**UNIVERSIDAD RAFAEL LANDÍVAR**

FACULTAD DE INGENIERÍA

ANÁLISIS DE DATOS

SECCIÓN 1 VESPERTINA

MGTR. JAIRO OMAR SALAZAR CHÁVEZ

**EXAMEN PARCIAL 2**

**Julio Anthony Engels Ruiz Coto 1284719**

GUATEMALA DE LA ASUNCIÓN, ABRIL 26 DE 2023

CAMPUS CENTRAL

**Serie 1 (30 puntos)**

1. Desarrolle con sus palabras, ¿cuál es la diferencia entre aprendizaje supervisado y el no supervisado?

*R// La diferencia es que el aprendizaje supervisado se orienta en predecir una variable especial que tiene relación con datos descriptivos, caso contrario el no supervisado se orienta en buscar información, estructuras o patrones en los datos en este aprendizaje no existe una variable especial que haya que predecir.*

2. ¿En qué se basa el algoritmo de kmeans para determinar a qué clúster debe de pertenecer cada una de las observaciones?

*R// El algoritmo es de aprendizaje no supervisado este se basa en elegir de manera aleatoria las coordenadas de los centroides (k) del conjunto de datos, los centroides son los puntos que marcan el centro de cada agrupación de datos, cada punto se agrupa con el centroide más cercano para eso se utiliza una medida de distancia llamada distancia euclidiana.*

3. Desarrolle con sus palabras, ¿Qué acciones se pueden tomar si tengo datos incompletos en un set de datos?

*R// Las acciones que se pueden tomar son determinar si la cantidad de datos faltante es mayor al 80% entonces se procede a eliminar los datos de dicha variable caso contrario, se rellenan los datos faltantes con el valor promedio del dato anterior y el dato posterior, otra alternativa rellenar con el valor promedio de los datos existentes,*

4. Si tuviera un set de datos con variables categóricas, ¿qué acción tomaría para poder utilizar estos datos en el entrenamiento?

*R// La acción que tomaría para este tipo de caso es, convertir esas variables categóricas en un formato numérico en la que pueda ser entendido por el modelo, una de las técnicas que se pueden aplicar es el one hot encoding este crea un nueva columna binaria para cada categoría donde 1 dato presente y 0 dato ausente. Así también el label encoding consiste en asignar a dichas categorías un valor numérico, es una manera simple, pero si no se tiene un orden puede causar problemas ya que el modelo puede entender una relación ordinal entre las categorías que no existe.*

5. ¿por qué es importante “normalizar” las características numéricas para efectuar un entrenamiento?

*R//Es recomendable hacerlo cuando los datos no tienen la misma escala, la normalización limita los valores entre [0 y 1] además el método de K-means es altamente sensible a los outliers.*

**Serie 2 (70 puntos) En la siguiente serie deberá utilizar las fuentes de datos indicadas para analizar la información usando R.**

Como primer paso importo el dataset a R y librerías a utilizar.

Texto

Descripción generada automáticamente



Tabla

Descripción generada automáticamente

Luego genero la estadística general de los datos usando la función summary.



Texto

Descripción generada automáticamente

Acá elimino la columna CustomerID porque es un identificador único para cada persona y no aporta una información útil para el agrupamiento, ya que en el clustering buscamos en agrupar datos con características similares y no se relaciona con las otras características del conjunto de datos.



Luego aplico una codificación binaria para la columna Genre por que como tiene dos categorías Male, Female y se puede convertir en una variable binaria asignando 1 como Male y 0 para Female con esto facilito el análisis y el procedimiento de los datos por parte del algoritmo de aprendizaje, en este paso se usa la librería dplyr para el uso del if\_else.

Imagen que contiene Texto

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

Luego hago una verificación rápida si hay valores nulos, en este caso no presenta.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ya con esto empiezo a entrenar el algoritmo de K-means.

Imagen que contiene Calendario

Descripción generada automáticamente

Realizo el diagrama con centros y se ve la distribución de los clúster.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Luego aplico la regla del codo, el número de clúster con el error.

Tabla

Descripción generada automáticamente

Por último, se realiza la gráfica de la regla del codo y se determina que con un clúster de 5 es el valor óptimo para entrenar el algorimo de clustering.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Y procedo a transformar la columna de annual\_income\_k aplicando la fórmula de estandarización.

Tabla

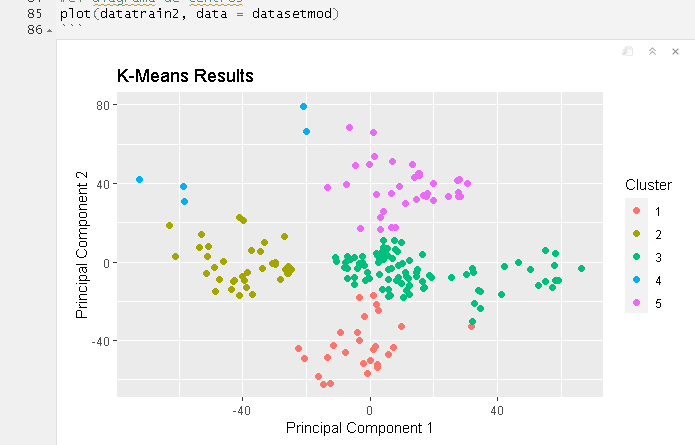
Descripción generada automáticamente

Se repite los pasos para entrenar el algoritmo con datatrain2, income modificado.

Calendario

Descripción generada automáticamente

El diagrama de centros con Income modificado, se puede ver como hay datos que tienen cierta relación por la forma de su agrupamiento cercano de los centroides con cada uno de los datos como es el clúster 1,3,5.



Luego aplico la regla del codo nuevamente, el número de clúster con el error.

Tabla

Descripción generada automáticamente

Luego, se realiza la gráfica de la regla del codo y se determina como queda esta grafica con los valores estandarizados con el clúster igual a 5.

Gráfico

Descripción generada automáticamente

Por último, realizo la gráfica de resultados utilizando la función plot() de R, dicha grafica se observa la clasificación de los datos según la cantidad optima de clústers, usando la regla del codo las estrellas que se observan son los centros de los clústers y los círculos de colores las observaciones.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente