Universidad de los Andes
Facultad de Ingeniería

Universidad de los Andes
MINERÍA DE DATOS (MIIA 4200)
Profesor: Rafael Amaya Gómez
Fecha: 12 de Febrero 2024

RECUERDE:
El link para enviar los archivos de soporte estará habilitado hasta las 11:59p.m. del 20 de febrero, cualquier entrega posterior tendrá una calificación sobre 4.5
La entrega del taller deberá hacerse en un informe autocontenido incluyendo las gráficas, interpretación y conclusiones. Adicionalmente debe entregarse el código utilizado en un archivo R, de considerar necesario comentar todas las partes del código que crean necesarias para el entendimiento de este.
Para todos los literales que lo requieran, se recomienda utilizar una significancia del 5% (confianza del 95%). En caso de utilizar otro nivel de significancia, hacerlo explícito en el informe que se debe entregar en PDF adicional al script de R.

TAREA 1

PUNTO 1 (40%)

El crecimiento poblacional, la urbanización y el cambio climático han aumentado la conciencia de las personas sobre el impacto de las actividades humanas en el medio ambiente y en los recursos naturales disponibles, como los recursos hídricos. En este contexto, la gestión sostenible de los sistemas hídricos es crucial para evitar la escasez de agua o el agotamiento de las fuentes disponibles, por los cual es fundamental identificar patrones en las curvas de consumo de cada uno de los usos de la infraestructura de una cuidad. Identificando variables que puedan explicar el comportamiento de las curvas de demanda.

Tabla 1. Áreas características con

DMA	Area characteristics
A	Hospital district
B	Residential district in the countryside
C	Residential district in the countryside
D	Suburban residential/commercial district
E	Residential/commercial district close to the city centre
F	Suburban district including sport facilities and office buildings
G	Residential district close to the city centre
H	City centre district
I	Commercial/industrial district close to the port
J	Commercial/industrial district close to the port

1. Se cuenta con dos dataset. Como primer objetivo se deben unir las bases de datos, identificando si hay fechas repetidas en cuyo caso se debe quedar con el primer registro. Por último, se debe usar una técnica de imputación (asignación o eliminación) para los datos faltantes.

2. Se desea identificar si se tiene un patrón dentro de las series de tiempo, para esto se debe usar la técnica de media móvil variando la ventana de tiempo que se usa para cada corrida. Se le indica que debe usar 5 ventanas de tiempo diferentes para comparar los resultados y así poder concluir sobre el patrón en cada serie de tiempo.
3. Asimismo, se le pide que busque variables que tenga una relación con el consumo de cada una de las áreas mostradas en la Tabla 1.

Bono (5%): El dataset incluye posibles outliers. Identifique, elimine y compare los resultados obtenidos.

PUNTO 2 (40%)

Spotify es una empresa para la reproducción de música vía streaming que se ha venido consolidando como una de las plataformas más relevantes a nivel mundial. Considere que usted se ha conectado con la API de Spotify y ha podido descargar la información de la banda **"Los Planetas"**, una banda española de indie rock que ha estado activa desde 1993

Dentro de la información disponible descargada se encuentra nombres de canciones, nombres de los álbumes, duración o tempo. Adicionalmente, tiene unos descriptores en términos a qué tan positiva, bailable, enérgica, acústica, instrumental o en vivo fue cada canción. Con base en esta base de datos, se le pide lo siguiente:

1. Realice un análisis de PCA que permita describir el tipo de canciones. Por ejemplo, ¿Cómo se relaciona las canciones bailables con las de tipo instrumental? ¿Las canciones con larga duración con las enérgicas? ¿Las canciones en vivo con las acústicas?
2. Con base en las nubes de individuos y variables, describan los álbumes de esta banda. ¿Cuáles podrían ser más enérgicos, positivos, entre otras características?
3. Explique el funcionamiento de los ejes de variables. ¿Qué separa el primer y el segundo eje?
4. ¿La proyección es suficientemente buena para poder interpretar los datos para todas las variables? Justifique su respuesta

PUNTO 3 (20%)

La reducción de dimensionalidad no se centra únicamente en las aproximaciones que proyectan los datos en una dimensión diferente como en el caso de PCA, MCA o KPCA. También se incluye la selección o eliminación de las variables menos representativas por lo que hay herramientas como la selección recursiva hacia adelante o hacia atrás. Sin embargo, hay otras alternativas que puede usar en este sentido como los métodos Wrapper como el caso de la eliminación recursiva de rasgos (RFE) o filtros como la importancia de variables.

Con base en lo anterior:

1. Investigue cuáles son los métodos más usados para hacer selección de rasgos y describa su funcionamiento a través de R con un ejemplo con la base de datos de expertos de Vino disponible en Bloque Neón con nombre "data_PCA_ExpertWine.csv"
2. Explique cuáles son sus principales características y cuáles ventajas o desventajas trae su implementación. ¿Cuál contemplaría en su análisis y por qué?