Serie 1 (30 puntos)

1. Desarrolle con sus palabras, ¿cuál es la diferencia entre aprendizaje supervisado y el no supervisado?
    a. El aprendizaje supervisado es un tipo de técnica de aprendizaje automático donde se proporciona al modelo un conjunto de datos etiquetados previamente, es decir, que ya se conoce la salida deseada. El objetivo es que el modelo aprenda a predecir la salida correcta a partir de nuevas entradas desconocidas. En cambio, el aprendizaje no supervisado no utiliza datos etiquetados previamente, en su lugar, busca patrones en los datos sin una salida específica en mente. El objetivo es agrupar o clasificar los datos en función de su similitud o diferencia.
2. ¿En que se basa el algoritmo de kmeans para determinar a que cluster debe de pertenecer cada una de las observaciones?
    a. El algoritmo de K-means se basa en la distancia Euclidiana para determinar a qué cluster pertenece cada observación. En pocas palabras, el algoritmo comienza por seleccionar k centroides iniciales de manera aleatoria. Luego, calcula la distancia entre cada observación y los centroides y las asigna al cluster cuyo centroide esté más cercano. A continuación, se recalculan los centroides de cada cluster utilizando la media de las observaciones que lo componen. Este proceso se repite iterativamente hasta que los centroides ya no cambian de posición y los clusters son estables.
3. Desarrolle con sus palabras, ¿Qué acciones se pueden tomar si tengo datos incompletos en un set de datos?
    a. Algunas acciones que se pueden tomar si se tienen datos incompletos en un set de datos son: eliminar las observaciones con datos faltantes, reemplazar los valores faltantes con la media o mediana de la variable, utilizar técnicas de interpolación para estimar los valores faltantes o utilizar técnicas más avanzadas de imputación de datos para llenar los valores faltantes.

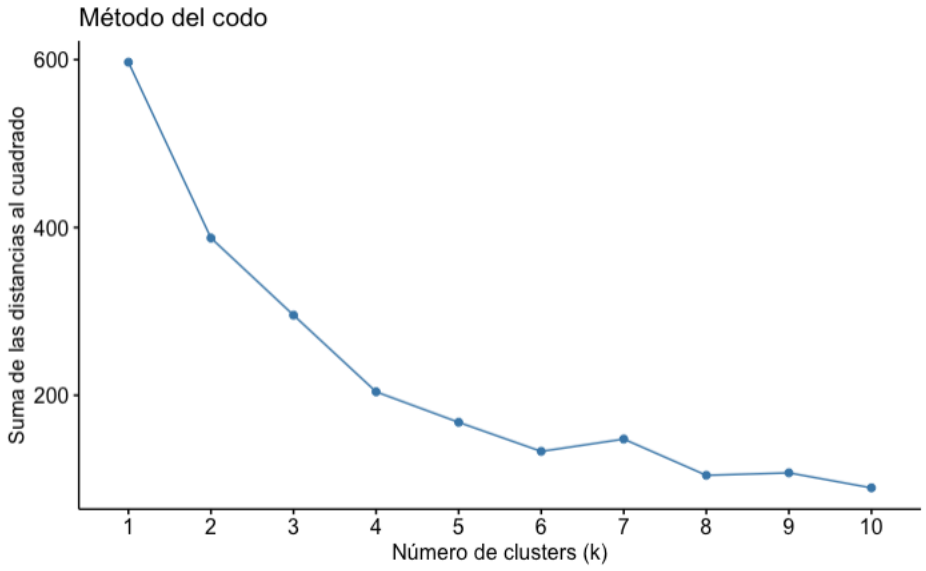
4. Si tuviera un set de datos con variables categoricas, ¿qué accion tomaria para poder utilizar estos datos en el entrenamiento?
    a. Si tuviera un set de datos con variables categóricas, una acción que podría tomar para utilizar estos datos en el entrenamiento es convertir las variables categóricas en variables numéricas mediante técnicas de codificación, como la codificación one-hot o la codificación ordinal. Esto permitirá que el modelo aprenda de las variables categóricas y las utilice en su entrenamiento.
5. ¿por qué es importante “normalizar” las caracteristicas numericas para efectuar un entrenamiento
    a. Es importante normalizar las características numéricas antes del entrenamiento porque los algoritmos de aprendizaje automático suelen funcionar mejor con características en la misma escala. Si una característica tiene un rango mucho mayor que otra, esto puede afectar la contribución relativa de cada característica al modelo. Además, la normalización ayuda a evitar que los algoritmos de aprendizaje automático sean demasiado sensibles a las características con valores atípicos o extremos.

Serie 2 (70 puntos)

En la siguiente serie deberá utilizar las fuentes de datos indicadas para analizar la información usando R.

Un mall en Estados unidos ha ido recopilando una base de datos de los clientes que llegan a sus instalaciones. Se desea efectuar un estudio para determinar y clasificar los segmentos de clientes que visitan el centro comercial. Se le ha contratado para efectuar este estudio:

a) De la base de datos llamada “Mall_Customers 5.csv”
    a. Genere la estadística General de los datos. (se recomienda usar la funcion summary de R).
    b. Efectue la limpieza de los datos según lo visto en las secciones de feature engineering en clase
b) Determinar por medio del metodo del codo (elbow method) la cantidad de clusters optima para efectuar la segmentacion. (presente grafica con sus conclusiones de la cantidad de segmentos).
c) En base a la recomendación del inciso anterior, efectue la clasificacion de acuerdo a la cantidad de cluster optima, presentando el grafico correspondiente.
    a. Según la gráfica del método del codo, el punto de inflexión se encuentra en 6 clusters, lo que indica que esa es la cantidad óptima de clusters para la segmentación de los datos en este caso.



Diseño:

- Exclusion de columnas
    - Podemos eliminar la columna "ID" ya que no es relevante para nuestro análisis
- Limpieza de datos
    - El genero se paso de string a factor
- Se utilizara la librería ("factoextra") para el metodo de el codo.