

3. DATASET Y HOJAS DE CALCULO

Introducción


La ciencia de datos es la clave del futuro de la inteligencia artificial. Pueden hacerse realidad muchos de los conceptos que no era posible hasta hace poco.

La creación de soluciones de almacenamiento de datos era el objetivo principal. El enfoque ha cambiado al procesamiento de estos datos resuelto con éxito el problema del almacenamiento. Por esto para la manipulación de datos surgen herramientas y métodos de organización de estos para generar resultados que potencien la toma de decisiones o la generación de nuevas soluciones.

Iniciamos con 2 elementos básicos y primordiales para el procesamiento y análisis, los cuales serán de uso cotidiano en el proceso de ciencia de datos y su uso en inteligencia artificial, estos son: los Dataset y las hojas de cálculo.

3.1 DATASET

Un dataset es una colección de datos estructurados que se utiliza para analizar información y entrenar modelos en ciencia de datos. Los datasets pueden ser de diversos tipos, incluyendo tablas, imágenes, textos, etc. Aquí te presentamos algunos conceptos fundamentales sobre los datasets orientados a aprendices en ciencia de datos.



Description	Deadline	Status	Amount	Add row
Alphabet puzzle	2016-01-15	Done	1500	+ -
Layout for poster	2016-01-31	Planned	1854	+ -
Image creation	2016-01-23	To Do	1500	+ -
Create font	2016-02-06	Done	1200	+ -
Sticker production	2016-02-18	Planned	2100	+ -
Quoting poster	2016-03-17	To Do	850	+ -
Beer label	2016-05-28	Confirmed	2400	+ -
Shop sign	2016-06-19	Other	1099	+ -
XMAS decoration	2016-10-31	Confirmed	1150	+ -
Wallpaper photo	2016-08-12	Planned	600	+ -
Reading announcement	2016-07-09	To Do	299	+ -
Member passport	2016-06-02	Other	149	+ -

3.1.1. Estructura del Dataset

Los dataset generalmente están organizados en forma de tablas, donde:

Filas (Registros): Cada fila representa una instancia o un ejemplo del conjunto de datos relacionados a una misma entidad.

- **Columnas** (Características): Cada columna representa una característica, atributo o variable de los datos.

Los datos deben ir acompañados de una descripción detallada de su contenido, lo que llamamos Metadata.

METADATA:

La metadata de un dataset se refiere a la información descriptiva sobre los datos contenidos en él. Es como una etiqueta que proporciona contexto y detalles sobre cómo se recopilaron, qué significan y cómo están organizados los datos.

Antes de manipular un dataset debemos tener claro:

a) Entendimiento Inicial:

La metadata nos ayuda a comprender la naturaleza de los datos. ¿Son mediciones, texto, imágenes o algo más? ¿Cuál es la fuente?, cantidad, tamaño etc. Esto es crucial para abordar la manipulación de manera adecuada.

b) Calidad y Fiabilidad:

La metadata revela si los datos son confiables, completos o si hay valores faltantes. Esto afecta la calidad de los análisis y modelos que se construirán.

c) Formato y Estructura:

La metadata indica el formato (CSV, JSON, etc.) y la estructura (columnas, tipos de datos) del dataset. Esto guía la limpieza y transformación.

FORMATO USUAL DE LOS DATASET

Generalmente son generados desde bases de datos y se les exporta a formatos como: CSV, JSON, XML, o TXT

- Librerías útiles para la importación y manipulación de datos con Python son Panda y Numpy entre otras

Ejemplo:

ID	Nombre	Edad	Género	Ingresos
1	Juan	28	M	30000
2	María	34	F	40000
3	Carlos	22	M	25000
4	Ana	30	F	35000

3.1.2. Tipos de Datos

Dentro de un dataset, las columnas pueden contener diferentes tipos de datos:

- **Numéricos:** Datos que representan cantidades o valores continuos, como la edad o los ingresos.
- **Categoricos:** Datos que representan categorías o grupos, como el género o el estado civil.
- **Fechas y Tiempos:** Datos que representan fechas y tiempos, como la fecha de nacimiento o la hora de una transacción.
- **Texto:** Datos que contienen texto libre, como comentarios o descripciones.

3.1.3. Fuentes de Datasets

Los dataset pueden provenir de diversas fuentes, como:

- **Bases de datos:** Colecciones de datos organizados que pueden ser consultadas.
- **Archivos CSV,TXT/Excel:** Archivos tabulares donde los datos se almacenan en formato delimitado por comas o en hojas de cálculo.
- **APIs:** Interfaces de programación que permiten acceder a datos desde servicios web.
- **Scraping web:** Proceso de extraer datos de sitios web.

3.1.4. Importancia de los Datasets

En ciencia de datos, los datasets son fundamentales porque se utilizan para:

- **Entrenamiento de Modelos:** Los datos se utilizan para entrenar algoritmos de aprendizaje automático.
- **Validación y Pruebas:** Los datos permiten evaluar y validar el rendimiento de los modelos.
- **Análisis Exploratorio:** Los datos ayudan a entender patrones, tendencias y relaciones en la información.

3.1.5. Limpieza y Preparación de Datos

Antes de utilizar un dataset, es crucial realizar tareas de limpieza y preparación, que incluyen algunas tareas como por ejemplo:

- **Manejo de valores nulos:** Imputar o eliminar datos faltantes.

- **Normalización y Estandarización:** Ajustar los valores de los datos a un rango o escala común. (Valores, fechas, formatos)
- **Codificación de Datos Categóricos:** Convertir datos categóricos en numéricos utilizando técnicas como One-Hot Encoding.
- Eliminación de duplicados
- Corrección valores Null

3.1.6. Ejemplo Práctico: Dataset de Iris

Un ejemplo clásico de dataset utilizado en ciencia de datos es el Dataset de Iris. Este dataset contiene medidas de distintas características de flores y se utiliza comúnmente para problemas de clasificación. Existen varios sitios para descargar este dataset, uno de ellos es:

<https://www.kaggle.com/datasets/vikrishnan/iris-dataset>

Estructura del Dataset de Iris:

Largo_Sépalo	Ancho_Sépalo	Largo_Pétalo	Ancho_Pétalo	Especie
5.1	3.5	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica

3.1.7. Herramientas para Trabajar con Datasets

Los científicos de datos utilizan diversas herramientas para manipular y analizar datasets, tales como:

- Python: Con librerías como Pandas, NumPy y Scikit-learn.
- R: Un lenguaje de programación especializado en análisis estadístico.

- **SQL:** Para consultar y manipular bases de datos relacionales.
- **Herramientas de Visualización:** Como Matplotlib, Seaborn, y Tableau.
- **Hojas de calculo:** Indiscutiblemente es una herramienta muy valiosa en varios procesos de ciencia de datos.

Otras alternativas para obtener dataset

Hojas de cálculo: Programas como Microsoft Excel o Google Sheets son útiles para organizar y almacenar datos estructurados en forma de tablas. Permiten ingresar, editar y manipular datos de manera intuitiva.

Bases de datos: Se utilizan sistemas de gestión de bases de datos (SGBD) como MySQL, PostgreSQL, MongoDB, entre otros, para almacenar y gestionar grandes volúmenes de datos. Proporcionan capacidades de consulta y acceso eficiente a los datos.

Web scraping: Es el proceso de extracción automatizada de datos de sitios web. Se utilizan herramientas como BeautifulSoup, Selenium o Scrapy para recopilar datos estructurados o no estructurados desde páginas web.

APIs: Las interfaces de programación de aplicaciones (APIs) permiten la extracción de datos de servicios en línea, como redes sociales, plataformas de datos abiertos, servicios de clima, entre otros. Estas APIs proporcionan acceso programático a los datos y se pueden utilizar con lenguajes de programación como Python o JavaScript.

Herramientas de extracción de datos: Existen herramientas especializadas para extraer datos de diferentes fuentes, como herramientas de extracción de datos de documentos PDF, herramientas de OCR (reconocimiento óptico de caracteres) para extraer texto de imágenes escaneadas, etc.

Anotación de datos: Para construir data sets etiquetados o anotados, se utilizan herramientas de anotación, como Labelbox, RectLabel, VGG Image Annotator (VIA) o Prodigy, que permiten etiquetar imágenes, segmentar objetos, clasificar texto, entre otros.

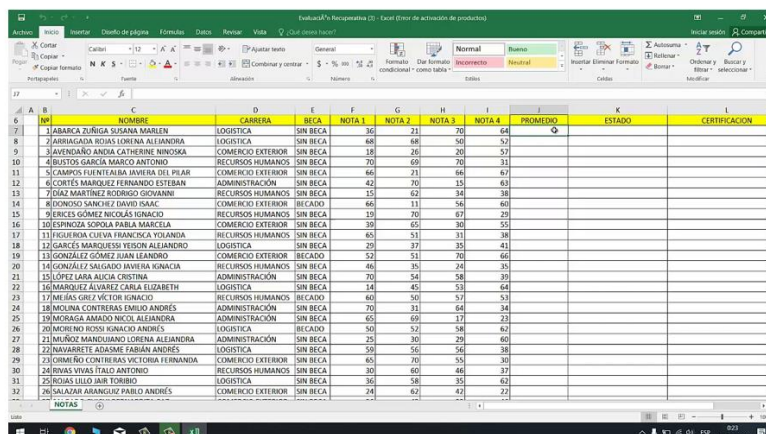
Aplicaciones personalizadas: En algunos casos, se desarrollan aplicaciones personalizadas para recopilar y almacenar datos de acuerdo con los requisitos específicos del proyecto. Esto puede implicar el diseño y la implementación de formularios personalizados o interfaces de usuario para capturar datos.

Datos Públicos: Cuando la información publica se presenta como datos y se los distribuye en formatos digitales abiertos, ya sea en archivos de texto separados por coma, hojas de cálculo u otros formatos abiertos, decimos que son datos públicos en formato abierto. Estos datos deben tener cumplir cierto parámetros como:

Los datos deben ser:
completos, primarios, oportunos, accesibles, procesables por máquinas, acceso no discriminatorio, formatos estándar, sin restricciones.

Sitios web para descargar datasets (Kaggle, UCI Machine Learning Repository, etc.)

3.2. HOJAS DE CALCULO



ID	NOMBRE	CARRERA	BECA	NOTA 1	NOTA 2	NOTA 3	NOTA 4	PROMEDIO	ESTADO	CERTIFICACION
1	BARCA ZUÑIGA SUSANA MARLEN	LOGISTICA	SIN BECA	35	21	70	64			
2	ARRIAGA ROJAS LORENA ALEJANDRA	LOGISTICA	SIN BECA	68	68	50	52			
3	AVENDANO ANDA CATHERINE WINDSA	COMERCIO EXTERIOR	SIN BECA	18	26	20	57			
4	BUSTOS GARCIA MARCO ANTONIO	RECURSOS HUMANOS	SIN BECA	70	69	70	31			
5	CAMPOS FUENTEALBA JAVIERA DEL PILAR	COMERCIO EXTERIOR	SIN BECA	66	71	66	67			
6	CORTES MARIQUEZ FERNANDO ESTEBAN	ADMINISTRACION	SIN BECA	42	70	15	63			
7	DIAZ MARTINEZ RODRIGO GIOVANNI	RECURSOS HUMANOS	SIN BECA	15	62	34	38			
8	DOMINGO SANCHEZ DAVID ISAAC	COMERCIO EXTERIOR	BECA	66	11	56	60			
9	FERNANDEZ GOMEZ NICOLAS IONICO	RECURSOS HUMANOS	SIN BECA	19	70	67	29			
10	ESPINOZA SOROLA PAULA MARCELA	COMERCIO EXTERIOR	SIN BECA	39	65	30	55			
11	FUENTES CUBIA FRANCISCA YOLANDA	RECURSOS HUMANOS	SIN BECA	65	51	31	38			
12	GARCIA MARIQUEZ YESSEN ALEJANDRO	LOGISTICA	SIN BECA	29	37	30	41			
13	GONZALEZ GOMEZ JUAN LEANDRO	COMERCIO EXTERIOR	BECA	52	51	70	66			
14	GONZALEZ SALGADO JAVIERA KINACIA	RECURSOS HUMANOS	SIN BECA	46	35	24	35			
15	LOPEZ LABA AILICA CRISTINA	ADMINISTRACION	SIN BECA	70	54	58	39			
16	MARIQUEZ ALVAREZ CARLA ELIZABETH	LOGISTICA	SIN BECA	14	45	53	64			
17	MELIAS GARCIA VICTOR IONICO	RECURSOS HUMANOS	BECA	60	50	57	53			
18	MOLINA CONTRERAS ENRIQUE ANDRES	ADMINISTRACION	SIN BECA	70	31	64	34			
19	MORAGA AMADO NICOL ALEJANDRA	ADMINISTRACION	SIN BECA	65	69	17	23			
20	MONTANO ROSA RINACIO ANDRES	LOGISTICA	BECA	50	52	56	67			
21	MUNOZ MANDUANO LORENA ALEJANDRA	ADMINISTRACION	SIN BECA	25	30	29	60			
22	NAVARRETE ADASME FABIAN ANDRES	LOGISTICA	SIN BECA	59	56	56	38			
23	OMAR RO CONTRERAS VICTORIA FERNANDA	COMERCIO EXTERIOR	SIN BECA	68	70	55	30			
24	RIVAS VIVAS ITALO ANTONIO	RECURSOS HUMANOS	SIN BECA	39	60	66	37			
25	ROJAS LILLO JAIRO TORIBIO	LOGISTICA	SIN BECA	36	58	35	62			
26	SALAZAR ARANGUIZ PABLO ANDRES	COMERCIO EXTERIOR	SIN BECA	24	62	41	22			



Las hojas de cálculo son herramientas esenciales en la ciencia de datos para organizar, analizar y limpiar datos.

Una hoja de cálculo esta organizada en forma matricial, cada matriz es una hoja: columnas (verticales) filas (horizontales). Este concepto es importante porque los datos se ajustan a este tipo de estructuras.

Filas: representan la información relacionada a una misma entidad

Columnas: son los identificadores o variables que se relacionan a una entidad en una fila.

3.2.1 Operaciones básicas con Excel

Excel es una de las hojas de cálculo más populares. Para manipular Excel es indispensable conocer las acciones de desplazamiento en cada hoja:

- Moverse entre celdas continuas

- Ir a una referencia absoluta

- Ir al principio y fin de la hoja o de los datos

- Ir al principio y final de las filas y columnas con o sin datos

3.2.2 Funciones básicas

Las funciones de mas uso frecuente y algunas que se pueden utilizar en ciencia de datos tenemos algunas:

PROMEDIO: Calcula el promedio de un conjunto de valores.

MEDIANA: Devuelve el valor medio de un conjunto de valores.

MÁXIMO: Devuelve el valor máximo de un conjunto de valores.

MÍNIMO: Devuelve el valor mínimo de un conjunto de valores.

CONTAR: Cuenta el número de celdas que contienen valores.

SUMA: Suma un conjunto de valores

1. Función **SUMA:**

Uso Básico: para sumar un rango específico de celdas. Por ejemplo, =SUMA(A1:A10) suma los valores en las celdas A1 a A10.

Suma Automática: Exploraremos cómo Excel puede sugerir automáticamente el

2. Función **PROMEDIO:** Para calcular el promedio de un rango de celdas. Por ejemplo, =PROMEDIO(A1:A10) calcula el promedio de los valores en las celdas A1 a A10.

Decimales y Precisión: Veremos cómo ajustar la precisión del resultado, incluyendo la cantidad de decimales que deseas mostrar.

3. Funciones **MAX y MIN:**

Función MAX: para encontrar el valor más grande en un rango de celdas. Por ejemplo, =MAX(A1:A10) devuelve el valor más grande en las celdas A1 a A10.

Función MIN: para encontrar el valor más pequeño en un rango de celdas. Por ejemplo, =MIN(A1:A10).

Otro grupo de funciones primarias:

En ciencia de datos: permiten realizar el análisis estadístico de información, ya que este requiere de fórmulas para obtener la media, varianza mediana, desviación estándar y otras. Las principales funciones estadísticas comúnmente utilizadas en Excel son:

PROMEDIO, CONTAR, FRECUENCIA, MAX, MEDIANA, MIN y MODA

3.2.2 Algunas operaciones con hojas de calculo

1. **Filtros:** Para explorar datos, aplicar filtros en las columnas en una o varias, con criterios simples o expresiones complejas.
2. **Ordenar:** Si necesitas ordenar datos alfabéticamente o numéricamente, ascendente o descendente, por una o varias columnas. Muy importante para realizar posterior análisis y agrupamiento.
3. **Buscar / Reemplazar:** Esta función te permite buscar un valor específico en una columna y reemplazarlo.
4. **Convertir datos en TABLA** para que Excel le asocie operaciones de manipulación más precisas.

Operaciones de búsqueda en Excel

*Funciones de búsqueda

1. VLOOKUP (Buscar y reemplazar valor): Busca un valor en una tabla y devuelve el valor correspondiente en una columna específica.

2. INDEX-MATCH (Buscar y reemplazar valor): Busca un valor en una tabla y devuelve el valor correspondiente en una columna específica.

3. SEARCH (Buscar texto): Busca un texto dentro de una celda o rango de celdas.

4. FIND (Buscar texto): Busca un texto dentro de una celda o rango de celdas.
Funciones de reemplazo

1. REPLACE (Reemplazar texto): Reemplaza un texto por otro en una celda o rango de celdas.

2. SUBSTITUTE (Reemplazar texto): Reemplaza un texto por otro en una celda o rango de celdas.

3. LEN (Longitud de texto): Devuelve la longitud de un texto en una celda o rango de celdas. #LARGO

Ejemplos

1. Buscar y reemplazar un valor:

Supongamos que tienes una tabla con los nombres de los empleados y sus respectivos departamentos. Quieres buscar el nombre "John" y reemplazarlo por "Juan".

=VLOOKUP(A2,B:C,2,FALSE) # BUSCARV, BUSCAR

Donde A2 es la celda con el nombre "John", B:C es la tabla con los departamentos, y 2 es la columna que contiene los departamentos.

2. Buscar un texto:

Supongamos que tienes una celda con un texto y quieres buscar la palabra "Hola".

=SEARCH("Hola",A1)

Donde A1 es la celda con el texto.

3. Reemplazar un texto:

Supongamos que tienes una celda con un texto y quieres reemplazar la palabra "Hola" por "Adiós".

=REPLACE(A1,"Hola","Adiós") #REEMPLAZAR

Donde A1 es la celda con el texto.



Referencias:

Dataset

<https://keepcoding.io/blog/que-son-datasets/>

<https://archive.ics.uci.edu/dataset/53/iris>

<https://openwebinars.net/blog/datasets-que-son-y-como-acceder-a-ellos/>

<https://thedataschools.com/que-es/data-set/>

Excel

<https://dominaexcel.org/lo-basico-que-debes-saber-de-excel/>

[Lo básico que debes saber de Excel: guía completa para principiantes - Domina Excel](#)

<https://academiaaprenderhaciendo.com/introduccion-a-excel/#:~:text=Conceptos%20Clave%3A%201.%20Celdas%20y%20Rangos%3A%20,F%C3%B3rmulas%20y%20Funciones%3A%203%203.%20Hojas%20de%20C%C3%A1lculo%3A>

<https://dominaexcel.org/funciones-de-analisis-de-datos-en-excel/>

<https://excel-dashboards.com/es/blogs/blog/excel-tutorial-data-cleaning#:~:text=Tutorial%20de%20Excel%3A%20c%C3%B3mo%20hacer%20la%20limpieza%20de,filtrar%20para%20identificar%20y%20limpiar%20datos%208%20Conclusi%C3%B3n>

Fecha Creación	Enero 25 2024
Responsable	Plinio Neira Vargas
Revisado por	Sonia Escobar
Fecha Revisión	Febrero 10 2024