Informática y comunicaciones

ies azarquiel

Curso especialización en inteligencia artificial y big data

Análisis de datos

Profesor: Sebastián Rubio Valero

Septiembre 2025

Big Data

Es un componente elemental en la Ciencia de Datos. En este capítulo, exploraremos tres pilares fundamentales para llevarlo a cabo: el Análisis Exploratorio de Datos (EDA), la Limpieza de Datos (Data Cleaning) y la Manipulación de Datos (Data Wrangling). Además, abordaremos un cuarto aspecto esencial, el Profiling. Mientras se realiza el análisis y modelado de datos, se emplea una considerable cantidad de tiempo en preparación de los mismos. A menudo estas tareas le ocupan al analista más del 80% de su tiempo. Utilizaremos mucho técnicas descriptivas para averiguar en que estado se encuentran los datos. El resultado final de un preprocesamiento de datos es que es de obtener unos datos tabulares (datos en formato Tidy) que sean consistentes para el entrenamiento del modelo o de los modelos

1 Roles de la Ciencia de Datos

Vamos a echar un vistazo a todos los roles que estan amparados bajo el paraguas del término Ciencia de Datos

1.1 Científico de Datos

Una definición de científico de datos es la de alguien que sabe más programación que un estadístico y más estadística que un ingeniero de software. Los científicos de datos se encargan de afinar los modelos estadísticos y matemáticos que se aplican a los datos. Esto podría implicar la aplicación de su conocimiento teórico en estadística y algoritmos para encontrar la mejor solución a un problema con datos.

1.2 Analista de Datos y Analista de Negocio

Los analistas de datos se dedican a filtrar la información y generar informes y visualizaciones para revelar las ideas ocultas en los datos. La persona que ayuda a otros empleados de la empresa a entender consultas específicas mediante gráficos está ejerciendo el rol de analista de datos (o analista de negocio). En algunos aspectos, se les puede ver como científicos de datos junior o como el primer escalón hacia un puesto en ciencia de datos.
Los analistas de datos examinan los datos y proporcionan informes y visualizaciones para explicar qué ideas (o insights) ocultan esos datos. Cuando alguien ayuda a personas de toda la empresa a comprender consultas específicas con gráficos, está desempeñando el rol de analista de datos (o analista de negocio). En cierto modo, se les puede considerar como científicos de datos junior o el primer paso en el camino hacia un puesto de ciencia de datos.

1.3 Ingeniero de Datos

Los ingenieros de datos son ingenieros de software que manejan grandes cantidades de datos y, a menudo, sientan las bases y la infraestructura para que los científicos de datos puedan realizar su trabajo de manera efectiva. Son responsables de gestionar los sistemas de bases de datos, escalar la arquitectura de datos a múltiples servidores y escribir consultas complejas para examinar los datos. También pueden limpiar conjuntos de datos e implementar solicitudes complejas que provienen de los científicos de datos, por ejemplo, toman el modelo predictivo del científico de datos y lo implementan en código listo para producción. Conoce las tecnología de Hadoop,Spark, MapReduce, Hive, Pig,etc.

2 Exploración de datos básico

Es una etapa muy importante en cualquier proyecto de Análisis de Datos o Data Science. Es el proceso de usar herramientas estadísticas y visualizaciones para entender qué información aporta el conjunto de datos. Durante EDA, puedes encontrar patrones, identificiar anomalías, probar hipótesis o verificar suposiciones (Ejemplo del rompecabezas). El EDA, la limpieza de datos y manipulación están interconectados, cada uno tiene objetivos diferentes y requiere de distintas técnicas.

2.1 Estadística Descriptiva

Las variables numéricas son aquellas que representan cantidades numéricas y pueden tomar valores enteros o decimales. Las variables categóricas son aquellas que contienen valores que representan categorías. Estas variables pueden ser nominales (sin un orden inherente) o ordinales (con un orden lógico).

2.2 Limpieza de nulos

La eliminación de filas que contienen nulos no suele ser una práctica ideal, ya que podría resultar la pérdida de datos potencialmente valiosos.

2.3 Matriz de correlación

La Matriz de Correlación es una tabla que muestra los coeficientes de correlación entre varias variables numéricas. Cada celda de la matriz muestra el coeficiente de correlación entre dos variables. El Coeficiente de Correlación, típicamente el Coeficiente de Correlación de Pearson, mide la relación lineal entre dos variables numéricas.

- 1 indica una correlación positiva (a medida que una variable aumenta, la otra también lo hace en propoción directa)
- -1 indica una correlación negativa perfecta (a medida que una variable aumenta, la otra disminuye en proporción directa)
- 0 indica que no hay correlación lineal entre las variables

2.4 Histogramas

Cada barra del histograma representa la frecuencia que los valores de datos caen dentro del rango específico Lo ideal es realizar un histograma de cada una de las

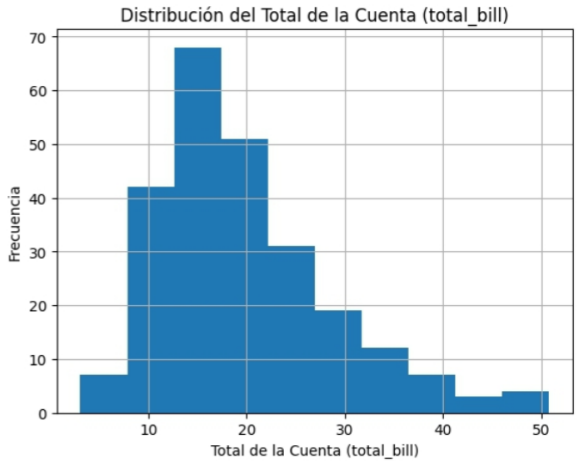

Figure 1: Enter Caption

características de nuestros datos. De un primer vistazo vamos a comprobar hacia donde se desplazan los datos (izquierda o derecha) en los casos de una distribución de Gauss, o bien si un distribución exponencial, en cuyo caso debemos aplicar una transformación para conseguir un distribución Gaussiana,esta distribución es señal de que hay un sesgo en esa característica. Una distribución Gaussiana, indica que esta bien representada una población con ese dato. Otra información que se puede extraer son los outliers

2.5 Diagrama de cajas

Es una representación visual que describe varias características importantes de un conjunto de datos, como la mediana,los cuartiles y los valores atípicos (outliers).

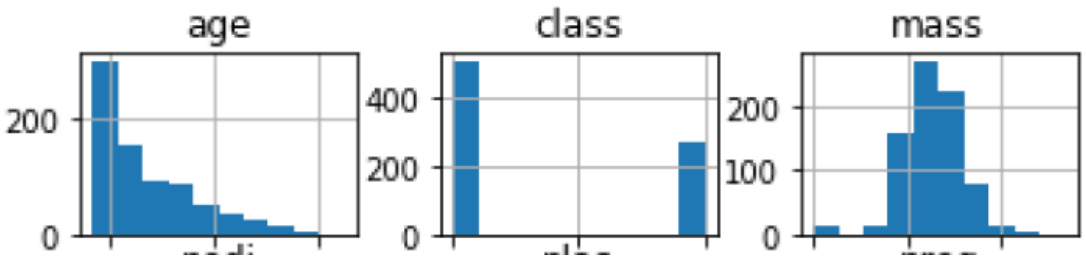

Figure 2: Enter Caption

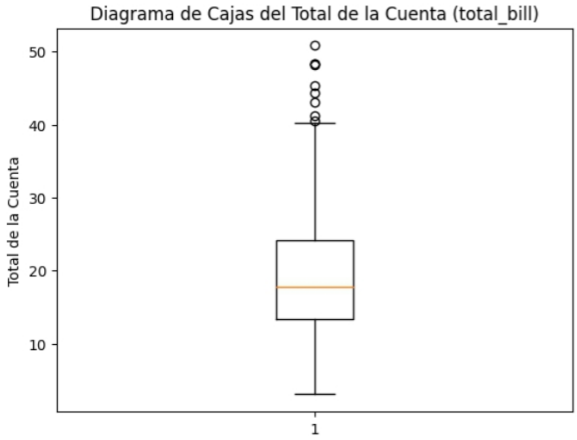

Figure 3: Enter Caption

2.6 Gráficos de dispersión

Es una representación visual que utiliza puntos para mostrar la relación entre dos o más variables.

3 Data Cleaning:Limpieza de Datos

El proceso de limpieza de datos implica identificar y corregir errores o inconsistencias en tus datos, como valores faltantes,duplicado, o incorrectos. El objetivo es asegurar que tus datos sean precisos y coherentes antes de comenzar el análisis. Con la libreria pandas, todas las estadísticas descriptivas dejan fuera a los datos ausentes de forma predeterminada.

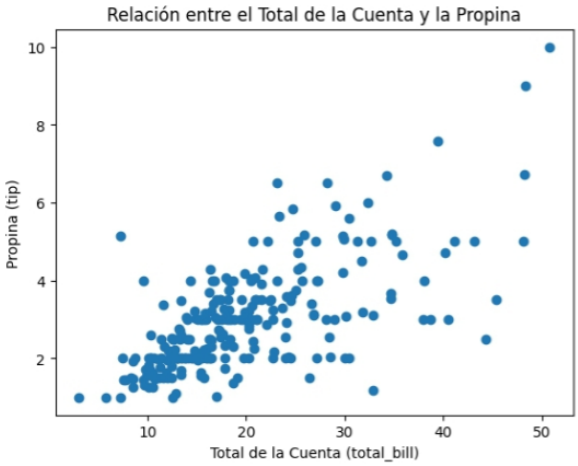

Figure 4: Enter Caption

3.1 Identificación de valores faltantes (Missing values

Hay que contabilizar el numero de valores faltantes por cada característica, en función del resultado se puede tomar varias soluciones

- Rellenar con la mediana
- Rellenar con la moda
- Eliminar la fila correspondiente

3.2 Eliminación de columnas

La eliminación de una o más columnas, es una operación que se utiliza una vez que se ha comprobado que esas columnas no forman parte del entrenamiento del modelo.

3.3 Eliminación de duplicados

La eliminación de filas duplicadas es otra operación que se realiza en ciertas ocasiones.

4 Manipulación de Datos (Data Wrangling)

La transformación o manipulación de datos, puede incluir cambiar el formato de los datos, combinar datos de diferentes fuentes, o modificarlos de tal manera que sean más útiles para su análisis.

4.1 Creación de variables

La necesidad de crear variables surge principalmente cuando las variables originales no son suficientes para capturar el conocimiento o la relación que se quiere modelar, o cuando los modelos estadísticos o de Machine Learning requieren datos en un formato específico, por ejemplo supongamos que un data set tenemos dos características el precio-unidad, y unidades-compradas, en este caso podría ser conveniente en crear un nuevo campo que sea el total-pagar.

4.2 Agrupamientos de datos

Es fundamental en el análisis de datos, se realiza por dos razones: reducir la complejidad y descubrir información oculta. Se realiza bajo dos premisas fundamentales:

- Cuando el volumen de datos es demasiado grande y necesitas una representación concisa y eficiente para la visualización, reportes y la reducción de la carga de procesamiento.
- Cuando buscas descubrir patrones, estructuras y grupos naturales en datos no etiquetado, para obtener una ventaja competitiva o identificar riesgos y oportunidades

4.3 Renombrado de columnas

No se trata solo de algo estético, sino de garantizar la calidad, le interoperabilidad y la legibilidad del conjunto de datos. En Big Data los datos viene de diferentes fuente es importante tener un formato uniforme. Aunque no sea propiamente un renombrado de columnas, pero cuando la característica que vamos a usar como etiqueta no se encuentra en la última columna, es mejor colocarla al final

4.4 Reemplazo de valores

Es una fase crítica y se hace necesario en tres escenarios principales:

- valores faltantes, estos huecos se reemplazan por algún valor estimado
- reemplazo por valores atípicos (outliers), si el error es claramente por error de tipeo, se debe reemplazar por un valor conocido (media/mediana) o eliminarlo
- Si se va usar un modelo sensible a los outliers estos casos se reemplaza para reducir su impacto sin perder observación, por ejemplo un percentil 95 o 99. Esto limita la influencia del extremo sin eliminar el registro. En otros caso es conveniente eliminar el registro.

- los datos que son categóricos se pueden reemplazar por un valor numérico, por ejemplo si=1, no=0

En algunos casos tendremos que usar un algoritmo de ML para que nos diga con qué valor debo rellenar el valor faltante. Podríamos pensar que los datos faltantes siempre los vamos a rellenar con la media, pero si el numero de valores faltantes es grande, esto puede provocar un sobreajuste (overfiting) del modelo.¿Por qué no podemos tomar un valor comprendido entre el rango de la característica?.

5 Profiling

El profiling es una fase esencial en el Análisis de Datos que te ayuda a conocer a fondo la naturaleza y la calidad de tus datos. En el contexto de la ingeniería y la calidad de datos (especialmente en Big Data), el Data Profiling es el proceso de examinar y analizar los datos de un conjunto de datos existente para recopilar metadatos y estadísticas informativas sobre su calidad, estructura y contenido.
El objetivo del perfilado es obtener una visión completa de los datos antes de analizarlos o usarlos en un modelo, ayuda a:

- Evaluar la Calidad de los Datos: Detectar problemas de integridad, consistencia y completitud.
- Comprender la Estructura: Saber si los datos están bien formateados para su uso.
- Facilitar la Toma de Decisiones: Asegurar que los análisis posteriores y las decisiones de negocio se basen en información precisa y confiable.

El profiling responde a las siguientes preguntas

Categoría	**Pregunta Clave**	**Resultado del Profiling**
Completitud	¿Cuántos valores faltan?	Porcentaje de valores nulos.
Estadísticas	¿Cómo se distribuyen los valores?	Mínimo, Máximo, Media, Mediana, Desviación Estándar.
Unicidad	¿Qué tan variados son los datos?	Conteo de valores únicos y porcentaje de filas con el mismo valor.
Formato	¿Qué tipo de datos contiene?	Tipo de dato (ej. `String`, `Integer`, `Fecha`).
Frecuencia	¿Cuáles son los valores más comunes?	Los 10 valores más frecuentes y su respectivo conteo.
Anomalías	¿Existen errores o valores atípicos?	Detección de *Outliers* o faltas de ortografía.

Table 1: Resultados del Profiling de Datos