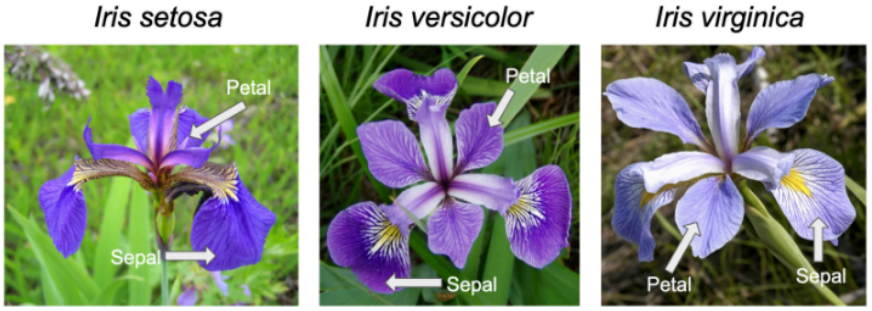
TAREA FINAL - Introducción a la Ciencia de Datos 2024

Agustina Añasco, Horacio Solé

Maestría en Bioinformática

Para este proyecto se utilizó el dataset IRIS, un pequeño conjunto de datos disponible en la *UC Irvine Machine Learning Repository*. Está organizado en un formato tabular, donde cada fila representa una instancia (el registro de una flor de iris) y cada columna representa un atributo (número de la flor, largo del sépalo, ancho del sépalo, largo del pétalo y ancho del pétalo, todos ellos medidos en centímetros, y la columna *target* cuyos valores van del 0 al 2, siendo cada uno un código de una de las tres clases: *setosa* (0), *versicolor* (1) y *virginica* (2). En total, el conjunto de datos presenta 150 instancias (filas, numeradas del 0 al 149) y 5 columnas (atributos).



En este dataset, todos los atributos son de tipo categórico ya que se trata de medidas ancho y largo, siendo estas medidas de tipo continuas.

A partir de conocer este pequeño dataset, surge la siguiente pregunta:

> **¿Es posible diferenciar las especies de iris con alguna de las características que aparecen en este dataset?**

Para resolver esto, lo primero que realizamos fue un análisis exploratorio inicial de los datos.

1. Análisis exploratorio de datos (EDA)

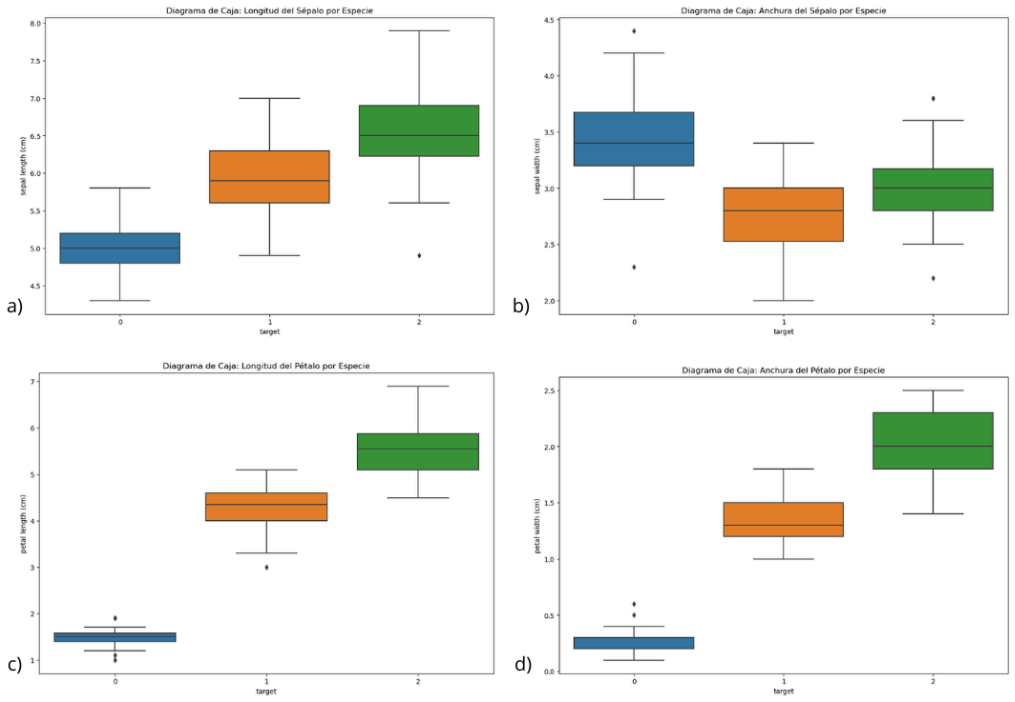
Estadísticas descriptivas

Podemos ver que las medidas de longitud y ancho del pétalo tienden a ser más variables (por su mayor desviación estándar) que las medidas del sépalo, lo que sugiere una mayor diversidad morfológica en el pétalo entre las diferentes especies de Iris. Además, se marca una alta correlación entre la longitud y la anchura del pétalo con las especies, por lo que suponemos que podrían tratarse de potenciales índices para diferenciarlas. La tabla no presenta valores nulos y al parecer, tampoco redundantes.

```
:Summary Statistics:

============== ==== ==== ======= ===== ====================
                Min  Max   Mean    SD   Class Correlation
============== ==== ==== ======= ===== ====================
sepal length:   4.3  7.9   5.84   0.83    0.7826
sepal width:    2.0  4.4   3.05   0.43   -0.4194
petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)
petal width:    0.1  2.5   1.20   0.76    0.9565  (high!)
============== ==== ==== ======= ===== ====================
```

Se realizaron los diagramas de caja para cada uno de los atributos por especie:



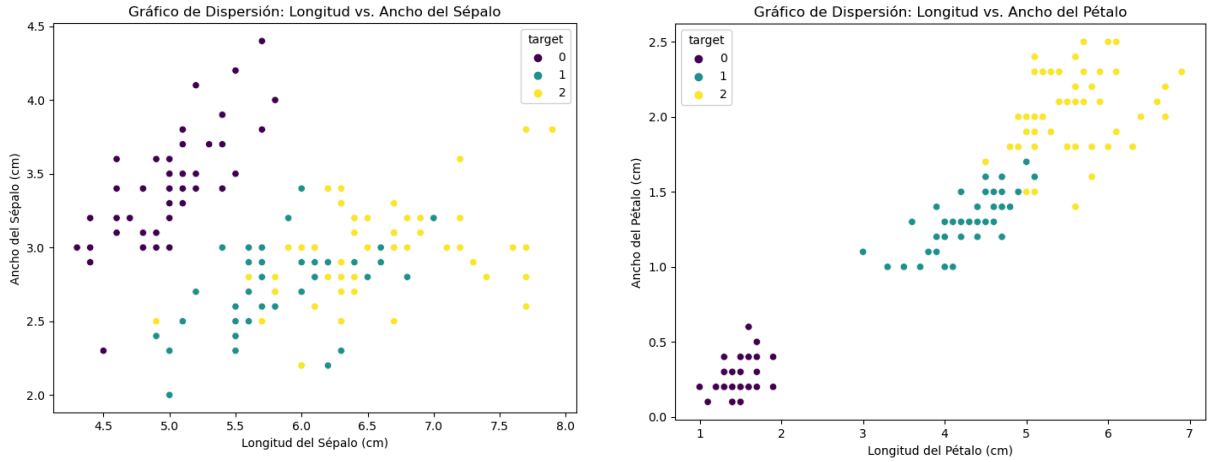
a) longitud de los sépalos, b) ancho de los sépalos, c) longitud de los pétalos, d) ancho de los pétalos por especie.

A grandes rasgos, viendo la gráfica a) podemos notar que existe cierta superposición de la longitud de los sépalos entre las tres especies, pero la especie *setosa* tiende a tener sépalos más cortos y la versicolor los más largos. Si bien la clase *setosa* presenta una mediana alrededor de 5,0 cm, la *versicolor* cercana a 6,0 cm y la *virginica* 6,5 cm, se observan algunos valores atípicos en estos registros.

La gráfica b) muestra que la especie *setosa* presenta los sepalos más anchos (mediana: 3,4 cm), sin embargo, presenta valores dispersos. Los sépalos de la especie *versicolor* son los más finos (mediana: 2,8 cm), siendo bastante cercanos a los *virginica* (mediana: 3,0 cm).

Las gráficas c) y d) pertenecen al largo y ancho de los pétalos, respectivamente. Estas dos son características muy distintivas entre las tres clases, *setosa* tiene pétalos mucho más cortos y anchos que las otras dos especies, lo que se observa como una notoria separación de las cajas en cada gráfico.

Diagrama de dispersión correspondiente a la longitud y ancho de los sépalos y pétalos por especies:



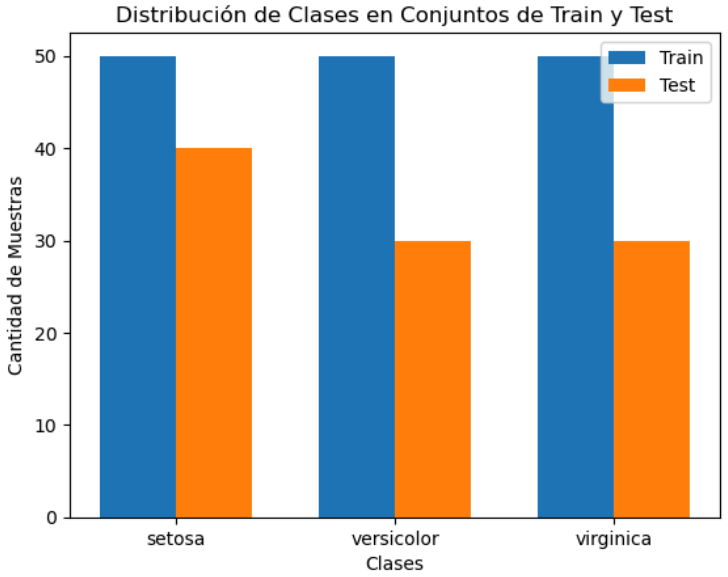
En la gráfica de la izquierda, observamos cómo las plantas de la especie *setosa* (0) se diferencian del resto de las especies, teniendo sépalos más cortos y anchos, mientras que las especies *versicolor* (1) y *virginica* (2) se superponen en la zona del gráfico donde el sépalo mide entre 4,5 y 7,0 cm de largo y entre 2,0 y 3,5 cm de ancho, sin embargo, algunas plantas de la especie *virginica* presentan sépalos más largos y anchos.

En la gráfica de la derecha, al analizar los pétalos, se observa que las *setosas* presentan pétalos mucho más pequeños tanto en largo como ancho que las otras especies, seguida por los Iris *versicolor* y por último los de especie *virginica*, que presentan los pétalos más grandes. Verificamos que las medidas de los pétalos son una característica bastante eficiente para hacer distinciones entre las diferentes especies, tal como se vió en las gráficas de caja anteriormente.

2. Aprendizaje Automático - Uso de Random Forest

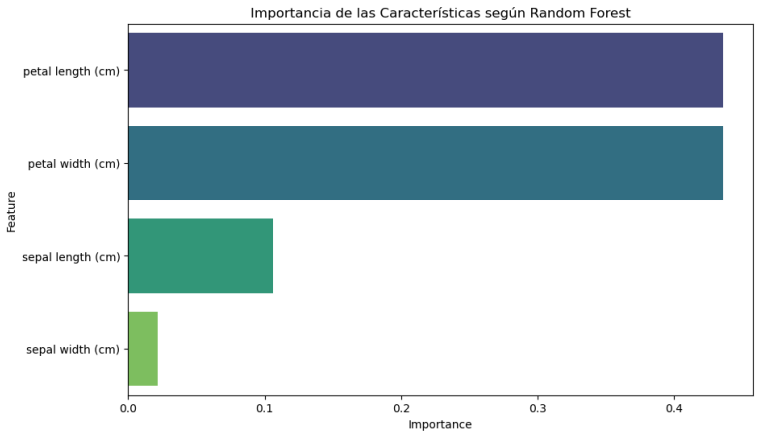
Distribución de las clases

En el dataset hay exactamente 50 muestras para cada una de las tres clases, por lo que las clases están balanceadas, lo que en la gráfica se ve representado como “Train” (Conjunto de entrenamiento). En el “Test” o conjunto de prueba se utilizaron valores ejemplo de 40, 30 y 30 para cada una de las clases.

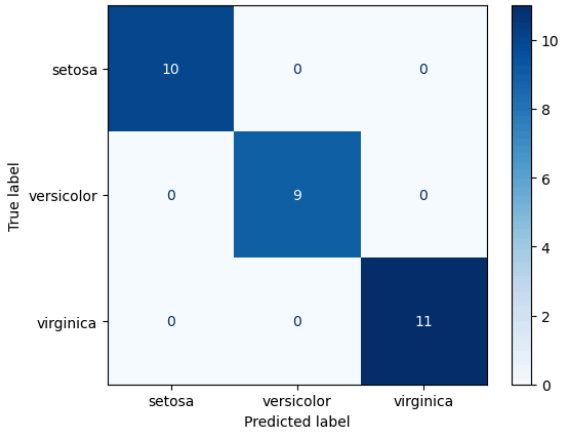


Análisis con algoritmos de Aprendizaje Automático

Debido a las características del conjunto de datos Iris, se eligió el modelo Random Forest. Este modelo es adecuado para analizar múltiples características, como la longitud y el ancho de los pétalos, ya que puede capturar relaciones complejas entre estas variables. Además, nos permite evaluar la importancia de los atributos en la clasificación, lo que es muy útil en este caso. Otra ventaja es que, debido a que el conjunto de datos Iris no presenta grandes desequilibrios entre las clases, Random Forest no es tan propenso a sesgos y sobreajuste como otros modelos. Sin embargo, para asegurar la efectividad del análisis, también se aplicaron algunas técnicas adicionales de evaluación y validación.



En concordancia con lo obtenido en resultados anteriores, vemos que el largo y ancho de los pétalos son las características que más contribuyen a la diferenciación entre las especies de iris (ambas contribuyen en un 43,6% a la predicción).
Los resultados de los análisis indican que la precisión del modelo es del 100% (*accuracy:* 1,00), lo cual sugiere que el modelo puede generalizar bien sobre nuevos datos. Debido a que los resultados de *precisión* y *recall* para las tres clases fueron de 1,00 podemos decir que todas las predicciones fueron correctas. Estos resultados indican, como se evidenció anteriormente, que el dataset está balanceado, facilitando al modelo distinguir entre las clases sin sesgos hacia una clase particular y a su vez, que el modelo es eficiente.



Accuracy del modelo: 1.00

	precision	recall	f1-score	support
setosa	1.00	1.00	1.00	10
versicolor	1.00	1.00	1.00	9
virginica	1.00	1.00	1.00	11
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

Se realizaron además los resultados mediante validación cruzada para tener una idea del rendimiento del modelo, utilizando 5 particiones del conjunto de datos. Observando esto en conjunto con la curva de aprendizaje obtenida, vemos que la precisión del conjunto de entrenamiento es del 100% y el promedio de los puntajes de precisión del conjunto de prueba fue del 96,67% en las diferentes particiones, indicando cierta consistencia en los datos, con valores muy cercanos entre sí, por lo que consideramos a Random Forest un modelo preciso en general, y no hay un sobreajuste. En la curva, de color verde, los reustados de la precisión en el conjunto de Validación cruzada convergen con la precisión del conjunto de entrenamiento, sugirieindo un buen ajuste del modelo y que hay sobreajuste ni subajuste, lo que refuerza la conclusión de que el modelo generaliza bien los datos.