

CLUSTERING – MODELAMIENTO NO SUPERVISADO

Prof. Adrian Armando Araneda Toro.

Instrucciones.

1. La Evaluación es Individual.
Las respuestas escritas realícelas en el Script con la enumeración correcta de la pregunta a la cuál responde. No se aceptará como respondida aquellas preguntas que el profesor reciba en formato manuscrito, a mano, escaneado o fotografiado.
2. Escriba su nombre, Apellido y RUN en el comienzo de su Script .R
3. Recuerde que al profesor debe ejecutar con éxito su sintaxis línea por línea. Si la pregunta X depende de la ejecución exitosa de la línea anterior y esta no se ejecuta, su respuesta X estará errónea.
4. Asuma que el profesor sólo hará “Ctrl+Enter” por cada línea de código. No corregirá código para que sea ejecutado exitosamente.
5. No asuma que el profesor posee los paquetes y librerías ya instalados(as) y ejecutados(as), debe enunciarlos en el script para habilitar por primera vez la función que desea utilizar.
6. El Script debe ser enviado a través de un solo correo electrónico (no dos, tres, etc.). En el caso que el alumno envíe mas de un correo electrónico dentro del horario consignado para el desarrollo y entrega de la evaluación con elementos adjuntos repetidos en un correo anterior, el profesor seleccionará aquellos elementos del último correo electrónico recibido dentro del horario de entrega.
7. En dicho correo electrónico se debe adjuntar el desarrollo del Script solicitado, o adjuntar un archivo comprimido con dicho Script.
8. Enviar sólo por U-Cursos hasta las 13.30 hrs. No se recibirán evaluación ni elementos desde las 13.31 hrs.
9. No se aceptarán enmendaciones de las entregas enviadas fuera de la hora indicada.

Observación: Revise que su Script se haya adjuntado al enviar la evaluación. Hágalo 15 minutos antes de la hora de cierre.

10. Desde su Script, cargue el siguiente set de Datos:

```
USArrests <- USArrests
```

Este conjunto de datos contiene estadísticas, en arrestos por cada 100.000 residentes por agresión, asesinato y violación en cada uno de los 50 estados de EE. UU. en 1973. También se proporciona el porcentaje de la población que vive en áreas urbanas.

# [,1]	Asesinato	Arrestos por asesinato (por 100.000)
# [,2]	Agresión	Detenciones por agresión (por 100.000)
# [,3]	UrbanoPop	Porcentaje de población urbana
# [,4]	Violación	Detenciones por violación (por 100.000)

Para mayor información del datasets, consulte:

Fuentes: <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/USArrests.html>

PREGUNTAS

1. Al set de datos cargado genere una variable denominada “nueva1”, la cuál debe ser el resultado de la suma de cada observación del vector 2 con el tercer número de su RUN (leyéndolo de izquierda a derecha. Si este es cero, pase al siguiente dígito en el mismo orden). **(0.2 pts.)**
2. En el mismo set de datos, genere una variable denominada “nueva2” la cuál debe ser el resultado de la multiplicación de cada observación del vector 3 con el cuarto número de su RUN (leyéndolo de izquierda a derecha. Si este es cero, pase al siguiente dígito en el mismo orden). **(0.2 pts.)**
3. En el set de datos anterior, entregue en un vector nuevo resultante de la división de la dimensión 1 con la 3. **(0.2 pts.)**
4. Por instrucción de Negocio (gerencia), usted debe trabajar con variables correlacionadas. No obstante, el equipo de analítica le solicita que de todas maneras identifique cuáles son los pares de variables correlacionadas fuertemente (utilizar criterio enseñado por el profesor). **(0.8 pts.)**

5. Responda sí o no si su set de datos posee outliers. Para esto, realice una demostración a través de la **aproximación visual** más completa enseñada en clases para responder si su set de datos posee outliers o no. **(0.7 pts.)**
6. Según las técnicas vistas en clases, proponga un número óptimo de K y genere los clusters con el algoritmo enseñado en clases sensible a los outliers, y con 100 iteraciones. **(0.6 pts.)**
7. Traiga los clusters al dataframe original. **(0.25 pts.)**
8. El equipo de Analítica le encomienda eliminar los registros que hayan quedado en el clusters con la menor cantidad de elementos. Asumiendo que son outliers. Y le solicita que vuelva a generar los clusters con el mismo algoritmo sensible a los outliers, y definiendo por usted, según las técnicas que está en conocimiento, cuál sería el mejor óptimo de K. **(1.25 pts.)**
9. Finalmente, el equipo de analítica le solicita que convenza a Negocio con dos tipos de argumentos, técnicos y económicos. Para esto, analítica le solicita que utilice la librería `clValid`, evaluando la primera clusterización que realizó. Versus la segunda clusterización que realizó. **(1.8 pts.)**

Con argumentos técnicos, usted entiende lo enseñado por el profesor para los parámetros de `clValid`.

Con argumentos económicos, usted asumirá que el costo de cualquier intervención para los estados son los que se consignan en el set de datos adjunto, llamado “Costo_por_Estado_”. Tendrá que responder con evidencia en costos, cuál de las dos clusterizaciones es la más eficiente.