Autor: Joaquín León Martínez

1. **Descargue el dataset IRIS**
   Done, archivo adjunto.

2. **Describir la base de datos: información, variables o atributos, variable objetivo y número de patrones por cada clase. Puede realizar una búsqueda bibliográfica.**

   En este dataset podemos encontrar un total de 5 atributos si tenemos en cuenta la variable objetivo. Son los siguientes:
   - sepal.length: Indica la longitud del sépalo del individuo.
   - sepal.width: Indica la anchura del sépalo del individuo.
   - petal.length: Indica la longitud del pétalo del individuo
   - petal.width: Indica el ancho del pétalo.
   - variety: Indica la variedad o tipo de flor de la que se trata.

   El conjunto Iris dataset ha sido ampliamente utilizado en la comunidad de aprendizaje automático y ha servido como un problema de clasificación de referencia para evaluar algoritmos y técnicas de clasificación.

   Lo que lo hace ideal es que se trata de un conjunto de datos relativamente pequeño pero desafiante, lo que lo hace ideal para la experimentación y la enseñanza de conceptos básicos de clasificación de patrones. Como hemos podido ver en la descripción de los atributos, la mayoría de estos son numéricos excluyendo la variable objetivo. Esto hace que conceptualmente sea muy fácil de entender.

   En el conjunto de datos Iris, hay un total de 150 patrones distribuidos en tres clases o especies de Iris. Cada clase contiene 50 patrones. (Esto también se indica en la pregunta 7)

3. Por cada uno de los atributos, calcular: (Esta todo en el Excel adjunto aunque se explique aqui)

a. Media: Para calcular la media se ha realizado la suma de todos los valores y se ha dividido entre el numero de valores. Ejemplo: =SUM(A2:A151)/COUNT(A2:A151)

b. Desviación típica: Este ha sido un poco mas complicado, en primer lugar hace falta calcular la varianza para después realizar la raíz cuadrada. Para calcular la varianza he creado una columna mas por cada atributo, en estas columnas se ha almacenado la diferencia de este valor con respecto a la media al cuadrado. Ejemplo: (=(A2-H$2)^2) (Siendo A2 el valor de la entrada en concreto y H$2 la celda con el valor de la media)

c. Mediana: En este caso ha sido necesario ordenar los datos de menor a mayor, tras esto se ha calculado la posición del medio dividiendo el numero de datos+1 entre 2 y consultando dicha posición en la columna con los datos ordenados.

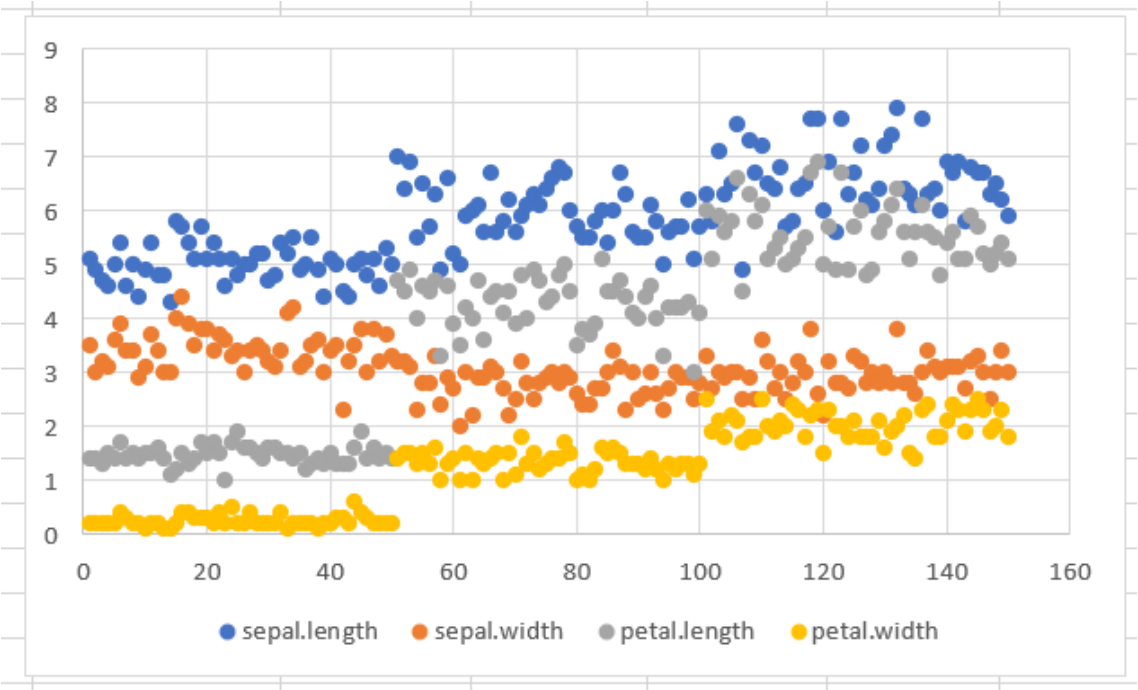
d. Cuartiles: Aprovechando que tenemos los datos ordenados del apartado anterior el proceso ha sido el mismo, para el primer cuartil en lugar de dividir entre 2 se ha multiplicado por 0.25. Para el segundo cuartil he reutilizado la media y para el tercer cuartil he multiplicado por 0.75.

e. Mínimo y máximo: He utilizado las funciones Large y Small, aunque realmente teniendo los datos ordenados basta con obtener el primero y el ultimo respectivamente.

A continuación, se muestra una captura de pantalla con los resultados obtenidos (se puede consultar también el Excel adjunto)

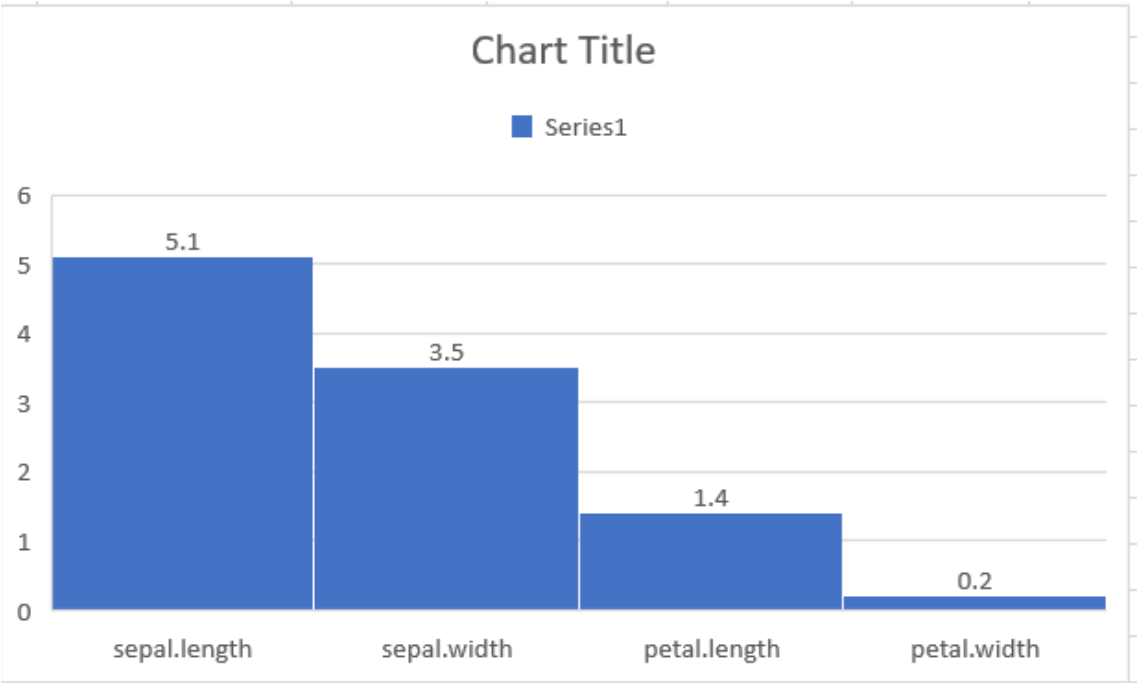
	F	G	H	I	J	K	L	
ty			sepal.length	sepal.width	petal.length	petal.width		Para ca
a		Media	5.84	3.06	3.76	1.20		
a		Varianza	0.69	0.19	3.12	0.58		
a		Desviacion tipica	0.83	0.44	1.77	0.76		
a		Mediana	6.20	3.00	4.30	1.30		
a		Cuartil 1	5.30	2.80	1.40	0.30		
a		Cuartil 2	6.20	3.00	4.30	1.30		
a		Cuartil 3	6.30	3.30	5.60	1.80		
a		Minimo	4.3	2	1	0.1		
a		Maximo	7.9	4.4	6.9	2.5		
a								
a								

4. Visualice la relación entre atributos. ¿Hay alguna relación que sea visualmente significativa?



Quizá se podría decir que hay una relación directa entre el petal length y el petal width.

5. Realice un histograma de los atributos de entrada y la variable objetivo.



No acabo de comprender el sentido de realizar un histograma en este caso en concreto. Tampoco he sabido como añadir la variable objetivo dentro del histograma, ya que no tiene valores numéricos.

6. ¿Qué técnica utilizaría para recuperar los valores perdidos? En caso de que existan, aplíquela.

En este caso no hay datos perdidos, pero una opción seria intentar identificar patrones o simplemente calcular la media o moda de los datos que no están perdidos, de esta forma el resultado final no debería verse demasiado afectado.

7. ¿Es un dataset balanceado?

Si, ya que de los tres posibles valores que puede haber hay un tercio de los datos para cada uno. (50 de cada tipo, haciendo un total de 150 muestras)