

## CLUSTERING – MODELAMIENTO NO SUPERVISADO

Prof. Adrian Armando Araneda Toro.

### Instrucciones.

1. La Evaluación es Individual.
2. Las respuestas escritas realícelas en el Script con la enumeración correcta de la pregunta a la cuál responde.
3. No se aceptará como respondida aquellas preguntas que el profesor reciba en formato manuscrito, a mano, escaneado o fotografiado.
4. Escriba su nombre, Apellido y RUN en el comienzo de su Script .R
5. Recuerde que el 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 correctamente, su respuesta X estará errónea.
6. Asuma que el profesor sólo hará “Ctrl+Enter” por cada línea de código. No corregirá código para que sea ejecutado exitosamente.
7. 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. Si su desarrollo es correcto, pero arroja error dado que usted no habilitó la librería correspondiente, el desarrollo no tendrá puntaje.
8. El Script debe ser enviado a través de un solo correo electrónico (no dos, tres, etc.). En el caso que el alumno envíe más 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.

9. En dicho correo electrónico se debe adjuntar el desarrollo del Script solicitado, o adjuntar un archivo comprimido con dicho Script.
10. Enviar sólo por U-Cursos hasta las 23.59 hrs. No se recibirán evaluación ni elementos desde las 00.00 hrs. o 24 hrs.
11. No se aceptarán enmendaciones de las entregas enviadas fuera de la hora indicada.
12. 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>

**Observación 1:** Revise que su Script se haya adjuntado correctamente al enviar la evaluación. Hágalo 15 minutos antes de la hora de cierre.

**Observación 2:** Observe la asignación de puntaje por cada pregunta.

## 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 segundo 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 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.)**
3. En el set de datos anterior, genere 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 (mencione) cuales son los pares de variables correlacionadas fuertemente (utilizar criterio enseñado por el profesor).

Así también negocio le pide que le explique no en más de 5 líneas, como podría influir en la calidad de los clústers trabajar con variables correlacionadas.

**(0.4 pts.)**

5. Responda si su set de datos posee outliers hipotéticos. Para esto realice solo una demostración a través de la **aproximación visual** más completa enseñada en clases para responder si su set de datos posee outliers hipotéticos o no. **(0.7 pts.)**
6. Según las técnicas vistas en clases,
  - a) Proponga un número óptimo de K utilizando la librería `clValid` y una técnica adicional enseñada en clases. Fundamente técnicamente en virtud de todos los parámetros vistos en clases.
  - b) Genere los clusters con el algoritmo enseñado en clases que es altamente sensible a los outliers, y con 100 iteraciones.
  - c) Plotee en 3D los clusters pero solo con las tres primeras PCA.

**(1 pts.)**

7. Traiga los clusters obtenidos al dataframe original. **(0.25 pts.)**
8. El equipo de Analítica le encomienda:
  - a.) Eliminar del modelo de datos los registros que hayan quedado en el clusters con la menor cantidad de elementos, asumiendo que son outliers.
  - b.) Le solicita que vuelva a recomendar un óptimo de K utilizando la librería `clValid` y una técnica adicional enseñada en clases, y volver a generar los clusters según ese óptimo con el mismo algoritmo sensible a los outliers.
  - c.) Plotee en 3D los clusters pero solo con las tres primeras PCA.

**(1.25 pts.)**

9. Asuma que el costo de intervención en seguridad y delincuencia para los estados (o comunas de un país) que se consignan en el set de datos adjunto llamado "Costo\_por\_Estado\_", varían de estado a estado.

En virtud de los hallazgos de la letra b de la pregunta 8, indique solo dos técnicas de Clustering que usted recomendaría (distinto K y distinto tipo de algoritmo) y lo más importante: Negocio le solicita que demuestre comercialmente qué técnica implicaría mayores costos de intervención para la empresa de seguridad o institución policiaca; ¿la técnica recomendada A o la técnica recomendada B sería la más eficiente?

**(1.8 pts.)**