INGENIERÍA EN DESARROLLO Y TECNOLOGÍAS DE SOFTWARE

[Proyecto Final: Kmeans-Clustering]

K-means Clustering: Es un método de aprendizaje automático no supervisado que se utiliza para descubrir patrones en un conjunto de datos. Los autores Hastie et al lo definen como "un algoritmo de agrupamiento popular que tiene como objetivo dividir puntos de datos en k grupos distintos y no superpuestos, donde cada punto de datos pertenece al grupo con la media más cercana". (Hastie, Tibshirani y Friedman, 2009).

Propósito: El propósito del presente proyecto final de la UC: BigData, es aplicar el uso de algoritmo de aprendizaje no supervisado denominado K-means; con la finalidad de analizar un conjunto de datos de consumidores frecuentes relacionados con la empresa WallCityTap S.A. Con la finalidad de dar respuesta a una serie de cuestionamientos e identificar patrones de comportamiento que coadyuven al proceso de toma de decisiones de los directivos; para ello, se realizará el desarrollo de un Dashboard de indicadores utilizando técnicas de visualización de datos y el framework Shiny for R or Python.

El conjunto de datos se encuentra disponible para descarga en el siguiente repositorio:

https://github.com/christianmce/unach/blob/master/BigData/WallCityTap_Consumer.csv

El Diccionario del conjunto de datos, se describe a continuación:

- **CustomerID**: un identificador único para cada cliente.
- **Gender**: El género del cliente.
- **Age**: La edad del cliente.
- **Annual_Income**: El ingreso anual del cliente (en miles de dólares).
- **Spending_Score (1–100)**: Puntuación asignada al cliente en función de sus hábitos de gasto. La puntuación varía de 1 a 100, y una puntuación más alta indica un cliente que gasta más.
- **Payment_Methods**: El tipo de medio de pago que más utiliza el consumidor.
- **Online_Purchases**: Indicador de compras por mes mediante uso del E-commerce.

Paso 1.- Realizar la carga del conjunto de datos hacia un dataframe

```
df = pd.read_csv("WallCityTap_Consumer.csv")          (Python)
df <- read.csv("WallCityTap_Consumer.csv");           (R)
```

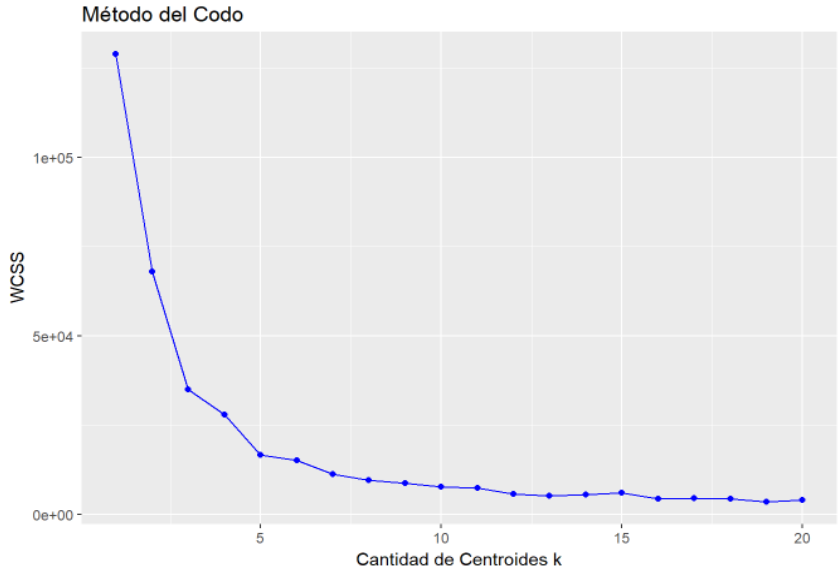
Paso 2.- Identificar el Valor K (Centroides) utilizando el método del codo

El método de K-Medios basa su funcionamiento en agrupar los datos de entrada en un total de k conjuntos definidos por un centroide, cuya distancia con los puntos que pertenecen a cada uno de los datos es la menor posible. En términos generales, el algoritmo puede resumirse como:

- Definir un total de k centroides al azar.
- Calcular las distancias de cada uno de los puntos de entrada a los k centroides, y asignar cada punto al centroide cuya distancia sea menor.
- Actualizar la posición de los k centroides, calculando la posición promedio de todos los puntos que pertenecen a cada clase.

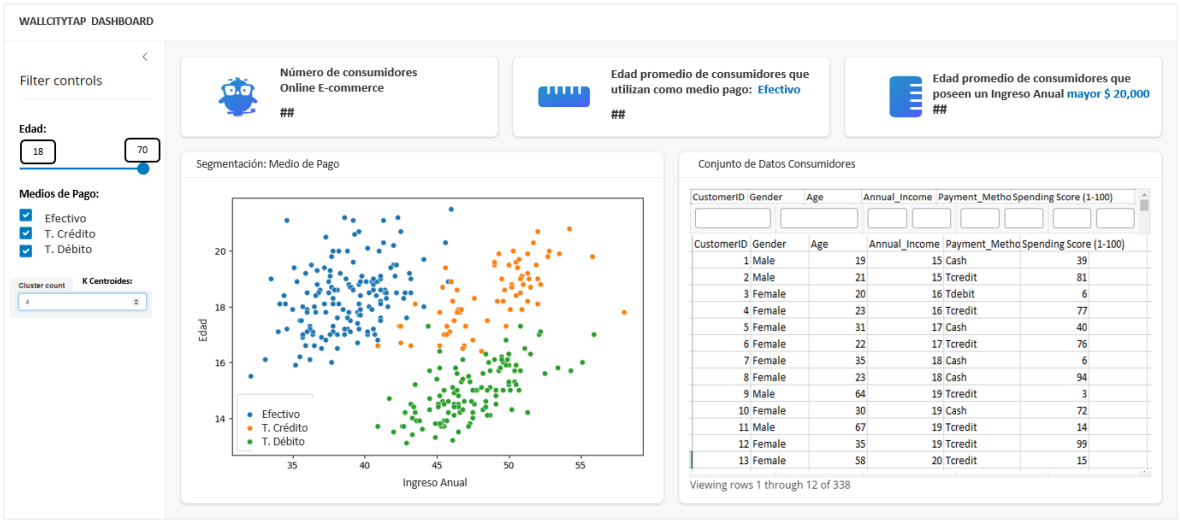
Repetir los pasos 2 y 3 hasta que los centroides no cambien de posición y, por lo tanto, las asignaciones de puntos entre clases no cambie.

La **cantidad óptima de centroides** k a utilizar no necesariamente se conoce de antemano, por lo que es necesario aplicar una técnica conocida como el **Método del Codo o Elbow Method** a fin de determinar dicho valor. Básicamente, este método busca seleccionar la cantidad ideal de grupos a partir de la optimización de la WCSS (Within Clusters Summed Squares).



A partir de la curva obtenida podemos ver cómo a medida que se aumenta la cantidad de centroides, el valor de *WCSS* disminuye de tal forma que la gráfica adopta una forma de *codo*. Para seleccionar el valor óptimo de *k*, se escoje entonces ese punto en donde ya no se dejan de producir variaciones importantes del valor de *WCSS* al aumentar *k*. En este caso, vemos que esto se produce a partir de *k* >= 7, por lo que evaluaremos los resultados del agrupamiento, por ejemplo, con los valores de 7, 8 y 9 a fin de observar el comportamiento del modelo.

Paso 3.- Mockup del Dashboard de Visualización de Datos (Shiny for R or Python)



Para diseñar el Dashboard dinámico, se sugiere preferentemente utilizar el framework Shiny optando por el lenguaje de programación R o Python; el cual debe visualizar el conjunto de datos en un componente DataTable, visualizar la aplicación del algoritmo K-means Clustering mediante un componente Plot, con sus respectivos controles de manipulación de entrada de datos.

Interpretación de Dashboard para toma de decisiones

Los directivos desear dar respuesta a los siguientes cuestionamientos, con la finalidad de formular una próxima campaña de nuevos productos y ofertas para sus consumidores frecuentes.

Cuestionamientos de los Directivos de WallCityTap S.A:

1. ¿Es viable financieramente lanzar una promoción del 2 X 1 + meses a interés para los consumidores con una edad mayor a 50 años?

2.- ¿Cuál es el medio de pago preferido de nuestros consumidores?

3.- Identificar y describir los consumidores potenciales, para generar mayores ventas y obtener mayores ingresos.

PRODUCTOS ENTREGABLES:

- Informe del proyecto final en formato .pdf
- Dashboard (Shiny) para visualización de datos

EQUIPO: Integrado por un máximo de 3 estudiantes.

El **Informe final** deberá contener:

- Hoja de presentación (nombre del proyecto, estudiantes, fecha, etc.)
- Introducción
- Código del proyecto y URL del repositorio Github
- Resultados (Captura de pantallas del funcionamiento)
- Interpretación de resultados y cuestionamientos del caso de estudio.

FECHA DE ENTREGA Y PRESENTACIÓN: **24 DE MAYO DE 2024**

Referencias y tutoriales:

https://shiny.posit.co/r/gallery/start-simple/kmeans-example/

https://github.com/posit-dev/py-shiny-templates/tree/main/dashboard

https://towardsdatascience.com/k-means-data-clustering-bce3335d2203

https://www.aprendemachinelearning.com/k-means-en-python-paso-a-paso/

https://rpubs.com/rdelgado/399475

https://rpubs.com/Dariel1102/1046632

Dr. Christian Mauricio Castillo Estrada <cmce@unach.mx>
Profesor de tiempo completo
Ingeniería en Desarrollo y Tecnologías de Software
Facultad de Negocios Campus IV UNACH