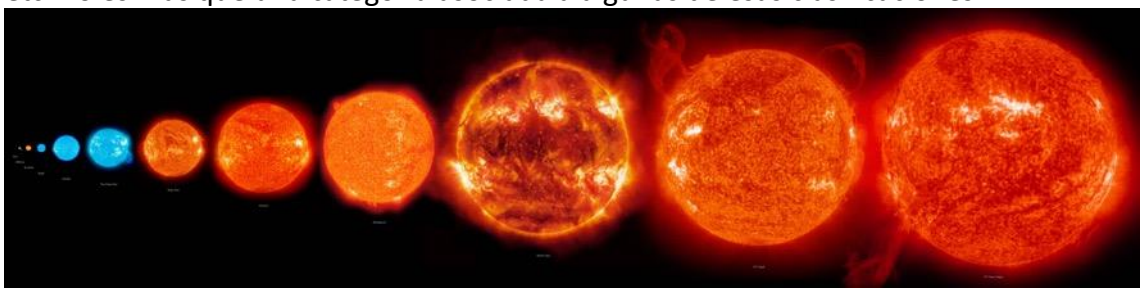
Práctica 2
DETERMINACIÓN DE TIPOS DE ESTRELLAS

2,5 puntos

INTRODUCCIÓN

En astronomía la clasificación de estrellas en tipos permite realizar hipótesis acerca de su dinámica, establecer cómo y cuándo se formaron, su composición y cuál será su evolución y fin. El nombrar a las estrellas como enanas blancas, tipo sol, supergigantes, etc. no es más que una categoría asociada a algunas de esas clasificaciones.



Se propone en este trabajo obtener los tipos de estrellas utilizando técnicas de aprendizaje automático no supervisado. Para lo cual, se tiene un conjunto de datos de 240 estrellas con los siguientes atributos:

- **Temperature**: Temperatura promedio de la superficie en grados K
- **L**: Luminosidad comparada con la del Sol.
- **R**: Radio comparado con la del Sol.
- **A_M**: Magnitud absoluta (brillo aparente de la estrella si estuviera a 10 parsec de distancia)
- **Spectral_Class**: Clasificación espectral: es un valor que identifica la presencia elementos químicos en el espectro de la estrella. Es una secuencia (O, B, A, F, G, K, M) que se asocia a las estrellas desde las más calientes O, hasta las más frías M
- **Color**: Color principal del espectro

CONSIDERACIONES GENERALES

1. Para realizar la práctica, los estudiantes emplearán un repositorio de código en GitHub. Para ello, cada grupo debe crear un repositorio de código privado y agregar como «colaborador» al profesor de prácticas (que indicará a los estudiantes su nombre de usuario en GitHub). Se espera que cada grupo haga un commit semanal del código de la práctica. Esta parte de la práctica se valorará con 0.5 puntos. Además, también habrá que entregar el cuaderno (notebook) final a través de Aula Global.
2. Se proporciona un fichero de datos: “starts2.csv”

3. Los resultados deben ser reproducibles. Por lo tanto, hay que fijar la semilla de números aleatorios en los lugares adecuados. Se usará como semilla el NIA de uno de los miembros del grupo o bien el número del grupo de prácticas. Si hay que utilizar más semillas se usan números consecutivos al NIA de base.

QUÉ HACER

1. Programar k-means, utilizando buenas prácticas de programación, es decir, debe ser una función y comparar los resultados y la eficiencia con la implementación de k-means en scikit-learn.
2. Tenemos variables categóricas (Color, Clase Espectral). Tenemos dos posibilidades: (a) codificar con one-hot-encoding; (b) codificar como una variable ordinal. Tener en cuenta que el color está asociado a la cantidad de energía, y algo parecido con la clase espectral.
3. Aplicar diferentes algoritmos de clustering comparando y discutiendo los resultados que se obtienen de ellos (**al menos dos métodos**)
4. Discutir los resultados que se obtienen si a los atributos categóricos se les aplica one-hot-encoding o se asigna un valor numérico a las secuencias (variable ordinal).
5. A partir de los resultados obtenidos, ¿qué pipeline de clustering, es decir, qué transformaciones de datos, algoritmo, con sus hiperparámetros, transformación de datos y análisis de resultados recomendaría realizar?

Estas son clases y atributos asociados a las estrellas que la astronomía utiliza habitualmente:

Clase	Temperatura	L	R	A_M	Color	Clase Espectral
Enana roja	3.000	$7{,}0 \cdot 10^{-4}$	$1{,}0 \cdot 10^{-1}$	+17.5	rojo	K-M
Enana marrón	3.300	$5{,}5 \cdot 10^{-3}$	$3{,}5 \cdot 10^{-1}$	+12.5	rojo	M
Enana blanca	14.000	$2{,}5 \cdot 10^{-3}$	$1{,}0 \cdot 10^{-2}$	+12.6	blanca	B-G
Estrella en secuencia principal	16.000	$3.2 \cdot 10^{4}$	4,4	-0.4	blanca-amarilla	B-M
Super gigante	15.000	$3.0 \cdot 10^{5}$	$5{,}0 \cdot 10^{1}$	-6.4	blanca-amarilla	B-M
Hiper gigante	11.000	$3.0 \cdot 10^{5}$	$1{,}4 \cdot 10^{3}$	-9.6	amarilla	B-M

6. ¿Hay similitudes con los grupos obtenidos en el punto 4? Explicar

QUÉ ENTREGAR

Entregar un único notebook con:

1. La implementación propia de k-means y discusión de la comparación con k-means de scikit-learn
2. Comparación y discusión de diferentes algoritmos de clustering. Comparación de los resultados obtenidos con las dos formas de tratar los atributos categóricos
3. Pipeline recomendado para realizar el clustering de los datos.

4. Discusión de la comparación de los grupos obtenidos del punto 3 y los utilizados en astronomía