*Validación de Modelos – Métricas*

**¿Qué buscábamaos al aplicar un algoritmo de Clustering?**

**Diagrama

Descripción generada automáticamente**

Esto significa que:

* Las observaciones dentro de un cluster sean similares entre sí.
* Las observaciones entre clusters sean lo más distintas posibles.

Resulta importante destacar, que los algoritmos de Clustering siempre **“buscan una forma natural de agrupar los datos”**. Veamos algunos ejemplos!

**Ejemplo 1:**

Un dibujo de un perro

Descripción generada automáticamente con confianza media

Un dibujo de un perro

Descripción generada automáticamente con confianza media

**Ejemplo 2:**

Un dibujo de un perro

Descripción generada automáticamente con confianza media Dibujo animado de un personaje de caricatura

Descripción generada automáticamente con confianza media

**Ejemplo 3:**

Un dibujo de un perro

Descripción generada automáticamente con confianza media

Un dibujo de un muñeco de peluche

Descripción generada automáticamente con confianza media

***¿Cómo se analiza un cluster?***

Por cada modelo de cluster hay que describirlo, si tengo 3 clúster no habría demasiado problema, pero si tengo muchos clústers, el proceso se vuelve mucho más complejo!.

**Ejemplo:** Clúster 1 – Clientes nuevos que miran Netflix.

Antigüedad 4 años

Hacen 48 clicks

Miran 5 videos

Otra opción muy habitual, es calcular el promedio de las variables de interés por cada clúster para posteriormente realizar una comparación entre ellos.

*IMPORTANTE!*

No hay una forma natural de agrupar los datos, el clustering es subjetivo, la mejor elección de grupos depende de qué le queremos preguntar a los datos.

***Métricas de calidad en clustering***

Lo primero que tenemos que saber, es que no hay una métrica objetiva como el accuracy o la curva ROC. Esto significa, que el data scientist es el principal “juez” de un algoritmo de Clustering y dicho juicio se va a encontrar claramente condicionado por el problema a resolver.

***Índice de Silhouette***

Silhouette analysis (ó análisis de la silueta) es una métrica que nos ayuda a conocer si un caso está bien asignado o no.

Es un índice y por lo tanto, tiene un rango de valor que puede asumir. La métrica de silhoutte varía entre -1 a +1. Esto significa que entre más cercano a 1 la observación ha sido bien asignada, en consecuencia más cercano a -1 la observación no ha sido bien asignada dentro del cluster.

Gráfico, Gráfico de embudo

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente