesar 1 millón de facturas en segundos.
6. Visualización	La comunicación: convierte números en historias.	Seaborn, Tableau, PowerBI	Un mapa de calor que muestre zonas de fraude.
7. Mentalidad hacker	La creatividad: resuelve problemas con ingenio.	Scraping, Regex, APIs	Obtener precios de la competencia automáticamente.
8. Dominio del negocio	El contexto: entiende el sector (finanzas, salud).	Experiencia, KPIs	Saber que en Navidad las ventas suben por estacionalidad.

Pro-Tip

No necesitas ser un genio en las 8 disciplinas para empezar. La mayoría de profesionales destacan en 2 o 3 y tienen conocimientos básicos del resto. Este módulo se centra principalmente en **Programación**, **Estadística**, **Visualización** e **Ingeniería de Datos**.

Recuerda

Conexión con UT2: En la próxima unidad (**NumPy**) aprenderás cómo Python gestiona millones de números de forma instantánea. Sin NumPy, gran parte de la ciencia de datos moderna sería imposible por lentitud.

Pregunta de Reflexión

Reflexión sobre las disciplinas:

1. Si una empresa tiene datos perfectos pero no entiende su sector (Dominio del Negocio), ¿qué disciplinas están fallando y cuál es el impacto en sus resultados?

2. ¿En qué se diferencia la labor de un Ingeniero de Datos de la de un Analista al enfrentarse a una base de datos "sucia" o desordenada?
3. ¿Por qué el **Método Científico** es el "pegamento" que une a las otras 7 disciplinas?
4. Compara la **Visualización** con la **Estadística**: ¿puede existir una buena comunicación de hallazgos sin rigor matemático? Justifica tu respuesta.
5. ¿Cómo puede la **Mentalidad Hacker** ayudar a un analista cuando los datos no están disponibles en un formato estándar (como un CSV) y hay que extraerlos de una web compleja?

Manos a la Obra

1.5.2. Autoevaluación de las 8 Disciplinas

Objetivo: Identificar tus fortalezas y áreas de mejora en las 8 disciplinas.
Instrucciones:

1. Para cada disciplina, puntúa tu nivel actual del 1 al 5 (1=ninguno, 5=avanzado)
2. Identifica las 2 disciplinas donde te sientes más fuerte
3. Identifica las 2 disciplinas donde necesitas más trabajo

Criterio de éxito:

- Tabla completada con 8 puntuaciones
- Reflexión escrita de 3-4 líneas sobre tu perfil

1.6. Casos de Éxito que Cambiaron Industrias

Los siguientes casos demuestran cómo la ciencia de datos ha transformado industrias completas. No son ejemplos teóricos: son historias reales de empresas que usaron datos para obtener ventajas competitivas extraordinarias.

1.6.1. Caso 1: Moneyball - Datos vs. Tradición en el Béisbol

El Problema: El equipo de béisbol Oakland Athletics tenía uno de los presupuestos más bajos de la liga (41 millones de dólares) frente a gigantes como los New York Yankees (125 millones).

La Solución de Datos: En lugar de confiar en la "sabiduría" subjetiva de los ojeadores veteranos, el manager Billy Beane y su analista Paul DePodesta usaron un enfoque estadístico riguroso llamado **sabermetría**.

El Insight Clave: Descubrieron que la métrica **OBP** (On-Base Percentage - porcentaje de veces que un bateador llega a base) era mejor predictor del éxito ofensivo que el tradicional **BA** (Batting Average - promedio de bateo). El mercado infravaloraba a jugadores con alto OBP.

Comparativa Real (Temporada 2002):

Tabla 1.3.: Comparativa de Métricas de Bateo (Temporada 2002)

Equipo	Batting Average (BA)	On-Base Percentage (OBP)	Presupuesto	Resultado
Yankees	.275 (1º AL)	.354 (1º AL)	$125M	Playoffs
A's (Oakland)	.261 (9º AL)	.339 (1º AL)	$41M	Playoffs (103 victorias)

Como puedes ver, los Athletics eran mediocres en bateo tradicional (BA), pero lideraban la liga en capacidad de llegar a base (OBP). Compraban victorias baratas analizando datos que otros ignoraban.

El Resultado: En 2002, con un presupuesto 3 veces menor, los Athletics lograron 20 victorias consecutivas y llegaron a playoffs. La historia se convirtió en libro (Lewis, 2003) y en la película *Moneyball* (Miller, 2011), revolucionando la gestión deportiva mundial.

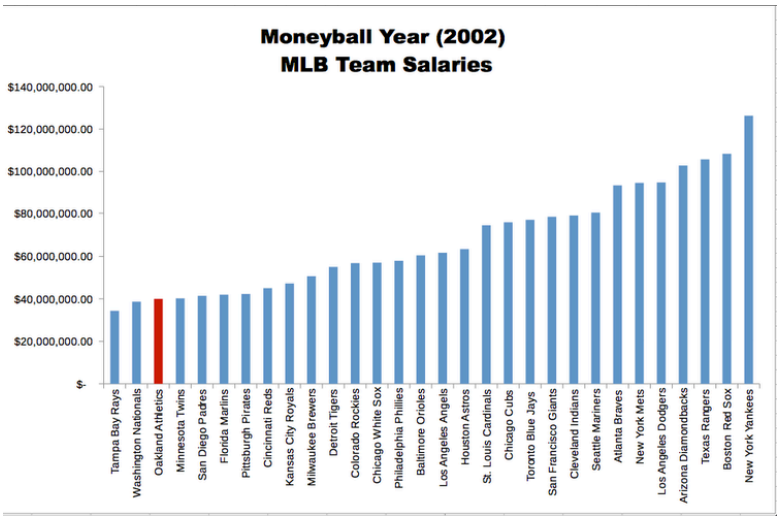

Figura 1.5.: Relación entre presupuesto y victorias en la MLB (Moneyball)

> **Contexto Profesional**
>
> **Aplicación laboral:** El enfoque de Moneyball se ha expandido a todos los deportes (fútbol, baloncesto, F1) y también a recursos humanos. Empresas como Google usan "People Analytics" para contratar y retener talento basándose en datos, no en corazonadas.

1.6.2. Caso 2: Netflix - El Motor Oculto del Entretenimiento

Netflix no es solo una plataforma de streaming. Es una **empresa de datos** que produce contenido. Cada decisión está respaldada por análisis masivo.

Pregunta de Negocio 1: *¿Cómo garantizar streaming de alta calidad sin agotar la tarifa de datos?*

- **Solución:** Modelos que optimizan la **compresión de vídeo** en tiempo real según el dispositivo y la conexión.

Pregunta de Negocio 2: *¿Qué serie original debemos producir y cuántas temporadas tendrá éxito?*

- **Solución:** Análisis predictivo de **patrones de visionado** para tomar decisiones millonarias. *House of Cards* se produjo porque los datos mostraban que los usuarios que veían películas de Kevin Spacey también veían series políticas británicas y películas de David Fincher.

Pregunta de Negocio 3: *¿Qué carátula mostramos a cada usuario para maximizar que vea un título?*

- **Solución: Personalización de marketing** mediante test A/B a gran escala. Una misma película puede tener docenas de carátulas diferentes, y cada usuario ve la que tiene más probabilidad de atraerle.

Cifras de Impacto:

- **80% del contenido** visto en Netflix proviene de su motor de recomendaciones.
- La personalización ahorra a Netflix **1.000 millones de dólares al año** en reducción de bajas (*churn*).
- Los experimentos de portadas personalizadas aumentan el visionado entre un **20% y 30%**.

> **Contexto Profesional**
>
> **Retención de Clientes:** En cualquier servicio de suscripción (SaaS), la métrica clave es el *Churn* (tasa de cancelación). Como analista, tu trabajo será identificar patrones que indiquen que un usuario va a darse de baja antes de que lo haga, para que el equipo de marketing pueda actuar.

1.6.3. Caso 3: Detección de Fraude Bancario

El Problema: Los bancos necesitan detectar transacciones fraudulentas entre millones de operaciones legítimas, de forma instantánea y precisa. Un falso positivo molesta al cliente; un falso negativo cuesta dinero.

La Solución de Datos:

1. **Aprender la "Normalidad":** Los algoritmos de Machine Learning analizan millones de transacciones históricas para aprender cuál es el comportamiento normal de cada usuario.
2. **Detectar la Anomalía:** Cuando una nueva transacción rompe ese patrón (compra inusualmente grande, ubicación extraña, hora atípica), el modelo la marca como sospechosa.
3. **Asignar Puntuaciones de Riesgo:** El sistema asigna una **puntuación de riesgo** a cada operación en tiempo real, permitiendo bloquear las peligrosas y verificar las dudosas.

> **Contexto Profesional**
>
> **Análisis de Riesgos:** El sector financiero es uno de los mayores empleadores de analistas de datos. Más allá del fraude, se usan modelos para decidir si se concede un crédito o para predecir movimientos de bolsa. Es un entorno de alta precisión y responsabilidad.

> **Aviso**
>
> **Contexto ético:** Los sistemas de detección de fraude son un ejemplo de cómo la ciencia de datos puede usarse para proteger a las personas. Sin embargo, los mismos algoritmos pueden generar sesgos si no se diseñan cuidadosamente. Este es un tema que exploraremos en unidades posteriores.

1.6.4. Caso 4: Tinder - Algoritmos de Conexión Humana

El caso de Tinder es especialmente interesante porque muestra cómo la ciencia de datos se aplica a un ámbito muy personal: las relaciones.

El modelo antiguo: ELO Score

El sistema original asignaba a cada perfil una puntuación de "deseabilidad" en función de quién le daba "me gusta", mostrando después perfiles con puntuación similar. La crítica fue rápida: el modelo era superficial, creaba "ligas" de usuarios y fomentaba el *swipe* masivo sin intención real.

El modelo actual: enfoque predictivo (según fuentes externas)

Tinder no ha publicado los detalles de su algoritmo actual, por lo que lo que sigue se basa en análisis de ingeniería inversa y declaraciones parciales de la empresa. La hipótesis más extendida es que el objetivo ha evolucionado: ya no se trata de puntuarte, sino de **predecir la probabilidad de una interacción exitosa**. La métrica clave pasaría a ser si dos personas llegan a mantener una conversación real, penalizando el volumen de *likes* masivos a favor de la calidad de la interacción.

Principios que el algoritmo (probablemente) valora:

Tabla 1.4.: Principios del Algoritmo Predictivo de Tinder

Principio	Qué mide	Por qué importa
Actividad de calidad	Sesiones que generan matches y conversaciones	Usuarios que crean experiencias positivas son prioritarios
Selectividad	Ratio matches/likes enviados	Swipe masivo = "datos sucios", ratio bajo = bandera roja
Engagement post-match	Tiempo hasta primer mensaje, tasa de respuesta	Un match sin conversación es un "match fallido"
Optimización de perfil	Cambios que mejoran likes recibidos/impresiones	Cada cambio es un experimento A/B
Sincronización	Actividad en horas de máxima concurrencia	Matches en tiempo real son más valiosos

Contexto Profesional

Algoritmos de Recomendación: El motor de Tinder es un sistema de recomendación, similar al que usa Amazon para sugerirte productos o YouTube para vídeos. Dominar cómo funcionan estos motores es una de las habilidades más valoradas en empresas de producto digital.

Manos a la Obra

1.6.5. Analiza un Caso de Éxito

Objetivo: Aplicar el pensamiento analítico a un caso real.
Instrucciones: Elige UNO de los casos anteriores (Moneyball, Netflix, Fraude o Tinder) y responde:

1. ¿Cuál era el **problema de negocio** original?

2. ¿Qué **datos** se necesitaron para resolverlo?
3. ¿Cuál fue el **insight clave** que cambió la situación?
4. ¿Qué **métricas** se usaron para medir el éxito?
5. ¿Cómo podrías aplicar un enfoque similar en **otro sector**?

Criterio de éxito:

- Respuestas de 2-3 líneas para cada pregunta
- Propuesta de aplicación a otro sector coherente

1.6.6. Caso 5: Spotify Wrapped - Personalización a Gran Escala

Cada diciembre, millones de usuarios comparten en redes sociales su "Spotify Wrapped": un resumen personalizado de su año musical. Lo que parece un simple resumen es en realidad una obra maestra de ciencia de datos y marketing.

El Problema de Negocio: ¿Cómo convertir datos de uso en una herramienta de marketing viral que los propios usuarios quieran compartir?

La Solución de Datos:

1. **Recolección masiva:** Spotify registra cada canción, cada skip, cada repetición, cada hora del día, cada estado de ánimo inferido.
2. **Análisis individual:** Para cada uno de sus 500+ millones de usuarios, calcula:
   - Total de minutos escuchados
   - Artistas y canciones más reproducidos
   - Géneros dominantes
   - Patrones temporales ("eres un night owl musical")
   - Comparaciones con otros usuarios ("estás en el top 1% de fans de X")
3. **Visualización atractiva:** Los datos se presentan en formato de "historias" optimizado para compartir en Instagram y TikTok.

El Resultado: Wrapped genera millones de posts orgánicos (publicidad gratuita), refuerza la lealtad del usuario ("Spotify me conoce") y diferencia la plataforma de competidores como Apple Music.

Contexto Profesional

Aplicación laboral: La estrategia de Spotify Wrapped se ha convertido en referencia para el marketing basado en datos. Empresas de todos los sectores buscan formas de convertir los datos de uso de sus clientes en experiencias personalizadas que generen engagement y viralidad.

Pregunta de Reflexión

Reflexión sobre casos de éxito:

1. En el caso **Moneyball**, ¿por qué crees que los veteranos del béisbol se resistían a usar nuevas métricas como el OBP? ¿Qué sesgos humanos estaban en juego?
2. Si Netflix dejara de usar datos para producir sus series originales, ¿cómo crees que afectaría a su modelo de negocio a largo plazo?
3. ¿Cuál consideras que es el principal dilema ético en el algoritmo de **Tinder** al predecir si una interacción será exitosa?
4. ¿Cómo ha cambiado **Spotify Wrapped** la percepción del usuario sobre la privacidad? ¿Por qué estamos dispuestos a "regalar" nuestros datos de escucha a cambio de este resumen?
5. Imagina un sistema de detección de fraude que bloquea el 100% de los fraudes pero también el 50% de las compras legales. ¿Es un éxito o un fracaso profesional? Justifica.

1.7. El Flujo de Trabajo del Analista

1.7.1. El Ciclo de Vida de un Proyecto de Datos

Todo análisis profesional sigue un mapa, un proceso iterativo que se detalla en la Figura 1.6. No importa si trabajas en una startup o en una multinacional: este flujo es universal.

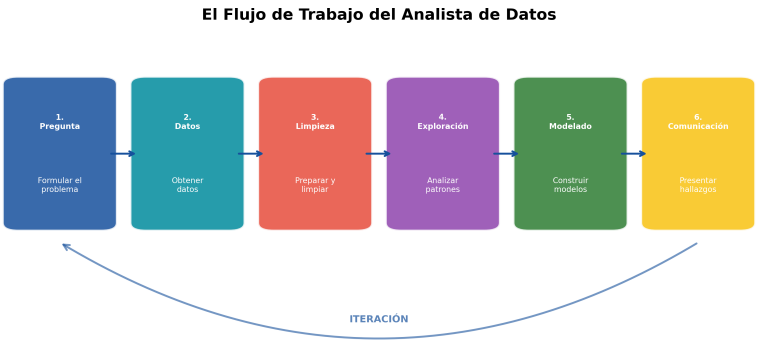

Figura 1.6.: El flujo de trabajo del analista de datos

Tabla 1.5.: Fases del Ciclo de Vida de un Proyecto de Datos

Fase	Descripción	UT del módulo
1. Formular la Pregunta	Traducir un problema de negocio en una pregunta que los datos puedan responder	UT1 (esta unidad)
2. Obtener los Datos	Importar datos desde ficheros, bases de datos y APIs	UT4
3. Limpiar y Ordenar	Asegurar la calidad y formato de los datos. La fase más crítica (80% del tiempo)	UT5
4. Explorar y Entender	Usar estadística y visualización para encontrar patrones	UT6
5. Modelar (Opcional)	Construir modelos predictivos con Machine Learning	UT8
6. Comunicar Hallazgos	Presentar resultados de forma clara y accionable	UT7, UT9

Recuerda

Herramienta Estrella (Pandas): En la **UT3** empezaremos a usar **Pandas**, la librería que convierte a Python en una hoja de cálculo programable con capacidad de análisis industrial. Es la herramienta que más usarás en tu día a día como analista.

Importante

Concepto crítico: El flujo de trabajo NO es lineal. Es iterativo. Mientras limpias datos (fase 3) descubrirás que necesitas más datos (volver a fase 2). Mientras exploras (fase 4) reformularás la pregunta original (volver a fase 1). Esta iteración es normal y deseable.

Manos a la Obra

1.7.2. Mapea un Problema al Flujo de Trabajo

Objetivo: Practicar la descomposición de un problema real en las 6 fases del flujo.
Escenario: Trabajas en el departamento de marketing de una tienda online de ropa. El director te dice: “Nuestros clientes compran una vez y no vuelven. Quiero saber por qué y cómo mejorar la retención.”
Instrucciones: Para cada fase del flujo, describe qué harías:

1. **Formular la Pregunta:** ¿Cuál es la pregunta específica que los datos pueden responder?
2. **Obtener los Datos:** ¿Qué datos necesitas? ¿De dónde los sacarías?
3. **Limpiar y Ordenar:** ¿Qué problemas de calidad podrías encontrar?
4. **Explorar y Entender:** ¿Qué patrones buscarías?
5. **Modelar:** ¿Sería útil un modelo predictivo? ¿Para qué?
6. **Comunicar:** ¿Cómo presentarías los hallazgos al director?

Criterio de éxito:

- Una respuesta de 2-3 líneas para cada fase
- Las respuestas deben ser específicas para el escenario (no genéricas)

Pregunta de Reflexión

Reflexión sobre el método:

1. ¿Por qué decimos que la **Limpieza de Datos** consume el 80% del tiempo? ¿Es una ineficiencia o es una parte necesaria de la creación de valor?
2. ¿Qué riesgos corres si saltas directamente de la "Pregunta" al "Modelado" sin pasar por la fase de "Exploración"?
3. ¿Cómo influye la fase de **Comunicación** en la reformulación de la pregunta inicial para un segundo ciclo del proyecto?
4. En un equipo pequeño, ¿quién crees que debería liderar la fase de "Obtención de Datos" y por qué?
5. Pon un ejemplo de un problema real de tu entorno que NO necesite la fase de "Modelado" (Machine Learning) para ser resuelto con éxito.

1.8. Nuestro Laboratorio: Google Colab

1.8.1. ¿Qué es un Notebook?

Un **notebook** es un cuaderno interactivo que se ejecuta en un navegador web. Es la herramienta estándar en la industria del análisis de datos porque permite combinar en un solo documento:

- **Código** ejecutable (Python en nuestro caso)
- **Texto** formateado (explicaciones, notas, conclusiones)
- **Resultados** visuales (tablas, gráficos)

Información

¿Por qué Google Colab?

- **Cero instalación:** Solo necesitas un navegador y una cuenta de Google

- **Entorno preconfigurado:** Las librerías fundamentales (Pandas, NumPy, Matplotlib) ya están instaladas
- **Potencia en la nube:** Accedes a máquinas potentes sin saturar tu ordenador
- **Colaboración fácil:** Puedes compartir notebooks como si fueran documentos de Google Drive

1.8.2. Los Dos Idiomas del Notebook

Un notebook tiene dos tipos de **celdas**:

Celdas de Código:

- Escribes y ejecutas código Python
- Atajo para ejecutar: `Shift + Enter`
- Los resultados aparecen justo debajo

Celdas de Texto (Markdown):

- Documentas tu trabajo con texto formateado
- Usas un lenguaje simple llamado **Markdown**
- Permiten títulos, listas, negritas, enlaces, imágenes...

Pro-Tip

Principio fundamental: Un buen notebook se lee como un informe técnico o una historia coherente, no como un script de código desordenado. La documentación es tan importante como el código. Un notebook sin explicaciones es un notebook que nadie (incluido tu yo del futuro) podrá entender.

1.8.3. Markdown: El Lenguaje de la Documentación

Markdown es un lenguaje de formato ligero que se convierte en texto formateado. Es el estándar en notebooks, GitHub, documentación técnica y muchas otras herramientas.

Tabla 1.6.: Sintaxis Básica de Markdown para Analistas

Sintaxis	Resultado
`# Título`	Título grande
`## Subtítulo`	Subtítulo
`**negrita**`	**negrita**
`*cursiva*`	*cursiva*

Sintaxis	Resultado
`- elemento`	Lista con viñetas
`1. elemento`	Lista numerada
`` `código` ``	Código en línea
`[texto](url)`	Enlace

Pro-Tip

Documentación Profesional: En el mundo del software y los datos, el archivo `README.md` es la puerta de entrada a cualquier proyecto. Aprender Markdown no es solo para notebooks; es la forma en que comunicarás tus proyectos en plataformas como GitHub o GitLab.

Manos a la Obra

1.8.4. Tu primer notebook en Google Colab

Objetivo: Configurar tu entorno de trabajo y crear tu primer notebook funcional.
Instrucciones:

1. Abre Google Colab[a]
2. Crea un nuevo notebook: Archivo → Nuevo cuaderno
3. Renómbralo: "UT01_MiPrimerNotebook"
4. Crea una celda de texto con:
   - Un título con tu nombre
   - Una lista de 3 cosas que esperas aprender en este módulo
5. Crea una celda de código con:

```
print("Hola, soy un analista de datos en formación!")
2 + 2
```

6. Ejecuta la celda con `Shift + Enter`

Análisis de los resultados: Este primer bloque confirma que tu entorno de ejecución está activo. El mensaje impreso verifica que Python responde correctamente, y el resultado `4` demuestra que Colab puede realizar cálculos matemáticos inmediatos, funcionando como una calculadora científica avanzada integrada en tu informe.
Criterio de éxito:

- Notebook creado y renombrado correctamente
- Celda de texto con título y lista formateados
- Celda de código ejecutada mostrando el mensaje y el resultado 4

[a] https://colab.research.google.com

Manos a la Obra

1.8.5. El arte de la documentación (Markdown)

Objetivo: Practicar la creación de documentación estructurada usando Markdown.
Instrucciones: En una nueva celda de texto en tu notebook de Colab, crea una "Ficha de Investigación" para el caso Moneyball que incluya:

1. Un **título principal** (#)
2. Un **subtítulo** (##) con el nombre del equipo
3. Una **lista con viñetas** de los 3 descubrimientos clave
4. Una palabra en **negrita** y otra en *cursiva*
5. Un **enlace** a la página de Wikipedia de Billy Beane

Criterio de éxito:

- Estructura jerárquica clara (títulos y subtítulos)
- Lista correctamente formateada
- Enlace funcional

1.8.6. Gestión de ficheros: Google Drive

El entorno de ejecución de Colab es **efímero**: cualquier fichero que subas directamente a la sesión se borrará cuando esta se cierre. Para trabajar de forma profesional, necesitamos conectar Colab con Google Drive.

Aviso

Sesiones efímeras: No confundas subir un archivo a la barra lateral de Colab con guardarlo permanentemente.
Si cierras la pestaña del navegador o pasan unas horas de inactividad, esos archivos desaparecerán. Usar Google Drive es la única forma de garantizar que tus datasets y resultados estén disponibles mañana.

Método recomendado: Montar Google Drive

```
from google.colab import drive
drive.mount('/content/drive')
```

Al ejecutar este código, Colab te pedirá autorización para acceder a tu Drive. Una vez montado, puedes acceder a tus ficheros como si fueran carpetas locales:

```
# Leer un CSV desde Drive
import pandas as pd
df = pd.read_csv(
↪  '/content/drive/MyDrive/PYAD/datasets/mi_archivo.csv')
```

Aviso

Estructura de carpetas recomendada:
Crea en tu Google Drive una carpeta para el módulo con esta estructura:

```
PYAD/
|-- datasets/       # Ficheros CSV, Excel, JSON...
|-- notebooks/      # Tus cuadernos Jupyter
`-- output/         # Gráficos, informes generados
```

Mantener una estructura ordenada desde el principio te ahorrará muchos problemas.

Manos a la Obra

1.8.7. Configura tu entorno profesional

Objetivo: Montar Google Drive y cargar tu primer dataset (conjunto de datos).
Instrucciones:

1. Crea la estructura de carpetas en tu Google Drive (PYAD/datasets, notebooks, output)
2. Descarga el dataset Titanic desde el repositorio del libro[a] y guárdalo en `datasets/`
3. En tu notebook, monta Google Drive
4. Carga el dataset con Pandas:

```
import pandas as pd
df = pd.read_csv(
↪  '/content/drive/MyDrive/PYAD/datasets/titanic.csv')
df.head()
```

5. Ejecuta `df.info()` para ver información básica del dataset

Análisis de los resultados: Al ejecutar `df.head()`, has validado visualmente que los datos se han leído correctamente desde Drive. Por otro lado, la salida de `df.info()` es tu primer diagnóstico técnico: confirma que tienes 891 registros y te revela qué columnas son numéricas y cuáles de texto, además de detectar si faltan datos (valores nulos), paso fundamental antes de cualquier análisis.
Criterio de éxito:

- Estructura de carpetas creada en Drive
- Dataset Titanic cargado correctamente
- `df.head()` muestra las primeras 5 filas
- `df.info()` muestra 891 filas y 12 columnas

[a] https://raw.githubusercontent.com/jgarcia314/analisis-datos-python-fp/main/data/raw/titanic.csv

Pregunta de Reflexión

Reflexión sobre las herramientas:

1. ¿Cuáles son las limitaciones de **Google Colab** para un proyecto profesional que maneje datos médicos altamente confidenciales?
2. Compara un **Notebook** con un script tradicional de Python (.py). ¿En qué escenarios es mejor usar cada formato?
3. ¿Por qué el lenguaje **Markdown** es vital para garantizar la reproducibilidad en la ciencia de datos?
4. ¿Qué pasaría con tu trabajo en Colab si Google Drive dejara de funcionar por mantenimiento durante 2 horas? ¿Cómo podrías mitigar ese riesgo?
5. ¿Crees que Google Colab es una herramienta definitiva para una gran multinacional o solo una plataforma de prototipado? Justifica.

1.9. Roadmap del módulo

1.9.1. Tu viaje de aprendizaje

Ahora que entiendes el contexto, veamos el mapa completo del módulo. Cada unidad temática (UT) construye sobre la anterior:

Tabla 1.7.: Roadmap de Unidades Temáticas y su Conexión con el Flujo de Trabajo

UT	Título	Qué aprenderás	Conexión con el flujo
UT1	Introducción (esta unidad)	Contexto, flujo de trabajo, Colab	Fase 1: Formular preguntas
UT2	NumPy	Arrays y computación numérica eficiente	Base técnica
UT3	Pandas Básico	DataFrames, selección, filtrado	Fases 2-3
UT4	Adquisición de Datos	CSV, APIs, web scraping	Fase 2: Obtener datos

UT	Título	Qué aprenderás	Conexión con el flujo
UT5	Preprocesamiento	Limpieza, valores nulos, transformación	Fase 3: Limpiar datos
UT6	Análisis Exploratorio (EDA)	Estadística descriptiva, patrones	Fase 4: Explorar
UT7	Visualización	Gráficos con Matplotlib y Seaborn	Fases 4 y 6
UT8	Machine Learning Básico	Clasificación, regresión, evaluación	Fase 5: Modelar
UT9	Dashboards y Comunicación	Streamlit, storytelling con datos	Fase 6: Comunicar
UT10	Proyecto Final	Integración de todo lo aprendido	Todas las fases

Recuerda

Al terminar UT2 (NumPy), tendrás las bases para entender cómo Python maneja datos eficientemente. Al terminar UT3 (Pandas), ya podrás hacer análisis reales. Cada unidad desbloquea nuevas capacidades.

1.9.2. Competencias profesionales que desarrollarás

Al completar este módulo dominarás las competencias clave del analista de datos:

Tabla 1.8.: Competencias profesionales del módulo y unidades donde se desarrollan

Competencia	Qué serás capaz de hacer	Unidades
Gestión de Datos	Cargar, limpiar y normalizar datasets heterogéneos	UT3, UT4, UT5
Análisis de Calidad	Identificar y resolver inconsistencias, nulos y outliers	UT5
Investigación Exploratoria	Descubrir patrones y tendencias ocultas en los datos	UT6

Competencia	Qué serás capaz de hacer	Unidades
Comunicación Visual	Diseñar gráficos profesionales que apoyen decisiones	UT7, UT9
Modelado Predictivo	Desarrollar y evaluar modelos de predicción	UT8, UT9

1.9.3. Perfiles profesionales de salida

Al completar este módulo, estarás preparado para:

Tabla 1.9.: Perfiles profesionales de salida y sectores de inserción laboral

Perfil	Descripción	Sectores típicos
Junior Data Analyst	Analiza datos, crea informes y dashboards	Retail, banca, marketing
Business Intelligence Analyst	Traduce datos en insights de negocio	Consultoría, grandes empresas
Data Quality Analyst	Asegura la calidad de los datos	Cualquier sector con datos críticos
Marketing Analyst	Analiza campañas, segmentación, ROI	Agencias, e-commerce, startups

Contexto Profesional

Siguiente paso profesional: Este módulo te da las bases sólidas de analista. Si quieres especializarte en ciencia de datos avanzada (Machine Learning, deep learning), tendrás los cimientos necesarios para continuar con formación especializada.

1.10. Fuentes y lecturas recomendadas

Recuerda

¿Quieres profundizar más? Consulta la bibliografía detallada, los enlaces a la documentación oficial y los recursos de aprendizaje para esta unidad en el **Apéndice B: Fuentes y lecturas recomendadas** al final de este libro.

1.11. Resumen de la unidad

> **! Importante**
>
> **Mensaje clave:** La ciencia de datos no es solo programar. Es una forma de pensar: observar problemas, formular preguntas, buscar evidencia y comunicar hallazgos. Las herramientas (Python, Pandas, SQL) son medios para un fin, no el fin en sí mismo.

En la próxima unidad (UT2: NumPy), empezaremos a construir las bases técnicas. Aprenderemos cómo Python maneja datos de forma eficiente usando arrays, lo que será fundamental para todo lo que viene después.

¡Manos a la obra!