Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Jairo Salazar

Examen Parcial No.2

Lea con atención las instrucciones que se le presentan. De manera individual, desarrolle las siguientes actividades. Tiene de 17:30 a 19:00 para responder, utilice su criterio y conocimientos desarrollados durante el curso para ejecutar cada actividad. Al finalizar el examen, deberá enviar por la plataforma Moodle un zip que contenga los elementos siguientes:

1. Word con respuestas de Serie 1
2. Script de R con el código generado (no olvide comentar su procedimiento)
3. Presentación de PowerPoint requerida.

Importante: Por favor no se arriesguen ayudando a otros compañeros o pidiendo ayuda a los demas, se sancionara drasticamente al que provea y al que copie el examen de otro compañero.

Serie 1 (25 puntos)

1. Desarrolle con sus palabras, ¿cuál es la diferencia entre aprendizaje supervisado y el no supervisado?
2. ¿En que se basa el algoritmo de kmeans para determinar a que cluster debe de pertenecer cada una de las observaciones?
3. Desarrolle con sus palabras, ¿Qué acciones se pueden tomar si tengo datos incompletos en un set de datos?
4. Si tuviera un set de datos con variables categoricas, ¿qué accion tomaria para poder utilizar estos datos en el entrenamiento?
5. ¿por qué es importante “normalizar” las caracteristicas numericas para efectuar un entrenamiento?

Serie 2 (75 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”
    i. Genere la estadística General de los datos. (se recomienda usar la funcion summary de R).
    ii. Efectue la limpieza de los datos según lo visto en las secciones de feature engineering en clase.

Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Jairo Salazar

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.

Entregables:

1. Su codigo en un R notebook.
2. Pdf explicando sus decisiones en relacion al numero de clusters y alguna otra decision de diseño que haya podido tomar (exclusion de columnas, limpieza de datos, imputacion de datos faltantes, normalizacion de los datos, etc).