**Introduccion**

El término "k-medias" fue utilizado por primera vez por James MacQueen en 1967, aunque la idea se remonta a Hugo Steinhaus en 1957.

​ El algoritmo estándar fue propuesto por primera vez por Stuart Lloyd en 1957 como una técnica para modulación por impulsos codificados, aunque no se publicó fuera de los laboratorios Bell hasta 1982.

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características.

El **clustering**es una tarea que tiene como finalidad principal lograr el agrupamiento de conjuntos de objetos no etiquetados, para lograr construir subconjuntos de datos conocidos como Clusters.

El **Clustering** o la clusterización es un proceso importante dentro del **Machine learning**. Este proceso desarrolla una acción fundamental que le permite a los algoritmos de **aprendizaje automatizado**entrenar y conocer de forma adecuada los datos con los que desarrollan sus actividades.

Esta técnica no supervisada se basa en identificar grupos en los datos de tal manera que todos los datos del grupo (clúster) son datos con características similares mientras que los datos de los otros grupos son diferentes.

**Que es**

K-Means (traducido como K-Medias) es un método de agrupamiento es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos.

Para saber si los datos son parecidos o diferentes el algoritmo K-medias utiliza la distancia entre los datos. Las observaciones que se parecen tendrán una menor distancia entre ellas. En general, como medida se utiliza la distancia euclidiana, aunque también se pueden utilizar otras funciones.

### **Clasificación**

La clasificación automática de objetos o datos es uno de los objetivos del aprendizaje de máquina. Podemos considerar tres tipos de algoritmos:

* **Clasificación supervisada:** disponemos de un conjunto de datos (por ejemplo, imágenes de letras escritas a mano) que vamos a llamar datos de entrenamiento y cada dato está asociado a una etiqueta (a qué letra corresponde cada imagen). Construímos un modelo en la fase de entrenamiento (training) utilizando dichas etiquetas, que nos dicen si una imagen está clasificada correcta o incorrectamente por el modelo. Una vez construído el modelo podemos utilizarlo para clasificar nuevos datos que, en esta fase, ya no necesitan etiqueta para su clasificación, aunque sí la necesitan para evaluar el porcentaje de objetos bien clasificados.
* **Clasificación no supervisada:** los datos no tienen etiquetas (o no queremos utilizarlas) y estos se clasifican a partir de su estructura interna (propiedades, características).
* **Clasificación semisupervisada:** algunos datos de entrenamiento tienen etiquetas, pero no todos. Este último caso es muy típico en clasificación de imágenes, donde es habitual disponer de muchas imágenes mayormente no etiquetadas. Estos se pueden considerar algoritmos supervisados que no necesitan todas las etiquetas de los datos de entrenamiento.

### **Funcionamiento paso a paso de K-Means**

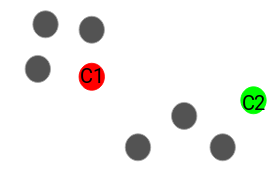
Suponiendo que tenemos los datos de la imagen de abajo, los pasos de ejecución del algoritmo son los siguientes:

1. **Elección del número de clústers k**El primer paso siempre es elegir en cuantas agrupaciones queremos segmentar los datos.



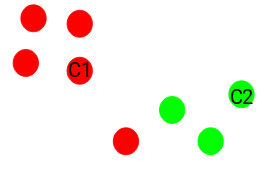
1. **Inicializar las coordenadas de los centroides**

Los centroides se inicializan en coordenadas aleatorias. Suponiendo que tenemos k=2, iniciamos dos centroides, uno rojo y otro verde, en puntos aleatorios de los datos.



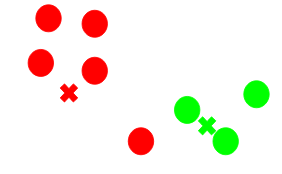
1. **Asignamos cada punto a un clúster**

Se calcula la distancia de cada punto a cada centroide, y se agrupa con aquel centroide más próximo

.

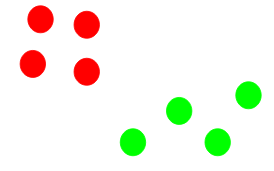
1. **Se recalculan los centroides de los clústers**

Una vez tenemos todos los puntos asignados a un clúster, se recalculan los centroides de manera que vuelven a ser los centros de cada clúster.



1. **Se repiten los pasos 3 y 4 hasta que se llega al criterio de parada.**

El proceso de asignar cada punto a un clúster y calcular los centros se repite hasta que se cumple el criterio de parada estipulado.



## Desventajas del K-Means

Ya hemos visto la potencia que tiene este algoritmo. Por lo sencillo que es de aplicar y la valiosa información sobre nuestros datos que nos aporta. Como no es oro todo lo que reluce, tengo que comentaros también las desventajas que ofrece:

* **Tenemos que elegir *k***nosotros mismos. Es muy posible que nosotros cometamos un error, o que sea imposible escoger una k óptima.
* Es **sensible a outliers**. Los casos extremos hacen que el clúster se vea afectado. Aunque esto puede ser algo positivo a la hora de detectar anomalías.
* Es un algoritmo que sufre de la [**maldición de la dimensionalidad**](https://www.themachinelearners.com/tradeoff-bias-variance/).

### Ejemplos: cuándo se puede utilizar K-Medias

Este tipo de algotimo de aprendizaje no supervisado es útil para explorar, describir y resumir datos de una forma distinta. Utilizar este agrupamiento de datos nos puede servir para confirmar (o rechazar) algún tipo de clasificación previa. También nos puede ayudar a descubrir patrones y relaciones que desconocíamos.

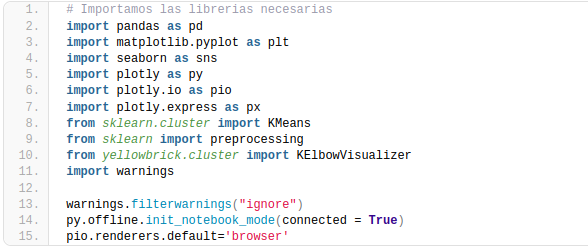
Por ejemplo, podemos aplicar K-means en:

* Segmentación de clientes
* Agrupación de textos que hablan de temas similares
* Geoestadística
* Comunidades de redes sociales

El clustering se utiliza principalmente en la fase de exploración de datos. Si buscamos ejemplos sobre aplicaciones en el campo de inversión cuantitativa hay varias ideas sobre aplicaciones interesantes:

Ejemplo:

librerias



Se obtienen los datos del documento

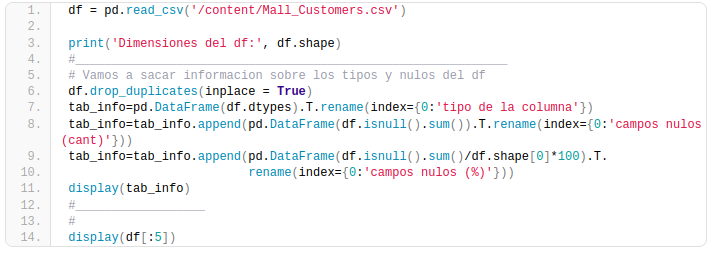
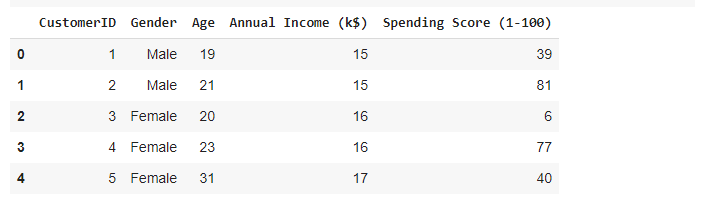
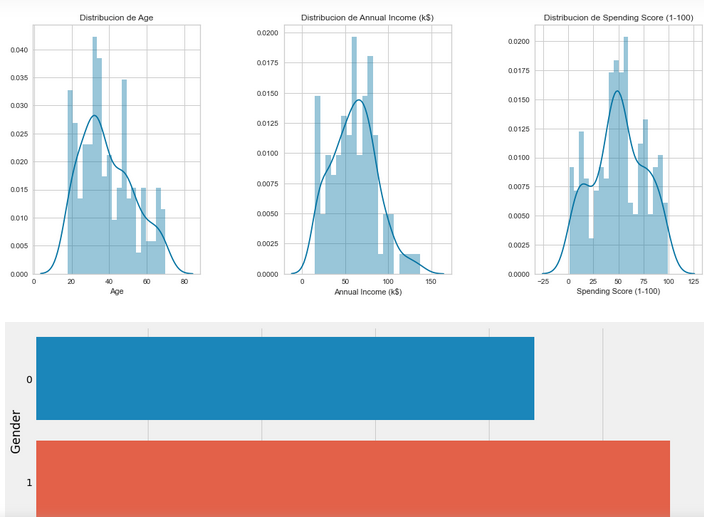


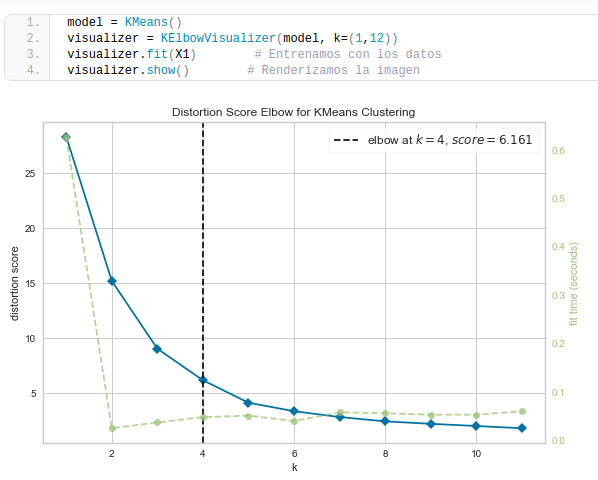
tabla obtenida



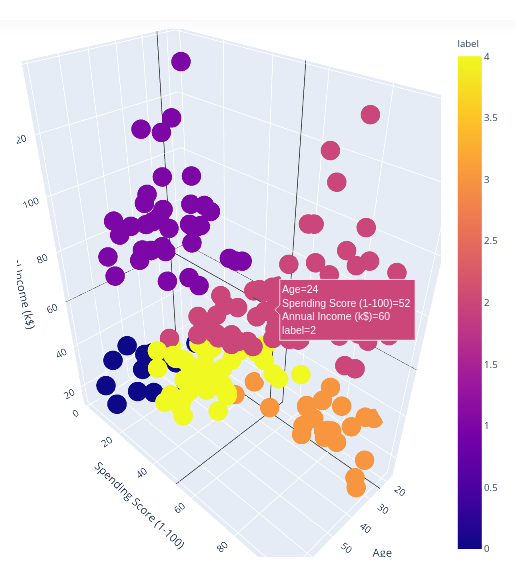
distribuciones



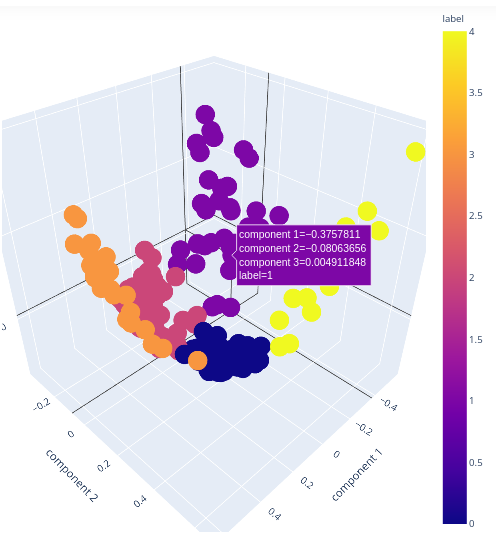
numero de clusters



resultado sin ordenar



datos ordenados



Conclusión

El clústering es una técnica muy popular para problemas sin etiqueta y también para tareas de EDA. El K-Means es el rey de esta técnica por su sencillez tanto de entender como de aplicar. Se basa en agrupar los datos según la distancia entre ellos.

Sanz, F. (2021, 14 mayo). *✅ Algoritmo K-Means | Clustering de forma sencilla*. The Machine Learners. Recuperado 8 de mayo de 2022, de https://www.themachinelearners.com/k-means/

O. (2020, 7 septiembre). *¿Qué es el Clustering?* GraphEverywhere. Recuperado 8 de mayo de 2022, de https://www.grapheverywhere.com/que-es-el-clustering/