**PREWORK**

**SESIÓN 05**

**Objetivo:**

En esta sesión trataremos acerca de nuestro primer algoritmo supervisado de clasificación: Árboles de decisión y Random Forests. Para ello, necesitaremos estudiar los siguientes pre-requisitos

**Pre-requisitos:**

En ciencias computacionales, un árbol es una herramienta que nos permite separar información en dos o más partes. Una vez dividida la información en varias partes, es más sencillo atacar cada una por separado que atacar toda la información junta. Esta clase de problemas se le conoce como problemas de *dividir-y-conquistar* (*divide-and-conquer*).

Por ejemplo: Supongamos que queremos encontrar un dato en una lista enorme de datos que se encuentran ordenados. Por ahora, imaginaremos que la lista es:

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

Y el dato que queremos encontrar es 3.

Si la lista es enorme, es una mala idea ir uno por uno de los datos, nos vamos a tardar muchísimo computacionalmente hablando, en cambio, lo que podemos hacer es dividir la lista en dos partes iguales:

Lista de datos <= 5 es: [1, 2, 3, 4, 5]

Lista de datos > 5 es: [6, 7, 8, 9, 10]

Como sabemos dónde dividimos la lista (en el numero 5) podemos saber que lista descartar y con qué lista seguir (como 3 < 5, usamos la lista de datos <= 5). Con esto podemos dividir la lista una y otra vez hasta que nos quedemos con pocos datos y podamos buscar el dato rápidamente.

La división puede hacerse con poner cualquier condición, por ejemplo, decir que en una lista pondremos los números pares, y en otra los nones:

Lista de pares: [2, 4, 6, 8, 10]

Lista de nones: [1, 3, 5, 7, 9]

Todo depende de cual sea el criterio de división, sin embargo un árbol trata de dividir lo más posible los datos que tiene a la mano para reducir el problema a algo mucho más pequeño o más tratable.

**Quiz:**

**¿Qué significa dividir y conquistar?**

1. Separar un resultado en varias partes.
2. Separar un problema en varias computadoras.
3. **Separar un problema en varias partes más sencillas.**
4. Separar los datos en varios sistemas.

**Si tenemos la lista [a, b, c, d, e, f], ¿cuál sería una buena división?**

1. [a, c, e, f] , [b, d]
2. [a], [b], [c, d, e, f]
3. **[a, b, c], [d, e, f]**
4. Todas las anteriores.

**¿Por qué es una mala idea buscar un dato uno por uno?**

1. Utiliza demasiada memoria.
2. **Utiliza demasiado tiempo.**
3. Es demasiado difícil de programar.
4. Es muy aburrido ir dato por dato.

**¿Qué es un criterio de división?**

1. Una forma de separar los datos en entrenamiento y prueba
2. Una manera de saber cómo elegir los datos
3. Una forma de hacer una división entre dos números
4. **Una forma en la cual podemos dividir los datos en partes iguales.**