ACTIVIDAD #3

Tipo actividad: Exploración de datos

Actividad para el desarrollo de la sesión

En esta sesión se realizarán las explicaciones de cada fase junto con una pequeña práctica a efectuar utilizando Google sheets o Excel como herramienta de análisis y de visualización de la información.

Para iniciar necesitamos el archivo de datos de flores iris. Este conjunto de datos consta de 150 observaciones de iris, un tipo de flor distribuida en tres especies diferentes, las cuales son: iris setosa, iris versicolor e iris virginica. Cada especia está representada por 50 muestras, Para cada flor los investigadores registraron cuatro características distintivas de las flores que son:

Longitud del sépalo en centímetros.

Ancho del sépalo en centímetros.

Longitud del pétalo en centímetros.

Ancho del pétalo en centímetros.

Es uno de los conjuntos de datos más comunes ya que es fácil trabajar con el para el estudio de conceptos de ciencias de datos.

El conjunto de datos se puede obtener desde:

https://www.kaggle.com/datasets/arshid/iris-flower-dataset/download?datasetVersionNumber=1

Es un archivo en formato .csv.

Exploración de Datos:

La exploración de datos es la primera etapa esencial en cualquier proyecto de análisis de datos. En las actividades de ejemplo, nos sumergimos en el conjunto de datos de los Mundiales de Fútbol para entender su estructura y características clave. Comenzamos identificando las variables presentes, analizando los tipos de datos que contienen y examinando la distribución general. La visualización juega un papel crucial en esta exploración, permitiéndonos obtener una visión inicial de patrones, tendencias y posibles anomalías. A través de gráficos y diagramas, como histogramas y diagramas de dispersión, podemos revelar información valiosa sobre la naturaleza de los datos y plantear preguntas específicas que guiarán el análisis subsiguiente.

Antes de realizar cualquier proceso de analítica de datos es crucial el análisis exploratorio de los datos para conocer qué información hay disponible, que variables están involucradas y qué modelan los datos existentes. También es importante contar con asesoría de expertos en el dominio de los datos. Es decir, si se están analizando datos financieros de una empresa, el contexto de las operaciones y las decisiones que muestran los datos se puede obtener indagando al experto en el tema. Con eso se pueden identificar características en los datos y se puede tener una idea previa antes de la exploración.

Exploración del dataset irirs: Como parte de la actividad se debe descargar el dataset de irirs (archivo .zip) y descomprimirlo para poder obtener un archivo en formato csv. Un archivo CSV (Comma-Separated Values, por sus siglas en inglés) es un formato de archivo de texto que se utiliza para almacenar datos tabulares. En un archivo CSV, cada línea representa una fila de datos, y los valores de cada columna están separados por comas u otro delimitador especificado, como punto y coma o tabulación. Este formato es ampliamente utilizado debido a su simplicidad y facilidad de lectura y escritura tanto para humanos como para máquinas.

Los archivos CSV son compatibles con una variedad de aplicaciones y herramientas de software, como hojas de cálculo (como Microsoft Excel y Google Sheets), bases de datos y programas de análisis de datos. Debido a su

simplicidad y portabilidad, los archivos CSV son una elección común para compartir y transferir conjuntos de datos entre diferentes plataformas y sistemas

Para abrir el conjunto de datos y realizar un análisis exploratorio se debe ir a un navegador y colocar la dirección web sheets.new, esta página creará una nueva hoja de cálculo de Google sheets (ver figura 2)

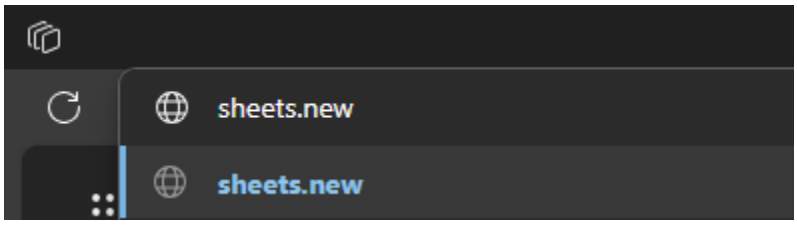


Figura 2: acceso a Google sheets

Al abrir un nuevo libro de Google sheets, se va al menú archivo y luego escogemos la opción importar como se ve en la figura 2:

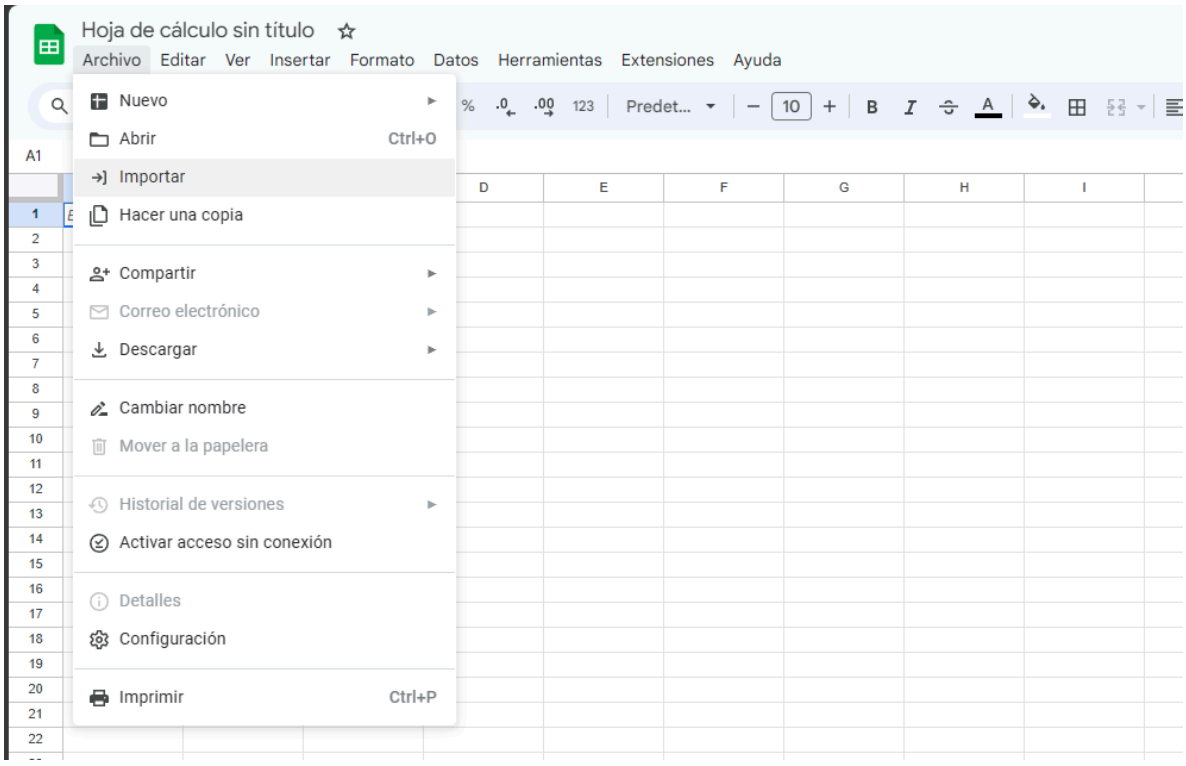


Figura 2: importar datos a Google sheets.

Posteriormente se selecciona la pestaña subir y se navega hasta encontrar el archivo descomprimido llamado iris.csv (figura 3).

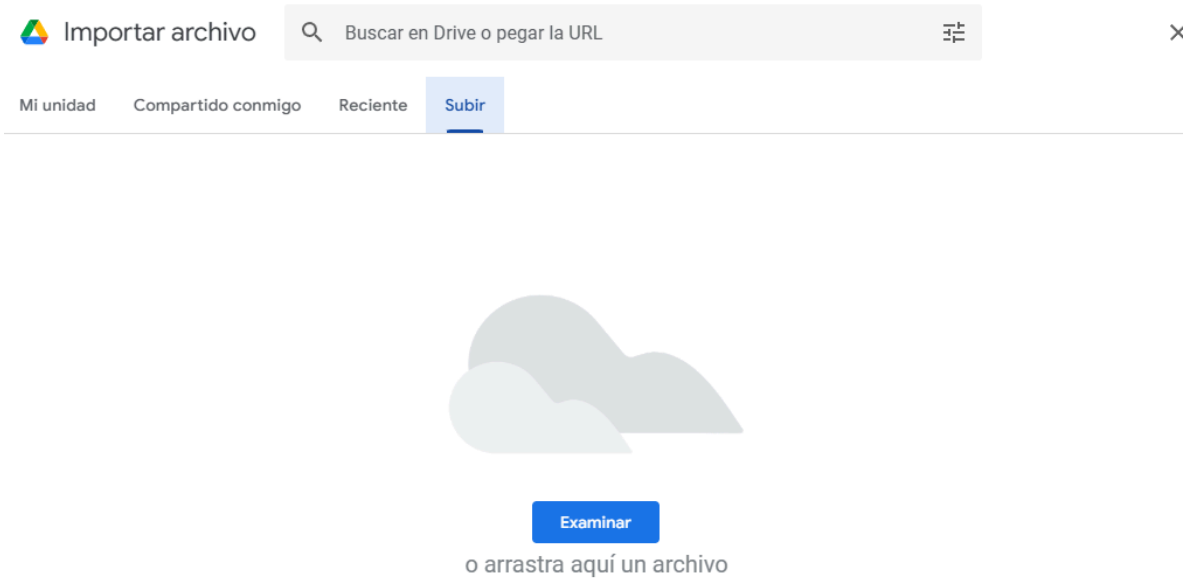
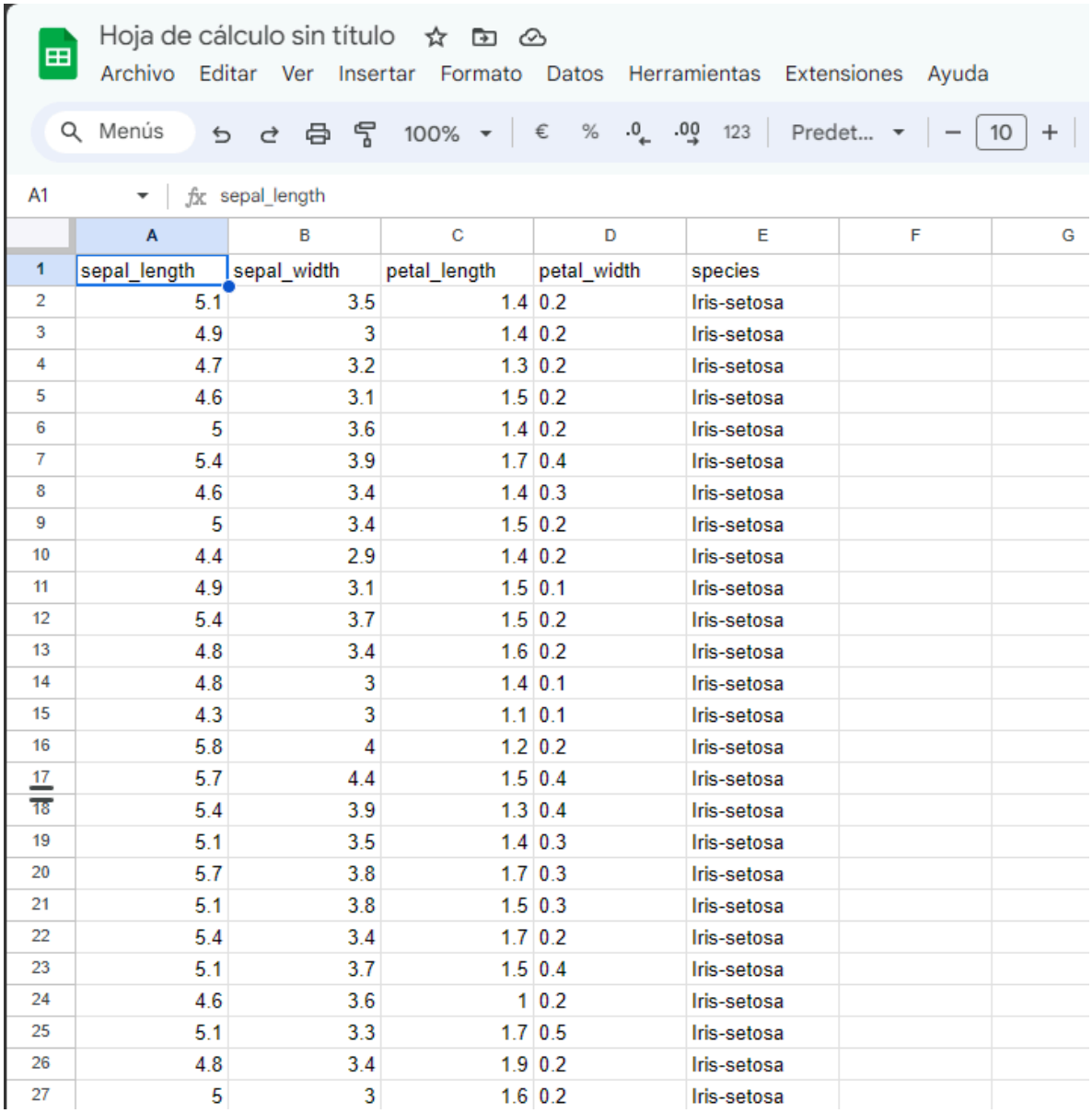


Figura 3: selección del archivo a cargar.

Una vez el archivo es cargado, se debe permitir que Google sheets haga el preprocesamiento de datos, aceptando las opciones por defecto (figura 4).

Figura 4: opciones de preprocesamiento de datos.

Al hacer clic en el botón importar datos de la figura 4, se deben ver los datos en forma de hoja de datos, como se aprecia en la figura 5.



	A	B	C	D	E	F	G
1	sepal_length	sepal_width	petal_length	petal_width	species		
2	5.1	3.5	1.4	0.2	Iris-setosa		
3	4.9	3	1.4	0.2	Iris-setosa		
4	4.7	3.2	1.3	0.2	Iris-setosa		
5	4.6	3.1	1.5	0.2	Iris-setosa		
6	5	3.6	1.4	0.2	Iris-setosa		
7	5.4	3.9	1.7	0.4	Iris-setosa		
8	4.6	3.4	1.4	0.3	Iris-setosa		
9	5	3.4	1.5	0.2	Iris-setosa		
10	4.4	2.9	1.4	0.2	Iris-setosa		
11	4.9	3.1	1.5	0.1	Iris-setosa		
12	5.4	3.7	1.5	0.2	Iris-setosa		
13	4.8	3.4	1.6	0.2	Iris-setosa		
14	4.8	3	1.4	0.1	Iris-setosa		
15	4.3	3	1.1	0.1	Iris-setosa		
16	5.8	4	1.2	0.2	Iris-setosa		
17	5.7	4.4	1.5	0.4	Iris-setosa		
18	5.4	3.9	1.3	0.4	Iris-setosa		
19	5.1	3.5	1.4	0.3	Iris-setosa		
20	5.7	3.8	1.7	0.3	Iris-setosa		
21	5.1	3.8	1.5	0.3	Iris-setosa		
22	5.4	3.4	1.7	0.2	Iris-setosa		
23	5.1	3.7	1.5	0.4	Iris-setosa		
24	4.6	3.6	1	0.2	Iris-setosa		
25	5.1	3.3	1.7	0.5	Iris-setosa		
26	4.8	3.4	1.9	0.2	Iris-setosa		
27	5	3	1.6	0.2	Iris-setosa		

Figura 5: datos importados para explorar.

Como actividad se propone realizar la exploración previa de los datos, separando manualmente los 150 registros en tres grupos por especie, cada uno en una hoja diferente.

Una vez separados, se propone como actividad calcular el promedio de cada una de las cuatro características y comparar tales promedios entre las tres especies. Posteriormente discutir los resultados con los estudiantes.

Con esta información contestar a las siguientes preguntas.

1. ¿Qué pasos debería tener el análisis exploratorio del dataset IRIS?
2. ¿Qué diferencia existe entre las medidas promedio de las tres especies de iris?
3. ¿Si se saben las medidas de una flor iris y conociendo los promedios, se podría determinar a qué especie pertenece esa flor?
4. ¿Si para alguna flor faltara alguno de los datos cómo se puede manejar esta situación?

Discutir con los estudiantes las respuestas orientando sus análisis a las observaciones del conjunto de datos.